Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Міністерство освіти і науки України

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Міністерство освіти і науки України

Кваліфікаційна наукова праця

На правах рукопису

УДК 004.852

ЖУК ІВАН СЕРГІЙОВИЧ

ДИСЕРТАЦІЯ

МАТЕМАТИЧНІ МОДЕЛІ І МЕТОДИ ВИЯВЛЕННЯ ЗА ПУБЛІЧНО ДОСТУПНИМИ ДАНИМИ ПІДОЗРІЛИХ НА ФІКСОВАНИЙ РЕЗУЛЬТАТ ФУТБОЛЬНИХ МАТЧІВ

11 – Математика та статистика

113 – Прикладна математика

Подається на здобуття наукового ступеня доктора філософії.

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

_____ І. С. Жук

Науковий керівник

Чертов Олег Романович, доктор технічних наук, професор.

Київ 2023

АНОТАЦІЯ

Жук І. С. Математичні моделі і методи виявлення за публічно доступними даними підозрілих на фіксований результат футбольних матчів. — Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії з галузі знань 11 Математика та Статистика за спеціальністю 113 Прикладна математика. — Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, 2023.

Метою роботи є підвищення ефективності виявлення підозрілих на фіксований результат футбольних матчів на базі обробки виключно загальнодоступних публічних даних за результатами сезону футбольного турніру.

Однією з найважливіших проблем футболу, що зіставна з проблемою допінгу, є договірні матчі. Результати таких матчів або певний перебіг подій в них є наперед визначеними, тобто фіксованими. У договірних матчах, пов'язаних з виграшом за ставками, завданням є отримання результату, відмінного від очікуваного. Тому такі результати можна розглядати як нетипові, аномальні.

Для перевірки поточного матчу на фіксований результат використовують математичні методи футбольної аналітики, такі як: прогнозування результату матчу, аналіз ставок або дій учасників матчу протягом всієї гри. Їх перевагою є оперативність прийняття рішень, а недоліком — необхідність використання дуже великої кількості даних, які, зазвичай, не є публічно доступними. Альтернативним може розглядатись підхід, коли рішення щодо фіксованості матчу приймається за результатами усього сезону. При цьому загально доступною є публічна інформація щодо результатів проведених ігор усіх команд, що дозволяє формалізувати пошук

матчів, підозрілих на фіксований результат як задачу виявлення контекстуальних аномалій.

Найбільш адекватними розглянутій задачі виявлення підозрілих на фіксований результат матчів на основі доступної публічної інформації є статистичні непараметричні гістограмні методи. Це обумовлено тим, що вхідні дані характеризуються малою кількістю дискретних числових значень і їх закони розподілу ймовірностей є невідомими. Водночас, ефективність використання цих методів залежить від об'єму вибірки. Разом з цим, перспективним є математичний апарат конформних предикторів та степеневих мартингалів, який не вимагає знання законів розподілу даних, використовує інформацію про міру неконформності даних та може бути використаний для вирішення задач класифікації даних.

Тому актуальною науковою задачею є розробка методів виявлення підозрілих на фіксований результат матчів з використанням апарату конформних предикторів і степеневих мартингалів на базі обробки виключно загальнодоступних публічних даних за результатами сезону футбольного турніру.

В *першому розділі* розглянуто проблематику договірних футбольних матчів і обґрунтовано актуальність досліджень, спрямованих на пошук матчів, потенційно підозрілих на фіксованість результату, на базі обробки виключно загальнодоступних публічних даних. Показано, що застосування математичних методів футбольної аналітики, таких як прогнозування результату матчу, аналізу ставок або дій учасників матчу протягом всієї гри щодо виявлення підозрілих на фіксований результат матчів вимагає великої кількості даних, які не завжди доступні для аналізу. Відмічено, що задача виявлення підозрілих щодо фіксованого результату матчів за результатами сезону футбольного турніру відноситься до класу задач виявлення контекстних аномалій, які вирішуються в області інтелектуального аналізу даних. Проведено аналіз статистичних методів, а також методів машинного навчання, які використовуються для виявлення аномалій. Особливу увагу

присвячено методам класифікації на основі конформних предикторів, які запропоновано використати для підвищення ефективності виявлення потенційно підозрілих договірних матчів. Сформульовано мету і наукові завдання дисертаційного дослідження.

Другий розділ присвячено розробці імітаційної моделі футбольного сезону з матчами з фіксованим результатом. Для визначення контекстуального атрибуту «сила команди» проведено групування команд методами *K*-середніх та Гаусівських сумішей за ознаками кількість очок та різницею між забитими і пропущеними м'ячами в одновимірному та двовимірному просторах. Визначено початкові дані, обмеження та формули розрахунку ймовірностей забиття голів командами під час гри на основі реальних даних сезону для побудови імітаційної моделі. Розроблено імітаційну модель футбольного сезону та проведено її аналіз шляхом статистичного моделювання. Розроблено алгоритм моделювання договірних матчів, пов'язаних із заробітком на ставках.

Третій розділ присвячено розробці методів виявлення підозрілих щодо фіксованості результату футбольних матчів за наявності даних про весь сезон. Запропоновано міри неконформності поточного матчу. На основі аномального конформного детектору розроблено метод виявлення підозрілих щодо фіксованості результату футбольних матчів, в якому прийняття рішень відбувається за пороговим правилом. Проведено порівняльний аналіз методів виявлення матчів, підозрілих на фіксований результат на основі експертно визначеного порогу відхилення і конформного аномального детектору. З використанням степеневого і інтегрального мартингалів розроблено методи виявлення підозрілих щодо фіксованого результату футбольних матчів, в яких прийняття рішення відбувається при зростанні значення степеневого мартингалу для поточного спостереження по відношенню до значення цього ж мартингала для попереднього спостереження.

В *четвертому розділі* шляхом імітаційного моделювання розглянуто аналіз особливостей розроблених методів за даними окремих класів

модельного сезону. Проведено порівняльний аналіз розроблених методів та відомого гістограмного методу за даними модельного сезону з використанням метрики точності (precision, P), повноти (recall, R) і міри F_1. Запропоновані методи також використано для виявлення матчів, які вважаються договірними, в сезоні 2014–2015 рр. Серії *B* Італії.

Наукова новизна одержаних результатів полягає у наступному:

1. Розроблено **новий метод** виявлення підозрілих щодо фіксованості результату футбольних матчів, який відрізняється від відомих застосуванням конформного аномального детектора із запропонованою мірою неконформності поточного матчу, що забезпечує можливість визначення порогу прийняття рішення у відповідності до заданого значення апріорної ймовірності появи аномальних даних.

2. Розроблено **новий метод** виявлення підозрілих щодо фіксованості результату футбольних матчів, який відрізняється від відомих застосуванням степеневого мартингалу і правилом прийняття рішення на основі порівняння поточного значення степеневого мартингалу з попереднім, що дозволяє за рахунок зміни параметра чутливості налаштовувати степеневий мартингал на виявлення аномалій відповідного рівня і знаходити їх.

3. Розроблено **новий метод** виявлення підозрілих щодо фіксованості результату футбольних матчів, який відрізняється від відомих застосуванням інтегрального мартингалу і правилом прийняття рішення на основі порівняння поточного значення інтегрального мартингалу з попереднім, що дає змогу виявляти аномальні матчі без налаштування параметрів.

4. Доведені **нові властивості** степеневого мартингалу:

- за яких завгодно малих значень ступеня конформності (p-value) поточного спостереження значення степеневого мартингала для поточного спостереження є більшим за значення цього ж мартингала для попереднього спостереження;

- збільшення значення степеневого мартингала для поточного спостереження по відношенню до попереднього еквівалентно виконанню

правила конформного аномального детектора зі значенням рівня аномальності, який дорівнює $\eta^{\frac{1}{1-\eta}}$, де η — параметр чутливості степеневого мартингалу $M_k^{(\eta)}$.

5. **Отримала подальший розвиток імітаційна модель** футбольного сезону, яка на відміну від існуючих враховує розбиття матчів на класи за контекстуальними атрибутами «сила команди» і «тип гри» — домашня або виїзна, що забезпечує моделювання договірних матчів з фіксованим результатом, які мають аномальний характер.

6. **Удосконалено метод кластеризації** на основі Гаусівських сумішей в частині регуляризації недіагональних елементів коваріаційних матриць, що дало змогу зменшити чутливість до початкових умов і отримувати кластери еліпсоподібної форми, які враховують неочевидні зв'язки між точками набору даних.

Практичне значення одержаних результатів полягає у тому, що:

1. Розроблена імітаційна модель футбольного сезону забезпечує подібність змодельованого сезону з реальним за типами результатів матчів як за всіма класами матчів, так і в цілому на рівні значущості 0,001 за критерієм Колмогорова-Смирнова.

2. Запропоновані методи виявлення на основі конформного аномального детектора, степеневого мартингалу й інтегрального мартингалу на модельних даних забезпечили підвищення ефективності виявлення підозрілих на фіксований результат футбольних матчів у порівнянні з відомим гістограмним методом на 3-13 % за метрикою точності, 11-30% — за метрикою повноти і 10-18% — за метрикою F_1.

3. Запропоновані методи на основі конформного аномального детектора, степеневого мартингалу й інтегрального мартингалу виявили 4 з 5 матчів сезону 2014–2015 рр. Серії *B* Італії, які вважаються договірними за інформацією від офіційних правоохоронних органів Італії.

Результати роботи впроваджено у навчальний процес кафедри прикладної математики Національного технічного університету України

«Київський політехнічний інститут імені Ігоря Сікорського» в рамках нормативної дисципліни «Машинне навчання».

Розроблені методи також напряму можуть бути використані для виявлення підозрілих на фіксований результат матчів у змаганнях з інших видів спорту, таких як: хокей, волейбол, бейсбол, баскетбол, кіберспорт тощо.

Більше того, за відповідного переформулювання і підбору адекватної міри неконфорності запропоновані в дисертаційному дослідженні методи можуть бути використані для пошуку широкого кола контекстних аномалій (нетипові транзакції по банківському рахунку, проникнення до закритої мережі, аномальна кількість повідомлень в соціальних мережах на певну тематику тощо).

Ключові слова: математичне моделювання, ймовірнісне моделювання, інтелектуальний аналіз даних, машинне навчання, самоконтрольоване навчання, глибинне навчання з підкріпленням, штучний інтелект, мультиагентна система, кластерний аналіз, функція відстані, найближчий сусід, класифікація даних, регресійна модель, факторний аналіз, метрика, регуляризація, міра неконформності, конформний аномальний детектор, мартингал, функція ранжування, спортивна подія.

ABSTRACT

Zhuk I. S. Mathematical models and methods for detecting football match-fixing using publicly available data. — Qualifying scientific work, the manuscript.

PhD thesis in the field of knowledge 11 Mathematics and Statistics in speciality 113 Applied Mathematics. — National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, 2023.

The aim of the thesis is to increase the efficiency of identifying suspects for fixed results of football matches based on the processing of exclusively publicly available public data based on the results of the football tournament season.

One of the most important problems of football, comparable to the problem of doping, is fixed matches. The results of such matches or a certain sequence of events in them are predetermined, that is, fixed. In fixed matches related to winning bets, the task is to obtain a result that is different from the expected one. Therefore, such results can be considered atypical, or abnormal.

Mathematical methods of football analytics are used to check the current match for a fixed result, such as prediction of the match result, and analysis of bets or actions of the match participants throughout the game. Their advantage is the promptness in decision-making, and their disadvantage is the need to use a very large amount of data, which, as a rule, is not publicly available. As alternative there can be considered an approach where the decision regarding the match fixedness is made based on the results of the whole season. At the same time, public information about the results of the games played by all teams is publicly available, which allows formalizing the search for matches suspicious of a fixed result as a task of detecting contextual anomalies.

Statistical non-parametric histogram methods are the most adequate for the considered task of identifying matches suspicious for a fixed result based on available public information. This is due to the fact that the input data are characterized by a small number of discrete numerical values and their probability

distribution laws are unknown. At the same time, the effectiveness of using these methods depends on the size of the sample. Along with this, the mathematical apparatus of conformal predictors and power martingales is promising, which does not require knowledge of data distribution laws, uses information about the degree of non-conformity of data and can be used to solve data classification problems.

Therefore, an urgent scientific task is the development of methods for detecting suspicious matches with a fixed result using the apparatus of conformal predictors and power martingales based on the processing of exclusively publicly available public data based on the results of the football tournament season.

In *the first section*, the problems of contractual football matches are considered and the relevance of research aimed at finding matches potentially suspicious of fixed results based on the processing of exclusively publicly available public data is substantiated. It is shown that the application of mathematical methods of football analytics, such as predicting the result of the match, analysis of bets or actions of the participants of the match during the entire game to identify matches suspicious for match-fixing requires a large amount of data that is not always available for analysis. It is noted that the task of identifying suspicious data regarding the fixed result of matches based on the results of the football tournament season refers to the detection of contextual anomalies, which is solved in the field of intelligent data analysis. An analysis of statistical methods, as well as machine learning methods used to detect anomalies, was carried out. Special attention is paid to classification methods based on conformal predictors, which are proposed to be used to increase the effectiveness of detecting potentially suspicious contractual matches. The goal and scientific tasks of the dissertation research are formulated.

The second section is devoted to the development of a simulation model of a soccer season with fixed-score matches. To determine the contextual attribute "team strength", the teams were grouped using the *K*-means and Gaussian mixtures methods based on the number of points and the difference between scored and conceded goals in one-dimensional and two-dimensional spaces. The initial data, limitations and formulas for calculating the probabilities of scoring goals by teams

during the game based on real data of the season for building a simulation model are determined. A simulation model of the football season was developed and its analysis was carried out by means of statistical modeling. An algorithm for modeling contractual matches related to earnings on bets has been developed.

The third section is devoted to the development of methods for detecting football matches suspicious for fixed results in the presence of data for the entire season. Measures of non-conformity of the current match are proposed. On the basis of the anomalous conformal detector, a method of detecting football matches suspicious for fixed results has been developed, in which decision-making takes place according to the threshold rule. A comparative analysis of methods for detecting matches suspicious for a fixed result based on an expertly defined deviation threshold and a conformal anomaly detector was conducted. With the use of power and integral martingales, methods have been developed for detecting football matches suspicious for fixed results, in which the decision is made when the value of the power martingale for the current observation increases in relation to the value of the same martingale for the previous observation.

In *the fourth section*, the analysis of the features of the developed methods based on the data of individual classes of the model season is considered by means of simulation modeling. A comparative analysis of the developed methods and the known histogram method was conducted based on model season data using precision (precision, P), completeness (recall, R) and F_1 measures. The proposed methods have also been used to detect matches considered to be fixed in the 2014–2015 Italian Serie *B* season.

The scientific novelty of the obtained results is as follows:

1. A **new method** of detecting football matches suspicious for fixed results has been developed, which differs from the known ones by the use of a conformal anomaly detector with a proposed measure of non-conformity of the current match, which provides the possibility of determining the decision threshold in accordance with the given value of the a priori probability of the appearance of anomalous data.

2. A **new method** of detecting football matches suspicious for fixed results has been developed, which differs from the known ones by using a power martingale and a decision-making rule based on a comparison of the current value of the power martingale with the previous one, which allows, by changing the sensitivity parameter, to adjust the power martingale to detect anomalies of the appropriate level.

3. A **new method** of detecting football matches suspicious for fixed results has been developed, which differs from the known ones by the use of an integral martingale and a decision rule based on the comparison of the current value of the integral martingale with the previous one, which makes it possible to detect anomalous matches without adjusting the parameters.

4. **New properties** of the power martingale are proved:

- for any small values of the *p*-value of the current observation, the value of the power martingale for the current observation is greater than the value of the same martingale for the previous observation;

- increasing the value of the power martingale for the current observation in relation to the previous one is equivalent to fulfilling the rule of the conformal anomaly detector with the value of anomaly threshold equals to $\eta^{\frac{1}{1-\eta}}$, where η is the sensitivity parameter of the power martingale $M_k^{(\eta)}$.

5. The simulation model of the football season **got further development**, which, unlike the existing ones, takes into account the grouping of matches into classes according to the contextual attributes "team strength" and "type of game" - home or away, which provides simulation of contractual matches with a fixed result, which are anomalous.

6. The method of clustering based on Gaussian mixtures **has been improved** in terms of regularization of off-diagonal elements of covariance matrices, which made it possible to reduce the sensitivity to initial conditions and to obtain clusters of an elliptical shape that take into account non-obvious connections between the points of the data set.

The practical significance of the obtained results is that:

1. The developed simulation model of the football season ensures the similarity of the simulated season with the real one in terms of types of match results both for all classes of matches and in general at the significance level of 0.001 according to the Kolmogorov-Smirnov test.

2. The proposed methods of detection based on the conformal anomaly detector, power martingale and integral martingale on model data ensured an increase in the effectiveness of detecting suspicious football matches with a fixed result compared to the known histogram method by 3-13% according to the accuracy metric, 11-30% — according to the completeness metric and 10-18% — according to the F_1 metric.

3. The proposed methods based on conformal anomaly detector, power martingale, and integral martingale detected 4 out of 5 matches of the 2014–2015 Serie *B* season in Italy, which are considered to be fixed according to information from official Italian law enforcement agencies.

The results of the work are implemented in the educational process of the Department of Applied Mathematics of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" within the framework of the normative discipline "Machine Learning".

The developed methods can also be directly used to identify matches suspicious for match-fixing in other sports competitions, such as: hockey, volleyball, baseball, basketball, e-sports, etc.

Moreover, with appropriate reformulation and selection of an adequate measure of non-conformity, the methods proposed in the dissertation research can be used to search for a wide range of contextual anomalies (atypical bank account transactions, penetration of a closed network, anomalous number of messages in social networks on a certain topic, etc.).

Keywords: mathematical modeling, probabilistic modeling, data mining, machine learning, self-supervised learning, deep reinforcement learning, artificial intelligence, multi-agent system, cluster analysis, distance function, nearest

neighbor, data classification, regression model, factor analysis, metric, regularization, nonconformity measure, conformal anomaly detector, martingale, ranking function, sports event.

Список публікацій здобувача

1. Zhuk, I., & Chertov, O. (2023). Framework based on conformal predictors and power martingales for detection of fixed football matches. *Eastern-European Journal of Enterprise Technologies*, *2*(4 (122), 6–15. https://doi.org/10.15587/1729-4061.2023.276977 [Scopus Q3].

2. Chertov, O., & Zhuk, I. (2023). Detection of fixed football matches based on the theory of conformal predictors using the modified Stepanets indicator function. *Eastern-European Journal of Enterprise Technologies*, *3*(4 (123), 22–32. https://doi.org/10.15587/1729-4061.2023.282645 [Scopus Q3].

3. Чертов, О. Р., & Жук, І. С. (2022). Імітаційна модель футбольного сезону з матчами з фіксованим результатом, *Наукові вісті КПІ, 1–2*, с. 82–94, 2022. https://doi.org/10.20535/kpisn.2022.1-2.287916

4. Chertov, O. R., & Zhuk, I. S. (2023). Clusterization of soccer season teams using K-means and GMM methods. *Intelligent Solutions-S: Proceedings of the International Symposium, September 28, 2023, Kyiv-Uzhorod, Ukraine / Ministry of Education and Science of Ukraine, Taras Shevchenko National University of Kyiv and* [etc]; Vitaliy Ye. Snytyuk (Editor) (pp. 14-15).

5. Жук, І. (2020). Застосування конформних предикторів і степеневих мартингалів для виявлення підозрілих матчів футбольних турнірів. *Прикладна математика та комп'ютинг. ПМК, 2022* : п’ятнадцята наук. конф. магістрантів та аспірантів, Київ, 16—18 лист. 2022 р. : зб. тез доп. / [редкол.: Дичка І. А. та ін.]. — К. : Просвіта, 2022. — 368 с. ISBN 978-617-7010-23-3 С.24 – 29.

6. Chertov, O., Zhuk, I., & Serdyuk, A. (2021). Search of the Deviation from

the Natural Process Using Stepanets Approach for Classification of Functions. *2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)* (pp. 720-724), doi: 10.1109/IDAACS53288.2021.9660997. [Scopus].

7. Жук, І. (2020). Виявлення зовнішнього впливу в інформаційних потоках мережі Internet як проблема ідентифікації образу. *Філософія і науково-технічна творчість у хронотопі технічного університету: Матеріали III Міжнародної науково-практичної конференції.* – 410 с. С.144 – 147

8. Жук, І. С., & Чертов, О. Р. (2020). Використання математичного апарату наближень Степанця для виявлення штучних втручань у сигналах різної природи. *Інтегровані інтелектуальні робототехнічні комплекси (ІІРТК-2020). Тринадцята міжнародна науково-практична конференція 19-20 травня 2020 р.*, Київ, Україна. – К.: НАУ, 2020. – 305 с. С.276 – 278.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ ТА ПОЗНАЧЕНЬ

ANN – Artificial Neural Network, штучна нейронна мережа

FIFA – Fédération Internationale de Football Association, Міжнародна федерація футбольних асоціацій

GMM – Gaussian Mixture Model, модель Гаусівських сумішей

HMM – Hidden Markov Model, прихована Марківська модель

IBIA – International Betting Integrity Association, Міжнародна асоціація із забезпечення чесності спортивних ставок

ICSS – International Centre for Sport Security, Міжнародний центр з безпеки спорту

KDE – Kernel Density Estimation, ядерна оцінка щільності розподілу

RIPPER – Repeated Incremental Pruning to Produce Error Reduction, повторюване нарощуване скорочення для зменшення помилки

SVM – Support Vector Machine, метод опорних векторів

UEFA – Union of European Football Associations, Союз європейських футбольних асоціацій

UNODC – United Nations Office on Drugs and Crime, Управління ООН з наркотиків та злочинності

WADA – World Anti-Doping Agency, Міжнародна антидопінгова агенція

НІР – незалежні і ідентично розподілені

ООН – Організація Об'єднаних Націй

США – Сполучені Штати Америки

ФФУ – Федерація футболу України

a_k – міра неконформності об'єкта z_k

$\Gamma^{\varepsilon}(z_1, \dots, z_i, \dots, z_k)$ – конформний предиктор на рівні значущості ε для об'єкта z_k

F_1 – метрика бінарної класифікації F_β при $\beta = 1$

$\mathbf{M}(X|Y)$ – умовне математичне сподівання випадкової величини X за умови відомої випадкової величини Y

$M_k^{(\eta)}$ – степеневий мартингал для об'єкта z_k

ε – рівень значущості (для конформного предиктора) або поріг аномальності (для конформного аномального детектора), $\varepsilon \in [0; 1]$

η – параметр чутливості степеневого мартингала, $\eta \in [0; 1]$

P – метрика точності бінарної класифікації

p_k – ступінь конформності (p-value) об'єкта z_k

R – метрика повноти бінарної класифікації

d(*t*) – кількість нічиїх команди *t* під час сезону

w(*t*) – кількість перемог команди *t* під час сезону

s(*t*) – загальні очки команди *t* під час сезону

gf(*t*) – кількість забитих голів командою *t* під час сезону

ga(*t*) – кількість пропущених голів командою *t* під час сезону

$gd(t)$ – різниця між *gf*(*t*) та *ga*(*t*)

c_k – центроїда кластера C_k

$J(i; K)$ – значення цільової функції задачі кластеризації з кількістю кластерів *K* на *i*-*й* ітерації

$\theta_n^{(i)}(k)$ – ступінь належності точки x_n до кластера з номером *k* на *i*-й ітерації

μ_k – вектор середніх значень *k*-ї Гаусіани

Σ_k – коваріаційна матриця *k*-ї Гаусіани

w_k – ваговий коефіцієнт *k*-ї Гаусіани

γ – коефіцієнт регуляризації коваріаційної матриці

p_A – ймовірність появи аномальної різниці голів, рівень аномальності

$avg(i, j)$ – середнє значення різниці голів в класі матчів (i, j)

$round(x)$ – округлене число x за арифметичними правилами до цілого значення

$z_i = (x_i, y_i)$– *i*-тий об'єкт вибірки *Z*, що складається з вектору ознак x_i та мітки класу y_i

ВСТУП

Актуальність роботи. Футбол — це величезна індустрія, порівнянна з традиційними економічними галузями. Однією з найважливіших проблем, з якою стикається ця галузь, є договірні матчі. За негативним ефектом такі явища співмірні з проблемою допінгу. Саме тому FIFA і ООН об'єднали свої зусилля щодо боротьби з договірними матчами у футболі. На сьогодні такі матчі кваліфікуються як кримінальний злочин.

Договірний матч характеризується тим, що його результат або певний перебіг подій є наперед визначеними, тобто фіксованими. У договірних матчах, пов'язаних з виграшом за ставками, завданням є отримання результату, відмінного від очікуваного. Тому такі результати можна розглядати як нетипові, аномальні.

Для перевірки поточного матчу на фіксований результат використовують математичні методи футбольної аналітики такі, як: прогнозування результату матчу, аналіз ставок або дій учасників матчу протягом всієї гри. Їх перевагою є оперативність прийняття рішень, а недоліком — необхідність використання дуже великої кількості даних, які, зазвичай, не є публічно доступними. Альтернативним може розглядатись підхід, коли рішення щодо фіксованості матчу приймається за результатами усього сезону. При цьому загально доступною є публічна інформація щодо результатів проведених ігор усіх команд, що дозволяє формалізувати пошук матчів, підозрілих на фіксований результат, як задачу виявлення контекстуальних аномалій. На сьогодні, не існує універсальних методів вирішення цієї задачі, а результати роботи існуючих методів суттєво залежать від вхідних параметрів, структури даних, природи їх походження. Більшість методів машинного навчання, що використовуються для пошуку аномалій в задачах інтелектуального аналізу даних, орієнтовані на роботу з неперервнозначними даними в багатомірних ознакових просторах. Також вони вимагають для навчання значних об'ємів вхідних даних.

Особливістю задачі виявлення підозрілих на фіксований результат матчів за результатами усього сезону є відсутність розмітки нормальних і аномальних класів даних, що обумовлює необхідність її розгляду як задачі неконтрольованого навчання. Найбільш адекватними розглянутій задачі виявлення підозрілих на фіксований результат матчів на основі доступної публічної інформації є статистичні непараметричні гістограмні методи. Це обумовлено тим, що вхідні дані характеризуються малою кількістю дискретних числових значень. Також закони розподілу ймовірностей вхідних значень є невідомими. Водночас, ефективність використання цих методів залежить від об’єму вибірки.

Новим перспективним напрямом пошуку аномалій у даних є математичний апарат конформних предикторів і мартингалів. Переваги даного математичного апарату полягають, з одного боку, у поєднанні процесу навчання і прогнозування в одну стадію, а з іншого боку — у незалежності від ймовірнісного розподілу, з якого генеруються дані. Також цей підхід дозволяє вводити оцінки гарантованої точності для отриманих рішень.

Тому актуальною науковою задачею слід вважати розробку методів виявлення підозрілих на фіксований результат матчів з використанням апарату конформних предикторів і степеневих мартингалів на базі обробки виключно загальнодоступних публічних даних.

Зв'язок роботи з науковими програмами, планами, темами. Робота виконувалася на кафедрі прикладної математики Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» Дослідження даної роботи проводились згідно з планами науково-дослідних робіт кафедри прикладної математики факультету прикладної математики:

- в рамках науково-дослідної роботи № 2310п «Інформаційно-аналітична система для математичного моделювання та управління соціальними ризиками з застосуванням у техніці та медицині» (номер державної реєстрації — 0120U102216);

- в рамках міжнародного наукового проєкту «Cyber Rapid Analysis for Defense Awareness of Real-time Situation» (CyRADARS) / «Оперативний аналіз кіберзагроз для володіння ситуацією в умовах реального часу» за програмою NATO SPS (номер проєкту: SPS G5286).

Мета та задачі дослідження. Метою роботи є підвищення ефективності виявлення підозрілих на фіксований результат футбольних матчів на базі обробки виключно загальнодоступних публічних даних за результатами сезону футбольного турніру.

Для досягнення вказаної мети у даній дисертаційній роботі вирішуються такі задачі:

1. Розробка процедури групування команд футбольного сезону.
2. Розробка імітаційної моделі футбольного сезону з матчами з фіксованим результатом.
3. Розробка методу виявлення підозрілих футбольних матчів з фіксованим результатом за допомогою конформного аномального детектора.
4. Розробка методу виявлення підозрілих футбольних матчів з фіксованим результатом за допомогою степеневого мартингалу.
5. Розробка методу виявлення підозрілих футбольних матчів з фіксованим результатом за допомогою інтегрального мартингалу.
6. Дослідження й оцінка ефективності розроблених методів виявлення підозрілих на фіксованість результату футбольних матчів на основі метрик класифікації, які є базовими в машинному навчанні на модельних та реальних сезонах.

Об’єктом дослідження є виявлення матчів, підозрілих щодо фіксованого результату в футбольних турнірах.

Предметом дослідження є моделі та методи виявлення потенційно підозрілих договірних матчів у футбольних турнірах.

Методи дослідження. Апаратом дослідження є математичний аналіз, теорія ймовірності і математична статистика, методи машинного навчання, теорія конформних предикторів, методи імітаційного моделювання.

Для оцінювання ефективності розроблених методів використовувались метрики точності (precision, P), повноти (recall, R) і міра F_1.

Наукова новизна отриманих результатів. Наукова новизна одержаних результатів полягає у такому:

1. Розроблено **новий метод** виявлення підозрілих щодо фіксованості результату футбольних матчів, який відрізняється від відомих застосуванням конформного аномального детектора із запропонованою мірою неконформності поточного матчу, що забезпечує можливість визначення порогу прийняття рішення у відповідності до заданого значення апріорної ймовірності появи аномальних даних.

2. Розроблено **новий метод** виявлення підозрілих щодо фіксованості результату футбольних матчів, який відрізняється від відомих застосуванням степеневого мартингалу і правилом прийняття рішення на основі порівняння поточного значення степеневого мартингалу з попереднім, що дозволяє за рахунок зміни параметра чутливості налаштовувати степеневий мартингал на виявлення аномалій відповідного рівня і знаходити їх.

3. Розроблено **новий метод** виявлення підозрілих щодо фіксованості результату футбольних матчів, який відрізняється від відомих застосуванням інтегрального мартингалу і правилом прийняття рішення на основі порівняння поточного значення інтегрального мартингалу з попереднім, що дає змогу виявляти аномальні матчі без налаштування параметрів.

4. Доведені **нові властивості** степеневого мартингалу:

- за яких завгодно малих значень ступеня конформності (p-value) поточного спостереження значення степеневого мартингала для поточного спостереження є більшим за значення цього ж мартингала для попереднього спостереження;

- збільшення значення степеневого мартингала для поточного спостереження по відношенню до попереднього еквівалентно виконанню

правила конформного аномального детектора зі значенням рівня аномальності, який дорівнює $\eta^{\frac{1}{1-\eta}}$, де η — параметр чутливості степеневого мартингалу $M_k^{(\eta)}$.

5. **Отримала подальший розвиток імітаційна модель** футбольного сезону, яка на відміну від існуючих враховує розбиття матчів на класи за контекстуальними атрибутами «сила команди» і «тип гри» — домашня або виїзна, що забезпечує моделювання договірних матчів з фіксованим результатом, які мають аномальний характер.

6. **Удосконалено метод кластеризації** на основі Гаусівських сумішей в частині регуляризації недіагональних елементів коваріаційних матриць, що дало змогу зменшити чутливість до початкових умов і отримувати кластери еліпсоподібної форми, які враховують неочевидні зв'язки між точками набору даних.

Практичне значення одержаних результатів полягає у тому, що:

1. Розроблена імітаційна модель футбольного сезону забезпечує подібність змодельованого сезону з реальним за типами результатів матчів як за всіма класами матчів, так і в цілому на рівні значущості 0,001 за критерієм Колмогорова-Смирнова.

2. Запропоновані методи виявлення на основі конформного аномального детектора, степеневого мартингалу й інтегрального мартингалу на модельних даних забезпечили підвищення ефективності виявлення підозрілих на фіксований результат футбольних матчів у порівнянні з відомим гістограмним методом на 3-13 % за метрикою точності, 11-30% — за метрикою повноти і 10-18% — за метрикою F_1.

3. Запропоновані методи на основі конформного аномального детектора, степеневого мартингалу і інтегрального мартингалу виявили 4 з 5 матчів сезону 2014–2015 рр. Серії *B* Італії, які вважаються договірними за інформацією від офіційних правоохоронних органів Італії.

Результати роботи впроваджено у навчальний процес кафедри прикладної математики Національного технічного університету України

«Київський політехнічний інститут імені Ігоря Сікорського» в рамках нормативної дисципліни «Машинне навчання».

Розроблені методи також напряму можуть бути використані для виявлення підозрілих на фіксований результат матчів у змаганнях з інших видів спорту, таких як: хокей, волейбол, бейсбол, баскетбол, кіберспорт тощо.

Більше того, за відповідного переформулювання і підбору адекватної міри неконформності запропоновані в дисертаційному дослідженні методи можуть бути використані для пошуку широкого кола контекстних аномалій (нетипові транзакції по банківському рахунку, проникнення до закритої мережі, аномальна кількість повідомлень в соціальних мережах на певну тематику тощо).

Особистий внесок здобувача. Усі основні результати дисертаційного дослідження, представлені до захисту, одержані автором особисто. У публікаціях у співавторстві, здобувачеві належать такі результати: в [1] розроблені методи виявлення підозрілих на фіксований результат футбольних матчів на основі конформного предиктора і степеневого мартингала, які застосовано до даних реального футбольного сезону 2013-2014 років Ліги 2 Франції; в [2] розроблено удосконалений метод виявлення підозрілих на фіксований результат футбольних матчів та застосовано запропоновані методи до даних реального футбольного сезону 2014-2015 років Серії *B* Італії; в [3] розроблено алгоритм формування імітаційної моделі футбольного сезону з договірними матчами.

Апробація результатів дисертації. Результати та основні положення роботи подавалися та обговорювалися на таких конференціях:

- III International Scientific Symposium «Intelligent Solutions» (2023), м. Київ, Україна [4];
- 2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (доповідь на цій конференції проіндексована в наукометричній базі Scopus) [5];

- «Прикладна математика та комп'ютинг. ПМК, 2022 : п’ятнадцята наук. конф. магістрантів та аспірантів», м. Київ, Україна [6];

- Філософія і науково-технічна творчість у хронотопі технічного університету (2021), м. Київ, Україна [7];

- Інтегровані інтелектуальні робототехнічні комплекси (ІІРТК-2020), м. Київ, Україна [8].

Публікації. За результатами досліджень опубліковано 8 наукових робіт, в тому числі **3 статті у наукових фахових виданнях України,** зокрема **2 статті** опубліковано в Східно-Європейському журналі передових технологій (Eastern-European Journal of Enterprise Technologies), який включено до списку міжнародної наукометричної бази **Scopus з квартилем Q3**.

Структура та обсяг дисертації. Дисертаційна робота складається із вступу, чотирьох розділів, висновків, списку використаних джерел із 166 найменувань. Загальний обсяг роботи складає 187 сторінок, з яких 164 сторінки основного тексту та 20 сторінок використаних джерел. Робота містить 42 рисунка, 21 таблицю.

РОЗДІЛ 1
АКТУАЛЬНІСТЬ ЗАДАЧІ ТА АНАЛІЗ ІСНУЮЧИХ РІШЕНЬ

1.1 Актуальність задачі виявлення футбольних матчів, підозрілих на фіксований результат

Договірні матчі, тобто матчі з фіксованим результатом (match-fixing), поруч з допінгом називають раковою хворобою спорту [9]. Для боротьби з допінгом в 1999 році була створена спеціалізована міжнародна організація WADA (World Anti-Doping Agency) та використовуються високотехнологічні тести. Ці тести дозволяють з майже абсолютною точністю виявляти наявність заборонених препаратів в організмі спортсмена.

На відміну від ситуації з допінгом, успіхи у боротьбі з договірними матчами значно скромніші. Хоча поєдинки з фіксованим результатом, мабуть, існують стільки, скільки й самі спортивні змагання. Принаймні, маємо історичні відомості, що такі ганебні вчинки мали місце ще на стародавніх Олімпійських іграх в Греції [10].

Європейська комісія дає означення фіксації результату як маніпулюванню спортивними результатами, яке охоплює домовленості про перебіг або результат спортивного змагання чи будь-якої його окремої події (наприклад, матчу, гонки) з метою отримання фінансової вигоди для себе чи для інших, та з метою повністю або частково усунути невизначеність, яка зазвичай пов’язана з результатами змагань [11].

Незважаючи на те, що досі немає єдиного авторитетного визначення договірних матчів, у своїй основній формі їх можна визначити як програш або гру до заздалегідь визначеного результату в спортивних матчах шляхом незаконного маніпулювання результатами на свою користь [12].

Договірний матч характеризується тим, що його результат і/або певний перебіг подій (призначення пенальті, отримання гравцем попередження чи вилучення тощо) є наперед визначеними, тобто фіксованими.

Проведене 2013 року розслідування Інтерполу та Європолу показало масштаби корупції у світовому футболі на той час[1]: під підозру слідчих тоді потрапило 680 ігор, 380 з них було організовано у Європі — у 13 чемпіонатах, єврокубках, відбіркових турнірах чемпіонатів світу та Європи, товариських матчах. Ще 300 — в Азії, Америці та Африці. Як виявилося, злочинний синдикат підкуповував команди і ставив на відомий результат у азіатських букмекерів. Тоді там приймали ставки, не звертаючи особливої уваги на розміри максимальних сум та особи гравців. На підготовку одного матчу з фіксованим результатом було задіяно до 50 фігурантів з 10 різних країн. Як заявив генеральний секретар Інтерполу Рональд Ноубл (Ronald K. Noble)[2], прибуток організаторів договірних матчів був порівнянний з прибутком компанії Coca-Cola, яка у 2012 році заробила 9 млрд доларів США.

З того часу ситуація з договірними матчами в спортивній галузі тільки погіршилася. За даними Міжнародної асоціації із забезпечення чесності спортивних ставок (IBIA), кількість підозрілої активності у ставках на спортивні події збільшилася на 14% у період з 2021 по 2022 рік [13]. Найбільше шахрайських дій було зафіксовано в тенісі (102) та футболі (67).

Французька спортивна агенція Spotradar, яка спеціалізується на моніторингу спортивних подій, у своєму щорічному звіті за 2022 рік [14] зазначила, що найбільш вразливою до договірних матчів є сфера футболу: за 2022 рік по всьому світу було виявлено 775 договірних футбольних матчів, що становить 64% від усіх матчів з фіксованим результатом за різними видами спорту. Водночас, перелік перших трьох видів спортивних змагань за кількістю договірних матчів за даними цієї організації залишається незмінним

[1] Europol (2013). Update—Results from the largest football match-fixing investigation in Europe. (n.d.). Europol. Retrieved 3 September 2023, from https://www.europol.europa.eu/media-press/newsroom/news/update-results-largest-football-match-fixing-investigation-in-europe

[2] RG NEWS. Interpol integrity in sport video report – match fixing: the ugly side of the beautiful game. Retrieved 3 September 2023, from https://www.responsiblegambling.eu/videos/interpol-integrity-in-sport-video-report-match-fixing-the-ugly-side-of-the-beautiful-game/

вже декілька років: перше місце займає футбол, друге — баскетбол, третє — волейбол. При цьому, у кожному з цих видів змагань зафіксовано збільшення кількості встановлених договірних матчів у порівнянні з 2021 роком. Скажімо, у футболі кількість договірних матчів збільшилася на 10,8% (з 695 до 775). Найбільшу кількість договірних спортивних змагань зафіксовано у Європі (630 матчів), Азії (240 матчів) та Південній Америці (225 матчів). До рейтингу країн з найбільшою кількістю договірних спортивних змагань увійшли: Бразилія (152), Росія (92), Чехія (56), Казахстан (43), Китай (41), Греція (40), Аргентина (39), Філіппіни (37), Польща (36), Таїланд (35).

Враховуючи наявні загрози для чесного проведення спортивних змагань, Управління ООН з наркотиків та злочинності (UNODC) і Міжнародний центр з безпеки спорту (ICSS) у 2015 р. уклали угоду, покликану зміцнити співпрацю щодо боротьби з організацією договірних матчів та іншими формами протиправного впливу на результати змагань[3]. В цьому документі стверджувалося, що протиправний вплив на результати спортивних змагань — це не просто порушення правил, а кримінальний злочин та причина підриву довіри.

У 2016 році цими ж організаціями спільно було підготовлено «Довідкове керівництво про передові методи розслідування договірних матчів»[4] як практичне керівництво для надання посадовим особам підтримки у виявленні та розслідуванні договірних матчів. У відповідності до цього керівництва, до появи договірних матчів призводять такі фактори: жадоба до особистої наживи, слабкі керівні структури в галузі спорту, легкодоступні світові букмекерські платформи, відкриті для широкого користування, недостатня увага правоохоронних органів до проблеми договірних матчів як загрози,

[3] DOHA: UN anti-crime agency unveils new partnership to tackle match-fixing, sports betting | UN News. (2015, April 15). Retrieved 3 September 2023, from https://news.un.org/en/story/2015/04/496032

[4] ICSS. Good Practices in the Investigation of Match Fixing. (n.d.). Retrieved 3 September 2023, from https://theicss.org/2018/12/31/good-practices-in-the-investigation-of-match-fixing/

а також використання спорту організованою злочинністю для досягнення своїх власних інтересів. Наголошується, що поєднання усіх цих факторів призводить до того, що загроза розповсюдження договірних матчів постійно зростає.

Українська асоціація футболу є членом Міжнародної футбольної асоціації (FIFA) із 1992 року. У 2015 році Україна підписала конвенцію Ради Європи щодо боротьби з договірними матчами. Метою цієї конвенції є запобігання, виявлення та покарання договірних результатів у спорті[5].

Також у 2015 році Верховна Рада України прийняла Закон про запобігання впливу корупційних правопорушень на результати офіційних спортивних змагань, який покликаний унеможливити зазначений вплив [15]. Закон передбачає штраф від 200 до 1000 неоподатковуваних мінімумів доходів, обмеження волі на строк один-три роки або позбавлення волі до трьох років зі спецконфіскацією за підкуп, примус або вступ у змову для досягнення результатів з метою отримання неправомірної вигоди для себе чи третьої особи. Закон стосується як спортсменів, так і осіб, які беруть участь у організації змагань, а також спортивних функціонерів.

Про актуальність проблеми договірних матчів для українського футболу свідчить заява голови Комітету ФФУ з етики та чесної гри Ігоря Кочетова у 2016 році, в якій він наголосив, що половина клубів Першої ліги могли бути учасниками договірних матчів[6]. Під підозру потрапило близько 100 футболістів, арбітри, тренери.

Напередодні фіналу Ліги Чемпіонів UEFA, який відбувся 26 травня 2018 року в Києві, в українському футболі спалахнув найбільший корупційний

[5] Chart of signatures and ratifications of Treaty 215. (n.d.). Retrieved 3 September 2023, from https://www.coe.int/en/web/conventions/full-list?module=signatures-by-treaty&treatynum=215

[6] Половину клубів Першої ліги України запідозрили в договірних матчах (5 грудня 2016). (n.d.). Retrieved 3 September 2023, from https://sport.nv.ua/ukr/football/vitse-prezident-ffu-zajaviv-shcho-polovina-klubiv-pershoji-ligi-pidozrjujutsja-u-dogovirnih-matchah-309028.html

скандал в історії[7]. 35 футбольних клубів із чотирьох ліг — а це 67% усіх професійних команд України — виявилися причетними до організації договірних матчів із свідомо фіксованим результатом. Заздалегідь знаючи результат гри, учасники злочинних груп робили ставки у букмекерських конторах, зареєстрованих у країнах Азії, та завдяки анонімності, непрозорості та безконтрольності тоталізаторів у цих країнах отримували до 5 млн доларів США щорічно.

Отже, задача виявлення договірних матчів є актуальною як у світі, так і в Україні. Причому найбільший інтерес представляють футбольні змагання як найпопулярніші в нашій країні та закордоном.

Типи договірних матчів виділяються за двома критеріями [16] (табл. 1.1): вид організації тих, хто стоїть за договірним матчем, та наявність залучення фінансів для взаєморозрахунків учасників змови. Серед видів організації виділяють внутрішню та комбіновану. До внутрішньої організації належать люди, які мають відношення безпосередньо до гри або організації та проведення цього матчу. Наприклад, це можуть бути гравці, директор клубу, тренер команди, яка бере участь у договірному матчі, а також інші співробітники спортивної сфери, задіяні у організації або проведенні матчу. Комбінована організація додатково включає в себе учасників, які є зовнішніми по відношенню до матчу або до спорту в цілому, але при цьому зацікавлені у результаті цього матчу. Якщо для організації договірного матчу було задіяно комбіновану організацію учасників із залученням фінансових взаєморозрахунків, такий матч відноситься до організаційних договірних матчів (тип I). Якщо в комбінованій організації зловмисники не використовують для взаєморозрахунків безпосередньо гроші, то такі договірні матчі відносять до типу II. Якщо ж організація є внутрішньою із задіянням фінансових потоків, то проведення таких договірних матчів призводить до

[7] Ігри без правил. Всі подробиці скандалу навколо договірних матчів в Україні (23 травня 2018). (n.d.). Retrieved 3 September 2023, from https://nv.ua/ukr/ukraine/events/ihri-bez-pravil-vsi-podrobitsi-skandalu-navkolo-dohovirnikh-matchiv-v-ukrajini-2471458.html

лише вирішення питання забезпечення додатковими засобами для існування учасників команди або організаторів матчу (тип III). Якщо організація є внутрішньою і при цьому фінанси не задіються, то проведення таких договірних матчів має на меті покращення показників успішності команди (тип IV).

Таблиця 1.1 – Класифікація договірних матчів [16]

		Залучення фінансів	
		Присутнє	Відсутнє
Вид мережі зловмисників	Комбінована	**I тип** Організаційні договірні матчі	**II тип** Договірні матчі, які базуються на взаємовідносинах учасників
	Внутрішня	**III тип** Договірні матчі, засновані на засобах для існування	**IV тип** Договірні матчі, які базуються на голах

Нам для подальшого буде зручніше об’єднати договірні матчі першого і третього типу в групу «договірні матчі з метою виграшу за ставками чи іншою фінансовою вигодою», а матчі другого та четвертого типу — в групу «договірні матчі, спрямовані на досягнення спортивного результату».

Договірні матчі першої об’єднаної групи можуть організовувати особи, які безпосередньо не приймають участі у спортивному змаганні, з метою отримання незаконного фінансового прибутку, поєднуючи легальні та нелегальні спортивні букмекерські платформи та розподіляючи частину прибутку серед тих, хто був безпосередньо пов'язаний із залученими до угоди особами-учасниками змагань.

Також, така форма договірного матчу може бути організована та контролюватися самими учасниками спортивного змагання, які роблять ставки самі або через підставну особу, що фактично діє від їхнього імені.

У договірних матчах, пов’язаних з виграшом за ставками, основним завданням є отримання результату матчу, відмінного від очікуваного, щоб максимально заробити на ставці. Чим неймовірнішим буде результат, тим потенційно більший прибуток можна отримати на ставці.

Угоди за договірними матчами також можуть бути пов'язані з іншими формами злочинної діяльності, включаючи відмивання коштів, торгівлю людьми, ухилення від сплати податків, загрози фізичної розправи та насильства, шахрайство, хабарництво та здирництво[16]. Вони також є джерелом фінансування організованих злочинних угруповань з метою здійснення інших, більш прибуткових/тяжких злочинів. Важливо зазначити, що саме цей тип договірних матчів пов'язаний із криміналом, незаконним збагаченням і викликає максимальну тривогу FIFA та ООН [8].

Договірні матчі другої об’єднаної групи, тобто спрямовані на досягнення спортивного результату, не пов'язані зі ставками та з меншою ймовірністю залучені до кримінальної діяльності. Однак, і в цьому випадку зазвичай існує щонайменше непряма фінансова вигода від угоди.

Мотиви для договірних матчів цього типу можуть бути різними. Наприклад, типовими ситуаціями є такі:

а) команди домовляються виграти по одному матчу в очних зустрічах, забезпечуючи собі по три очка — це більше, ніж за дві нічиї. При цьому не доводиться ризикувати, намагаючись здобути чотири або шість очок;

б) лідер пари, страхуючи себе від випадкової втрати очок і бажаючи поберегти гравців, обіцяє нічию на виїзді в обмін на гарантовану перемогу вдома. Аутсайдер з такою пропозицією погоджується, адже теоретично йому

[8] ICSS. Good Practices in the Investigation of Match Fixing. (n.d.). Retrieved 3 September 2023, from https://theicss.org/2018/12/31/good-practices-in-the-investigation-of-match-fixing/

важко було б набрати навіть одне очко. Така ситуація може мати місце у матчах суперників різного рівня;

в) команді, яка вже вирішила всі свої турнірні завдання, запропонують у цьому сезоні програти команді з нижчим рейтингом, з якою буде проведений матч, щоб та не опустилася до нижчої ліги, взамін гарантувавши перемогу в наступному сезоні. Ця типова ситуація може відбуватися наприкінці чемпіонату.

Договірні матчі другої об'єднаної групи, крім того, можуть організовуватися для підвищення позиції команди в турнірній таблиці.

Типовою ситуацією є також покупка результату через хабар керівництву клубу, команді, кільком гравцям (а часом — лише воротареві) або арбітру поєдинку. Матч «Металіст» – «Карпати» (2008 рік), визнаний Федерацією футболу України договірним[9], був організований саме за такою схемою: гравці львівської команди отримали 110 тис. доларів США, забезпечивши харків'янам перемогу 4:0 і відповідне підвищення в турнірній таблиці.

Разом з тим слід зазначити, що особи, які знають про факт договірного матчу, пов'язаного зі спортивною мотивацією, можуть використовувати цю «внутрішню інформацію» на біржах ставок з метою отримання прибутку [16].

Загалом, виявлення договірних матчів, спрямованих на досягнення спортивного результату, є складнішим завданням, оскільки за відсутності доказів існування ставок набагато складніше довести сам факт шахрайства. Тут у сторін, що домовляються, є завдання максимально приховати договірний матч під звичайний і при цьому вирішити свою проблему.

Іншим наслідком договірних матчів такого типу є те, що їхні учасники можуть згодом стати жертвами шантажу з боку інших зловмисників, які дізнаються про домовленості договірного матчу [16].

[9] Як Металіст і Карпати покарали за договірний матч, зіграний у 2008 році (16 Червня 2020). (n.d.). Retrieved 3 September 2023, from https://sport.ua/uk/news/491446-kak-metallist-i-karpaty-nakazali-za-dogovornoy-match-sygranniy-v-2008-godu

Припустимо, що домовленість на конкретний підсумковий результат матчу чи певні його характеристики (кількість забитих м’ячів, різниця м’ячів тощо) є метою отримання неправомірної вигоди на ставках у букмекерських компаніях. В такому випадку, це потенційно можливо відслідкувати через фінансові транзакції або неприродний розподіл обсягів ставок на відповідні результати матчу [17, 18]. У випадку, якщо угоди по договірних матчах укладаються через офіційні букмекерські контори або ж у країнах, де існує жорстке та ефективне регулювання, слідчим значно простіше отримати інформацію та докази про такі угоди. Це пояснюється тим, що авторизовані букмекерські оператори повинні відповідати встановленим стандартам з погляду прозорості та цілісності ставок.

Однак, дані про кількість, розмір та час відповідних ставок є непублічною інформацією, а відомі методи розкриття прихованої інформації [19] тут малозастосовні.

Проте, навіть наявність даних від букмекерських компаній нічим не допоможе в певних ситуаціях. Таке можливо, якщо є кулуарна домовленість про програш однієї команди іншій з певною фінансовою чи іншою компенсацією через незалежні канали [10]. Також це можливо, коли є матеріальна стимуляція третьою стороною однієї команди проти іншої. Такі факти встановити напевне майже нереально, принаймні, без проведення спеціальної поліцейської операції. Справа в тому, що сама природа спортивних змагань гарантує, що не завжди перемагає сильніша команда, і цьому можна навести багато відповідних прикладів у будь-якому виді спорту. Тому в літературі [10, 17, 18] говорять лише про те, що результат певних матчів є підозрілим, нетиповим, алогічним, аномальним тощо. Але й така інформація є корисною та важливою, оскільки виступає додатковим індикатором того, що певна команда може мати стосунок до порушення спортивних принципів чесної гри.

В світі кожного дня проводяться сотні футбольних матчів, існує сотні чемпіонатів в різних країнах, тому виникає гостра потреба в автоматизації

процесу виявлення підозрілих щодо фіксованого результату матчів по великому обсягу даних.

Наразі майже всі відомі [17, 18] випадки виявлення договірних матчів пов'язані з використанням даних від букмекерів. Однак не на всі матчі в спортивних турнірах приймаються ставки, а якщо й приймаються, то інформація про них є комерційною таємницею, або ставки взагалі можуть прийматися на нелегальних платформах спортивних ставок (illegal sports betting platforms). Тому ***актуальними*** слід вважати дослідження, спрямовані на пошук потенційно підозрілих результатів матчів спортивних змагань на базі обробки виключно загальнодоступних публічних даних.

1.2 Аналіз існуючих рішень щодо виявлення матчів, підозрілих на фіксований результат

Проблематика дослідження договірних матчів є багатоаспектною і включає в себе пошук причин їх виникнення та поширення, врахування національних особливостей, економічне підґрунтя, юридичні питання, зв'язок з кримінальним світом, аналіз особливостей в різних видах спорту тощо [9, 10]. В останнє десятиріччя до обговорення цих питань активно долучилися й соціологи [20, 21]. Змагання, де гра не вважається автентичною, не можуть очікувати, що вони будуть продовжувати отримувати громадську підтримку [22]. Сама мережа учасників, втягнутих до сфери організації договірних матчів, може мати різну структуру: від горизонтальної між слабко зв'язаними тимчасовими учасниками (як в Koriopolis, коли щонайменше 40 матчів грецького чемпіонату з футболу в сезоні 2009-10 років були визнані договірними [23]) до строгої вертикальної ієрархії (як в «Şike Davası» скандалі під час сезону 2010-2011 років в турецькому футболі, де легальний бізнес футбольного клубу діяв за нелегальною схемою [24]).

Окремою категорією досліджень у сфері спортивних, зокрема футбольних, змагань є математичні методи виявлення матчів з фіксованим

результатом. Пошук таких аномалій у спортивних змаганнях, принаймні, у сфері футболу наразі зводиться до ***прогнозування результату матчу, аналізу ставок або дій учасників матчу протягом всієї гри*** [17, 22, 25–28] (рис. 1.1).

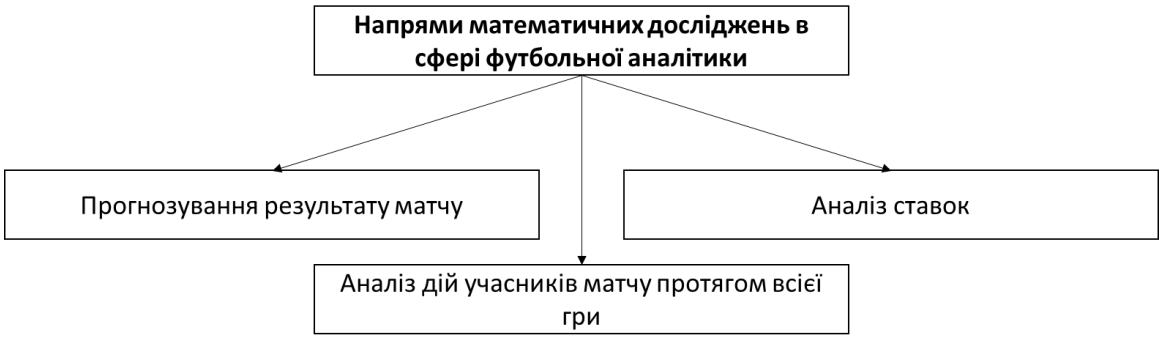


Рисунок 1.1 – Напрямки математичних досліджень у сфері футбольної аналітики

Прогнозування результату футбольного матчу вже досить давно є популярною дослідницькою проблемою. Методи отримання рішення для цієї проблеми включають статистичні методи [25] (зокрема, Байєсівські мережі) та методи машинного навчання [26]. В роботі [25] було запропоновано Байєсівську мережу для прогнозування результату матчу в термінах нічиєї або виграшу місцевої або гостьової команди, точність якої досягала 75 % і була на момент публікації результатів однією з найвищих при розв’язанні задач подібним способом. Для прогнозування результату модель використовує 18 атрибутів для кожного матчу. Перевагою цього методу є можливість чисельного ймовірнісного визначення впливу обраних атрибутів на результат матчу, що дає змогу сформулювати ймовірнісні правила визначення цього результату. Недоліком даного методу є необхідність використання суттєвого обсягу додаткових атрибутів про кожен матч для більш точного прогнозування результату.

У роботі [28] прогнозувався не рахунок футбольного матчу, а тільки результат — тобто яка команда переможе чи буде зафіксований нічийний результат. Для цього було використано два різних метода машинного

навчання: штучні нейронні мережі та дерева прийняття рішень. Результати матчів прогнозувались на основі загальної сили команд та результатів ігор, зіграних між двома командами у минулих сезонах. Для навчання і тестування вказаних вище методів було використано дані з трьох футбольних ліг: турецька суперліга, англійська прем'єр-ліга та бразильська Серія А.

У роботі [29] для прогнозування результатів футбольних матчів було використано чотири різних метода інтелектуального аналізу даних: дерева прийняття рішень, наївний Байєсівський класифікатор, дерева градієнтного бустінгу та випадковий ліс. Результати матчів прогнозувались у форматах «перемога-нічия-поразка» та кількості отриманих очок. Результати прогнозувались на основі як властивостей команд, так і показників успішності команд протягом попередніх матчів. Дані були зібрані з 10 сезонів 2007-2017 років турецької суперліги. Використанням вказаних моделей була досягнута акуратність (accuracy) прогнозування 74 % для результатів у форматі «перемога-нічия-поразка», і 86 % — по кількості набраних очок.

У роботі [26] запропонована нейронна мережа глибинного навчання для передбачення переможця футбольного матчу. Для прогнозу результату поточного матчу як вхідні дані використовується інформація про останні 35 матчів кожної з команд-учасниць матчу. З цієї інформації для кожного матчу розглядаються значення дев’яти атрибутів. По кожному з атрибутів на основі 35 останніх матчів формується шість статистичних індикаторів. Таким чином, для команди-учасниці матчу формується вектор з 54 координатами. Об’єднаний вектор з 108 координатами є вхідним набором даних для обраної нейронної мережі. Результатом роботи нейронної мережі є три значення з діапазону [0; 1], які в сумі дають одиницю й інтерпретуються як ймовірності того, що матч закінчиться нічиєю або виграшом однієї з команд. Використанням вказаної моделі була досягнута точність прогнозування 62 %.

В [30] для прогнозування результатів футбольних матчів реалізовано три алгоритми машинного навчання: ліс дерев прийняття рішень (Decision Forest), штучна нейронна мережа (Artificial Neural Network, ANN) та метод опорних

векторів (Support Vector Machine). Найкращий результат вийшов у алгоритму лісу дерев прийняття рішень, який досяг точності 89%, ANN досягла точності 72%, а SVM — 70%.

У [31] з використанням нейронних мереж були спрогнозовані результати сезону 2015/2016 років німецьких команд, що входять до «Бундесліги» — головної футбольної ліги Німеччини. Було виявлено чинники, що найбільше впливають на успішність виступу команди в сезоні: середній ріст гравців команди, кількість очок, набраних командою в попередньому сезоні, середній вік гравців.

В [32] наведено опис розробки нейромережевої системи для прогнозування результатів сезону італійської футбольної ліги «Серія А». Для підбору початкової множини були використані тематичні сайти, що містять повну статистику за необхідними характеристиками. Система заснована на вартісних характеристиках: має 12 вхідних параметрів, більшість з яких описують вартість окремих гравців та клубу в цілому, та один вихідний параметр, який показує місце команди в наступному сезоні. Система дозволяє виконувати оцінку виступу футбольної команди в сезоні в рамках ранжування від 1 до 5 позиції, де 1 — це 1-4 місця та вихід до Ліги чемпіонів, а 5 — команда залишає лігу. Визначено рівень значущості впливу кожного з вхідних параметрів на результат нейронної мережі. Досліджено вплив вхідних параметрів на підсумкове розташування команди в сезоні.

Нейронні мережі також широко використовуються для прогнозування результатів в інших видах спорту: бейсболі [33, 34], легкій атлетиці [35], волейболі [36] тощо.

Але запропоновані в [25-26, 29-32] підходи є малоприйнятними для виявлення договірних матчів, пов’язаних з виграшом за ставками, бо в таких матчах саме рахунок зазвичай є вирішальним, а не просто результат матчу. Недоліками вказаних моделей також є використання значної кількості додаткових атрибутів матчу, які не завжди є доступними, та відсутність можливості отримання аналітичних закономірностей для прогнозування

результату. Ці недоліки зумовлюють необхідність проведення досліджень у напрямку розробки алгоритмів, які використовують більш доступні атрибути матчу і дозволяють побудувати відповідні аналітичні закономірності для прогнозування не тільки результату, але й рахунку матча.

Для виявлення матчів з фіксованим результатом використовуються ***методи на основі аналізу ставок*** [22, 37]. У світі спортивних ставок склалися певні природні закономірності. Тому завдання пошуку договірного матчу зводиться до знаходження тих сигналів, які суперечать таким закономірностям. Наприклад, на матч за участю будь-якої команди проставляється приблизно однакова сума грошей. Але якщо обсяг ставок у кілька разів перевищує звичайний, це потужний індикатор того, що матч потрапляє до категорії договірних. Також оскільки зазвичай більшість гравців ставить на фаворита матчу, то і грошей на фавориті завжди значно більше, ніж на аутсайдері. Але якщо виявляється, що грошей на аутсайдера поставлено більше, ніж на фаворита, то це також індикатор того, що матч потрапляє до категорії договірних.

У [22] автори використовували спеціальні моделі для прогнозування обсягу ставок і порівняння реального обсягу ставок із прогнозованим за допомогою статистичних методів. Обсяги цих ставок порівнювалися в той час, коли йде матч. Моделі, які використовувалися для прогнозування, базуються на таких атрибутах як плин часу, присудження червоних карток і поточний рахунок. Якщо під час матчу різниця між реальним обсягом ставок і прогнозованим є статистично значущою, матч вважається фіксованим. У [37] введено модель з цілями та бронюваннями. Ця модель включає в себе тип команди (вдома, на виїзді), події, пов’язані із забиттям голу і отриманням гравцями жовтої або червоної картки. Модель розроблено на основі моделі подійного процесу Вейбулла. Ці моделі засновані на ймовірностях і спрямовані на прогнозування результату матчу за поставленими ставками. Однак інформація щодо ставок часто не є загально доступною і не всі договірні матчі пов’язані зі ставками.

Значного розвитку набули ***методи аналізу продуктивності гравця або команди в грі*** [27, 38-40]. Ці дослідження використовували дані про позиції гравців під час матчу, щоб оцінити їх результативність. Серед усіх командних видів спорту футбол посідає перше місце за складністю траєкторій руху гравців [40]. За кількістю гравців на полі, залучених до колективного руху, футбол значно перевершує хокей та баскетбол. З точки зору математичного моделювання це означає, що футбол має більше ступенів свободи, ніж інші види спорту, роблячи досить складним можливість оцінювання ігрової ситуації, використовуючи одну або невелику кількість метрик.

Під час футбольних матчів накопичується певна інформація, яку можна поділити на три типи даних [40]:

— статистика матчу — інформація верхнього рівня про матч (склади, розстановка гравців на полі, заміни, картки, голи тощо);

— дані про події — впорядкований набір даних, що описує послідовність дій гравців з м'ячом (паси, обведення, підкати, удари тощо);

— трекінгові дані — точні просторові координати всіх гравців та м'яча у кожний момент часу матчу.

Перелічені типи даних мають різний ступень доступності. Якщо загальна статистика матчу відома майже для всіх професійних та напівпрофесійних матчів по всьому світу, то трекінгові дані найвищого рівня деталізації сьогодні доступні лише для обмеженої кількості команд, в основному з найсильніших європейських футбольних ліг [40].

У роботах [27, 38] розглянуто питання відстеження просторово-часових даних гравця або команди під час гри, представлених у вигляді координат усіх гравців на полі та м'яча, впроваджено аналіз діаграми домінантної області з використанням діаграм Вороного, а в роботі [39] розглянуто аналіз траєкторій гравців з можливостями внутрішньокомандної взаємодії. У роботах [41, 42] розглянуто моделі, що враховують просторове положення тіла на основі аналізу відеозаписів.

Загалом, трекінгові дані можуть надавати значно більше інформації: можливість перехоплення пасів, розташування гравців протидіючих команд, ступінь контролю простору, швидкість гравців [40]. Наприклад, використовуючи трекінгові дані, модель в роботі [41] забезпечує комплексне представлення різних станів футбольного матчу. Модель поєднує ймовірність голу з конкретної точки поля, ймовірність команди контролювати цю точку та ймовірність того, що м'яч може бути доставлений у цю точку, надаючи на виході оцінку очікуваної довгострокової значущості від володіння м'ячом. Подальший розвиток цей підхід отримав в роботі [42].

Розглянуті методи спрямовані на оцінку продуктивності гравця та команди на основі дій гравця під час гри та його позиціонування. В рамках вказаних досліджень розглядаються траєкторії руху гравців протягом матчу. Для оцінки якості роботи гравця порівнюються траєкторії руху цього гравця протягом різних матчів, що починаються ним на одній і тій же ігровій позиції. На основі такого порівняння можна зробити припущення про фіксованість результату матчу, виходячи з істотної різниці в «роботі» гравця в цьому матчі й в інших матчах.

Основним недоліком розглянутих методів прогнозування результату матчу, аналізу ставок або дій учасників матчу протягом всієї гри є необхідність використання великої кількості даних, які не є публічно доступними, що в деяких ситуаціях суттєво зменшує практичну цінність цих підходів.

У договірних матчах, пов’язаних з виграшом за ставками, основним завданням є отримання результату матчу, відмінного від очікуваного, щоб максимально заробити на ставці. Тому ключовою характеристикою матчу з фіксованим результатом є ***різниця м’ячів***. Але нетиповість результату матчу залежить від різних факторів або контексту. Так виграш команди аутсайдера на виїзді з різницею в один м’яч розглядається як нетиповий аномальний результат. В той же час такий результат є цілком нормальним при зустрічі близьких за силою команд. Проте реальна сила команд стає зрозумілою лише по завершенню всього турніру.

Базовим показником сили команди є ***кількість набраних очок*** за результатами усього сезону. Також атрибут «сила команди» може визначатись за кількома показниками, наприклад: за кількістю набраних очок та різницею між забитими і пропущеними м’ячами команди в сезоні тощо [43].

За своєю силою команди можна поділити на ***групи*** близьких команд. На рис. 1.2, як приклад, наведено евристичне розбиття команд на чотири групи за кількістю набраних очок на прикладі сезону 2013-–2014 років Ліги II Франції: команди №№1–4 належать до кластеру №1, команди №№5–9 належать до кластеру №2, команди №№10–16 належать до кластеру №3, команди №№17–20 належать до кластеру №4. Таким чином, чим менший номер кластеру, тим вище місце в підсумковій таблиці зайняли команди цього кластеру. Тому логічно номер кластеру назвати рангом команд, що входять до нього. Для більш коректного (строгого) розбиття команд за силою на групи доцільно використовувати методи кластеризації.

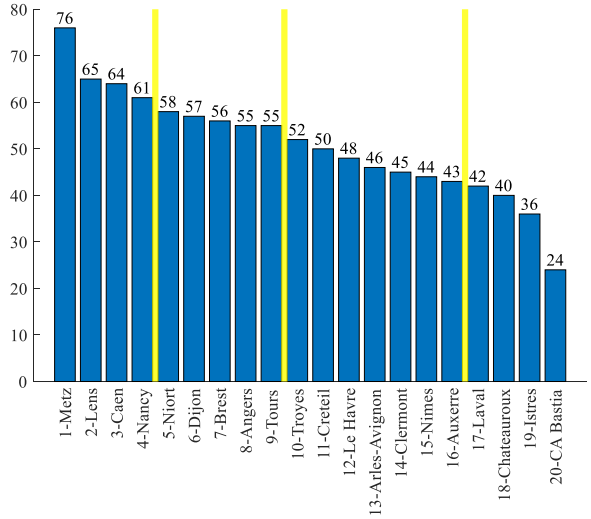


Рисунок 1.2 – Успішність команд сезону 2013-–2014 років Ліги II Франції

Проте наразі відсутні роботи, в яких на основі групування команд за їх силою всі матчі турніру ділилися б на певні класи, уже в межах яких би визначалася аномальність конкретного матчу.

1.3. Аналіз методів інтелектуального аналізу даних для пошуку аномалій. Використання контрольованого та неконтрольованого навчання, самоконтрольованого навчання, глибинного навчання з підкріпленням та мультиагентних систем

Як відмічалося в підрозділі 1.1, договірні матчі, пов’язані з виграшом на ставках, мають завдання отримати результат матчу, відмінний від очікуваного, щоб максимально заробити на ставці. Таку ж ознаку мають договірні матчі, спрямовані на досягнення спортивного результату, пов’язані з підкупом сильнішої команди слабкішою. В цьому випадку можна стверджувати, що результат певних матчів є аномальним. Тому доцільно провести аналіз методів пошуку аномалій у даних.

Аномальна поведінка процесів і об'єктів, що спостерігаються, може розглядатися як дивна, підозріла, незвичайна або несподівана і зустрічається в різних предметних областях. Так, як виявлення аномалій (Anomaly Detection) можна розглядати виявлення підозрілих банківських операцій [44-47], відмов при роботі різних датчиків та систем, подій у сенсорних мережах [48-49], а також у медичній діагностиці [50-56], у аналізі потокових даних [57-59], зокрема, з соціальних мереж [60-63] тощо.

Виявлення аномалій використовується при попередній обробці для видалення їх з набору даних, що часто призводить до статистично значущого підвищення точності. Своєчасне виявлення цих рідкісних подій має вирішальне значення для забезпечення можливості вжиття відповідних рішень та заходів і є актуальним напрямком сучасних наукових досліджень. Особливу роль завдання виявлення аномалій має в інтелектуальному аналізі даних (Data mining) [64-68], де використовується щодо виділення нової значимої інформації з великого обсягу даних, зокрема, при виявленні вторгнень у комп’ютерні мережі [64], аналізі зображень та різних одновимірних і багатовимірних даних [69-72].

Виявлення аномалій можна розглядати як завдання бінарної класифікації, де кожна точка даних належить або нормальному класу, або аномальному (ненормальному). Тому для виявлення аномальної поведінки можуть бути використані ***сигнатурні методи*** [64]. Вони припускають, що конкретні моделі незвичайної поведінки, такі як правила та шаблони, можуть бути визначені апріорі та можуть використовуватися для автоматичного розпізнавання образів у нових даних.

Однак самих сигнатурних методів зазвичай недостатньо, оскільки часто не вдається побудувати адекватні моделі нової, незвичайної поведінки. Це може бути пов'язане з відсутністю експертних знань та анотованих практичних даних. У таких випадках застосовуються методи, які дістали назву ***методи виявлення аномалій***, новизни або викидів [65]. Вони спрямовані на виявлення «дивних» та ненормальних фрагментів поведінки, які відхиляються від очікуваних чи «нормальних» моделей. Ці методи базуються на тому, що для отримання моделей, які описують нормальну поведінку, є зазвичай достатня кількість апріорних даних. Тому ці методи вимагають значно менше знань про аномальну поведінку.

У математичній статистиці методи виявлення викидів (Outlier Detection) відомі давно та широко використовуються. Викид визначається як спостереження, яке так сильно відрізняється від інших спостережень, що спричиняє підозру, що воно було створено іншим механізмом формування даних [73]. Зазвичай передбачається, що цей механізм визначається стаціонарним розподілом ймовірностей. Отже, виявлення викидів по суті включає визначення того, чи згенеровано конкретне спостереження відповідно до того ж розподілу, що й інші спостереження. Традиційно виявлення викидів у статистиці використовують для очищення наборів даних перед побудовою статистичних моделей; викиди вважаються шумом та видаляються для покращення якості статистичних моделей [74].

У галузі інтелектуального аналізу даних аномалії — це закономірності в даних, які не відповідають чітко визначеному поняттю нормальної

поведінки [65]. Поняття нормальної поведінки визначається моделлю нормальності, яка будується з урахуванням навчальних даних. Відповідно підходи до виявлення аномалій базуються на побудові моделі нормальних даних, а потім спробах виявлення відхилення від нормальної моделі в даних, що спостерігаються. На відміну від традиційних статистичних застосувань, при інтелектуальному аналізі даних виявлення аномальних спостережень розглядається як виявлення новизни (Novelty Detection), оскільки вони можуть відповідати цікавим та важливим подіям. Виявлення договірних матчів також відноситься до виявлення нових важливих закономірностей в отриманих даних, тому представляє інтерес аналіз методів виявлення аномалій у задачах інтелектуального аналізу даних.

Аномалії даних можуть бути віднесені до одного з трьох основних типів [65]: точкові, контекстуальні та колективні (рис. 1.3).

Найбільшого поширення набули точкові аномалії, для яких окремий екземпляр даних можна розглядати як аномальний стосовно інших даних. Даний вид аномалій найбільш легко розпізнається і більшість існуючих методів створено для розпізнавання саме точкових аномалій [45-48, 52, 53, 56, 57, 75-82].

Рисунок 1.3 – Типи аномалій даних

Контекстуальні аномалії спостерігаються, якщо екземпляр даних є аномальним лише у певному контексті [82-86]. Цей вид аномалій також називається умовним. Для визначення аномалій цього типу основним є

виділення контекстуальних і поведінкових атрибутів. Контекстуальні атрибути використовуються для визначення контексту (або оточення) кожного екземпляра. Аномальна поведінка визначається у вигляді значень поведінкових атрибутів, виходячи з конкретного контексту. Таким чином, екземпляр даних може бути контекстуальною аномалією за даних умов, але за таких же поведінкових атрибутів вважатися нормальним в іншому контексті.

Колективні аномалії виникають, коли послідовність зв'язаних екземплярів даних (наприклад, ділянка тимчасового ряду) є аномальною по відношенню до цілого набору даних [87, 88]. Окремий екземпляр даних у такій послідовності може не бути відхиленням, однак спільна поява таких екземплярів є колективною аномалією.

Пошук матчів, підозрілих на фіксований результат, можна розглядати як виявлення контекстуальних аномалій. Дійсно, як поведінковий атрибут матчу з фіксованим результатом можна використовувати різницю м’ячів, а за контекстуальні атрибути взяти силу команд та тип гри — виїзна або домашня. Таким чином, гра може бути контекстуальною аномалією за одних значень різниці голів, але при такому ж значенні цієї різниці (поведінкового атрибуту) вважатися нормальною в іншому контексті. У подальшому зосередимося на аналізі методів машинного навчання та інтелектуального аналізу даних, які застосовуються чи можуть бути застосовані для пошуку контекстуальних аномалій.

За апріорною інформацією методи виявлення аномалій класично поділяються на три класи [65], до яких варто додати новітні підходи на базі самоконтрольованого навчання (self-supervised learning) [89-94] (рис. 1.4):

1. Supervised anomaly detection (контрольоване виявлення аномалій) [50, 95-108] — методи, для роботи яких обов’язкова наявність навчальної вибірки, де аномальні дані мають відповідні мітки. Справжній клас точки даних зазвичай визначається експертом у предметній області з урахуванням його досвіду того, що є нормальним, а що — ненормальним. Методи реалізуються у два етапи: навчання та виявлення. На першому етапі будується модель,

з якою згодом порівнюються екземпляри, що не мають мітки. Найчастіше передбачається, що дані змінюють свої статистичні характеристики, інакше виникає необхідність змінювати класифікатор.

2. Semi-supervised anomaly detection (напівконтрольоване виявлення аномалій) [109-114] — на вхід подається вибірка, що складається лише з нормальних значень, без будь-яких відхилень. Основна ідея полягає в тому, що аномалії виявлятимуться на наступних етапах як результат відхилення від значень, що належать початковій вибірці.

3. Unsupervised anomaly detection (неконтрольоване виявлення аномалій) [46, 49, 55, 63, 75, 115-119] — випадок, коли відсутні апріорні дані, і методу, що застосовується, необхідно самостійно визначити, які дані є аномаліями. Зазвичай, методи розпізнавання при неконтрольованому навчанні базуються на припущенні про те, що аномальні екземпляри зустрічаються набагато рідше за нормальні. Ідея полягає у виявленні аномалій на основі внутрішніх властивостей даних. В більшості випадків використовується параметр distance («відстань») для ухвалення рішення, є це аномальним чи ні. Дані, визначені в результаті обробки як найбільш віддалені, ідентифікуються як аномалії.

4. Self-supervised anomaly detection (самоконтрольоване виявлення аномалій) [89-94, 120, 121] — методи, які працюють з вибіркою нерозмічених об’єктів і використовують її для вивчення властивостей нормальних об’єктів. Це вивчення відбувається за допомогою алгоритмів контрольованого машинного навчання, які застосовуються для вирішення визначених задач на окремих об’єктах вибірки або на парах цих об’єктів. Такі завдання спеціально налаштовані таким чином, щоб допомогти обраній моделі контрольованого навчання вивчити представлення об’єкта вибірки, яке створюється спеціально для задачі пошуку аномалій і має відмінності від представлення цього ж об’єкта для загальної задачі самоконтрольованого навчання.

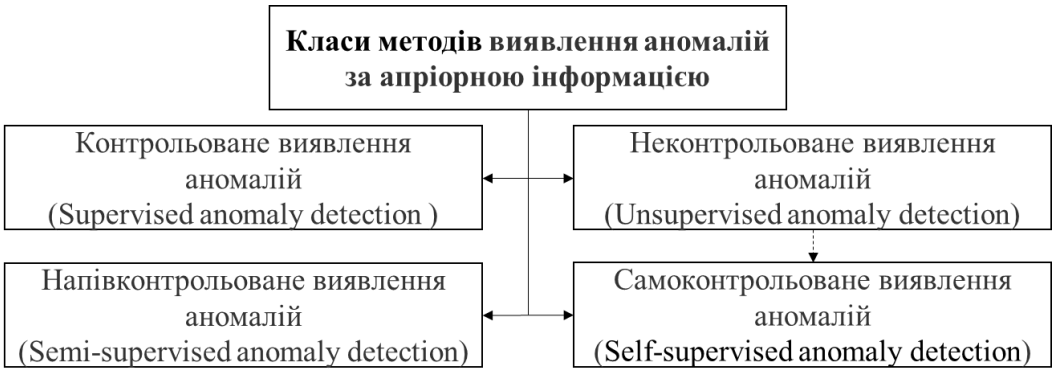


Рисунок 1.4 – Класи методів виявлення аномалій за апріорною інформацією

На сьогодні не існує універсальних методів виявлення аномалій, а результати роботи існуючих методів суттєво залежать від вхідних параметрів, структури даних, природи їх походження.

В області інтелектуального аналізу даних можна виділити такі основні групи методів виявлення аномалій [65] (рис. 1.5):

— статистичні методи (параметричні та непараметричні);

— методи, що ґрунтуються на класифікації;

— методи на основі глибинного навчання з підкріпленням;

— методи на базі мультиагентних систем;

— методи на основі найближчих сусідів (метричні методи класифікації);

— методи на основі кластеризації.

На рис. 1.5 штрихованими лініями на діаграмі показано зв’язки між окремими групами методів: метод, що належить групі, від якої проведено штриховану стрілку, може бути складовою частиною методу, що належить групі, до якої проведено таку стрілку. Наприклад, методи навчання з підкріпленням в якості складових частин можуть використовувати методи класифікації та методи глибинного навчання.

Статистичні методи виявлення аномалій можна в цілому розділити на параметричні та непараметричні [69]. Параметричні методи припускають, що модель процесу відома з точністю до набору параметрів. Найбільшого

поширення набули параметричні моделі на основі розподілу Гауса, Гаусівських сумішей розподілів (GMM), прихованих марківських моделей (HMM). Загальним недоліком параметричних методів є припущення, що модель існує і її можна точно оцінити. Але це сумнівно у багатьох застосуваннях, де справжня структура розподілу ймовірностей невідома і може бути дуже складною [69].

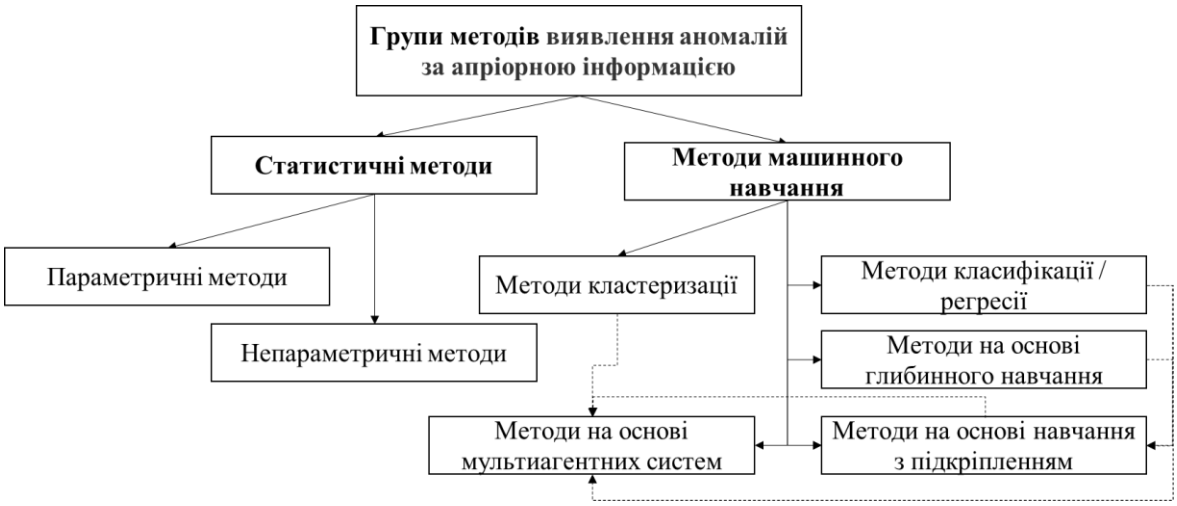


Рисунок 1.5 – Основні групи методів виявлення аномалій в області інтелектуального аналізу даних

Серед непараметричних методів найбільшого поширення набули гістограмні методи та ядерна оцінка щільності розподілу (Kernel Density Estimation, KDE) [65]. У методах, що ґрунтуються на гістограмах, виділяють два підходи до виявлення аномалій. При першому підході просто перевіряють, чи потрапляє нова точка даних у будь-яку область гістограми. Якщо так, точка даних класифікується як нормальна, інакше — вона аномальна [65]. Другий підхід відносить значення даних до аномального з урахуванням його частоти появи [65]. Якщо частота нижча за вказаний поріг, дані з даним значенням класифікуються як аномальні. Наприклад, якщо ввести певний поріг зустрічаємості різниці м’ячів матчу серед усіх матчів сезону, то поєдинки, остаточний рахунок в яких попадає у стовпчики гістограми, сума висот яких менше такого порогу, відносяться до матчів із фіксованим результатом.

На рис. 1.6 наведено приклад ряду ймовірностей різниці м’ячів футбольних матчів, а червоним кольором виділено значення різниць м’ячів матчу, сума ймовірностей яких не перевищує певний фіксований поріг, показаний штрихованою лінією.

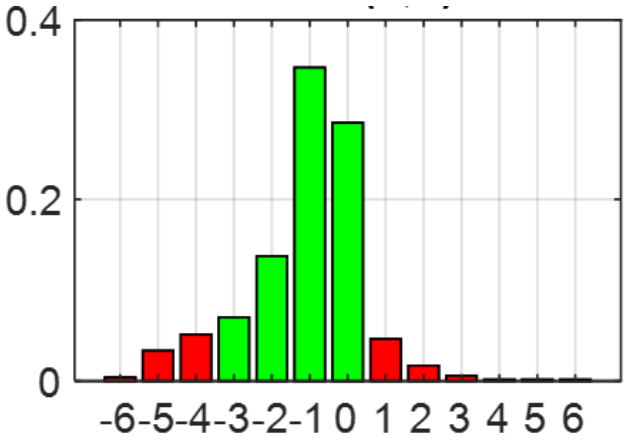


Рисунок 1.6 – Приклад гістограми різниці м’ячів і фіксованого порогу аномальності матчу

Недоліком методів гістограм є те, що неперервні вхідні дані мають бути дискретизовані. А також ці методи мають низьку ефективність при малій вибірці. Метод KDE використовується для аналізу неперевнозначних даних. Він дозволяє отримати неперервну оцінку щільності ймовірності на основі гістограми з використанням обраних у якості ядер функцій. Однак, KDE має щонайменше два недоліки: він досить чутливий до шуму в даних [69], а також вимагає відносно великої вибірки для точної оцінки щільності.

Наступні методи виявлення аномалій належать до методів машинного навчання. ***Методи, засновані на класифікації***, припускають, що межі рішення у просторі ознак можна визначити з навчального набору даних, отже, межі рішення розділяють нормальні та аномальні точки даних. Залежно від кількості нормальних класів даних детектори аномалій на основі класифікації можуть бути згруповані в однокласові або багатокласові детектори аномалій. Більшість класифікаційних методів виявлення аномалій базуються на методі опорних векторів (SVM) або методах, заснованих на правилах [65].

Метод опорних векторів (Support Vector Machine, SVM) застосовується для пошуку аномалій у системах, де нормальною поведінкою характеризується лише один клас [67, 68, 70]. Цей метод належить до алгоритмів контрольованого навчання. Він визначає межу у вигляді гіперплощини, яка поділяє об'єкти на два класи: нормальний та аномальний. Для кожного досліджуваного екземпляра визначається — в якій області він знаходиться, та приймається відповідне рішення.

Стандартний метод SVM є лінійним класифікатором. Головним принципом SVM є максимізація відступів, тобто, відділяюча гіперплощина будується так, щоб вона проходила максимально далеко від найближчих до неї точок, які називають опорними векторами [122]. Оптимальна розділова гіперплощина визначається шляхом вирішення задачі квадратичного програмування у опуклій області. Тому рішення завдання є єдиним. Перевагою SVM є те, що він забезпечує можливість проводити більш впевнену класифікацію, так як алгоритм визначає смугу максимальної ширини. До недоліків методу відноситься відсутність спільного підходу до вибору ядра у разі лінійної нероздільності класів, а також чутливість до шумів даних.

Метод опорних векторів одного класу (one-class SVM) є модифікацією класичного алгоритму методу опорних векторів [71, 72], який на основі тренувальної нерозміченої вибірки відокремлює таку область простору *S*, у якій концентрується більшість точок вибірки і при цьому забезпечується таке обмеження: ймовірність того, що нова точка тестової вибірки опиниться поза областю *S*, є певним апріорним значенням між 0 та 1. Зазначається [71, 72], що на відміну від стандартного методу опорних векторів в методі опорних векторів одного класу як ядра застосовуються лише радіальні базисні функції, інші нелінійні ядра показують гірші результати. Ефективність методу сильно залежить від наявної навчальної вибірки.

Ще один метод на основі класифікації ґрунтується на формуванні правил, які відповідають нормальній поведінці системи [65]. Екземпляр даних, який не відповідає цим правилам, розпізнається як аномальний. Алгоритм

складається із двох кроків. Перший крок: навчання правил вибірки за допомогою одного з алгоритмів, таких як RIPPER, Decision Trees тощо [65]. Кожне правило має своє значення достовірності, пропорційне відношенню між кількістю об’єктів тренувальної вибірки, правильно прокласифікованих даним правилом, до загальної кількості об’єктів тренувальної вибірки, охоплених правилом. Другий крок: пошук для кожного тестового об’єкта такого правила, яке найкраще підходить до цього екземпляра. Методи можуть розпізнавати як один, так і кілька класів даних.

В окрему групу можна виділити ***методи на основі глибинного навчання***, засновані на використанні різних типів нейронних мереж, які навчаються за тренувальною вибіркою прогнозувати рівень аномальності або цільові ознаки об’єкта, за якими приймається рішення про його аномальність [100, 123-127].

Необхідно відзначити, що нейронні мережі є також потужними класифікаторами, оскільки вони можуть апроксимувати довільні складні межі рішень у просторі ознак [66]. Проте вони призначені для визначення нових класів даних. Оскільки нейронні мережі не створюють закритих меж для класів, адаптація їх до виявлення аномалій (новинок) та забезпечення того, що узагальнена властивість мережі не повинна впливати на її здатність виявляти аномалії, є досить складним завданням [66]. Іншими словами, існує, зазвичай, ризик перенавчання моделі тренувальними даними. Також труднощі виникають при виборі відповідної структури нейронної мережі для конкретної задачі та при інтерпретації отриманих закономірностей, оскільки навчені нейронні мережі є моделями типу чорного ящика.

Налаштування коефіцієнтів нейронної мережі відбувається з використанням ітераційних методів багатовимірної оптимізації. Метою оптимізації є мінімізація цільової функції помилки прогнозування. В свою чергу, ***методи на основі глибинного навчання з підкріпленням*** [128, 129] додатково використовують поняття агента, в якості якого виступає нейронна мережа, яка навчається як на основі тренувальної вибірки, так і своїх

прийнятих рішеннях про аномальність об’єкта. Для оцінювання прийнятих агентом рішень вводиться функція нагороди, значення якої враховується агентом під час прийняття рішень по наступним об’єктам [128-130]. Узагальненням навчання з підкріпленням є ***мультиагентні системи*** [131, 132], які складаються з декількох агентів, що розв’язують загальну задачу паралельно і можуть комунікувати між собою. Тренування агентів відбувається як за рахунок комунікації, так і за рахунок тих же підходів, що й у випадку навчання з підкріпленням. Агенти можуть бути представлені як різними типами нейронних мереж, так й іншими алгоритмами машинного навчання [131-133].

Методи найближчих сусідів для виявлення аномалій засновані на припущенні, що нормальні дані зустрічаються у щільних областях вибірки, тоді як аномалії виникають далеко від своїх найближчих сусідів [65]. Це передбачає, що необхідно визначити поняття функції відстані (міри схожості) між точками даних [113]. Часто для цього використовується Евклідова відстань. Відповідно за методом найближчих сусідів визначається відстань або до одного, або до k найближчих сусіди та за значеннями цих характеристик приймається рішення про аномальність даних. До переваг методу найближчих сусідів можна віднести такі:

— процес навчання полягає у запам'ятовуванні навчальної вибірки;

— простота реалізації та можливість вводити додаткові параметри налаштування вибірки;

— легко інтерпретується логіка алгоритму.

Як недолік можна відзначити неефективне використання пам'яті, внаслідок необхідності збереження повної вибірки. Також велика кількість проведених операцій робить алгоритм трудомістким.

Методи кластеризації ґрунтуються на групуванні схожих точок даних у кластери та не вимагають знань про властивості можливих відхилень [134-137]. У випадку, коли аномальні екземпляри не є одиничними, їх

виявлення будується на припущенні про те, що нормальні дані утворюють великі щільні кластери, а аномальні — маленькі та розрізнені. Однією з найпростіших і найвідоміших реалізацій підходу на основі кластеризації є алгоритм *k*-середніх, описаний у роботах [134, 138].

Одними з загальновизнаних характеристик якості роботи класифікатора, які часто використовуються на практиці, є точність, повнота та їх середнє гармонічне (F-міра). Точність (precision) класифікації в межах класу — це частка знайдених класифікатором даних, які дійсно належать даному класу, відносно всіх даних, які система віднесла до цього класу. Повнота (recall) класифікації — це частка знайдених класифікатором даних, які дійсно належать даному класу, відносно всіх даних цього класу в тестовій вибірці. F_β-міра є зваженим гармонічним середнім точності і повноти, причому ваги цих характеристик у сумі мають дорівнювати $1 + \beta^2$, точність має вагу β^2, а повнота, відповідно, — 1. Таке зважування дає можливість регулювати вклад обох характеристик у значення F_β-міри. При $\beta = 1$ отримують, так звану, F_1-міру, яка є просто гармонійним середнім точності і повноти. Іншими словами, у F_1-мірі характеристики точності і повноти мають однакову вагу [139].

Розглянуті методи машинного навчання орієнтовані на роботу з неперервнозначними даними в багатовимірних ознакових просторах. Крім того, вони вимагають для навчання значних об’ємів вхідних даних, а їх результати не мають статистичного підкріплення у вигляді ймовірності аномальності даних. Аналіз розглянутих методів показує, що тільки один із них, а саме: статистичний непараметричний гістограмний метод в принципі дозволяє розв’язати задачу виявлення матчів з фіксованим результатом на основі **виключно загальнодоступної публічної інформації**. Для цієї задачі гістограмний метод зводиться до розрахунку полігону відносних частот, оскільки вхідні дані, породжуються дискретною випадковою величиною, що характеризується малою кількістю дискретних числових значень.

Вважається [14, 16, 17, 22, 65], що результати договірних матчів належать тільки області аномальних результатів. Таким чином, матчі, ***підозрілі*** на фіксований результат, складаються з двох множин:

1) звичайні матчі, але з аномальними результатами;
2) договірні матчі, результати яких є аномальними за визначенням.

Якщо гістограма ключових характеристик матчу, наприклад, різниці м’ячів матчів певної групи команд за умови відсутності договірних матчів, є відомою, то виявлення матчів, підозрілих на фіксований результат, не викликає труднощів і зводиться до визначення області аномальних результатів по заданій пороговій ймовірності. Отримані результати перевіряються на попадання в область аномальних результатів і класифікуються як нормальні або аномальні.

1.4. Конформні предиктори

Новим перспективним класом методів машинного навчання, які використовуються як для вирішення задач класифікації даних, так і виявлення аномалій, є методи на основі конформних предикторів [140]. Вони не вимагають знання законів розподілу даних, використовуються як з дискретнозначними, так і неперервнозначними даними. Зокрема, в [141] показано, як саме цей математичний апарат можна практично використати для покращення результатів класифікації методами опорних векторів і найближчих сусідів та для виявлення змін у потоках даних.

Фактично методи на основі конформних предикторів можна розглядати як певний розвиток гістограмних методів, оскільки вони надають статистичну оцінку аномальності об'єкта, використовуючи спеціальну характеристику, яка розраховується на основі гістограми значень міри неконформності.

Загалом, конформні предиктори — клас методів машинного навчання, які прогнозують належність об’єкту до певного класу на основі міри відмінності (конформності) поточного об’єкта від попередніх об’єктів [140].

Конформний аномальний детектор є методом на основі конформного прогнозування, що надає ймовірнісно подібну міру надійності для прогнозу аномальності об'єкта, який надходить від довільного методу виявлення аномалій [142].

У багатьох областях застосування машинного навчання важливі як самі передбачення, так й довірчі множини (інтервали) цих прогнозів. Підхід, заснований на конформних предикторах, дозволяє отримати не тільки точковий прогноз, але й довірчу множину спрогнозованих значень. Ще однією особливістю цього методу є можливість налаштовувати рівень довіри, що дозволяє уточнювати довірчі інтервали [144].

Розглянемо множину об'єктів $z_i \in Z$, де $z_i = (x_i, y_i)$, $x_i \in X$ є вектором ознак i-го об'єкта, $y_i \in Y$ є міткою класу i-го об'єкта, X є простором ознак об'єктів, Y є простором міток класів. Вважаємо, що кожен об'єкт z_i отримано з декартового добутку $Z = X \times Y$ на основі деякого ймовірнісного розподілу $P \in \mathcal{P}(\mathrm{Z})$, де $\mathcal{P}(\mathrm{Z})$ є множиною всіх можливих ймовірнісних розподілів на множині Z. Також вважаємо, що кожен об'єкт z_i отримано незалежно від усіх інших отриманих об'єктів. Ці припущення загалом називаються припущенням ***незалежності й ідентичності розподілу*** (HIP) [140]. Також вводиться ***припущення обмінюваності,*** за якого послідовність $(z_1, \dots z_k)$ генерується з ймовірнісного розподілу P^k на Z^k та має *властивість обмінюваності*: для будь-якої перестановки π множини $\{1, \dots, k\}$, ймовірнісний розподіл перемішаної послідовності $(z_{\pi(1)}, \dots, z_{\pi(k)})$ співпадає з розподілом P вихідної послідовності $(z_1, \dots z_k)$ [140].

Алгоритм конформного прогнозування у задачах класифікації визначає ймовірнісно подібні значення (***ступені конформності*** $p_k^{(y)}$) належності об'єкта z_k до кожного з класів $y \in Y$ за наявності вже класифікованих об'єктів $(z_1, \dots, z_{k-1})$. Відбувається це таким чином [140]:

1) для кожного об'єкту з послідовності $(z_1, \dots, z_i, \dots, z_k)$ обчислюється ***міра неконформності*** $(a_1^{(\hat{y}_k)}, \dots, a_i^{(\hat{y}_k)}, \dots, a_k^{(\hat{y}_k)})$ за умови, що об'єкт z_k

належить до класу $\hat{y}_k$:

$$a_1^{(\hat{y}_k)} = A_k(\{z_2, \dots, z_k\}, z_1),$$

…,

$$a_i^{(\hat{y}_k)} = A_k(\{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_k\}, z_i),$$

…,

$$a_k^{(\hat{y}_k)} = A_k(\{z_1, \dots, z_{k-1}\}, z_k),$$

де A_k є функцією, яка залежить від множини вигляду $\{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_k\}$ й об'єкта z_i, а у відповідність цим аргументам ставить дійсне число: $A_k: Z^{k-1} \times Z \to R$. Значення $a_i^{(\hat{y}_k)}$ характеризують міру того, наскільки кожен об'єкт є неконформним (відмінним) усій послідовності об'єктів $(z_1, \dots, z_i, \dots, z_k)$: чим більшим є значення $a_i^{(\hat{y}_k)}$, тим менш конформним є об'єкт z_i. Разом з цим, якщо функція міри неконформності A_k залежить від значення запропонованого класу $\hat{y}_k$, тоді значення $a_k^{(\hat{y}_k)}$ показує наскільки неконформною є умова, що об'єкт z_k належить до класу $\hat{y}_k$: чим більшим є значення $a_i^{(\hat{y}_m)}$, тим менш конформним є об'єкт z_k для класу $\hat{y}_k$;

2) для кожної окремої мітки $\hat{y}_m \in Y$ класу обчислюється ***ступінь конформності*** (p-value) $p_k^{(\hat{y}_m)}$ належності об'єкта z_k до класу $\hat{y}_k$:

$$p_k^{(\hat{y}_k)} = \frac{|\{\, j \mid 1 \le j \le k, a_j^{(\hat{y}_k)} \ge a_k^{(\hat{y}_k)}\}|}{k}.$$

Величина $p_k^{(\hat{y}_k)}$ приймає значення в діапазоні $[\frac{1}{k}; 1]$.

3) конформним предиктором $\Gamma^{\varepsilon}(z_1, \dots, z_i, \dots, z_k)$ є така множина прогнозованих міток класів для об'єкту z_k на рівні значущості $\varepsilon \in [0; 1]$:

$$\Gamma^{\varepsilon}(z_1, z_2, \dots, z_k) = \{\hat{y}_k \mid p_k^{(\hat{y}_k)} > \varepsilon\}.$$

Рівень значущості ε характеризує допустимість (надійність) отриманого конформного предиктора в такому сенсі: отримана множина прогнозованих міток $\Gamma^{\varepsilon}(z_1, z_2, \dots, z_k)$ тим більше є надійною, чим меншим є рівень значущості ε.

Також разом з рівнем значущості ε можливе використання рівня довіри $1-\varepsilon$: отримана множина прогнозованих міток $\Gamma^{\varepsilon}(z_1, z_2, \dots, z_k)$ тим більше є надійною, чим більшим є рівень довіри $1-\varepsilon$ [140].

Сімейство $\{\Gamma^{\varepsilon}(z_1, z_2, \dots, z_k) | \ \varepsilon \in [0;1]\}$ конформних предикторів $\Gamma^{\varepsilon}(z_1, \dots, z_i, \dots, z_k)$ для усіх можливих рівнів значущості ε називається ***предиктором довіри*** (довіреним предиктором, confidence predictor). Сімейство конформних предикторів є вкладеним в такому сенсі [142]: для будь-яких $0 < \varepsilon_1 < \varepsilon_2 < 1$

$$\Gamma^{\varepsilon_1}(z_1, z_2, \dots, z_k) \supseteq \Gamma^{\varepsilon_2}(z_1, z_2, \dots, z_k).$$

Однією з головних ознак якості конформного предиктора і предиктора довіри є ***валідність*** (достовірність): вона показує наскільки надійними є отримані конформні предиктори.

Конформний предиктор $\Gamma^{\varepsilon}(z_1, z_2, \dots, z_k)$ називається валідним (достовірним) на рівні значущості $\varepsilon \in [0;1]$, якщо незалежно від ймовірнісного розподілу P^k на Z^k, ймовірність того, що $y_k \notin \Gamma^{\varepsilon}(z_1, z_2, \dots, z_k)$ не перевищує ε. Предиктор довіри $\{\Gamma^{\varepsilon}(z_1, z_2, \dots, z_k) | \ \varepsilon \in [0;1]\}$ називається валідним, якщо кожен окремий конформний предиктор $\Gamma^{\varepsilon}(z_1, z_2, \dots, z_k)$ є валідним на своєму рівні значущості ε [140].

Доведено [140], що за припущення обмінюваності, ймовірність помилки $y_k \notin \Gamma^{\varepsilon}(z_1, z_2, \dots, z_k)$ не перевищує ε для будь-якого $\varepsilon \in [0;1]$ і конформного предиктора Γ і, таким чином, він має властивість валідності.

Рівень значущості ε використовується як рівень, що характеризує помилку класифікації: чим меншим є ε, тим теоретично меншою є ймовірність того, що для об'єкта z_k прогнозована мітка $\hat{y}_k$ не співпадає зі справжньою міткою y_k. Множина $\Gamma^{\varepsilon}(z_1, z_2, \dots, z_k)$ містить усі прогнозовані мітки класу для об'єкта z_k на рівні значущості ε. Важлива на практиці ситуація виникає, якщо рівень значущості вибрано досить низьким. Наприклад, нехай маємо $\varepsilon \leq 0{,}01$ та при цьому $\Gamma^{\varepsilon}(z_1, z_2, \dots, z_k) = \emptyset$. Остання рівність означає, що $\forall k \ p_k^{(\hat{y}_k)} < \varepsilon$, тобто об'єкт x_k є чимось новим та має зовсім мало спільних рис з кожним з

інших об'єктів $x_1, \dots, x_{k-1}$. Ситуація $\Gamma^{\varepsilon}(z_1, z_2, \dots, z_k) = \emptyset$ еквівалентна ситуації єдиної спрогнозованої мітки класу з достовірністю меншою за ε, що, в свою чергу, може означати, що тестовий об'єкт x_k не є представником навчальної вибірки $(z_1, z_2, \dots, z_{k-1})$ [144] і може розглядатися як аномальний.

Загалом, якщо новий об'єкт і навчальний набір є HIP або отримані з розподілу з властивістю обмінюваності, то за раніше наведеною властивістю валідності ймовірність помилки визначення мітки класу обмежена зверху ε. Таким чином, якщо об'єкт z_k класифікується як аномальний, коли відповідний набір прогнозів є невірним (порожнім або таким, який не містить справжньої мітки класу), тоді вказаний рівень значущості ε відповідає верхній межі ймовірності виявлення аномалії.

У задачах виявлення аномалій часто є важливим лише виявлення аномальних об'єктів z_k, а не визначення того, до якого нормального класу $\hat{y}_k$ належить об'єкт. У цьому випадку виходить розумніше оцінювати ступінь конформності *p-value* тільки для мітки, яка уже є у цього об'єкта, а не обчислювати *p-value* для всіх можливих міток $y \in Y$. У роботі [145] було запропоновано ***конформний аномальний детектор***, який є загальним алгоритмом для перевірки на аномальність об'єкта z_k на основі міри неконформності *A*, встановленого рівня значущості ε та тренувального набору об'єктів z_1,..., z_{k-1}. Правило роботи детектору: об'єкт вважається аномальним, якщо виконується умова

$$p_k < \varepsilon, \tag{1.1}$$

де $\varepsilon \in [0; 1]$ є порогом аномальності (anomality threshold).

Інакше, об'єкт вважається нормальним.

Множина всіх матчів, для яких виконується умова (1.1), називається ***конформним аномальним предиктором*** і позначається $\Gamma^{\varepsilon}(z_1, z_2, \dots, z_k,)$. Доведено, що за припущення обмінюваності або HIP об'єктів вибірки $z_1, \dots, z_k$ для будь-якої міри неконформності A_k і $k \geq 1$ ймовірність помилки у прийнятті рішення, що об'єкт є аномальним, не перевищує ε [146]. Параметр

ε регулює чутливість конформного аномального детектора до виявлення аномальних об'єктів [147].

Необхідно відмітити, що в роботах [145-147] конформний аномальний детектор був використаний для виявлення аномалій в даних в режимі on-line, тобто для перевірки поточних даних в реальному часі. При цьому при побудові міри неконформності використовувалися прогнозовані значення поточних даних на основі даних, отриманих до поточного моменту часу.

Задача, що розглядається, відрізняється наявністю всієї отриманої вибірки даних, по якій приймаються рішення по кожному матчу. Це також вимагає розробки міри неконформності, що враховує всю наявну інформацію.

Значення ступенів конформності, що отримані як результат конформного аномального детектора або конформного предиктора, можуть бути використані також для перевірки наявності у послідовності об'єктів властивості обмінюваності.

Послідовність випадкових величин $\{S_k\}_{k=1}^{\infty}$ називається ***мартингалом***, якщо [148-150]:

$$\mathbf{M}(S_n|S_{n-1}, S_{n-2}, \dots, S_1) = S_{n-1}, \qquad (1.2)$$

де $\mathbf{M}(X|Y)$ є умовним математичним сподіванням випадкової величини X за умови відомої випадкової величини Y.

Мартингали застосовуються для дослідження властивостей і прогнозування випадкових процесів [151, 152].

У роботі [148] було запропоновано використати випадкові змінні $M_0^{(\eta)}, M_1^{(\eta)}, \dots, M_k^{(\eta)}, \dots$, що називаються ***степеневим мартингалом*** [148, 150]:

$$M_k^{(\eta)} = \prod_{i=1}^{k} \left(\eta p_i^{\eta-1}\right),$$

де $\eta \in [0; 1]$, $M_0^{(\eta)} = 1$.

Послідовності випадкових змінних $M_0^{(\eta)}, M_1^{(\eta)}, \dots, M_k^{(\eta)}, \dots$, для усіх значень параметра $\eta \in [0; 1]$ називаються сімейством степеневих мартингалів.

Для нівелювання залежності від параметра ε використовують **інтегральний мартингал** [148, 150]:

$$M_k = \int_0^1 M_k^{(\eta)} d\eta.$$

В роботах [149, 152] доведено, що інтегральний мартингал M_k та степеневий мартингал $M_k^{(\eta)}$ задовольняють означенню мартингала (1.2).

Степеневий та інтегральний мартингал дають можливість перевірити гіпотезу про властивість обмінюваності для послідовності об'єктів $(z_1, z_2, \dots, z_k, \dots)$ поточково за рахунок властивості, яка доведена для мартингалів і має назву нерівності Дуба [145]:

$$P\left(\max_k M_k \geq \lambda\right) \leq \frac{1}{\lambda}.$$

Також у роботі [149] доведено, що степеневий та інтегральний мартингал з початковою умовою $M_k^{(\eta)} = 1$ є невід'ємним незростаючим випадковим процесом. Таким чином, вибравши значення порогу λ так, що значення $\frac{1}{\lambda}$ не перевищуватиме наперед заданий рівень значущості, виконання умови $M_k \geq \lambda$ вважається умовою того, що на об'єкті z_k порушується властивість обмінюваності для послідовності $(z_1, z_2, \dots, z_k, \dots)$. На практиці порушення цієї властивості означає такі можливі ситуації:

1) послідовність $(z_1, z_2, \dots, z_k, \dots)$ не є випадковою, тобто вона не є згенерованою випадковим розподілом;

2) точка z_k взята з випадкового розподілу, що відрізняється від випадкового розподілу, за яким була згенерована послідовність $(z_1, z_2, \dots, z_{k-1})$.

Як наслідок, степеневий та інтегральний мартингали було використано у задачі пошуку в потоці даних об'єкта, що характеризує момент зміни у властивостях середовища, в якому генеруються об'єкти [141]. Також мартингали було використано для задачі пошуку аномальних послідовностей точок на прикладі траєкторій польоту літаків [153]. Як міра неконформності

використовується відстань між положеннями поточної точки, спрогнозованими за нормальною (Гаусівською) регресійною моделлю, отриманою на основі тренувальної вибірки, та моделлю, побудованою за наявними на поточний момент точками траєкторії.

Переваги розглянутого математичного апарату полягають у поєднанні процесу навчання і прогнозування у одну стадію і, як наслідок, у використанні результатів класифікації попередніх об’єктів для прогнозування результату для поточного об’єкта.

Крім того, цей підхід дозволяє без значних обчислювальних витрат отримати непараметричні довірчі інтервали (довірчі множини) при вирішенні завдань класифікації та виявлення аномалій, що дозволяє вводити оцінки гарантованої точності для рішень.

1.5. Постановка задачі досліджень

Метою роботи є підвищення ефективності виявлення матчів, потенційно підозрілих на фіксованість результату, на базі обробки виключно загальнодоступних публічних даних про результати проведених матчів.

Для досягнення вказаної мети у даній дисертаційній роботі вирішуються такі задачі:

1. Розробка процедури групування команд сезону.
2. Розробка імітаційної моделі футбольного сезону з матчами з фіксованим результатом.
3. Розробка методу виявлення підозрілих футбольних матчів з фіксованим результатом за допомогою конформного аномального детектора.
4. Розробка методу виявлення підозрілих футбольних матчів з фіксованим результатом за допомогою степеневого мартингалу.
5. Розробка методу виявлення підозрілих футбольних матчів з фіксованим результатом за допомогою інтегрального мартингалу.

6. Дослідження й оцінка ефективності розроблених методів виявлення підозрілих на фіксованість результату футбольних матчів на основі метрик класифікації, які є базовими в машинному навчанні на модельних та реальних сезонах.

Об'єктом дослідження є виявлення матчів, підозрілих щодо фіксованого результату в футбольних турнірах.

Предметом дослідження є моделі та методи виявлення потенційно підозрілих на фіксований результат матчів у футбольних турнірах.

Висновки до розділу 1

1. Матчі з фіксованим результатом є реальною проблемою, яка загрожує основоположним принципам та авторитету футболу у багатьох країнах світу, у тому числі й в Україні. На сьогодні такі матчі кваліфікуються з юридичної точки зору як кримінальний злочин. Однак факти проведення договірних матчів встановити напевне майже неможливо, принаймні, без проведення спеціальної поліцейської операції, оскільки сама природа спортивних змагань гарантує, що не завжди перемагає сильніша команда. Тому можна говорити лише про те, що результат певних матчів є підозрілим, аномальним тощо. Але й така інформація є корисною та важливою, оскільки виступає додатковим індикатором того, що певна команда може мати стосунок до порушення спортивних принципів чесної гри.

2. У договірних матчах, пов'язаних з виграшом за ставками, завданням є отримати результат матчу, відмінний від очікуваного, щоб максимально заробити на ставці. Тому результати таких матчів можна розглядати як нетипові, аномальні, що дозволяє формалізувати пошук матчів з фіксованим результатом.

3. Використання методів прогнозування результату матчу, аналізу ставок або дій учасників матчу протягом всієї гри для виявлення підозрілих щодо фіксованого результату футбольних матчів вимагає великої кількості

даних, які не завжди доступні для аналізу. Це обумовлює доцільність проведення досліджень виявлення потенційно підозрілих щодо фіксованого результату матчів, які не вимагають великої кількості непублічно доступних даних.

4. Задача виявлення підозрілих щодо фіксованого результату матчів відноситься до виявлення контекстних аномалій, що вирішується в галузі інтелектуального аналізу даних. Більшість методів машинного навчання, що використовуються для пошуку аномалій в даних, орієнтовані на роботу з неперервнозначними даними в багатомірних ознакових просторах і вимагають великого об’єму вибірки, що обмежує можливість їх застосування для розв’язання поставленої в дисертаційному дослідженні задачі.

5. Конформні предиктори не вимагають знання законів розподілу даних, використовуються як з дискретнозначними, так і неперервнозначними даними. Переваги даного математичного апарату полягають у поєднанні процесу навчання і прогнозування у одну стадію і, як наслідок, у використанні результатів класифікації попередніх об’єктів для прогнозування результату для поточного об’єкта. Також цей підхід дозволяє без значних обчислювальних витрат отримати непараметричні довірчі інтервали (довірчі множини) при вирішенні завдань класифікації та виявлення аномалій, що дозволяє вводити оцінки гарантованої точності для рішень (прогнозів).

6. Актуальною науковою задачею є розробка методів виявлення підозрілих на фіксований результат матчів з використанням апарату конформних предикторів, степеневих та інтегральних мартингалів на основі використання даних лише про рахунки проведених матчів та місць їх проведення.

РОЗДІЛ 2
ФОРМАЛІЗАЦІЯ ОПИСУ ВХІДНИХ ДАНИХ, ГРУПУВАННЯ КОМАНД ТА РОЗРОБКА ІМІТАЦІЙНОЇ МОДЕЛІ ФУТБОЛЬНОГО СЕЗОНУ З МАТЧАМИ З ФІКСОВАНИМ РЕЗУЛЬТАТОМ

2.1. Групування команд за їхньою силою

Як зазначалося в підрозділі 1.3 ***матчі з фіксованим результатом відносяться до класу контекстних аномалій***. Для визначення аномалій цього типу одним із основних кроків є виділення контекстуальних і поведінкових атрибутів. Як поведінковий атрибут футбольного матчу будемо використовувати різницю м'ячів, бо вона дозволяє просто і однозначно встановити результат відповідного матчу. Проте матч може бути контекстуальною аномалією за одних значень різниці голів (скажімо, перемога слабкої команди над сильною з різницею у три м'ячі), але при такому ж значенні цієї різниці, тобто поведінкового атрибуту, вважатися нормальним у іншому контексті (при перемозі сильної команди над слабкою). Як контекстуальні атрибути візьмемо атрибути «сила команди» і «тип гри» — виїзна або домашня. За силою команди поділяються на групи. Групи визначаються шляхом проведення одновимірної або двовимірної кластеризації. Одновимірна кластеризація відбувається за кількістю набраних очок, а двовимірна — за кількістю набраних очок та різницею між забитими і пропущеними м'ячами команд у сезоні. Кластеризація дозволяє виділити ***групи однорідності команд за силою*** за результатами сезону. На основі контекстуальних атрибутів матчі турніру тоді можна буде розбити на ***класи*** і вже в кожному класі матчів за поведінковим атрибутом визначати аномальні матчі. Таким чином, важливою задачею є кластеризація команд за силою.

Одним зі стандартних і найбільш доступних показників успішності команд у сезоні є загальні очки, набрані кожною командою під час сезону: $s(t)=3w(t) + d(t)$, де $w(t)$ та $d(t)$ є відповідно кількості перемог та ігор у нічию для команди t. Крім цього показника, для аналізу і моделювання результатів футбольних матчів також зазвичай розглядають такі характеристики як кількість забитих $gf(t)$ та пропущених $ga(t)$ командою t голів, а також різницю цих величин $gd(t)=gf(t) – ga(t)$. Розглянуті характеристики наведено у табл. 2.1 для сезону 2013-2014 рр. Ліги 2 Франції. Цей сезон був обраний тому, що були публічні заяви про сфальшованість результатів деяких матчів команди Nîmes Olympique[10,11], щоправда потім цю інформацію відповідним органам не вдалося юридично довести.

Таблиця 2.1 – Показники успішності команд у сезоні 2013-2014 рр. Ліги 2 Франції

№	t	$s(t)$	$w(t)$	$d(t)$	$gf(t)$	$ga(t)$	$gd(t)$
1	Metz	76	22	10	55	28	27
2	Lens	65	17	14	58	40	18
3	Caen	64	18	10	65	44	21
4	Nancy	61	16	13	47	37	10
5	Niort	58	15	13	51	47	4
6	Dijon	57	14	15	53	42	11
7	Brest	56	15	11	38	32	6
8	Angers	55	14	13	46	45	1
9	Tours	55	15	10	63	56	7
10	Troyes	52	15	7	56	44	12
11	Creteil	50	12	14	57	58	-1
12	Le Havre	48	11	15	43	43	0
13	Arles-Avignon	46	10	16	36	38	-2
14	Clermont	45	10	15	31	38	-7
15	Nimes	44	10	14	49	54	-5
16	Auxerre	43	10	13	35	45	-10
17	Laval	42	10	12	44	52	-8
18	Chateauroux	40	10	10	43	59	-16
19	Istres	36	9	9	48	74	-26
20	CA Bastia	24	4	12	21	63	-42

[10] Le Monde. Matchs truqués présumés : ce que révèlent les écoutes (14 janvier 2015), from https://www.lemonde.fr/football/article/2015/01/15/les-pieds-nickeles_4557302_1616938.html

[11] L. Chami, Matchs truqués de Ligue 2 : 18 mois ferme pour les anciens dirigeants nîmois, leparisien.fr (Sep. 13, 2018), from https://www.leparisien.fr/sports/football/matchs-truques-de-l2-18-mois-ferme-pour-les-anciens-dirigeants-nimois-13-09-2018-7887090.php

На рис. 2.1 точками представлено положення команд за набраними очками в одновимірному просторі. На рис. 2.2 точками показано положення команд за кількістю набраних очок та різницею між забитими і пропущеними м’ячами команд у сезоні в двовимірному просторі. Параметри «кількість набраних очок» та «різниця між забитими і пропущеними м’ячами команд» обрані як найбільш інформативні параметри набору даних з табл. 2.1, оскільки інші параметри є складовими цих двох.

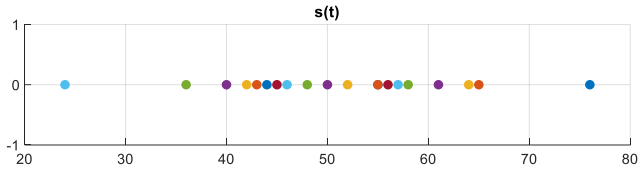


Рисунок 2.1 – Положення команд за набраними очками в одновимірному просторі

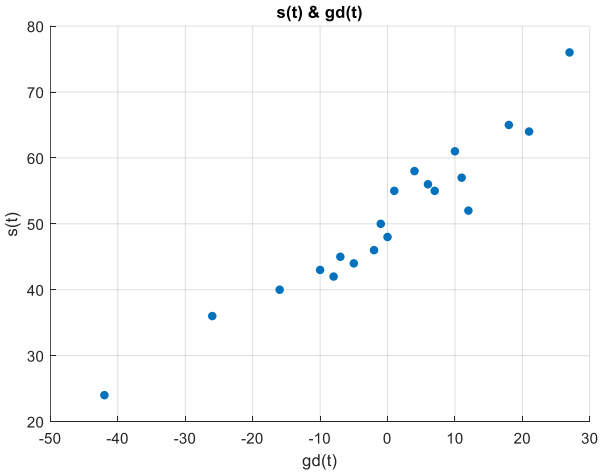


Рисунок 2.2 – Положення команд за кількістю набраних очок та різницею між забитими і пропущеними м’ячами команд у сезоні в двовимірному просторі

Спочатку для групування команд використаємо класичний метод кластеризації *K*-середніх [134-137]. Вхідними даними для методу *K*-середніх є матриця точок *X* розмірності (*N*, *D*), де *N* є кількістю точок вибірки, тобто

довжиною вибірки, а *D* — кількістю координат однієї точки (або розмірністю даних), та кількість кластерів *K*.

Результатом методу є вектор $G = (g_1, \dots, g_i, \dots, g_N)$ довжиною *N*, де координата g_i дорівнює номеру групи, до якої було віднесено точку x_i вибірки *X*. Також для визначення приналежності точки x_n до відповідного кластеру (групи) за методом *K*-середніх використовують вагову функцію $\theta_n(k)$ такого вигляду:

$$\theta_n(k) = \begin{cases} 1, \text{якщо точка } x_n \text{ належить кластеру } C_k, \\ 0, \text{якщо точка } x_n \text{не належить кластеру } C_k. \end{cases}$$

На початку роботи методу відбувається ініціалізація *K* точок c_k, які називаються центрами відповідних кластерів (c_k є центром *k*-го кластеру C_K). Після цього алгоритм групування точок за методом *K*-середніх ітеративно виконує два кроки:

1. Закріплення точок x_1, x_2,…, x_N за кластерами C_1, C_2,…, C_K за таким правилом:

$$x_n \in C_k \leftrightarrow k = \underset{j=1,\dots,K}{\operatorname{argmin}} ||x_n - c_j||,$$

де $||\cdot||$ є евклідовою нормою вектору, або, іншими словами, $||x_n - c_j||$ є евклідовою відстанню між точкою вибірки x_n і центроїдом c_j.

2. Оновлення центрів c_1, c_2,…, c_K кластерів C_1, C_2,…, C_K так, щоб центроїд c_k став центром мас кластера C_k, тобто:

$$c_k = \frac{1}{|C_k|} \sum_{x_n \in C_k} x_n,$$

де $|C_k|$ дорівнює кількості елементів у кластері C_k.

Після оновлення центрів відбувається перехід на наступну епоху роботи алгоритму, на якій знову повторюються наведені вище два кроки. Завершення роботи методу *K*-середніх відбувається за умови, що кластери після оновлення центрів не змінилися, тобто:

$$\forall k = 1, \dots, K \;\; C_k^{(E-1)} = C_k^{(E)},$$

$$\forall k = 1, \dots, K \;\; c_k^{(E-1)} = c_k^{(E)},$$

де E є номером поточної епохи роботи алгоритму.

Загальноприйнятою базовою рекомендацією початкового вибору центроїдів c_k є ініціалізація шляхом вибору випадкових точок серед точок вибірки *X*. Це дозволяє вибрати центроїди в залежності від розподілу точок у просторі, але не завжди забезпечує збіжність до якісного розв'язку: виникають ситуації, коли при такій ініціалізації в подальшому певні кластери можуть значно зменшитися, в той час як їх кластери-сусіди навпаки — масштабно розростися. Також випадкова ініціалізація центроїдів за методом *K*-середніх може призводити до отримання різних результатів кластеризації, тобто цей метод є нестійким до початкових умов.

Задача кластеризації набору точок *X* методом *K*-середніх розглядається як задача оптимізації такої цільової функції:

$$J(i;K) = \sum_{n=1}^{N}\sum_{k=1}^{K} \theta_n^{(i)}(k) \left\| x_n - c_k^{(i)} \right\|^2 . \qquad (2.1)$$

Для описаного алгоритму методу *K*-середніх ця функція є сумою квадратів відстаней між кожним елементом вибірки x_n і центроїдом $c_k^{(i)}$ кластеру, до якого цей елемент належить на момент епохи з номером *i* та розглядається як внутрішньокластерний квадрат відстані між елементом з вибірки *X* та центроїдом його кластеру у рамках епохи з номером *i*. Математично доведено [154], що для методу *K*-середніх значення цієї функції зменшується з плином часу (зі збільшенням номеру епохи) і процес її оптимізації завжди збігається, що в свою чергу є перевагою цього методу. З іншого боку, не гарантується збіжність процесу оптимізації до глобального мінімуму, що є недоліком методу.

Перевагою методу кластеризації *K*-середніх є об'єднання об'єктів у групи за близькістю до середнього значення, що у сенсі кластеризації за однією змінною може вважатися мірою близькості об'єктів: чим менший радіус кластера (відстань від центру до крайньої точки кластеру), тим менший діапазон значень об'єктів кластера, і тим ближчі вони між собою. Аналогічне визначення близькості об'єктів можна сформулювати і для випадку

двовимірної кластеризації: чим менший радіус кластера, тим меншим є окіл відносно центроїда, у якому знаходяться усі точки кластера, і тим ближчі вони між собою.

Також перевагою методу кластеризації *K*-середніх є визначення приналежності об'єктів кластеру на основі визначення відстані до центру кластеру без додаткового зазначення величини розкиду значень об'єктів або «розміру» кластера, тобто алгоритм не потребує додаткових фізичних обмежень на «об'єм» кластера.

Щодо недоліків цього методу, то до них також можна віднести необхідність визначення потрібної кількості кластерів *K*. В деяких задачах цей параметр може бути визначений наперед із емпіричних міркувань. Також цей параметр можна визначити наперед, якщо є можливість побудувати діаграму розсіяння для вибірки точок *X*, тобто, наприклад, у випадку, коли ми маємо справу з одновимірними, двовимірними або тривимірними точками, причому для останнього випадку, аналіз діаграми розсіяння є уже не настільки тривіальною процедурою як у перших двох випадках. Існують також евристичні підходи до визначення кількості кластерів. Наприклад, метод «ліктя», який полягає у визначенні оптимальної кількості кластерів шляхом аналізу графіку підсумкових значень (значень, отриманих на фінальних епохах) цільової функції (2.1) при різних значеннях параметра *K*, тобто за різної кількості кластерів. Оптимальною тоді вважається така кількість кластерів, за якої відбулося найбільш «стрімке» або швидке зменшення значення цільової функції. Існує багато інших подібних евристичних підходів [136].

Проте основним недоліком цього методу є те, що він дозволяє приналежність елементу вибірки лише одному кластеру. Однак на практиці таке обмеження може бути занадто жорстким, бо мають місце різні «прикордонні» ситуації, коли зіставлення окремого об'єкта і кластера не є однозначним.

Останнього недоліку позбавлений метод кластеризації на основі Гаусівських сумішей, який дозволяє формувати «м'які» рішення, що враховують можливість приналежності об'єкта даних одночасно декільком кластерам [137, 155-158].

Алгоритм групування команд на основі методу Гаусівських сумішей наведено нижче:

1) Маємо кількість груп K, множину точок $X = \{x_n\}$, яку потрібно розбити на групи, кожна з N точок x_n має d координат.

2) Формуємо множину Гаусіан $\{P(x_n|z_n = k; \mu_k, \Sigma_k)\}$, використовуючи метод Гаусівських сумішей. Змінна z_n є номером кластеру, до якої належатиме точка x_n. В загальній суміші кожна з Гаусіан матиме свою вагу $w_k = P_k(z_n; \mu_k, \Sigma_k)$. Тут під μ_k розуміється вектор математичних сподівань по кожній з d координат, Σ_k є коваріаційною матрицею розмірності dxd.

3) За налаштованими Гаусіанами для кожної точки x_n, використовуючи теорему Байєса [137, 159, 160], визначаємо номер кластера за одним із таких правил:

- чітким:

$$z_n = \underset{z_n}{\operatorname{argmax}} P(z_n = k|x_n; \mu_k, \Sigma_k);$$

- нечітким:

точка x_n належить до кластеру $z_n = k$ зі ступенем приналежності $\rho_n^{(k)} = P(z_n = k|x_n; \mu_k, \Sigma_k)$.

Функція $P(z_n = k|x_n; \mu_k, \Sigma_k)$ визначається за теоремою Байєса, використовуючи Гаусіани і їх ваги, отримані за методом Гаусівських сумішей:

$$P(z_n = k|x_n; \mu_k, \Sigma_k) = \frac{P(x_n|z_n = k; \mu_k, \Sigma_k)w_k}{P(x_n)}.$$

Функція $P(x_n|z_n = k; \mu_k, \Sigma_k)$ є щільністю нормального розподілу (Гаусіаною):

$$P(x_n|z_n = k; \mu_k, \Sigma_k) = \frac{1}{\sqrt{2\pi|\Sigma_k|}} e^{-\frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k)}.$$

Налаштування Гаусіан $P(x_n|z_n = k; \mu_k, \Sigma_k)$ відбувається у два етапи:

I) оновлення ступенів приналежності $\rho_n^{(k)}$:

$$\rho_n^{(k)} = \frac{w_k P(x_n|z_n = k; \mu_k, \Sigma_k)}{P(x_n)},$$

II) оновлення параметрів μ_k, Σ_k, w_k кожної з K Гаусіан у такій послідовності:

1) оновлення вектору математичних сподівань μ_k:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \rho_n^{(k)} x_n,$$

де $N_k = \sum_{n=1}^{N} \rho_n^{(k)}$;

2) оновлення коваріаційної матриці Σ_k:

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \rho_n^{(k)} (x_n - \mu_k)^T (x_n - \mu_k),$$

3) оновлення вагового коефіцієнта Гаусіани:

$$w_k = \frac{N_k}{N}.$$

Повторення наведених етапів відбувається до моменту, поки після епохи E кластери не зміняться:

$$\forall k = 1, \dots, K \quad \forall n = 1, \dots, N \;\; \rho_n^{(k)(E-1)} = \rho_n^{(k)(E)}.$$

Цільова функція задачі кластеризації методом Гаусівських сумішей має вигляд, подібний до цієї ж функції для методу K-середніх:

$$F(i; K) = \sum_{n=1}^{N} \sum_{k=1}^{K} \rho_n^{(k)(i)} \left\| x_n - \mu_k^{(i)} \right\|^2. \qquad (2.2)$$

Відмінність функції втрат для метода Гаусівських сумішей від цієї ж функції для методу K-середніх зумовлена тим, що належність елементів до кластерів у методі Гаусівських сумішей визначається через ступені приналежності $\rho_n^{(k)(i)} \in [0; 1]$, і таким чином певною мірою кожен елемент x_n належить до кожного кластера C_k.

Як початкові вектори математичних сподівань можна обирати випадкові точки з набору даних *X*. При цьому, для забезпечення спадання значення функції втрат рекомендується [161] обирати такі діагональні значення матриць Σ_k, які забезпечуватимуть якомога більш широке покриття всього набору даних початковими Гаусіанами.

За допомогою метода Гаусівських сумішей можна отримати кластери еліптичної або більш витонченої форми, ніж метод *K*-середніх. Це стає можливим, оскільки налаштування коваріаційної матриці Гаусіани дозволяє врахувати взаємну кореляцію точок кластеру. Це дозволяє більш точно враховувати особливості набору даних при формуванні окремих кластерів.

З іншого боку, цей процес також може призводити до появи вироджених кластерів або до нестабільної кластеризації. Для запобігання цьому нами запропоновано [4] процедуру регуляризації кореляційної матриці Σ_k шляхом множення її недіагональних елементів після кожного оновлення на коефіцієнт $\gamma < 1$. Це призводить до зменшення взаємної кореляції між значеннями контекстуальних атрибутів сили команд, які описуються умовною Гаусівською щільністю ймовірності. При моделюванні було використано значення $\gamma = 0{,}9$ і $\gamma = 0$.

Спочатку виконаємо групування команд на основі одновимірної кластеризації. При цьому буде використано лише метод *K*-середніх, оскільки в одновимірному випадку взаємна кореляція при використанні методу Гаусівських сумішей не розраховується і формування кластерів еліптичної форми можливо лише при розмірності простору ознак не менше двох.

Групування команд за показником успішності команд у сезоні дозволяє розподілити команди у групи близьких за успішністю учасників. У футболі, враховуючи кількість учасників у футбольних лігах (близько 20 команд-учасників), команди пропонується групувати у 4 кластери:

1) найкращі (перші *N* команд, які за правилами переходять у лігу вищого рівня чи стають призерами або приймають участь у єврокубках, а також

ще одна чи декілька команд, які є найближчими до найкращих за загальним результатом);

2) ті, хто ближчий до найкращих команд;
3) ті, хто ближчий до найгірших команд;
4) найгірші (останні L команд, які за правилами переходять у лігу нижчого рівня, а також ще одна чи декілька команд, які є найближчими до найгірших за загальним результатом).

Кластери №1 і №4 формуються фіксованим чином згідно з правилами футбольної ліги, яка розглядається. Кластери №2 і №3 є проміжними і потребують визначення.

Кластери 1 і 4 за правилами II Ліги Франції матимуть таке наповнення:

- кластер 1 – команди №№ 1-4;
- кластер 4 – команди №№ 17-20.

За близькістю значень та швидкістю спаду значень загальних результатів команд №№5-16 сезону 2013-2014 років Ліги II Франції візуально можна сформувати такі два кластери (що й було зроблено в дослідженні [1], рис. 2.3):

- кластер №2: команди №5 - №9;
- кластер №3: команди №10 - №16.

Значення цільової функції (2.1) для розглянутої вище кластеризації дорівнює $J = 71{,}7$.

Застосуємо класичний метод *K*-середніх для одновимірної кластеризації даних, що аналізуються.

Результати експерименту для загальних результатів команд сезону 2013-2014 років Ліги II Франції наведені в таблиці 2.2. Експеримент проводився за таких умов:

- кількість кластерів дорівнює 2;
- кількість запусків дорівнює 160.
- команди, для яких потрібно провести кластеризацію, – команди №№ 5-16.

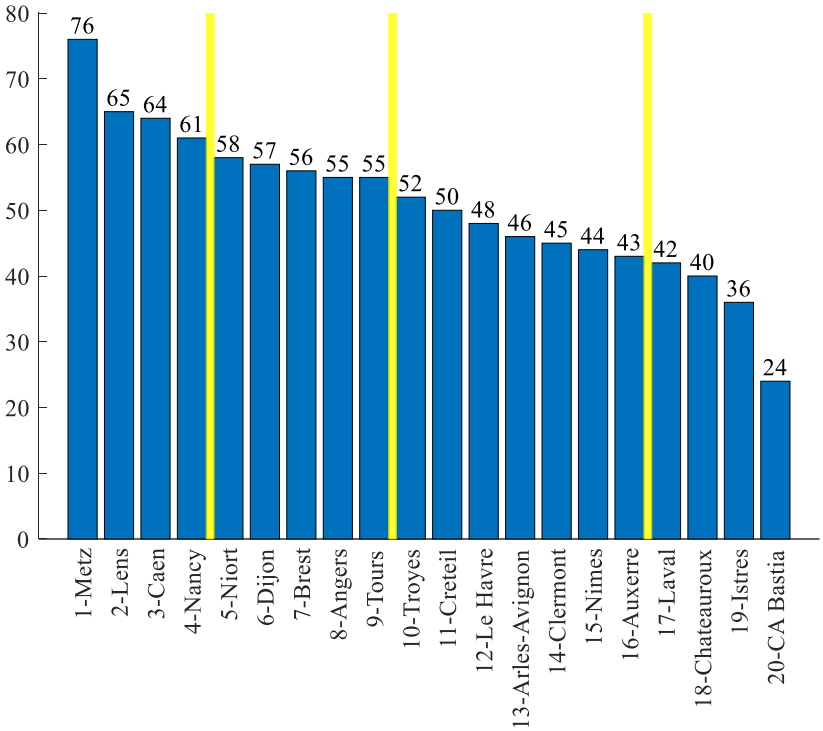


Рисунок 2.3 – Візуальна кластеризація загальних результатів команд

Таблиця 2.2 – Результати експерименту з кластеризації команд сезону 2013-2014 років Ліги II Франції

Результат кластеризації	Кількість появ в експерименті	Значення функції втрат J
{5-10; 11-16}	91	55,5
{5-11; 12-16}	69	62,2

Метод K-середніх в залежності від випадкових початкових умов збігався до двох можливих результатів. При цьому, частіше серед результатів з'являвся варіант {5-10; 11-16}, який має мінімальне значення цільової функції (2.1) серед усіх отриманих результатів. Тому як остаточний результат експерименту, за мінімумом функції втрат було обрано саме варіант групування {5-10; 11-16} (рис. 2.4):

- кластер №2 – команди №№ 5-10;
- кластер №3 – команди №№ 11-16.

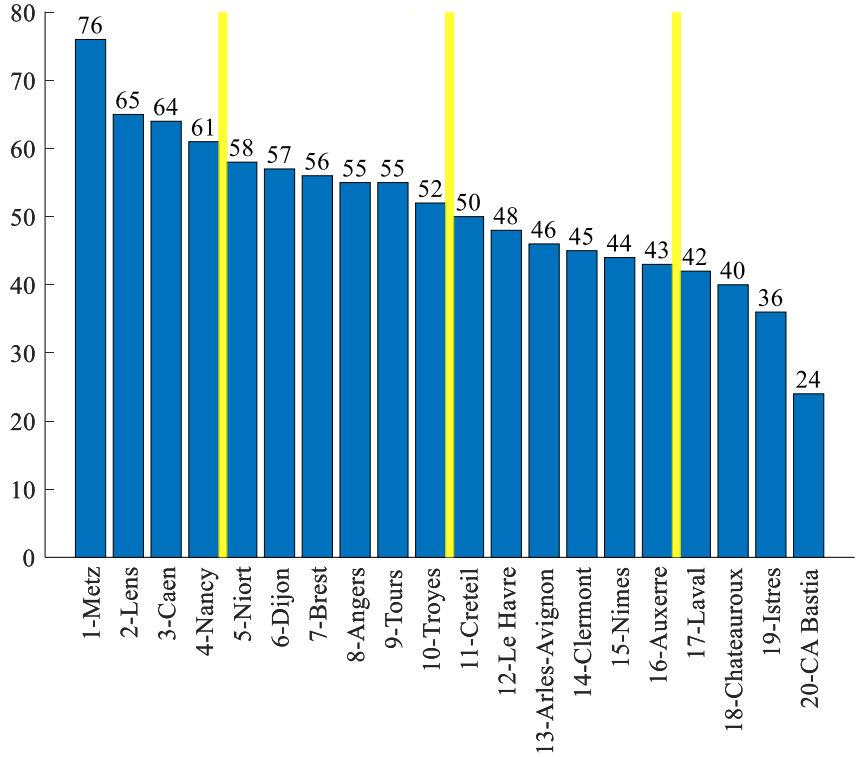


Рисунок 2.4 – Результуюче групування команд сезону 2013-2014 років Ліги II Франції

Двовимірна кластеризація проводиться на параметрах «кількість набраних очок» та «різниця між забитими і пропущеними м’ячами команд» за методами *K*-середніх та Гаусівських сумішей. В обох випадках множина точок *X* розбивалась на $K = 4$ групи. За рахунок того, що використовуються дві змінні (рис. 2.2), точки набору даних розташовані таким чином, що можлива поява вироджених кластерів не є такою очевидною, як це було у випадку одновимірної кластеризації (рис. 2.1). Тому у цьому випадку кластеризацію пропонується проводити без додаткових передумов (без фіксування першого і четвертого кластеру) одразу для пошуку усіх чотирьох кластерів.

На рис. 2.5 наведено можливі варіанти результатів кластеризації, отриманих за допомогою метода *K*-середніх. У заголовках окремих графіків цього рисунку для зручності наведено значення цільової функції (2.1). Найкращим результатом за значенням цільової функції (2.1) $J = 745{,}9$ є результат, зображений на рис. 2.5, а. Цей результат характеризується

компактними проміжними кластерами і невиродженістю крайніх кластерів. Оскільки у даних присутній лінійний тренд, отримані кластери також чітко піддаються ранжуванню: кожному отриманому кластеру можна присвоїти ранг від 1 до 4 за спаданням розглянутих параметрів сили команд. Найближчий до найкращого варіант кластеризації (рис. 2.5, б) програє внаслідок розширення одного з проміжних кластерів. Всі інші варіанти, крім розширюваності проміжних кластерів, мають також і вироджені кластери, що є небажаним при використанні групування команд для подальшої обробки даних.

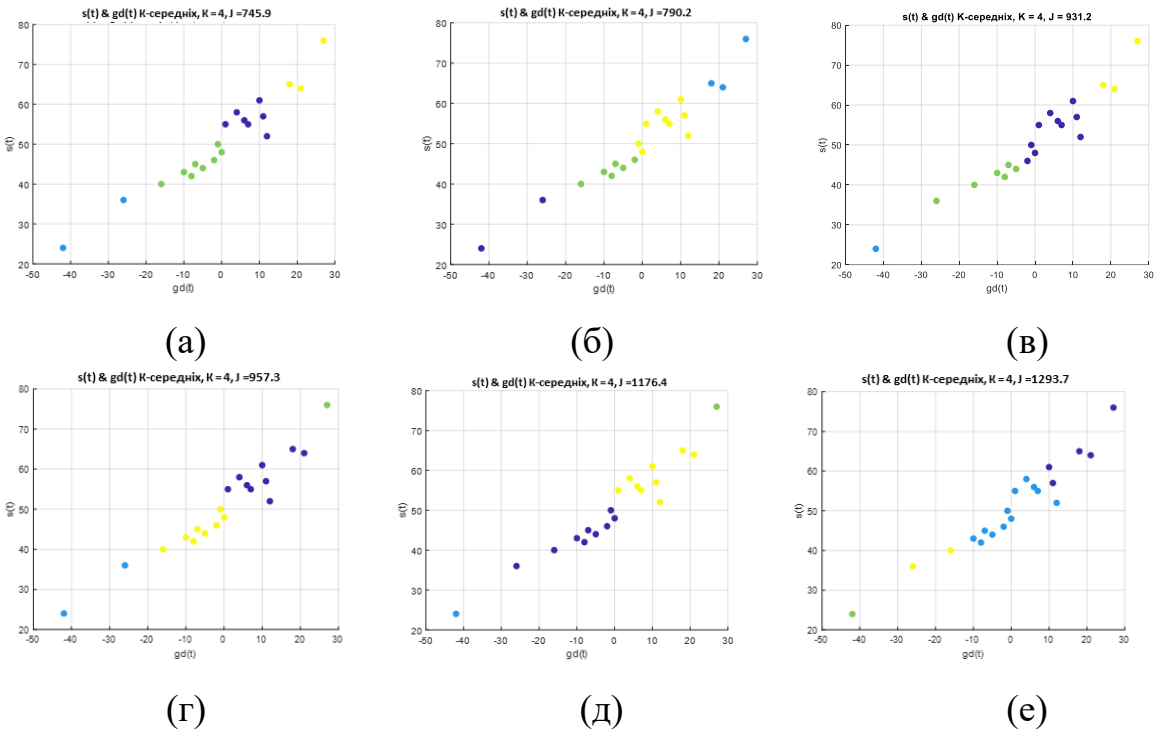
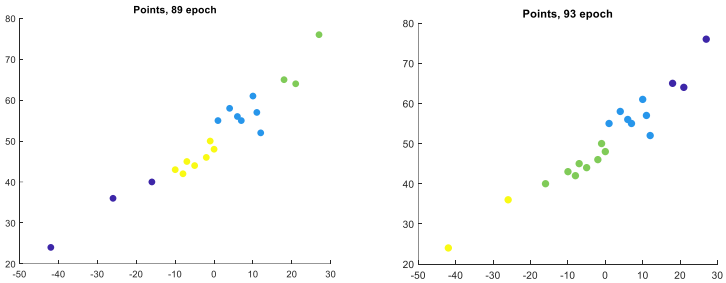


Рисунок 2.5 – Результати кластеризації команд методом *K*-середніх за змінними $s(t)$ та $gd(t)$

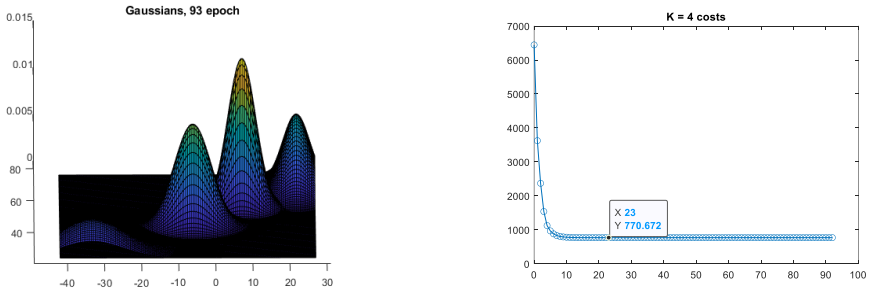
На рис. 2.6 наведено можливі варіанти результатів кластеризації за методом Гаусівських сумішей з параметром регуляризації $\gamma = 0$, тобто з повною регуляризацією недіагональних елементів коваріаційних матриць. Результат кластеризації, наведений на рис. 2.6, б збігається з найкращим результатом за методом *K*-середніх (рис. 2.5 а). На рис. 2.7, а і рис. 2.7, б для

результату кластеризації рис. 2.6, б наведено вигляд отриманої Гаусівської суміші і процес збіжності цільової функції відповідно. Кластери мають еліптичний вигляд. Мінімальне значення цільової функції (2.2) дорівнює 770,7.



(а) (б)

Рисунок 2.6 – Результати кластеризації команд методом Гаусівських сумішей при $\gamma = 0$ за змінними $s(t)$ та $gd(t)$



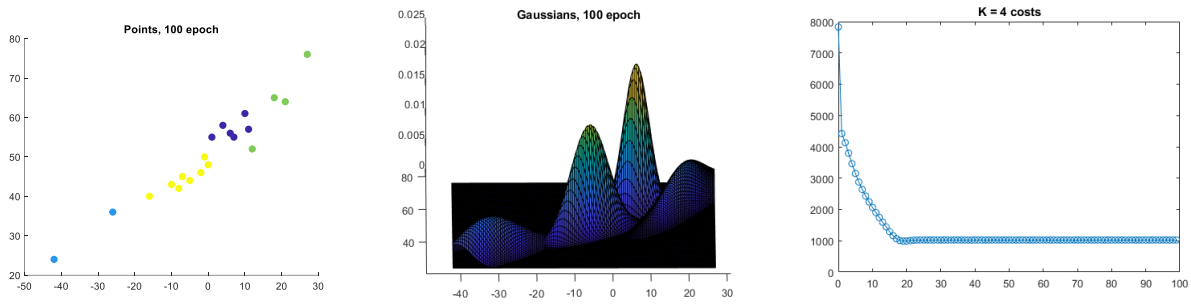
(а) (б)

Рисунок 2.7 – Гаусівська суміш і процес збіжності цільової функції з параметром регуляризації $\gamma = 0$

У той же час результат, зображений на рис. 2.6, а надає можливість отримати однакову кількість точок у крайніх кластерах. Значення цільової функції (2.2) для цього результату кластеризації дорівнює 914,3. Отже, за значенням цільової функції (2.2) найкращим результатом серед розглянутих для випадку $\gamma = 0$ є результат кластеризації, зображений на рис. 2,6, б.

Найкращий результат в даному випадку повністю співпав з найкращим результатом для методу *K*-середніх. При цьому, метод Гаусівських сумішей продемонстрував більш стабільну поведінку: можливих варіантів результату було отримано всього 2, у той час як за методом *K*-середніх було отримано 6 різних варіантів результату кластеризації.

На рис. 2.8, а наведено результат кластеризації за методом Гаусівських сумішей з параметром регуляризації $\gamma = 0{,}9$, тобто з частковою регуляризацією недіагональних елементів коваріаційних матриць. На рис. 2.8, б і рис. 2.8, в наведено вигляд отриманої Гаусівської суміші та процес збіжності цільової функції відповідно. У цьому випадку метод Гаусівських сумішей стабільно збігався до єдиного можливого результату кластеризації.



(а) (б) (в)

Рисунок 2.8 – Результати кластеризації команд методом Гаусівських сумішей при $\gamma = 0{,}9$

Запропоноване урахування недіагональних елементів коваріаційних матриць забезпечило вигляд кластерів еліптичної форми, що призвело до більш витончених результатів кластеризації, ніж раніше отримані результати, найкращі за значенням цільових функцій. Так, до найсильнішого кластера (зелений колір) додалась команда, яка є близькою до найкращих трьох команд за параметром різниці забитих та пропущених м'ячів $sd(t)$. Значення цільової функції (2.2) для отриманого результату кластеризації дорівнює 1024,0, а цільової функції (2.1) — 973,6. Наведений приклад показує, що запропонована

регуляризація може забезпечити отримання єдиного результату кластеризації незалежно від початкових умов, який водночас дозволяє врахувати неочевидні кореляційні зв’язки між точками з набору даних, що потенційно може покращити подальше дослідження наявних даних.

2.2. Формування початкових даних для побудови імітаційної моделі футбольного сезону

Для побудови та дослідження ефективності методів виявлення підозрілих щодо фіксованого результату матчів і їх порівняльного аналізу необхідно розробити імітаційну модель футбольного сезону, що враховує наявність таких матчів.

Розрізняють два типи матчів з фіксованим результатом: пов’язані з підкупом команд з метою заробітку на ставках у букмекерських конторах і ті, що переслідують турнірні цілі [14, 16, 17, 22]. Важливо зазначити, що саме перший тип договірних матчів пов'язаний із криміналом, незаконним збагаченням та викликає максимальну тривогу FIFA та ООН [16].

В подальшому у цьому розділі розглянемо алгоритм моделювання договірних матчів, пов’язаних із заробітком на ставках, з використанням якого формуються матчі, результати яких відмінні від очікуваних і можуть розглядатися як аномальні.

Початковими даними є дані реального сезону. Він складається з двох таблиць: таблиці матчів і таблиці команд. Опис параметрів цих таблиць наведено у табл. 2.3 і табл. 2.4 відповідно. У таблиці матчів дані упорядковані за датою проведення матчу D_k. У таблиці команд вони упорядковані за кількістю набраних очок $p(t_i)$. Деякі параметри табл. 2.4, а саме: загальні очки команд, кількості виграних, нічийних та програних матчів — можна порахувати із таблиці матчів (табл. 2.3).

Як реальний сезон використовувався сезон 2013–2014 років Ліги II Франції. Групування команд сезону показано на рис. 2.4.

Таблиця 2.3 – Опис параметрів таблиці матчів

k	Номер матчу
D_k	Дата проведення матчу
h_k	Ідентифікатор (або назва) команди-господарки матчу
o_k	Ідентифікатор (або назва) гостьової команди матчу
α_k	Кількість голів, забитих командою h_k
β_k	Кількість голів, забитих командою o_k

Таблиця 2.4 – Опис параметрів турнірної таблиці

i	Номер команди
t_i	Ідентифікатор (або назва) команди
$p(t_i)$	Загальні очки команди
$w(t_i)$	Кількість ігор команди, які завершились перемогою цієї команди
$d(t_i)$	Кількість ігор команди, які завершились нічиєю
$l(t_i)$	Кількість ігор команди, які завершились поразкою цієї команди
$group(t_i)$	Ранг команди (група, до якої належить команда)

На основі таблиці матчів і таблиці команд реального сезону обчислюються частоти голів, що забивають команди у матчі. Для кожної команди, обчислюється два типи частот. Перший тип частот (домашній) $\nu\left(X_i \middle| group\left(t_j\right) = g\right)$ — це частоти того, що команда t_i заб'є $X_i = x_i, x_i = \overline{0, K_i}$ голів на власному полі за умови, що команда супротивника t_j належить групі g. Другий тип частот (виїзний) $\nu\left(Y_j \middle| group(t_i) = d\right)$ — це частоти того, що команда t_j заб'є $Y_j = y_j$, $y_j = \overline{0, K_j}$ голів на виїзді за умови, що команда супротивника t_i належить групі d. У цих формулах під K_i і K_j розуміється максимальна кількість голів в аналізуємому реальному сезоні, які відповідно забила команда t_i в домашніх іграх та команда t_j у виїзних іграх. Таким чином, для кожної команди отримуємо чотири множини частот першого типу і

чотири множини частот другого типу. На основі отриманих частот розраховуються параметри теоретичних законів розподілу ймовірностей забиття голів командами під час їх гри з іншими командами відповідного рангу.

Вхідними даними алгоритму моделювання сезону є:

- N команд реального сезону з атрибутами, описаними в табл. 2.4;

- $P_g(X_i) = P\left(X_i \middle| group\left(t_j\right) = g\right),\ g = \overline{1,4}, i, j = \overline{1,N}$ — ймовірність того, що команда t_i заб'є $X_i = x_i$, $x_i = \overline{0,K_i}$ голів на власному полі при умові, що команда супротивника t_j належить групі g.

- $\check{P}_d\left(Y_j\right) = P\left(Y_j \middle| group(t_i) = d\right),\ d = \overline{1,4}, i, j = \overline{1,N}$ — ймовірність того, що команда t_j заб'є $Y_j = y_j$, $y_j = \overline{0,K_j}$ голів на виїзді при умові, що команда супротивника t_i належить групі d.

Результатом роботи алгоритму є модельний сезон, сформований на основі заданого реального сезону футбольного турніру. Модельний сезон складається з таблиці матчів (табл. 2.3) та турнірної таблиці сезону (табл. 2.4).

Припущення алгоритму є такими:

1) кількість голів, які команда-господарка заб'є під час гри, має розподіл Пуассона;

2) номер і група команди розглядаються як попередні оцінки її сили.

Розробку імітаційної моделі футбольного сезону розглянемо в такій послідовності, кожний пункт якої відповідає окремому підрозділу даного розділу:

1. Розрахунок ймовірностей забиття голів командами під час гри на основі реальних даних сезону.

2. Розробка імітаційної моделі футбольного сезону без договірних матчів і її аналіз шляхом статистичного моделювання.

3. Розробка алгоритму моделювання договірних матчів, пов'язаних із заробітком на ставках, і його аналіз.

Розроблені в даному розділі моделі та алгоритми, за необхідності та за наявності відповідних вхідних даних, можуть бути узагальнені для врахування

інших параметрів футбольних матчів, на які приймаються ставки, — кількість призначених пенальті, кількість попереджень і вилучень тощо.

2.3. Розрахунок ймовірностей забиття голів командами під час гри на основі даних реального сезону

Однією з найбільш розповсюджених моделей для прогнозування результату футбольних матчів є розподіл Пуассона [43]. Він використовується для розрахунку ймовірностей кількості голів, які команда забиває у матчі. При цьому використовується припущення, що гол є незалежною подією, оскільки він не впливає на ймовірність того, скільки голів буде забито у подальшому.

При розрахунку імовірностей голів, забитих командою, враховується тип гри — домашня або виїзна, а також сила команди супротивника, тобто до якої групи вона відноситься. Алгоритм побудови розподілів Пуассона числа голів команд сезону складається з таких етапів:

1. Формуємо множину значень голів, забитих командою t_i протягом ігор сезону. Отримуємо множину $G_i^{(g)} = \{k_i^{(g)},\ k_i^{(g)} + 1,\ \dots,\ K_i^{(g)}\}$, де $k_i^{(g)}$ та $K_i^{(g)}$ — відповідно мінімальна та максимальна кількість голів, забитих командою t_i протягом усіх її домашніх ігор сезону, у яких команда-суперник t_j належала групі g.

2. $\forall k \in G_i^{(g)}$ обчислюємо частоту появи кількості голів k як голів, забитих командою t_i протягом ігор сезону, у яких команда t_i була домашньою, а команда-суперник t_j належала групі g:

$$\nu_g(X_i = k) = \nu\left(X_i = k \middle| group(t_j) = g\right) = \frac{|\{m_{ij} | X_i = k, group(t_j) = g\}|}{|\{m_{ij} | group(t_j) = g\}|},$$

де m_{ij} — матч між командами (t_i, t_j), де t_i є домашньою командою матчу, а t_j є виїзною командою матчу; $\{m_{ij} | X_i = k, group(t_j) = g\}$ — множина, що містить лише ті домашні матчі команди t_i, у яких вона забила $X_i = k$ голів і

суперник t_j належав до групи g; $\{m_{ij}|group(t_j) = g\}$ — множина, що містить усі домашні матчі команди t_i, у яких суперник t_j належав до групи g, а знак модуля над множиною підраховує кількість елементів у цій множині.

Відносні частоти $\check{\nu}_d(Y_j = k)$ дискретної випадкової величини кількості голів Y_j, забитих командою t_j протягом ігор сезону, у яких команда t_j була виїзною, обчислюються аналогічно:

$$\check{\nu}_d(Y_j = k) = \nu(Y_j = k|group(t_i) = d) = \frac{|\{m_{ij}|Y_j = k, group(t_i) = d\}|}{|\{m_{ij}|group(t_i) = d\}|},$$

де $\{m_{ij}|Y_j = k, group(t_i) = d\}$ — множина, що містить лише ті виїзні матчі команди t_j, у яких вона забила $Y_j = k$ голів, і суперник t_i належав до групи d; $\{m_{ij}|group(t_i) = d\}$ — множина, що містить усі виїзні матчі команди t_j, у яких суперник t_i належав до групи d.

3. Згідно з оцінкою за методом максимальної правдоподібності [162], параметр $\lambda_i^{(g)}$ розподілу Пуассона для домашніх матчів команди t_i, у яких суперник належав групі g, визначається за формулою:

$$\lambda_i^{(g)} = \sum_{k=0}^{K_i} k \cdot \nu_g(X_i = k).$$

Параметр $\check{\lambda}_j^{(d)}$ розподілу Пуассона для виїзних матчів команди t_j, у яких суперник належав групі d, визначається за аналогічною формулою:

$$\check{\lambda}_j^{(d)} = \sum_{k=0}^{K_j} k \cdot \check{\nu}_{gd}(Y_j = k).$$

Розподіли Пуассона кількостей забитих голів командою t_i в її домашніх матчах і командою t_j в її виїзних матчах визначаються за формулами:

$$P(X_i = x_i) = \frac{e^{-\lambda_i^{(g)}} {\lambda_i^{(g)}}^{x_i}}{x_i!};$$

$$\check{P}_d\left(Y_j = y_j\right) = \frac{e^{-\check{\lambda}_j^{(d)}} \check{\lambda}_j^{(d)^{y_j}}}{y_j!}.$$

4. Для оцінювання наскільки гарною є відповідність між фактичним числом матчів з тією чи іншою кількістю голів та моделлю Пуассона використовуємо критерій Хі-квадрат. Для перевірки розподілу ймовірності Пуассона числа голів у домашніх матчах команди t_i, суперник якої належить групі g, статистика Хі-квадрат визначається за формулою [163]:

$$\chi^2_{i,g} = \sum_{k=0}^{n} \frac{(n_{ig}(k) - \tilde{n}_{ig}(k))^2}{\tilde{n}_{ig}(k)},$$

де $\tilde{n}_{ig}(k) = P_g(X_i = k)M_i^{(g)}$ та $n_{ig}(k) = \nu_g(X_i = k)M_i^{(g)}$ є, відповідно, прогнозованою та фактичною абсолютними частотами появи матчів команди t_i, у яких ця команда забила k голів та суперник t_j належав до групи g, $M_i^{(g)} = |\{m_{ij}|group(t_j) = g\}|$ — кількість усіх домашніх матчів команди t_i, у яких суперник t_j належав до групи g.

Статистика Хі-квадрат для виїзних матчів команди t_j, у яких суперник t_i належав до групи d, визначається аналогічно.

Критичне значення $\chi^2_\alpha(n)$ визначається на рівні значущості $\alpha = 0{,}05$.

Кількість степенів свободи $n = K_i^{(g)} - k_i^{(g)}$.

При виконанні умови $\chi^2_{i,g} \leq \chi^2_\alpha(n)$ вважається, що отриманий розподіл достатньо точно описує випадкову величину.

Якщо розподіл Пуассона не точно описує випадкову величину, можуть бути використані такі розподіли [43]:

- розподіл Пуассона з розширеною кількістю нулів;
- від'ємний біноміальний розподіл;
- геометричний розподіл;
- рівномірний розподіл.

Параметри вказаних розподілів можуть бути знайдені за методом найменших квадратів, методом максимальної правдоподібності або методом моментів. Адекватність отриманих розподілів також перевіряється за критерієм Хі-квадрат.

2.4. Імітаційна модель футбольного сезону та її аналіз

Побудову моделі футбольного сезону для імітації наявності договірних матчів, пов'язаних із заробітком на ставках, проведемо в три етапи.

І етап: отримання початкової реалізації модельного сезону

1. Використовуючи функції ймовірності $P_g(X_i)$ та $\check{P}_d(Y_j)$, моделюємо результат матчу (t_i, t_j) між командами t_i та t_j.

1.1. Генеруємо початкову кількість голів x_i, забиту командою t_i як значення випадкової величини з функцією ймовірності $P_g(X_i)$

$$S_{ix_i} \le r_i < S_{i(x_i+1)},$$

де

$$S_{in+1} = \sum_{k=0}^{n} P_g(X_i = k), S_{i0} = 0, S_{i(K_i+1)} = 1,$$

де S_{in} — ймовірність того, що випадкова величина R прийме значення $r_i < S_{in}$; початкове значення кількості голів x_i визначається за нижньою границею S_{ix_i} інтервалу, у який потрапляє випадкове число r_i.

1.2. Враховуючи частоту нічийних матчів домашньої команди в реальному сезоні, визначаємо — цей матч завершиться внічию чи ні. Для цього генеруємо бінарну випадкову величину, значення якої «1» відповідає події «матч закінчиться внічию», а «0» — «матч завершиться перемогою якоїсь команди», причому ймовірність події «1» для команди t_i розраховується як $d(t_i)/(w(t_i) + d(t_i) + l(t_i))$.

1.3. Якщо згенеровано значення «1», тоді матч закінчується нічиєю і кількість голів, забитих виїзною командою $y_j = x_i$.

1.4. Якщо отримуємо значення «0», то генеруємо початкову кількість голів y_j, забиту командою t_j як значення випадкової величини з функцією ймовірності $\check{P}_d(Y_j)$. Кількості голів y_j визначаються шляхом визначення інтервалу, у який потрапляє значення r_j рівномірно розподіленої на інтервалі [0; 1] випадкової величини *R*:

$$\check{S}_{jy_j} \le r_j < \check{S}_{j(y_j+1)},$$

де

$$\check{S}_{jn+1} = \sum_{k=0}^{n} \check{P}_d(Y_j = k), \check{S}_{j0} = 0, \check{S}_{j(K_j+1)} = 1,$$

де $\check{S}_{jn}$ — ймовірність того, що випадкова величина *R* прийме значення $r_j < \check{S}_{jn}$. Початкове значення кількості голів y_j визначається за нижньою границею $\check{S}_{jy_j}$ інтервалу, у який потрапляє випадкове числа r_j. Якщо при цьому $y_j = x_i$, тоді збільшуємо результат у сильнішої команди на 1 м'яч. Тобто, якщо $g \le d$, тоді збільшуємо x_i на 1, інакше — збільшуємо y_i на 1.

2. Повторюємо крок 1 для всіх $N(N-1)$ матчів, де *N* — це кількість команд, які приймали участь у цьому сезоні.

В підсумку, маємо модельний сезон футбольного турніру

$$M = \{m_{ij} = (x_i, y_j),\ i, j = 1, N,\ i \ne j\},$$

де x_i є кількістю голів, забитих командою t_i у матчі m_{ij}, y_j є кількістю голів, забитих командою t_j у матчі m_{ij}.

II етап: обчислення очок команд

1. Обчислюємо попередні очки кожної команди $p(t_i)$, $i = 1, N$:

$$p(t_i) = 3w(t_i) + d(t_i), w(t_i) = \left|\{m_{ij} | x_i > y_j\}\right|, d(t_i) = \left|\{m_{ij} | x_i = y_j\}\right|,$$

де $w(t_i)$, $d(t_i)$ є відповідно кількістю перемог та ігор у нічию для команди t_i.

III етап: групування команд за отриманими очками

Для групування команд модельного сезону використовується та сама методика, яка була застосована для групування команд реального сезону.

Проведено аналіз розробленого алгоритму шляхом статистичного моделювання і порівняння отриманих результатів з реальним сезоном. Спочатку було перевірено якість моделювання сезонів на основі етапів I та II розробленої моделі без спотворення результатів матчів. Для дослідження адекватності запропонованого методу імітаційного моделювання було змодельовано 100 сезонів.

На рис. 2.9 наведено абсолютні частоти типів результатів за класами матчів для реального сезону. Також на рис. 2.9 наведено середні значення абсолютних частот типів результатів по 100 змодельованих сезонам, згруповані за класами матчів. При цьому, в кожному модельному сезоні використовувалося групування команд, яке було отримано для команд з реального сезону. Як випливає з рис. 2.9, в змодельованих сезонах в цілому зберігається якісний характер залежностей між кількістю домашніх перемог, нічиїх та домашніх поразок за класами матчів реального сезону.

Також була здійснена перевірка достовірності різниці між типами результатів змодельованого і реального сезонів для кожного класу матчів. Перевірка здійснювалась за критерієм Колмогорова-Смирнова [164] на рівні значущості $\alpha = 0{,}001$. Критичне значення статистики $\lambda_\alpha = 1{,}95$. Результати застосування критерію для кожного класу матчів наведено в табл. 2.5. Критерій полягає в перевірці такої нульової гіпотези $H_0: F_r(x) = F_m(x)$, де $F_r(x)$ є емпіричною функцією розподілу вибірки типів результату матчів реального сезону, а $F_m(x)$ — відповідно змодельованого сезону. Гіпотеза H_0 виконується за умови $\lambda \leq \lambda_\alpha$. За обраного рівня значущості для усіх класів матчів різниця між типами результатів змодельованого і реального сезонів виявилася статистично незначущою.

На рис. 2.10 наведено абсолютні частоти типів результатів реального сезону. Також на рис. 2.10 наведено середні значення абсолютних частот типів результатів по 100 змодельованих сезонам. Сумарне відхилення від реального сезону за типами результату матчу у змодельованих сезонах становить 13%.

Перевіримо достовірність різниці двох наведених вибірок різниць голів за критерієм Колмогорова-Смирнова [164] з тим же рівнем значущості $\alpha = 0{,}001$, що й раніше. Для розглянутих вибірок статистика критерію $\lambda = 0{,}32$, критичне значення статистики $\lambda_\alpha = 1{,}95$. Отже, різниця між розглянутими загальними вибірками типів результату матчів є статистично незначущою на рівні значущості $\alpha = 0{,}001$. Таким чином, за розподілом типів результатів по сезону змодельований сезон є подібним до реального.

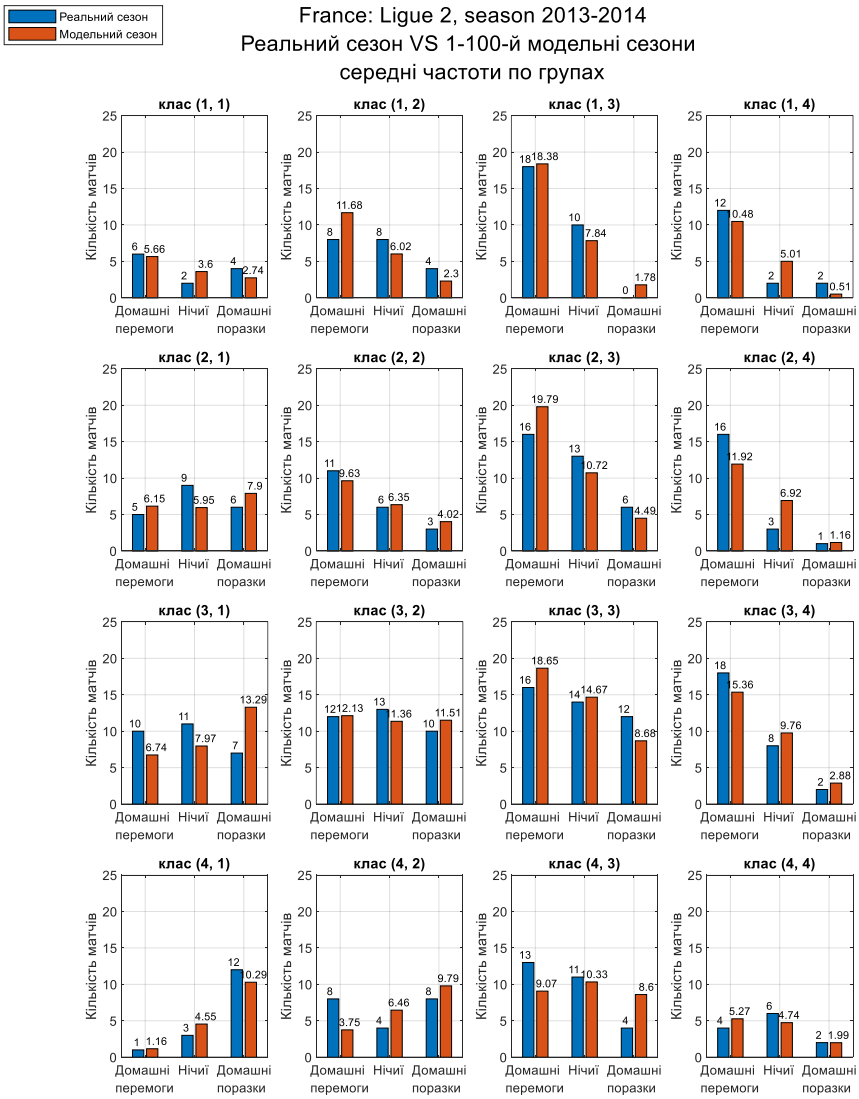


Рисунок 2.9 – Розподіл типів результатів матчів за класами матчів у змодельованих сезонах

Таблиця 2.5 – Результати перевірки критерію Колмогорова-Смирнова за рівня значущості $\alpha = 0{,}001$ для встановлення значущості різниці між типами результатів змодельованого і реального сезонів по класам матчів

Клас матчів	Значення статистики λ	Клас матчів	Значення статистики λ
(1,1)	0,339	(3,1)	1,910
(1,2)	0,598	(3,2)	0,311
(1,3)	0,876	(3,3)	0,442
(1,4)	0,102	(3,4)	0,749
(2,1)	1,319	(4,1)	1,824
(2,2)	0,732	(4,2)	0,777
(2,3)	0,077	(4,3)	0,317
(2,4)	0,495	(4,4)	0,073

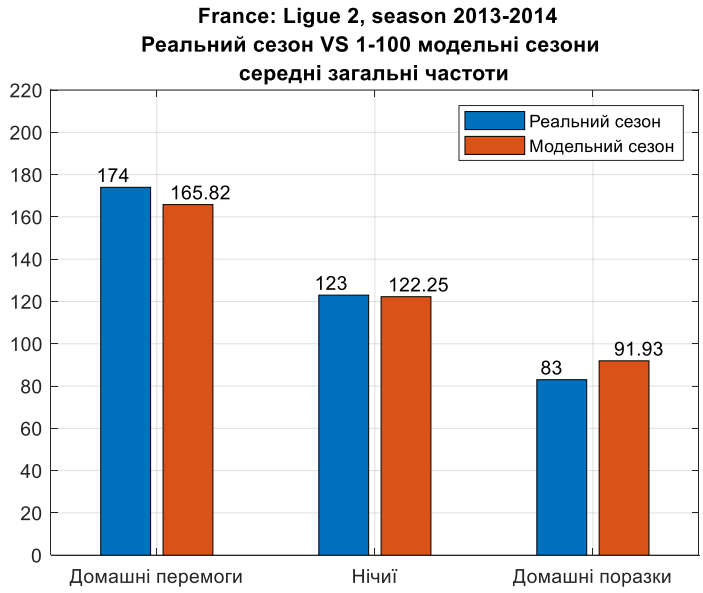


Рисунок 2.10 – Розподіл типів результатів матчів загалом у змодельованих сезонах

На рис. 2.11 наведено гістограми різниць голів у матчах реального і одного змодельованого сезонів. Перевіримо достовірність різниці двох

наведених вибірок різниць голів за критерієм Колмогорова-Смирнова [164]. За рівень значущості, як і раніше, візьмемо величину $\alpha = 0{,}001$. Критерій Колмогорова-Смирнова полягає в перевірці такої нульової гіпотези H_0: $F_r(x) = F_m(x)$, де $F_r(x)$ є емпіричною функцією розподілу вибірки різниць голів реального сезону, а $F_m(x)$ — відповідно змодельованого сезону. Для розглянутих вибірок статистика критерію $\lambda = 0{,}58$, критичне значення статистики $\lambda_\alpha = 1{,}95$.

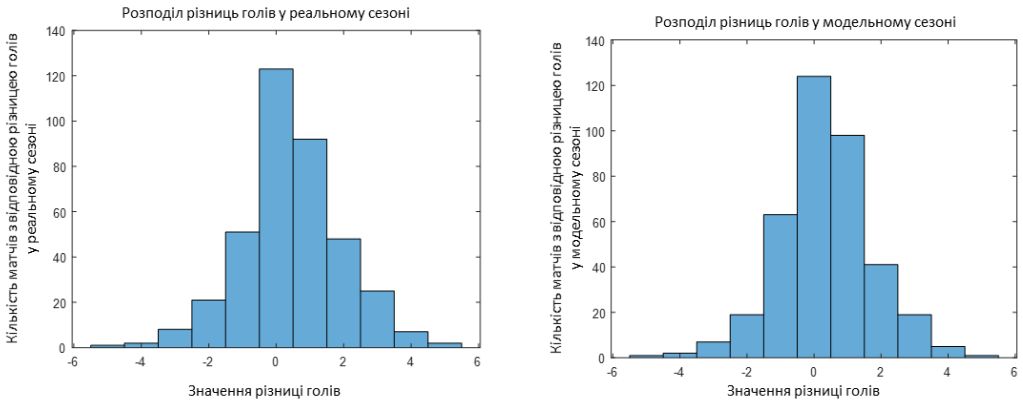


Рисунок 2.11 – Розподіл різниць голів реального і змодельованого сезонів

Отже, різниця між розглянутими вибірками різниць голів є статистично незначущою на рівні значущості $\alpha = 0{,}001$. Таким чином, за розподілом різниць голів усіх матчів змодельований сезон є подібним до реального сезону.

2.5 Алгоритм моделювання договірних матчів, пов'язаних із заробітком на ставках

Результат договірного матчу, щоб на ставках на ньому могли заробити зловмисники, повинен відрізнятися від очікуваного результату, тобто носити аномальний характер. Це його найважливіша властивість. Тому його значення не повинно міститися в області очікуваних значень результату матча, пов'язаних із силою команд. Запропоновано ввести порогове значення ймовірності p_A, яке не повинні перевищувати аномальні різниці голів матчів

класу і таким чином серед яких можуть знаходитися результати договірних матчів. Значення ймовірності p_A із евристичних міркувань запропоновано обирати в діапазоні $0 < p_A < 0{,}4$.

Алгоритм моделювання результатів договірних матчів, пов'язаних із заробітком на ставках, складається з таких етапів.

1. З використанням отриманих модельних сезонів турніру будуються гістограми різниць голів усіх матчів окремо для кожного класу. Приклади гістограм, отриманих по 100 модельних сезонах, наведено на рис. 2.12. Як випливає з цього рисунку, гістограми мають очікувані закономірності в результатах матчів (сильніші команди мають кращі результати в грі зі слабшими командами; вдома команди грають краще, ніж на виїзді).

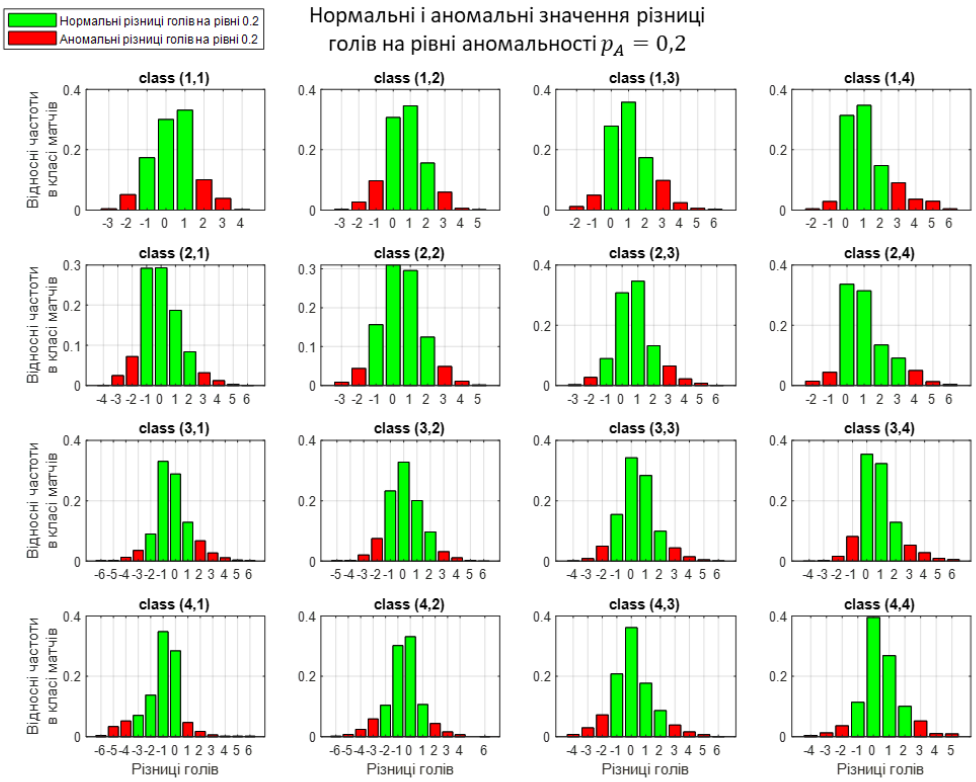


Рисунок 2.12 – Гістограми різниць голів по кожному класу матчів

2. Використовуючи гістограму різниць голів, збудовану за класами матчів за 100 змодельованими сезонами, визначимо множину нормальних

різниць голів $D_{ij}^{(N)}$ за допомогою ітераційного алгоритму, в якому спочатку вважатимемо, що $D_{ij}^{(N)}$ є порожньою множиною, тобто $D_{ij}^{(N)} = \emptyset$.

2.1. Обираємо значення різниці голів $\tilde{d}$, яке за гістограмою класу матчів (i, j) має найбільшу частоту появи h_d серед тих значень d, що $d \notin D_{ij}^{(N)}$.

2.2. Додаємо значення $\tilde{d}$ до множини $D_{ij}^{(N)}$.

2.3. Обчислюємо сумарну частоту появи усіх значень з множини $D_{ij}^{(N)}$:

$p_{ij}^{(N)} = \sum_{d \in D_{ij}^{(N)}} h_d$.

2.4. Якщо $p_{ij}^{(N)} \geq 1 - p_A$, переходимо на крок 2.5. Інакше, переходимо на крок 2.1.

2.5. Значення можливих різниць голів $d^* \notin D_{ij}^{(N)}$ утворюють множину $D_{ij}^{(A)}$ аномальних різниць класу матчів (i, j).

Значення множини $D_{ij}^{(A)}$ використовуються для моделювання результату договірного матчу.

Для прикладу на рис. 2.12 наведені нормальні (зелений колір) і аномальні (червоний колір) різниці голів по кожному класу матчів при рівні аномальності $p_A = 0{,}2$. Результати вибору значень за гістограмою з рис. 2.12 при рівні аномальності $p_A = 0{,}2$ відображено в табл. 2.6.

3. Вибирається N матчів сезону турніру випадковим чином, які будуть перетворюватися на договірні. Далі розглядатимемо всі дії на прикладі одного такого матчу.

4. Визначається клас матчів, до якого належить даний матч.

5. Якщо результат обраного матчу не міститься в області очікуваних значень результату матчів цього класу матчів (тобто цей результат і так схожий на договірний матч), то випадковим чином обираємо інший матч. Інакше переходимо до п. 6.

6. Різниця голів обирається випадково з множини аномальних значень різниць голів для відповідного класу матчів (табл. 2.6). При цьому випадковому виборі використовуються ймовірності, пропорційні частотам аномальних значень з гістограми різниць голів для відповідного класу матчів.

Таблиця 2.6 – Нормальні й аномальні різниці м’ячів

Клас матчів	Аномальні різниці м’ячів	Нормальні різниці м’ячів
(1,1)	{-3, -2, 2, 3, 4}	{-1, 0, 1}
(1,2)	{-3, -2, -1, 3, 4, 5}	{0, 1, 2}
(1,3)	{-2, -1, 3, 4, 5, 6}	{0, 1, 2}
(1,4)	{-2, -1, 3, 4, 5, 6}	{0, 1, 2}
(2,1)	{-4, -3, -2, 3, 4, 5, 6}	{-1, 0, 1, 2}
(2,2)	{-3, -2, 3, 4, 5}	{-1, 0, 1, 2}
(2,3)	{-3, -2, 3, 4, 5, 6}	{-1, 0, 1, 2}
(2,4)	{-2, -1, 4, 5, 6}	{0, 1, 2, 3}
(3,1)	{-6, -5, -4, -3, 2, 3, 4, 5, 6}	{-2, -1, 0, 1}
(3,2)	{-5, -4, -3, -2, 3, 4, 5, 6}	{-1, 0, 1, 2}
(3,3)	{-4, -3, -2, 3, 4, 5, 6}	{-1, 0, 1, 2}
(3,4)	{-4, -3, -2, -1, 3, 4, 5, 6}	{0, 1, 2}
(4,1)	{-6, -5, -4, 1, 2, 3, 4, 5, 6}	{-3, -2, -1, 0}
(4,2)	{-6, -5, -4, -3, 2, 3, 4, 6}	{-2, -1, 0, 1}
(4,3)	{-4, -3, -2, 3, 4, 5, 6}	{-1, 0, 1, 2}
(4,4)	{-4, -3, -2, 3, 4, 5}	{-1, 0, 1, 2}

7. Рахунок договірного матчу обирається випадково з відповідного стовпчика таблиці рахунків матчів (див. табл. 2.7), де номер стовпця таблиці дорівнює обраній різниці голів договірного матчу. При цьому випадковому виборі використовуються ймовірності на основі арифметичної прогресії з

$a_1 = 1$ і $a_n = 0{,}1, n \geq 2$ (рис. 2.13):

$$P(j) = \frac{a_j}{S} = \frac{a_1 + \frac{(a_n - a_1)}{(n-1)}(j-1)}{\frac{(a_1 + a_n)n}{2}} = \frac{1 + \frac{(0{,}1 - 1)}{(n-1)}(j-1)}{\frac{(1 + 0{,}1)n}{2}} =$$

$$= \frac{1 - \frac{0{,}9(j-1)}{(n-1)}}{0{,}55n} = \frac{(n - 1 - 0{,}9j + 0{,}9)}{0{,}55n(n-1)} = \frac{(n - 0{,}1 - 0{,}9j)}{0{,}55n(n-1)} =$$

$$= \frac{(n - 1 + 1 - 0{,}1 - 0{,}9j)}{0{,}55n(n-1)} = \frac{1}{0{,}55n} + \frac{(0{,}9 - 0{,}9j)}{0{,}55n(n-1)},$$

$$1 \leq j \leq n, n \geq 2;$$

$$P(1) = 1, n = 1,$$

де n — кількість варіантів рахунку у вибраному стовпчику таблиці рахунків матчів, j — порядковий номер рахунку у вибраному стовпчику таблиці рахунків матчів, S — сума перших n членів арифметичної прогресії.

Таблиця 2.7 – Теоретично можливі рахунки матчів для всіх класів матчів, коли однією командою було забито не більше 6 м’ячів

	Різниця голів матчу						
	0	1	2	3	4	5	6
Рахунок матчу	0:0	1:0	2:0	3:0	4:0	5:0	6:0
	1:1	2:1	3:1	4:1	5:1	6:1	
	2:2	3:2	4:2	5:2	6:2		
	3:3	4:3	5:3	6:3			
	4:4	5:4	6:4				
	5:5	6:5					
	6:6						

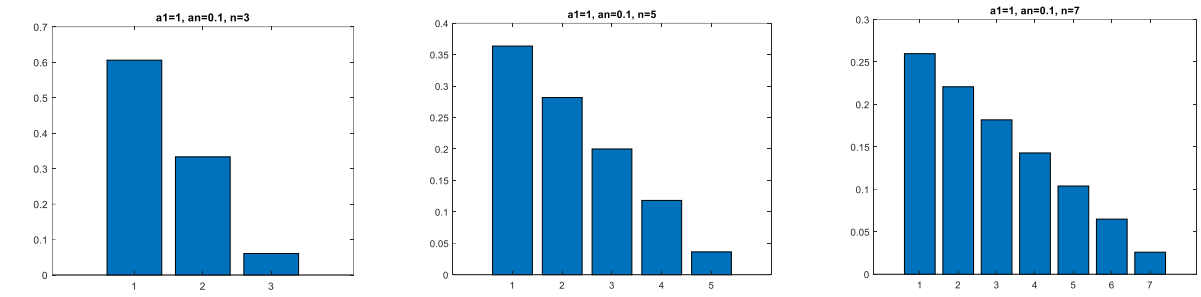


Рисунок 2.13 – Ймовірності на основі арифметичної прогресії з $a_1 = 1$ і $a_n = 0{,}1$ для $n \in \{3, 5, 7\}$

При n = 1 обрана функція ймовірності є сталою функцією зі значенням $P(1) = 1$. Для $n \geq 2$ обрана функція ймовірності є спадною ступінчатою функцією з початковим значенням ймовірності $P(1) = \frac{1}{0{,}55n}$ і кроком $\frac{0{,}9(1-j)}{0{,}55n(n-1)}$. Обрана функція ймовірності забезпечує поступове зменшення ймовірності зі збільшенням порядкового номеру результату матчу. Елементи стовпця таблиці рахунків матчів розташовані у порядку збільшення кількостей голів, забитих командами під час матчу. Отже, чим більше голів кожна команда заб'є під час гри, тим менша ймовірність обрання такого варіанту результату матчу як кандидата для формування фіксованого результату. Визначення апріорно невідомих ймовірностей як членів арифметичної прогресії використовується в задачах дослідження операцій як один з методів подолання апріорної невизначеності [165].

Якщо отриманого результату матчу не було в реальному сезоні, то відбувається повтор п. 6 та п. 7.

Результати моделювання матчів із фіксованим результатом продемонстровано на прикладі одного із змодельованих сезонів (див. табл. 2.8). Випадковим чином у змодельованому сезоні було обрано $T_s = 10$ матчів. Було використано рівень аномальності $p_A = 0{,}2$. Для кожного матчу сформовано результат за алгоритмом формування договірних матчів у імітаційній моделі у відповідності до аномальних різниць, визначених для кожного класу матчів у таблиці 2.6.

Таблиця 2.8 – Приклади утворених договірних матчів у модельному сезоні

Домашня команда	Виїзна команда	Оновлений результат	Початковий результат
team 3	team 13	4:1	0:0
team 2	team 15	4:1	3:2
team 4	team 19	0:1	3:2
team 9	team 7	0:2	2:2
team 11	team 7	4:0	2:3
team 16	team 19	3:0	1:1
team 17	team 2	4:1	0:0
team 17	team 7	5:2	1:2
team 17	team 15	3:0	4:4
team 17	team 19	4:0	4:3

Серед отриманих результатів зустрічаються такі, які можуть виглядати як очікувані, наприклад, 0:1 чи 0:2. Але для відповідних класів матчів, тобто у відповідному контексті, такі результати розглядаються як аномальні.

Висновки до розділу 2

1. Кластеризація дозволяє виділити групи однорідності команд за їх силою за результатами сезону. Тоді на основі контекстуальних атрибутів матчі турніру можна буде розбити на класи і вже в кожному класі матчів за поведінковим атрибутом визначати аномальні матчі. При одновимірній кластеризації команд враховані правила переходу у футбольні ліги вищого і нижчого рівнів і використано метод *K*-середніх, оскільки взаємна кореляція даних в одновимірному випадку відсутня. З двох отриманих при різних початкових умовах варіантів кластеризації обрано варіант групування з мінімальним значенням цільової функції, який також зустрічався найчастіше.

2. При двовимірній кластеризації команд результат, отриманий методом Гаусівських сумішей з параметром регуляризації $\gamma = 0$ збігається з найкращим за значенням цільової функції результатом, який найбільш часто отримується з використанням методу *K*-середніх при його запуску за різних початкових умов. При цьому чутливість методу Гаусівських сумішей до початкових умов значно нижче, а отримані кластери мають колоподібний вигляд.

3. Урахування запропонованих недіагональних елементів коваріаційних матриць в методі Гаусівських сумішей шляхом регуляризації з параметром $\gamma = 0{,}9$ забезпечує отримання єдиного результату кластеризації незалежно від початкових умов. При цьому кластери мають еліпсоподібний вигляд, що дозволяє врахувати неочевидні кореляційні зв'язки між точками з набору даних.

4. Особливістю розробленої імітаційної моделі футбольного сезону з матчами з фіксованим результатом є те, що команди поділяються на групи, які враховують їх силу за загальними очками в сезоні. Відповідно, ймовірність забиття голів командою під час матчу розраховується по групам, а не по усьому сезону. Також під час розрахунку цієї ймовірності враховується тип гри — домашня або виїзна. Це дозволяє врахувати особливості гри домашньої та виїзної команди.

5. Як випливає з розподілів типів результатів матчів, згрупованих за класами матчів, в змодельованих сезонах у середньому зберігається характер залежностей між кількістю домашніх перемог, нічиїх, виїзних перемог, притаманних класам матчів реального сезону. Сумарне відхилення від реального сезону за типами результату матчу у змодельованих сезонах не перевищує 13%.

6. За загальними розподілами типів результатів матчів та різницями голів усіх матчів змодельований сезон є подібним до реального сезону. За критерієм Колмогорова-Смирнова, різниця наведених розподілів на рівні

значущості 0,001 є статистично незначущою. Середнє значення статистики критерія для розподілів типів результату по класам матчів становить 0,71. На 12 з 16 класах матчів значення статистики критерія не перевищує 0,9, що у понад два рази менше за критичне значення статистики критерія на рівні значущості 0,001. Значення статистики критерія для розподілів різниць голів у сезоні становить 0,58, що у понад три рази менше за критичне значення статистики критерія на рівні значущості 0,001.

7. Гістограми різниць голів по кожному класу матчів мають очікувані закономірності в результатах матчів: сильніші команди мають кращі результати в грі зі слабшими командами; вдома команди грають краще, ніж на виїзді.

8. Розглянуту імітаційну модель може бути узагальнено для врахування інших параметрів футбольних матчів, на які приймаються ставки, — кількість призначених пенальті, кількість попереджень і вилучень тощо.

РОЗДІЛ 3
МЕТОДИ ВИЯВЛЕННЯ ПІДОЗРІЛИХ ЩОДО ФІКСОВАНОСТІ РЕЗУЛЬТАТУ ФУТБОЛЬНИХ МАТЧІВ ЗА НАЯВНОСТІ ДАНИХ ПРО ВЕСЬ СЕЗОН

3.1 Визначення міри неконформності поточного матчу

Базовим поняттям при використанні математичного апарату конформних предикторів і степеневих мартингалів є міра некомформності об'єкта даних. Її числове значення характеризує, наскільки кожен об'єкт є неконформним (відмінним) від усієї послідовності отриманих об'єктів. Від вибору міри неконформності залежить якість алгоритмів конформної класифікації та виявлення аномалій в даних [145].

Одним з основних показників футбольного матчу z є його результат у форматі $\alpha{:}\,\beta$, де число α дорівнює кількості голів, забитих командою-господаркою, а число β – відповідно гостьовою командою, під час матчу. Маючи результат матчу у такому форматі, легко підрахувати різницю м'ячів у цьому матчі як число $\alpha - \beta$. Ця характеристика є поведінковим атрибутом, оскільки у відповідності до визначеного контексту (сили команд і типу гри) за різницею м'ячів після завершення матчу опосередковано визначають чи є матч аномальним. Так, матчі з модулем різниці голів не більшим за 2 м'ячі, можуть викликати підозри, якщо слабка команда виграла у сильної або зіграла з нею у нічию. Також такі матчі можуть використовуватися командами для вирішення їхніх турнірних задач в сезоні. Якщо різниця м'ячів за модулем більше, ніж 2 м'ячі, такий матч викликає підозри щодо аномальності результату, коли сили команд близькі. Але такий матч може і не викликати підозри, якщо сили команд дійсно є кардинально різними. Проте, у цьому випадку такі матчі через неочікуваність результату (дуже великий рахунок) можуть бути договірними, бо дозволять шахраям отримати заробіток на ставках на такий результат.

На основі контекстуальних атрибутів, як було показано в розділі 2, матчі турніру розбиваються на ***класи.*** Кожному класу матчів притаманний власний розподіл ймовірностей різниці м'ячів, який залежить від сили команд учасниць і типу гри. Пошук аномальних матчів відбувається саме в межах визначеного класу матчів. Як прогнозоване значення результату матчу будемо використовувати математичне сподівання-різниці м'ячів матчів даного класу. Оскільки розподіл ймовірностей різниці м'ячів класу матчів вважається невідомим, використовується його оцінка, яка розраховується як середнє арифметичне значення різниці м'ячів класу матчів в сезоні, що розглядається.

Нехай маємо групування команд, тобто маємо функцію $group(t) = g$, яка команді з номером g у турнірній таблиці ставить у відповідність номер групи n, де $t = \overline{1,T}, g = \overline{1,K}, T$ є кількістю команд у сезоні, K є кількістю груп. На основі групування команд утворюється класифікація матчів в залежності від того, до яких груп належать команди — учасниці матчу. Класифікація матчів розглядається як функція $(i,j) = class(z)$, яка кожному матчу z ставить у відповідність упорядковану пару чисел (i,j), де $i = \overline{1,K}$ є номером групи, до якої належить домашня команда матчу, а $j = \overline{1,K}$ – номером групи виїзної команди матчу. Матчі z, для яких $class(z) = (i,j)$, утворюють множину матчів G_{ij} класу матчів (i,j). Множини G_{ij} є такими, які не перетинаються між собою і їх об'єднання утворює усю множину матчів сезону G:

$$\forall i \neq l, j \neq p \; G_{ij} \cap G_{lp} = \emptyset \, , \bigcup_{(i,j)\in C} G_{ij} = G,$$

де $C = \{(i,j) | i = \overline{1,K}, j = \overline{1,K}\}$.

Вважаємо, що матчі z є впорядкованими у множині G_{ij}, тобто, що всі матчі z_k є пронумерованими за номером $k = \overline{1, |G_{ij}|}$. Визначимо середню різницю голів по класу матчів (i,j):

$$avg(i,j) = \underset{class(z_l)=(i,j)}{\text{mean}} \{\alpha_l - \beta_l\} = \frac{1}{|G_{ij}|} \sum_{z_k \in G_{ij}} (\alpha_k - \beta_k), \qquad (3.1)$$

де k є номером поточного матчу в хронологічному порядку, l є номером іншого матчу, який відноситься до того класу матчів (i,j), що й поточний матч, i, j є відповідно номером групи домашньої і гостьової команди. Використовуючи раніше наведене припущення про групу команд, характеристика $avg(i,j)$ є середнім результатом групи спостережень, яка в подальшому використовується як прогнозоване значення результату гри у даному класі матчів (i,j). В свою чергу, відхилення фактичного результату матчу z_k від очікуваного результату можна розглядати як характеристику аномальності матчу по відношенню до класу матчів G_{ij}, до якого належить z_k. Тому використаємо цю різницю як міру неконформності a_k матчу z_k по відношенню до класу матчів G_{ij}:

$$a_k = |\alpha_k - \beta_k - avg(i,j)|. \tag{3.2}$$

Через те, що результат матчу $\alpha_k - \beta_k$ є цілим числом, а характеристика $avg(i,j)$ є раціональним числом, можливі відхилення від цілочисельного результату, що породжують похибки в обчисленні міри неконформності: відхилення у~0,5 за мірою неконформності призводять до того, що результати матчів, які відрізняються на один м'яч, в підсумку мають одне значення міри неконформності. Через це можливим доцільним рішенням є використання міри неконформності з округленою до цілого числа характеристикою $avg(i,j)$:

$$a_k = \left|\alpha_k - \beta_k - round\left(avg(i,j)\right)\right|. \tag{3.3}$$

З іншого боку, якщо крім підкреслення аномальності чисельного відхилення між фактичним і очікуваним результатами матчу виникає необхідність врахування аномальності матчів також за типом результату, то може бути використана така міра неконформності:

$$a_k = 1{,}5^{1-\mathrm{sgn}\left(avg(i,j)*(\alpha_k-\beta_k)\right)}\left|\left(\alpha_k-\beta_k\right)-avg(i,j)\right|, \tag{3.4}$$

де функція sgn(x) вказує знак числа x і дорівнює 1, якщо x>0, 0, якщо x=0 та –1, якщо x<0. Ця міра неконформності враховує як абсолютні результати

команд матчів, так і різницю у підсумках фактичного і спрогнозованого результатів, причому ця різниця має більш значний вплив, ніж абсолютні результати матчів. За рахунок цього, наприклад, матчі з фактичною перемогою і спрогнозованою поразкою є більш неконформними, ніж матчі з фактичною і спрогнозованою перемогою, які відрізняються за кількістю забитих м'ячів.

3.2 Метод виявлення підозрілих щодо фіксованості результату футбольних матчів з використанням конформного аномального детектора

В підрозділі 1.4 зазначалося, що важливим перспективним класом методів машинного навчання, які використовуються як для вирішення задач класифікації даних, так і виявлення аномалій, є методи на основі конформних предикторів. Конформний аномальний детектор є подальшим розвитком теорії конформного прогнозування [140, 142, 145]. Він використовує ймовірнісно подібну міру надійності для прогнозу аномальності об'єкта [140, 146].

Матчі з фіксованим результатом відносяться до класу контекстних аномалій. Визначення контекстуальних атрибутів і розбиття команд сезону на класи G_{ij} було розглянуто в розділі 2. Там же було показано, що виявлення матчів, підозрілих на фіксований результат, відбувається окремо в кожному класі матчів G_{ij} з використанням поведінкового атрибуту, за який береться різниця м'ячів матчу.

Одиницею вхідних даних є спостереження z_k, що описує матч футбольного сезону, k є порядковим номером матчу у сезоні. Спостереження z_k є набором значень $z_k=(i_k, j_k, \alpha_k, \beta_k, T_k)$, де i_k і α_k є відповідно групою (рангом) і результатом команди-господарки цього матчу, а j_k і β_k є групою (рангом) і результатом гостьової команди матчу, T_k є датою проведення матчу.

Виявлення матчів, підозрілих на фіксований результат з використанням конформного аномального детектора, складається з таких етапів:

1) для кожного матчу $z_k \in G_{ij}$ з послідовності $G_{ij} = (z_1, \dots, z_k, \dots, z_N)$ обчислюється міра неконформності $(a_1, \dots, a_i, \dots, a_N)$ по відношенню до всіх інших об'єктів:

$$a_1 = A_N(\{z_2, \dots, z_N\}, z_1),$$

…,

$$a_k = A_N(\{z_1, \dots, z_{k-1}, z_{k+1}, , \dots, z_N\}, z_k),$$

…,

$$a_N = A_N(\{z_1, \dots, z_{N-1}\}, z_N),$$

де A_N є функцією, яка залежить від множини вигляду $\{z_1, \dots, z_{k-1}, z_{k+1}, \dots, z_N\}$ і об'єкта z_k, і у відповідність цим аргументам ставить дійсне число: $A_N: \boldsymbol{Z}^{N-1} \times \boldsymbol{Z} \to \boldsymbol{R}$.

Міра неконформності a_k розраховується за однією з формул (3.2)-(3.4), що є першим етапом при розрахунку конформного предиктора.

2) використовуючи міри відмінності поточного *k*-го матчу і усіх інших матчів цього ж класу, обчислюється ступінь конформності (відмінності) (*p*-value) матчу від множини спостережень $\{z_1, \dots, z_k, \dots, z_{N-1}, z_N\}$:

$$p_k = p(z_1, z_2, \dots, z_k, \dots, z_N) = \frac{\#\{i : a_i \geq a_k, 1 \leq i \leq N\}}{N}, \qquad (3.5)$$

де операція *#A* повертає кількість елементів у множині *A*. Наприклад, для множини цілих чисел {1, 2, 5, 10, 15, 17} операція #{1, 2, 5, 10, 15, 17} = 6. У формулі (3.5) в чисельнику записана множина, що містить номери таких спостережень (матчів), міра відмінності яких є такою самою або більшою, ніж у поточного спостереження, включаючи й номер поточного спостереження. Тому кількість елементів у множині з чисельника цієї формули приймає значення в діапазоні [1; *N*]. Відповідно величина p_k приймає значення в діапазоні $[\frac{1}{N}; 1]$.

3) На основі ступеню конформності матчу p_k приймається рішення щодо класу об'єкта, що спостерігається, за таким правилом:

а) якщо

$$p_k < \varepsilon, \tag{3.6}$$

тоді об'єкт z_k вважається конформно аномальним;

б) якщо

$$p_k \geq \varepsilon, \tag{3.7}$$

тоді об'єкт z_k вважається нормальним, де $\varepsilon \in [0; 1]$ є ***порогом аномальності*** (anomality threshold).

Множина всіх матчів, для яких виконується умова (3.6) називається **конформним аномальним предиктором** і позначається як $\Gamma^{\varepsilon}(z_1, z_2, \ldots, z_k, \ldots, z_{N-1}, z_N)$.

Визначення конформної аномалії узгоджується зі статистичним визначенням викиду за Хокінсом [73]. Конформна аномалія є об'єктом z_k, який настільки відхиляється від z_1,..., $z_{k-1}, z_{k+1}, \ldots, z_N$ за мірою неконформності, що викликає підозру про те, що цей об'єкт було утворено механізмом, відмінним від того, за допомогою якого утворено інші об'єкти вибірки.

Показано [144, 146], що конформне прогнозування, а також його розширення у вигляді конформного аномального детектора, забезпечують гарантії покриття для ступеня конформності p_k , а саме: якщо виконується припущення обмінюваності або незалежності та ідентичності розподілу об'єктів вибірки $z_1, \ldots, z_N$ та умова, що на детектор потрапляє один об'єкт за одиницю часу, тоді для будь-якої міри неконформності A_N і $k \geq 1$ ймовірність помилки у прийнятті рішення, що об'єкт не є нормальним, не перевищує ε [146]:

$$P(p_k < \varepsilon) \leq \varepsilon. \tag{3.8}$$

Таким чином, параметр ε регулює чутливість конформного аномального детектора до виявлення аномальних об'єктів [147]: цей параметр є часткою

аномальних об’єктів, які виявлено як конформні аномалії. Налаштування цього параметру також впливає на точність виявлення, що дорівнює відносній кількості аномальних об’єктів серед тих, які виявлено як конформні аномалії. Високе значення параметра ε може збільшити чутливість детектора, але в той же час зменшить точність виявлення та збільшить частоту появи хибних виявлень. Хоча досягнення високої чутливості є важливим, стверджується, що обмежуючим фактором у виявленні аномалій є фактично зниження точності [166]. Ця проблема отримала назву омани базового рівня та полягає в тому, що точність виявлення починає поступатися частоті помилкових рішень про аномальність, що відбувається через низьку частоту аномальних об’єктів.

Отже, слід налаштовувати параметр ε залежно від рівня точності, прийнятного в конкретній прикладній задачі.

У випадку роботи конформного аномального детектора у неконтрольованому режимі, тобто режимі без вчителя, можна стверджувати, що *значення параметра* ε *слід встановити близько до апріорної ймовірності появи аномальних об’єктів* λ*, щоб досягти гарного балансу між чутливістю та точністю виявлення* [145]. Дійсно, припускаючи наявність такої ***ідеальної*** міри неконформності A_N, що $a_i > a_j$ для будь-яких об’єктів z_i та z_j, що належать до аномального та нормального класів відповідно, інтуїтивно зрозуміло, що встановлення параметра $\varepsilon = \lambda$ призведе до того, що точність виявлення буде близькою до 1.

Тим не менш, завжди слід уникати налаштування $\varepsilon < \frac{1}{N}$ незалежно від міри неконформності A_k, оскільки тоді чутливість до аномальних об’єктів буде нульовою. Щоб продемонструвати даний факт, припустимо, що ми спостерігаємо анормальний об’єкт z_k такий, що $a_k \gg a_i \ \forall i = 1,\ldots,N$. З формули для p_k випливає, що $p_k = \frac{1}{N}$. Отже, якщо $\varepsilon < \frac{1}{N}$, тоді об’єкт z_k не буде класифікуватися як аномальний, навіть якщо він виглядає дуже екстремальним за мірою неконформності.

3.3 Порівняльний аналіз методів виявлення матчів, підозрілих на фіксований результат, на основі експертно визначеного порогу відхилення і конформного аномального детектору

В підрозділі 2.5 для обґрунтування віднесення футбольних матчів до договірних використано статистичне визначення аномальних даних, яке ґрунтується на ймовірності їх появи p_A у відповідному класі матчів. Можливий також інший підхід для визначення аномалій, який ґрунтується на експертних даних. Для визначення аномального матчу експерту може бути простіше вказати не ймовірність p_A з якою з'являється аномальний результат, а визначити порогове відхилення результату матчу від деякого очікуваного значення, більше якого результат матчу може розглядатися як аномальний. Як аналогію можна навести оцінювання успішності студента, який відноситься до деякої категорії студентів. Тоді як очікуваний результат для цього студента може розглядатися середній бал навчання студентів даної категорії, а аномальними вважатимуться оцінки студента, що відхиляються від середнього балу більше визначеного порогового значення.

Враховуючи контекстну залежність аномальних матчів, як очікуване значення результату матча доцільно використати математичне сподівання різниці м'ячів класу матчів $m(i,j)$. Оскільки воно невідоме, то може бути використана його оцінка, отримана по певній кількості сезонів, наприклад, по 100 модельним сезонам. На рис. 3.1 наведені гістограми та математичні сподівання різниці м'ячів по кожному класу матчів (червона лінія) за 100 модельними сезонами. Розглянуті модельні сезони сформовано за алгоритмом, наведеним у розділі 2, використовуючи дані реального сезону 2013-2014 років Ліги II Франції.

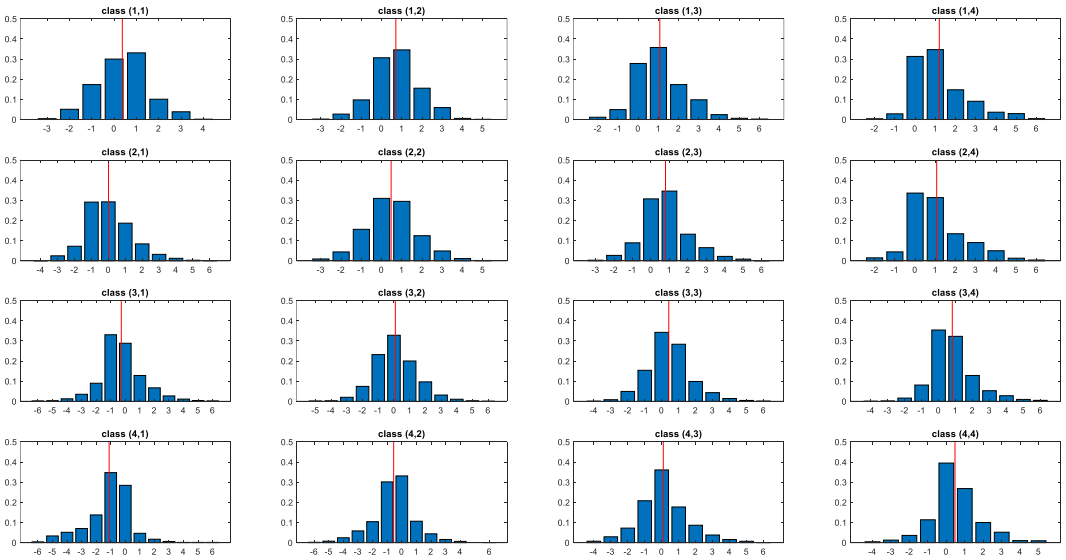


Рисунок 3.1 – Гістограми та математичне сподівання різниці м’ячів по класам матчів, сформовані за 100 модельними сезонами

Таким чином розмітку даних для модельного сезону можна виконати з використанням нерівності

$$|\alpha_k - \beta_k - m(i,j)| > \chi, \tag{3.9}$$

де χ — порогове значення відхилення результату матчу від очікуваного значення, що визначається експертом.

На практиці закон розподілу ймовірностей і математичне сподівання $m(i,j)$ різниці м’ячів класу матчів є невідомим. Тому, також як і при побудові міри неконформності (3.2), використовується оцінка цієї різниці м’ячів, яка розраховується як середнє арифметичне значення $avg(i,j)$ різниці м’ячів класу матчів в сезоні, що розглядається.

Таким чином, вирішальне правило прийняття рішення щодо аномального характеру матча може бути представлено у вигляді

$$|\alpha_k - \beta_k - avg(i,j)| > \chi. \tag{3.10}$$

Вирішальне правило (3.10) за структурою аналогічне правилу перевірки статистичних гіпотез. При цьому міра неконформності у лівій частині умови відповідає поняттю статистики, а порогове значення χ — критичному рівню. Помилки при прийнятті рішень за вирішальним правилом (3.10) можуть

виникати у випадку, коли отримане по реальній вибірці середнє значення $avg(i,j)$ різниці м’ячів класу буде відрізнятися від істинного математичного сподівання $m(i,j)$ не менше, ніж на один м’яч. Така ситуація може виникати у випадку малої вибірки результатів матчів класу.

Недоліком даного правила є те, що відсутня інформація щодо достовірності отриманих аномальних результатів як ймовірнісних аномалій. Але цей недолік можна усунути шляхом застосування математичного апарату аномального конформного детектору. Ліва частина вирішального правила (3.10) співпадає з мірою неконформності (3.2), що використовується при виявленні аномальних матчів з використанням конформного аномального детектора. Між значеннями міри неконформності і ступенем конформності (p-value) матчу існує однозначний зв'язок. Це призводить до того, що можна встановити значення порогу конформного аномального детектору, при якому він приймає ті ж самі рішення, що й на підставі вирішального правила (3.10). Для цього визначається максимальне значення міри неконформності, для якого не виконується умова (3.10). Відповідне йому значення p-value буде визначати значення порогу еквівалентного конформного детектора. Значення порогу ε аномального конформного детектора дозволяє визначити ступінь достовірності отриманих аномальних даних, а саме: ймовірність помилки у прийнятті рішення, що об’єкт не є нормальним, не перевищує ε.

Продемонструємо це більш детальніше на прикладі. Для наочності і без обмеження розгляду покладемо математичне сподівання $m(i,j) = 0$ і середнє значення $avg(i,j) = 0$. Враховуючи те, що різниця м’ячів матчу є цілим числом, поточна міра неконформності a_k також є цілим числом і породжує поділ усіх матчів класу на сегменти. Кількість цих сегментів дорівнює кількості унікальних значень міри неконформності. Окремому сегменту матчів відповідає одне унікальне значення міри неконформності a_k.

Розглянемо поділ матчів на сегменти на прикладі створення конформного аномального детектора для класу матчів одного із

змодельованих сезонів (рис. 3.2). Розглянуті модельні сезони сформовано за алгоритмом, наведеним у розділі 2, використовуючи дані реального сезону 2013-2014 років Ліги II Франції. Міра неконформності a_k для цього класу матчів приймає чотири значення. Оскільки існує чотири значення міри неконформності, то їм відповідають чотири можливих значення ступеня конформності $\varepsilon^{(I)} - \varepsilon^{(IV)}$. Як видно на рис. 3.2, одному й тому ж значенню міри неконформності відповідає одне й те ж значення ступеню конформності. Еквівалентний вирішальному правилу (3.10) алгоритм виявлення конформного аномального детектора описується виразом:

$$p_k < \varepsilon^{(t)}. \tag{3.11}$$

Таким чином, якщо поріг вирішального правила (3.10) $\chi = a_{\tilde{k}}$, то на основі значення $a_{\tilde{k}}$ міри неконформності однозначно визначається значення $\varepsilon^{(t)}$.

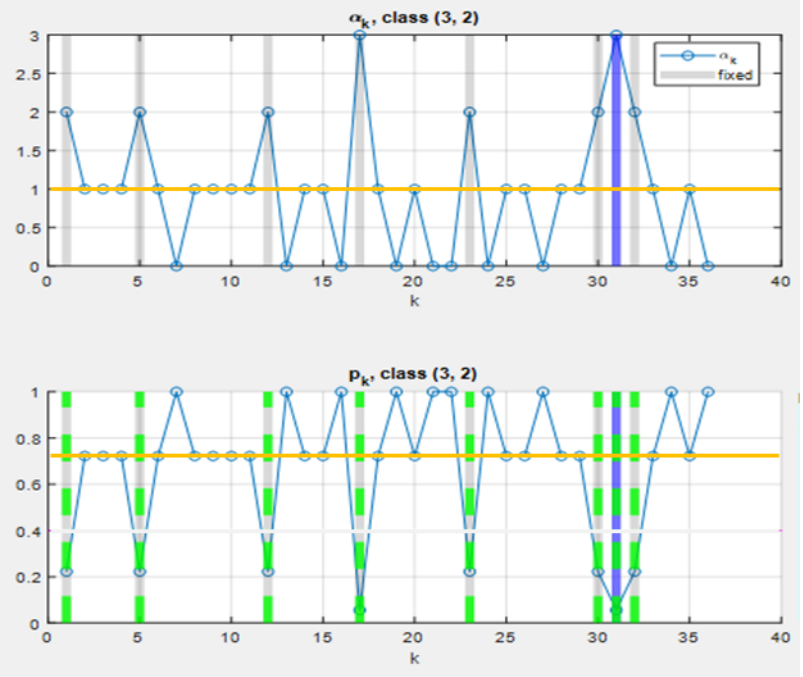


Рисунок 3.2 – Приклад графіків міри неконформності та p-value

Розглянемо інтерпретацію ступеню конформності p_k шляхом використання гістограми різниць голів матчів класу. У відповідності до

значень міри неконформності виділяється 4 сегменти матчів: I сегмент складається з матчів, для яких $a_k = 0$, II сегмент — матчі, для яких $a_k = 1$, III сегмент — матчі, для яких $a_k = 2$, IV сегмент — матчі, для яких $a_k = 3$.

На рис. 3.3 зображено гістограму різниць м'ячів матчів класу, що розглядається. Гістограма показує зв'язок між різницями м'ячів і відносною частотою їх появи у класі матчів. Можна встановити відповідність між сегментами матчів за ***мірою неконформності і стовпцями гістограми***. До I ***сегменту*** належать матчі, яким відповідає стовпець гістограми з аргументом 0: це стовпець, який відповідає різниці голів, яка дорівнює 0. До II сегменту належать матчі, яким відповідають стовпці гістограми з аргументами –1 і 1. До III сегменту належать матчі стовпців гістограми з аргументами –2 і 2. До IV сегменту — з аргументами –3 і 3.

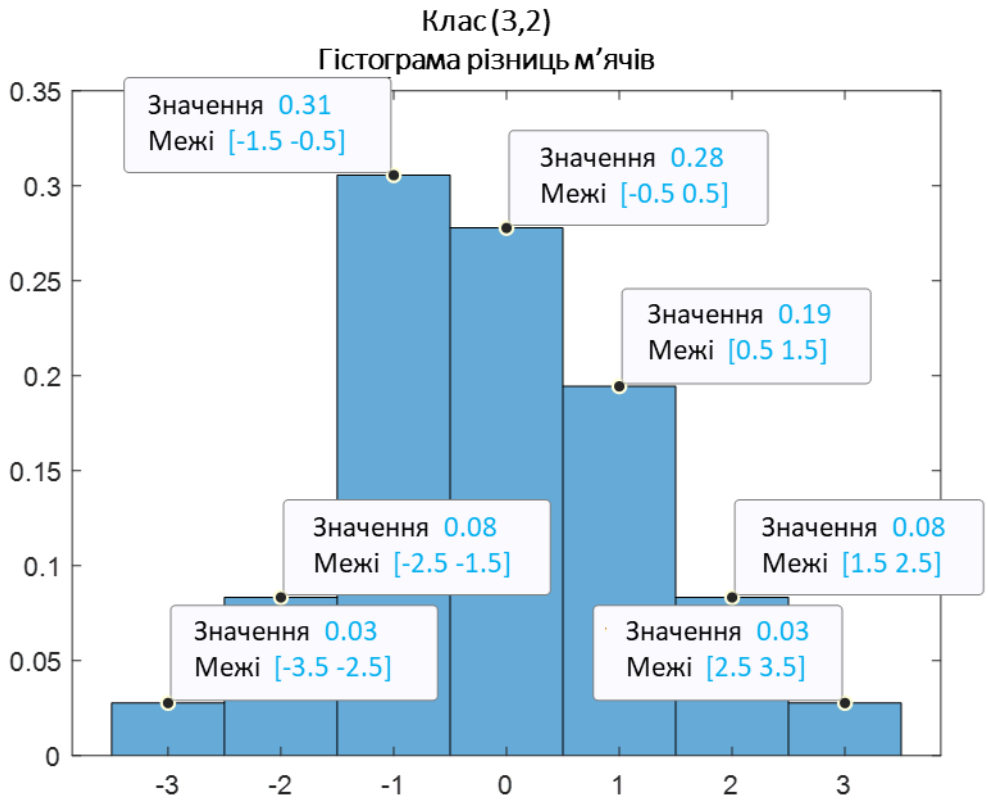


Рисунок 3.3 – Приклад гістограми різниці м'ячів для класу матчів

Тепер проведемо аналогічні міркування за ступенем конформності p_k. Оскільки існує чотири значення міри неконформності, то їм відповідають чотири можливих значення ступеня конформності $\varepsilon^{(I)} - \varepsilon^{(IV)}$. За значеннями ступеня конформності можна виділити вкладені множини (групи) матчів. ***Множина I*** матчів визначається нерівністю $p_k \leq \varepsilon^{(I)}$, де $\varepsilon^{(I)}$ є ступенем конформності, якому відповідає значення міри неконформності $a_k = 0$. Ця множина називається конформним предиктором з рівнем значущості $\varepsilon^{(I)}$ і позначається $\Gamma^{\varepsilon(\mathrm{I})}(z_1, z_2, \ldots, z_N)$. До множини I матчів входять ***усі матчі поточного класу матчів*** (матчі всіх сегментів I-IV), оскільки $a_k = 0$ є найменшим значенням міри неконформності. За визначенням ступеня конформності $\varepsilon^{(I)}$ дорівнює відносній частоті появи матчів, міра неконформності яких $a_k \geq 0$ і приймає значення $\varepsilon^{(I)} = 1$. Ця подія еквівалентна події появи будь-якого результату матча даного класу. Тому значення $\varepsilon^{(I)}$ також можна розрахувати як суму відносних частот голів за гістограмою рис. 3.3.

Наступна, ***множина* II** матчів визначається нерівністю $p_k \leq \varepsilon^{(II)}$ і є конформним предиктором з рівнем значущості $\varepsilon^{(II)}$ і позначається $\Gamma^{\varepsilon(\mathrm{II})}(z_1, z_2, \ldots, z_N)$. Вона включає всі матчі за виключенням тих, яким відповідає міра неконформності $a_k = 0$ або матчі сегментів II-IV. За визначенням ступеня конформності $\varepsilon^{(II)}$ дорівнює відносній частоті появи матчів, міра неконформності яких $a_k \geq 1$. Оскільки в множину матчів не входять матчі сегмента I, ця подія еквівалентна події появи будь-якого результату матча за виключенням 0. Тому, унікальне значення $\varepsilon^{(II)}$ можна визначити таким чином: $\varepsilon^{(II)} = \varepsilon^{(I)} - h_I = 1 - h_I$, де h_I дорівнює відносній частоті нульової різниці м'ячів, тобто нічийним результатам.

В даному прикладі $h_I = 0{,}28, \varepsilon^{(II)} = 1 - 0{,}28 = 0{,}72$. У результаті, $\varepsilon^{(II)}$ дорівнює відносній частоті різниць м'ячів, крім нульової або появи матчів, міра неконформності яких $a_k \geq 1$.

Множина III матчів за характеристикою p_k формується аналогічним чином і є конформним предиктором $\Gamma^{\varepsilon(\mathrm{III})}(z_1, z_2, \ldots, z_N)$. Їй належать матчі, для яких $a_k \geq 2$ або матчі сегментів III і IV. Отже, значення $\varepsilon^{(III)}$ визначається за схожим принципом: $\varepsilon^{(III)} = \varepsilon^{(II)} - h_{II}$, де h_{II} дорівнює сумі стовпців гістограми, які відповідають значенням різниць голів {–1, 1}. В даному прикладі, $h_{II} = 0{,}31 + 0{,}19 = 0{,}50$, $\varepsilon^{(III)} = 0{,}72 - 0{,}50 = 0{,}22$. У результаті, $\varepsilon^{(III)}$ дорівнює відносній частоті різниць м'ячів із значеннями {2, –2, 3, –3} або появи матчів, міра неконформності яких $a_k \geq 2$.

Множина IV за характеристикою p_k формується аналогічним чином і є конформним предиктором $\Gamma^{\varepsilon(\mathrm{IV})}(z_1, z_2, \ldots, z_N)$. Їй належать матчі, для яких $a_k \geq 3$ або матчі сегменту IV. Значення, $\varepsilon^{(IV)} = \varepsilon^{(III)} - h_{III}$, де h_{III} дорівнює сумі стовпців гістограми, які відповідають значенням різниць голів {–2, 2}. В даному прикладі, $h_{III} = 0{,}08 + 0{,}08 = 0{,}16$, $\varepsilon^{(IV)} = 0{,}22 - 0{,}16 = 0{,}06$. У результаті, $\varepsilon^{(III)}$ дорівнює відносній частоті різниць м'ячів із значеннями {3, –3} або появи матчів, міра неконформності яких $a_k \geq 3$.

Таким чином, конформний аномальний детектор за значенням ε формує вкладене сімейство конформних предикторів, для яких виконується умова: для будь-яких $0 < \varepsilon_1 < \varepsilon_2 < 1$

$$\Gamma^{\varepsilon_1}(z_1, z_2, \ldots, z_N) \supseteq \Gamma^{\varepsilon_2}(z_1, z_2, \ldots, z_N).$$

Як зазначалося в підрозділі 3.2, за припущення обмінюваності або HIP об'єктів вибірки $z_1, \ldots, z_N$ для будь-якої міри неконформності ймовірність помилки конформного аномального детектора у прийнятті рішення, що об'єкт не є нормальним, не перевищує ε. Отже, якщо об'єкт z_N насправді є нормальним, ймовірність неправильного рішення про аномальність об'єкта не перевищує ε. Тим самим, ми отримуємо оцінку достовірності прийнятих рішень як для конформного аномального детектора, так і для методу на основі експертно-визначеного порогу відхилення.

3.4 Методи виявлення підозрілих щодо фіксованості результату футбольних матчів з використанням степеневого та інтегрального мартингалів

Подальші методи виявлення підозрілих щодо фіксованого результату матчів розроблено на основі загального методу побудови конформного предиктора, запропонованого в [145], шляхом його наступної адаптації:

– введенням нової міри неконформності (відмінності), адаптованої під особливості рейтингування футбольних матчів і його учасників;

– використанням ступеня конформності p-value і степеневого або інтегрального мартингала;

– введенням вирішальних правил, за якими на основі використаних характеристик приймається рішення про потенційну підозрілість футбольного матчу.

На основі ступеня конформності можна сформувати характеристику, яка називається *степеневим мартингалом* [145] і використовується в теорії пошуку змін в потоках даних [141]:

$$M_k^{(\eta)} = \prod_{i=1}^{k}\left(\eta p_i^{\eta-1}\right) = \eta^k \prod_{i=1}^{k}\left(p_i^{\eta-1}\right) = \prod_{i=1}^{k}\left(\eta\left(\frac{1}{p_i}\right)^{1-\eta}\right), \qquad (3.12)$$

де $\eta \in [0;1]$ є параметром чутливості мартингалу. Степеневий мартингал є узагальненою статистикою, яка для k-го спостереження обчислюється шляхом піднесення «перегорнутої ймовірності» $\frac{1}{p_i}$ кожного спостереження від $i = 1$ до $i = k$ до степеня $1 - \eta$ та зважування кожного результуючого значення $\left(\frac{1}{p_i}\right)^{1-\eta}$ домноженням на η і добутком усіх окремих значень. Оскільки $0 < p_i \leq 1$, то $\frac{1}{p_i} \geq 1$. А оскільки $0 \leq \eta \leq 1$, то:

$$0 \leq 1 - \eta \leq 1,$$

$$1 \leq \left(\frac{1}{p_i}\right)^{1-\eta} \leq \frac{1}{p_i}.$$

Зважування $\left(\frac{1}{p_i}\right)^{1-\eta}$ шляхом домноження на η призводить до того, що

$$\eta \le \eta(1/p_i)^{1-\eta} \le \eta/p_i,$$

тобто значення $(1/p_i)^{1-\eta}$ після домноження на η може стати як меншим за 1, так і більшим. Якщо розглядати цю ситуацію відносно ступеня конформності p_k поточного (k-го) спостереження і степеневого мартингалу $M_{k-1}^{(\eta)}$ попереднього спостереження, то домноження $M_{k-1}^{(\eta)}$ на $\eta(1/p_k)^{1-\eta}$ може призвести до того, що значення $M_k^{(\eta)}$ може як зменшитися, так і збільшитися. На рис. 3.4 показано різні варіанти множника $\eta p_k^{\eta-1}$ як функції від η при значеннях $p_k \in \{0{,}1; 0{,}2; 0{,}3; 0{,}33; 0{,}4; 0{,}5\}$. З цього графіку можна зробити висновок, що на відрізку $\eta \in [0; 1]$ для будь-яких $0 < p_1 < p_2 < 1$ маємо $\eta p_1^{\eta-1} > \eta p_2^{\eta-1}$. Також видно, що при $p_k \ge 0{,}4$ функція $\eta p_k^{\eta-1} < 1$ на відрізку $\eta \in [0; 1]$. Починаючи зі значень $p_k = 0{,}33$ і менше, функція $f(\eta) = \eta p_k^{\eta-1}$ на відрізку $\eta \in [0; 1]$ при певних значеннях η починає приймати значення $f(\eta) > 1$. В цьому полягає *принцип чутливості степеневого мартингала*: за певних значень параметра η значення степеневого мартингалу $M_k^{(\eta)}$ починає зростати за умови низького значення ступеня конформності p_k.

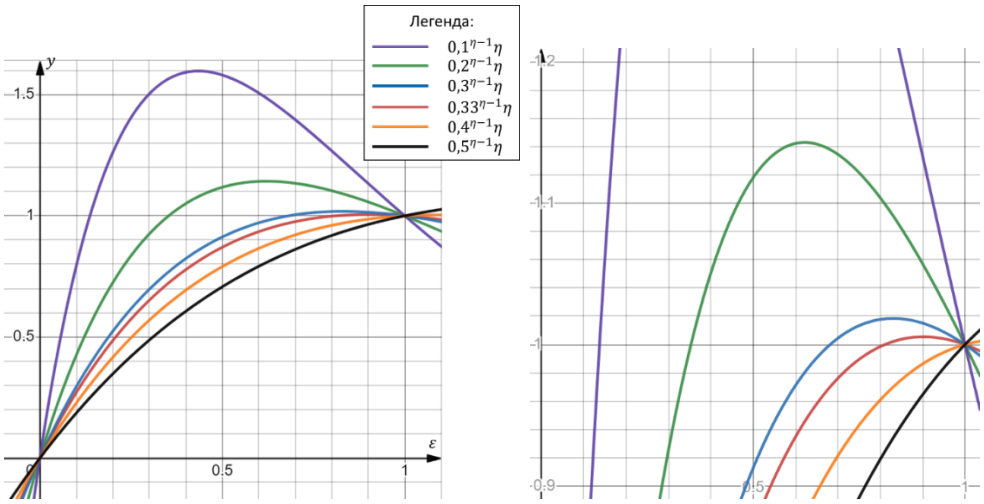


Рисунок 3.4 – Графік функції $f(\eta) = \eta p_k^{\eta-1}$ при $\eta \in [0; 1]$ для $p_k \in \{0{,}1;\ 0{,}2;\ 0{,}3;\ 0{,}33;\ 0{,}4;\ 0{,}5\}$

Щоб нівелювати продемонстровану залежність від значення параметра η, обчислюється інтегральний мартингал:

$$M_k = \int_0^1 M_k^{(\eta)} d\eta. \tag{3.13}$$

Знайдемо цей інтеграл:

$$\int_0^1 M_k^{(\eta)} d\eta = \int_0^1 \eta^k \prod_{i=1}^{k} \left(p_i^{\eta-1}\right) d\eta$$

за визначенням:

$$\int_0^1 \eta^k \prod_{i=1}^{k} \left(p_i^{\eta-1}\right) d\eta = \int_0^1 \eta^k \left(\prod_{i=1}^{k} p_i\right)^{\eta-1} d\eta = \int_0^1 \eta^k \left(\prod_{i=1}^{k} \frac{1}{p_i}\right)^{1-\eta} d\eta =$$

$$= \int_0^1 \eta^k e^{\ln\left(\prod_{i=1}^{k} p_i\right)^{\eta-1}} d\eta = \int_0^1 \eta^k e^{(\eta-1)\ln\left(\prod_{i=1}^{k} p_i\right)} d\eta =$$

= (позначимо через $c = -\ln\left(\prod_{i=1}^{k} p_i\right)$, оскільки $\forall i = \overline{1,k}\ 0 < p_i \leq 1$ та $\ln\left(\prod_{i=1}^{k} p_i\right) \leq 0$) =

$$= \int_0^1 \eta^k e^{-c(\eta-1)} d\eta = \int_0^1 \eta^k e^{-c\eta+c} d\eta = e^c \int_0^1 \eta^k e^{-c\eta} d\eta =$$

= (введемо заміну $c\eta = u$. Тоді $\eta = \frac{u}{c}, d\eta = \frac{du}{c}$. При $\eta = 0$ маємо $u = 0$, при $\eta = 1$ маємо $u = c$, тому $u \in [0, c]$) =

$$= \frac{e^c}{c} \int_0^c \left(\frac{u}{c}\right)^k e^{-u} du = \frac{e^c}{c^{k+1}} \int_0^c u^k e^{-u} du = \frac{e^c}{c^{k+1}} \int_0^c u^{(k+1)-1} e^{-u} du =$$

$$= \frac{e^c}{c^{k+1}} \gamma(k+1, c),$$

де $\gamma(k+1, c)$ є нижньою неповною гамма-функцією.

Отже,

$$\int_0^1 M_k^{(\eta)} d\eta = \int_0^1 \eta^k \prod_{i=1}^{k} \left(p_i^{\eta-1}\right) d\eta = \frac{e^c}{c^{k+1}} \gamma(k+1, c),$$

де $\gamma(k+1, c)$ є нижньою неповною гамма-функцією, $c = -\ln\left(\prod_{i=1}^{k} p_i\right)$.

При аналізі результатів застосування характеристик $M_k^{(\eta)}$ та M_k до окремих класів матчів, було виявлено такі нові властивості цієї характеристики:

Твердження. При $\eta \in (0;1)$ є справедливими такі твердження:

1) для спостережень z_k, для яких $p_k \to 0$, справедливо, що $M_k^{(\eta)} > M_{k-1}^{(\eta)}$;

2) якщо $M_k^{(\eta)} > M_{k-1}^{(\eta)}$, тоді $p_k < \eta$;

3) якщо $M_k^{(\eta)} > M_{k-1}^{(\eta)}$, тоді $p_k < \eta^{\frac{1}{1-\eta}}$.

Доведення. 1) Нехай $M_k^{(\eta)} > M_{k-1}^{(\eta)}$. Тоді:

$$\prod_{i=1}^{k}\left(\eta p_i^{\eta-1}\right) > \prod_{i=1}^{k-1}\left(\eta p_i^{\eta-1}\right).$$

Виокремимо спільний множник $\prod_{i=1}^{k-1}\left(\eta p_i^{\eta-1}\right)$ з обох сторін нерівності:

$$\eta p_k^{\eta-1}\prod_{i=1}^{k-1}\left(\eta p_i^{\eta-1}\right) > \prod_{i=1}^{k-1}\left(\eta p_i^{\eta-1}\right). \tag{3.14}$$

Оскільки $p_k \in (0;1]$, тоді $\forall x \in \boldsymbol{R}\ p_k^x > 0$. Звідки $p_k^{\eta-1} > 0$ і $\eta p_k^{\eta-1} > 0$. Тоді $\forall k \in \boldsymbol{N}\ \prod_{i=1}^{k}\left(\eta p_i^{\eta-1}\right) > 0$. Поділивши нерівність (3.14) на $\prod_{i=1}^{k-1}\left(\eta p_i^{\eta-1}\right)$, отримаємо $\eta p_k^{\eta-1} > 1$. Прологарифмувавши цю нерівність, отримаємо:

$$\ln \eta p_k^{\eta-1} > 0.$$

Скористаємося властивістю логарифмів $\ln(ab) = \ln a + \ln b$, коли $\forall a > 0, b > 0$:

$$\ln \eta + \ln p_k^{\eta-1} > 0.$$

Перенесемо

$$\ln \eta > -\ln p_k^{\eta-1}.$$

Скористаємося властивістю логарифмів $\forall a \in \boldsymbol{R}, b > 0\ a \ln b = \ln b^a$:

$$\ln \eta > (1-\eta)\ln p_k\ . \tag{3.15}$$

Розглянемо подвійну нерівність:

$$0 < \eta < 1.$$

Домножимо цю нерівність на –1:

$$-1 < -\eta < 0.$$

Додамо до цієї нерівності 1:

$$0 < 1 - \eta < 1.$$

Отже, на це число можна множити чи ділити нерівності, і знак при цьому не зміниться. Поділивши нерівність (3.15) на $(1-\eta)$, отримаємо:

$$\frac{\ln \eta}{1-\eta} > \ln p_k \ . \tag{3.16}$$

Оскільки $\eta < 1$ та $p_k < 1$, тоді $\ln \eta < 0$ та $\ln p_k < 0$.

Перейдемо до додатніх логарифмів. Для цього, скористаємося властивістю степеня $\forall x \in R\ x = \left(\frac{1}{x}\right)^{-1}$ та властивістю логарифмів $\forall a \in \boldsymbol{R}$, $b > 0\ a \ln b = \ln b^a$:

$$\frac{-\ln \frac{1}{\eta}}{1-\eta} > -\ln \frac{1}{p_k}.$$

Помножимо отриману нерівність на –1:

$$\frac{\ln \frac{1}{\eta}}{1-\eta} < \ln \frac{1}{p_k}. \tag{3.17}$$

Взявши границю при $p_k \to 0$ у нерівності (3.17), отримаємо:

$$\frac{\ln \frac{1}{\eta}}{1-\eta} < \infty.$$

Отримали нерівність, справедливу $\forall \eta \in (0;1)$.

Отже, при $p_k \to 0$ є справедливою нерівність $M_k^{(\eta)} > M_{k-1}^{(\eta)}$.

2) Скористаємося нерівністю (3.17). Перепишемо її у вигляді:

$$\frac{1}{1-\eta} \ln \frac{1}{\eta} < \ln \frac{1}{p_k}. \tag{3.18}$$

Оскільки $\eta \in (0;1)$, тоді вираз $\frac{1}{1-\eta}$ є повною сумою нескінченно спадної геометричної прогресії зі знаменником $0 < \eta < 1$:

$$\frac{1}{1-\eta} = 1 + \eta + \eta^2 + \cdots = \sum_{n=0}^{\infty} \eta^n.$$

Тому нерівність (3.18) перепишеться у вигляді:

$$(1 + \eta + \eta^2 + \cdots) \ln\frac{1}{\eta} < \ln\frac{1}{p_k}, \text{ тобто}$$

$$\ln\frac{1}{\eta}\sum_{n=0}^{\infty} \eta^n < \ln\frac{1}{p_k}. \tag{3.19}$$

Оскільки $0 < \eta < 1$, то $\eta + \eta^2 + \cdots = \sum_{n=1}^{\infty} \eta^n > 0$. Додавши до цієї нерівності 1, отримаємо $1 + \sum_{n=1}^{\infty} \eta^n > 1$, що еквівалентно нерівності $1 < \sum_{n=0}^{\infty} \eta^n$.

Помноживши цю нерівність на $\ln\frac{1}{\eta}$, отримаємо $\ln\frac{1}{\eta} < \ln\frac{1}{\eta}\sum_{n=0}^{\infty} \eta^n$. Враховуючи отримане і (3.19), можна побудувати такий ланцюг нерівностей: $\ln\frac{1}{\eta} < \ln\frac{1}{\eta}\sum_{n=0}^{\infty} \eta^n < \ln\frac{1}{p_k}$, з якого випливає, що $\frac{1}{\eta} < \frac{1}{p_k}$, що, в свою чергу, еквівалентно нерівності $p_k < \eta$, оскільки $\eta > 0$ та $p_k > 0$.

Отже, якщо для *k*-го спостереження виконується умова $M_k^{(\eta)} > M_{k-1}^{(\eta)}$, тоді автоматично виконується умова $p_k < \eta$.

3) Скориставшись нерівністю (3.16) та відомою властивістю логарифмів $\forall a \in \boldsymbol{R}, b > 0\ a \ln b = \ln b^a$, отримаємо:

$$\ln \eta^{\frac{1}{1-\eta}} > \ln p_k\ . \tag{3.20}$$

За рахунок властивості монотонності функції експонента і властивості логарифмів $\forall b > 0\ e^{\ln b} = b$, маємо:

$$\eta^{\frac{1}{1-\eta}} > p_k\ .$$

Перепишемо отриману нерівність як

$$p_k\ < \eta^{\frac{1}{1-\eta}}.$$

Отже, якщо для *k*-го спостереження виконується умова $M_k^{(\eta)} > M_{k-1}^{(\eta)}$, тоді автоматично виконується умова $p_k < \eta^{\frac{1}{1-\eta}}$.

Зворотне твердження також виконується, оскільки всі викладки здійснювались шляхом еквівалентних перетворень нерівностей.

Отже, умова $M_k^{(\eta)} > M_{k-1}^{(\eta)}$ є еквівалентною умові $p_k < \eta^{\frac{1}{1-\eta}}$.

Доведена властивість 3) за умови $M_k^{(\eta)} > M_{k-1}^{(\eta)}$ є уточненням обмеження зверху для ступеня конформності p_k по відношенню до обмеження, отриманого під час доведення властивості 2), оскільки $\forall \eta \in [0;1]\ \eta^{\frac{1}{1-\eta}} < \eta$. В свою чергу, обмеження зверху для p_k, отримане під час доведення властивості 2), є означенням поняття «конформний аномальний детектор».

Таким чином, доведено, що за справедливості умови $M_k^{(\eta)} > M_{k-1}^{(\eta)}$ виконується умова конформного аномального детектору з таким же значенням порогу аномальності ε, як і у параметра чутливості η степеневого мартингала $M_k^{(\eta)}$.

З іншого боку, доведено, що умова $M_k^{(\eta)} > M_{k-1}^{(\eta)}$ на значення степеневого мартингалу еквівалентна тому, що конформний аномальний предиктор має значенням порогу аномальності ε, який дорівнює значенню $\eta^{\frac{1}{1-\eta}}$, де η є значенням параметра чутливості степеневого мартингалу $M_k^{(\eta)}$.

Доведена властивість 1) показує, що умова $M_k^{(\eta)} > M_{k-1}^{(\eta)}$ автоматично виконується, коли $p_k \to 0$. З іншого боку, спостереження (матчі), для яких $p_k \to 0$, теоретично є аномальними. Це є так, оскільки за умовою (3.5) низькі значення ступеня конформності p_k можливі лише тоді, коли міра неконформності a_k поточного спостереження є близькою до максимально можливої, або іншими словами, якщо поточне спостереження є найбільш неконформним (невідповідним) до множини спостережень $\{z_1, z_2, \dots, z_k, \dots, z_N\}$.

Тому для визначення підозрілих на фіксований результат матчів за степеневим або інтегральним мартингалом пропонуються такі правила:

- множина $\Gamma^{\eta}(z_1, z_2, \dots, z_k, \dots, z_{N-1}, z_N)$ складається з матчів z_k, для яких справедлива така умова:

$$M_k^{(\eta)} > M_{k-1}^{(\eta)}; \tag{3.21}$$

- множина $\Gamma(z_1, z_2, \dots, z_k, \dots, z_{N-1}, z_N)$ складається з матчів z_k, для яких справедлива така умова:

$$M_k > M_{k-1}. \tag{3.22}$$

У правилі, описаному (3.21) або (3.22), аналізується зміна значення мартингалу в сусідніх спостереженнях, і тому в подальшому це правило називатиметься як ***правило мартингала.***

Також в ході досліджень корисною виявилася характеристика, схожа на вище наведений степеневий мартингал:

$$\widetilde{M}_{m,k}^{(\eta)} = \eta^m \prod_{i=1}^{k} \left(p_i^{\eta-1}\right), \tag{3.23}$$

де $m \in \boldsymbol{Z} \cap [0; k]$, зокрема при $m = 1$. У степеневому мартингалі параметр $m = k$. Інші значення параметра m дають можливість визначити, скільки перших m спостережень будуть додатково зважуватись шляхом множення на η після отримання значення $p_i^{\eta-1}$. Наприклад, значення m=1 буде проінтерпретовано таким чином: зважуватись множенням на η буде лише значення $p_1^{\eta-1}$, усі наступні $p_i^{\eta-1}$ будуть домножуватись на попереднє значення мартингалу без такого зважування. Множення $\widetilde{M}_{m,k}^{(\eta)}$ на $p_k^{\eta-1}$ за умови $m < k$ та $p_k < 1$ однозначно призводитиме до того, що $\widetilde{M}_{m,k}^{(\eta)}$ збільшуватиметься по відношенню до $\widetilde{M}_{m,k-1}^{(\eta)}$, а якщо $p_k = 1$, тоді буде $\widetilde{M}_{m,k}^{(\eta)} = \widetilde{M}_{m,k-1}^{(\eta)}$. Ця характеристика може використовуватись у випадку, коли будь-яке відхилення p_k від 1 означатиме, що поточне спостереження може бути нестандартним. При цьому усім нормальним спостереженням відповідатиме рівень ступеня конформності $p_k = 1$. Це можливо за умови, якщо введена міра неконформності всіх нормальних спостережень буде однакова. Наприклад, це можливо у випадку, коли міра неконформності

визначена як бінарний класифікатор, який попередньо класифікує об'єкти у два класи: нормальне або аномальне спостереження.

Сформулюємо тепер в узагальненому вигляді ***методи виявлення матчів, підозрілих на фіксований результат, з використанням степеневого та інтегрального мартингалів.***

Нехай маємо сезон футбольних матчів G, розділений на окремі класи матчів $\{G_{ij}\}$ у відповідності до використаного групування команд $group(t)$. Розглянемо клас матчів G_{ij}. Для кожного матчу $z_k \in G_{ij}$:

1) обчислюється міра відмінності a_k за однією з формул (3.2)-(3.4), що є першим етапом при розрахунку конформного предиктора;

2) використовуючи міри відмінності поточного k-го матчу і усіх інших матчів цього ж класу, обчислюється ступінь конформності (відмінності) (p-value) матчу від множини спостережень $\{z_1,\ldots, z_k,\ldots z_{N-1}, z_N\}$ за формулою (3.5);

3) підраховується значення степеневого мартингала (3.12) або інтегрального мартингала (3.13). Степеневий мартингал обчислюється за обраного значення параметру чутливості η;

4) перевіряється виконання умови (3.21) для степеневого мартингалу або (3.22) для інтегрального мартингалу відповідно. Якщо умова виконується, матч вважається підозрілим на фіксований результат і додається до множини таких матчів $\Gamma^{\eta}(z_1, z_2, \ldots, z_k, \ldots, z_{N-1}, z_N)$ або $\Gamma(z_1, z_2, \ldots, z_k, \ldots, z_{N-1}, z_N)$ відповідно.

В розділі 4 здійснюється порівняння запропонованих методів виявлення матчів, підозрілих на фіксований результат, з відомим гістограмним методом [122], який у даному випадку зводиться до перевірки на аномальність матчу на основі ***гістограми різниць м'ячів*** для поточного класу матчів (i, j) за рівнем аномальності p_A $(D_{ij}^{(N)} = \emptyset)$ за ряд послідовних кроків:

1) обирається значення різниці м'ячів $\tilde{d}$, яке за гістограмою поточного класу матчів (i, j) має найбільшу частоту появи h_d серед тих значень d, які не належать $D_{ij}^{(N)}$;

2) додається значення $\tilde{d}$ до множини $D_{ij}^{(N)}$;

3) обчислюється сумарна частота появи усіх значень з множини $D_{ij}^{(N)}$:

$$p_{ij}^{(N)} = \sum_{d \in D_{ij}^{(N)}} h_d ;$$

4) якщо $p_{ij}^{(N)} \geq 1 - p_A$, переходимо на крок 5, інакше — на крок 1;

5) значення можливих різниць м'ячів $d^* \notin D_{ij}^{(N)}$ утворюють множину $D_{ij}^{(A)}$ аномальних різниць класу матчів (i, j);

6) серед усіх матчів у поточному класі матчів визначаємо аномальні матчі за таким правилом:

– матч є аномальним, якщо різниця м'ячів у ньому $d \in D_{ij}^{(A)}$.

Висновки до розділу 3

1. Об'єднання матчів в класи за контекстними атрибутами дозволяє як прогнозне значення чисельного результату матчу використовувати середнє значення різниці м'ячів відповідного класу матчів. Відхилення фактичного результату матчу від очікуваного розглядається як характеристика аномальності матчу по відношенню до визначеного класу матчів (контексту). Також введення відповідної міри неконформності забезпечує можливість порівняння фактичного результату матчу з результатами усіх інших матчів групи і дозволяє враховувати як абсолютні результати команд, так і різницю у підсумках фактичного і прогнозованого результатів.

2. Розроблений на основі аномального конформного детектору метод виявлення підозрілих щодо фіксованості результату футбольних матчів дозволяє виявляти контекстні аномалії даних у класах матчів з використанням запропонованих мір неконформності шляхом порівняння ступеню конформності (p-value) матчу з пороговим значенням. Він відноситься до класу методів виявлення без вчителя і дозволяє вводити оцінки гарантованої

точності для отриманих рішень. Щоб досягти гарного балансу між чутливістю та точністю виявлення, значення порогу слід встановлювати близьким до апріорної ймовірності появи аномальних об’єктів.

3. Доведено такі властивості степеневого мартингалу.

– за яких завгодно малих значеннях ступеня конформності (p-value) поточного спостереження значення степеневого мартингала для поточного спостереження є більшим за значення цього ж мартингала для попереднього спостереження;

– при збільшенні значення степеневого мартингала для поточного спостереження по відношенню до попереднього автоматично виконується правило конформного аномального предиктору на поточне спостереження з таким же значенням рівня аномальності, як і у параметра чутливості η степеневого мартингала $M_k^{(\eta)}$;

– збільшення значення степеневого мартингала для поточного спостереження по відношенню до попереднього еквівалентно виконанню правила конформного аномальному детектору зі значенням рівня аномальності, який дорівнює значенню $\eta^{\frac{1}{1-\eta}}$, де η є значенням параметра чутливості степеневого мартингалу $M_k^{(\eta)}$.

4. У розробленому на основі степеневого мартингалу методі виявлення підозрілих щодо фіксованого результату футбольних матчів прийняття рішення відбувається при зростанні значення степеневого мартингалу для поточного спостереження по відношенню до значення цього ж мартингала для попереднього спостереження. При цьому зміна параметра чутливості дозволяє налаштовувати степеневий мартингал на виявлення аномалій відповідного рівня.

5. У розробленому на основі інтегрального мартингалу методі виявлення підозрілих щодо фіксованості результату футбольних матчів прийняття рішення відбувається при зростанні значення інтегрального мартингалу для поточного спостереження по відношенню до значення цього ж мартингала для

попереднього спостереження. При цьому використання інтегрального мартингала не вимагає налаштування параметрів.

6. Встановлено еквівалентність методу виявлення матчів, підозрілих на фіксований результат на основі експертно визначеного порогу відхилення та конформного аномального детектору, що дозволяє використовувати для опису достовірності отриманих результатів властивості конформного аномального детектора. Розраховане значення порогового рівня ε еквівалетного конформного аномального детектора дорівнює відносній частоті появи матчів у вибірці із значеннями міри неконформності, яка не нижче порогового рівня χ, що є важливим для розуміння отриманих даних як аномальних з позицій ймовірнісного підходу.

РОЗДІЛ 4
АНАЛІЗ МЕТОДІВ ВИЯВЛЕННЯ ПІДОЗРІЛИХ ЩОДО ФІКСОВАНОСТІ РЕЗУЛЬТАТУ МАТЧІВ ЗА НАЯВНОСТІ ДАНИХ ПРО ВЕСЬ СЕЗОН

4.1 Аналіз особливостей розроблених методів за даними окремих класів модельного сезону

Методи виявлення футбольних матчів, підозрілих щодо фіксованості результату, можна розглядати як бінарні класифікатори, які на виході дають значення 1, якщо матч є «потенційно підозрілим на фіксованість результату», і 0 — в протилежному випадку. Важливими для подальшого аналізу будуть такі елементи матриці невідповідностей (confusion matrix) бінарного класифікатора: кількість коректних спрацьовувань (true positives, TP), кількість хибних спрацьовувань (false positives, FP), кількість хибних пропусків (false negatives, FN). TP дорівнює кількості матчів, які є потенційно підозрілими і класифікатор їх виявив такими. FP дорівнює кількості матчів, які не є потенційно підозрілими, але класифікатор їх вважає такими. FN дорівнює кількості матчів, які є потенційно підозрілими, але класифікатор їх помилково пропустив. За цими характеристиками обчислюються метрики точності (precision, P), повноти (recall, R) та їх гармонічного середнього — міри F_1:

$$P = \frac{TP}{TP + FP}, \tag{4.1}$$

$$R = \frac{TP}{TP + FN}, \tag{4.2}$$

$$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2TP}{2TP + FP + FN}. \tag{4.3}$$

Обрані характеристики є базовими характеристиками аналізу ефективності алгоритмів, які застосовуються для розв'язання задач бінарної класифікації. Усі три характеристики приймають значення з діапазону [0; 1] і мають аналогічну інтерпретацію результатів: чим більш близьким до 1 є значення характеристики, тим алгоритм є більш ефективним з точки зору даної характеристики.

Продемонструємо роботу запропонованих методів на одному змодельованому сезоні. Визначення аномальних різниць м'ячів здійснювалося за гістограмами різниць м'ячів по кожному класу матчів на рівні аномальності даних p_A =0,2. Гістограми були побудовані за 100 модельними сезонами за методикою, розглянутою у розділі 2. Після визначення аномальних різниць м'ячів, в поточному сезоні було введено 10 договірних матчів за алгоритмом утворення договірних матчів, розглянутим у розділі 2. Усім введеним договірним матчам було присвоєно клас 2 у характеристиці «Потенційно підозрілий матч». Також на основі визначених аномальних різниць м'ячів було проведено розмітку матчів сезону на предмет їх аномальності. Усім матчам, які були утворені перед введенням договірних матчів і у яких різниця м'ячів була аномальною, було присвоєно клас 1 у характеристиці «Потенційно підозрілий матч». Приклад класу матчів після проведення розмітки і введення договірних показано в табл. 1 на прикладі класу матчів (1, 4). Матчі, які були введені як договірні, позначено синім кольором і в колонці «Потенційно підозрілий» вони мають значення 2. Матчі, які були змодельовані з аномальним результатом, у таблиці позначено сірим кольором і в колонці «Потенційно підозрілий» вони мають значення 1. Усі інші матчі, тобто матчі з очікуваним рахунком, в колонці «Потенційно підозрілий» мають значення 0.

Спочатку розглянемо роботу запропонованих методів виявлення матчів, підозрілих на фіксований результат, на класі матчів (1, 4) (табл. 4.1). В цю групу входять матчі, у яких команда-господарка належить до групи 1, тобто є однією з найуспішніших в цьому сезоні, а гостьова команда належить до групи 4, тобто характеризується одним з найнижчих значень успішності. Середній

результат по групі *avg(i, j)* дорівнює 1,125. Отже, очікуваним результатом матчу є виграш домашньої команди з різницею м'ячів на рівні 1 або 2 м'ячів.

Таблиця 4.1 – Матчі класу (1, 4) модельного сезону

№	Команда-господар	Гостьова команда	Результат	Потенційно підозрілий
1	team 1	team 15	0:0	0
2	team 2	team 15	6:2	2
3	team 3	team 15	3:2	0
4	team 4	team 15	1:2	1
5	team 1	team 17	2:2	0
6	team 1	team 18	2:1	0
7	team 1	team 20	1:1	0
8	team 2	team 17	2:0	0
9	team 2	team 18	2:2	0
10	team 2	team 20	2:2	0
11	team 3	team 17	2:2	0
12	team 3	team 18	6:1	1
13	team 3	team 20	4:1	1
14	team 4	team 17	2:0	0
15	team 4	team 18	3:3	0
16	team 4	team 20	2:1	0

Далі продемонстровано роботу методу виявлення матчів, підозрілих на фіксований результат на основі конформного предиктора і степеневого мартингала. Кожен футбольний матч є окремим спостереженням z_k, які послідовно оброблюються алгоритмом. Спочатку для поточного спостереження z_k обчислюється міра відмінності a_k за однією з формул (3.2)-(3.4). Розглянемо результати, отримані при обчисленні міри неконформності

за формулою (3.2) (рис. 4.1), тобто без округлення середнього результату по групі. Значення цієї міри показують наскільки результат матчу відрізняється за значенням від очікуваного результату, яким для цього методу є середній результат по класу матчів. Чим більшим є значення міри відмінності, тим більше цей матч виділяється з поміж інших за очікуваним результатом. На рис. 4.1 і усіх наступних стовпчиками сірого кольору виділені потенційно підозрілі матчі за принципом розмітки (ті матчі, для яких значення у стовпчику «Потенційно підозрілий» дорівнює 1), а синього кольору — договірні матчі, які були створені за методом з розділу 2 (ті матчі, для яких значення у стовпчику «Потенційно підозрілий» дорівнює 2). Міра відмінності кожного потенційно підозрілого матчу є більшою за міри відмінності інших матчів.

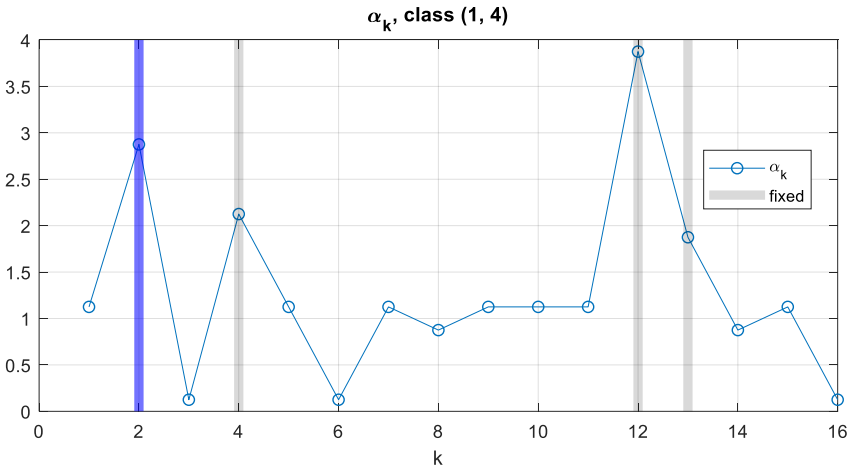


Рисунок 4.1 – Характеристика a_k, обчислена за формулою (3.2) для матчів класу (1, 4)

Далі за множиною матчів $\{z_1, z_2,\ldots, z_k,\ldots, z_N\}$ і отриманими значеннями міри неконформності a_k для кожного спостереження z_k обчислюється ступінь конформності p_k (рис. 4.2) за (3.5). Ця величина приймає значення в діапазоні $[1/k; 1]$ і характеризує частку таких матчів у множині $\{ z_1, z_2,\ldots, z_k,\ldots, z_N \}$, які є більш відмінними від поточного матчу, або такими ж, як поточний матч. Далі ця характеристика може бути проаналізована за ***правилом конформного аномального детектора*** (3.6). За цим правилом підозрілим є той матч, у якого

ступінь конформності p_k є меншим за поріг аномальності ε. На рис. 4.2 зображено результати виявлення підозрілих матчів для класу $(1, 4)$ при $\varepsilon = 0{,}2$. Також на цьому рисунку штрихованими лініями виділено матчі, які за (3.6) є правильними спрацьовуваннями (true positives, TP, зелений колір) та хибними пропусками (false negatives, FN, жовтий колір). На основі значень цих характеристик обчислюються метрики точності (4.1), повноти (4.2) і міра F_1 (4.3). Отже, для класу матчів $(1, 4)$ за (3.6) при $\varepsilon = 0{,}2$ конформний аномальний детектор за метрикою повноти (recall) спрацював на 75 %: виявлено більшість очікуваних підозрілих матчів. За метрикою точності (precision) алгоритм спрацював на 100 %: всі очікувані підозрілі на фіксованість результату матчі виявлено, й по іншим матчам помилок не було. Міра F_1 для класу $(1, 4)$ в цьому випадку дорівнює 0,86, тобто алгоритм загалом для класу $(1, 4)$ спрацював добре, але є можливості для покращення результату.

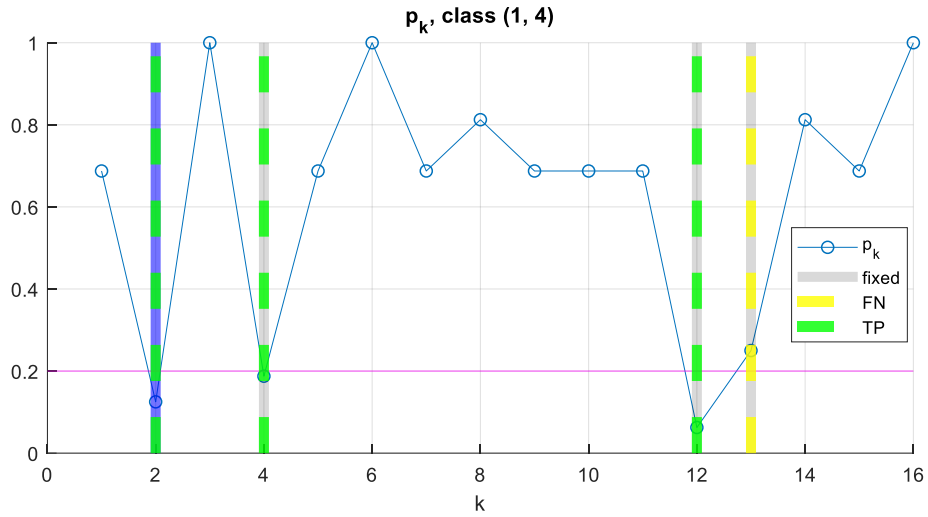


Рисунок 4.2 – Характеристика p_k для матчів класу $(1, 4)$, обчислена за мірою неконформності (3.2) і результати виявлення підозрілих матчів при $\varepsilon = 0{,}2$ за правилом конформного аномального детектора

На рис. 4.3 наведено результати обчислення ***степеневого мартингалу*** (3.12) з параметром чутливості $\eta = 0{,}8$ для класу матчів $(1, 4)$, обчисленим за раніше розглянутими значеннями ступенів конформності. Далі ця

характеристика аналізується за правилом (3.21): точки, у яких значення мартингалу зростає в порівнянні з попередніми, відповідають матчам, які є підозрілими на фіксованість їх результату. На рис. 4.3 також штрихованими лініями виділено матчі, які за (3.21) є правильними спрацьовуваннями (true positives, TP, зелений колір). На основі отриманих результатів обчислюються метрики точності (4.1), повноти (4.2) і міра F_1 (4.3). Отже, для класу матчів (1, 4) за методом виявлення на основі степеневого мартингалу (3.21) за метрикою повноти (recall) маємо спрацьовування на 100 %: виявлено всі очікувані підозрілі матчі. За метрикою точності (precision) алгоритм спрацював на 100 %: всі очікувані матчі виявлено, і хибних виявлень немає. Міра F_1 для класу (1, 4) дорівнює 1, тобто правило (3.21) для класу (1, 4) дало відмінні результати.

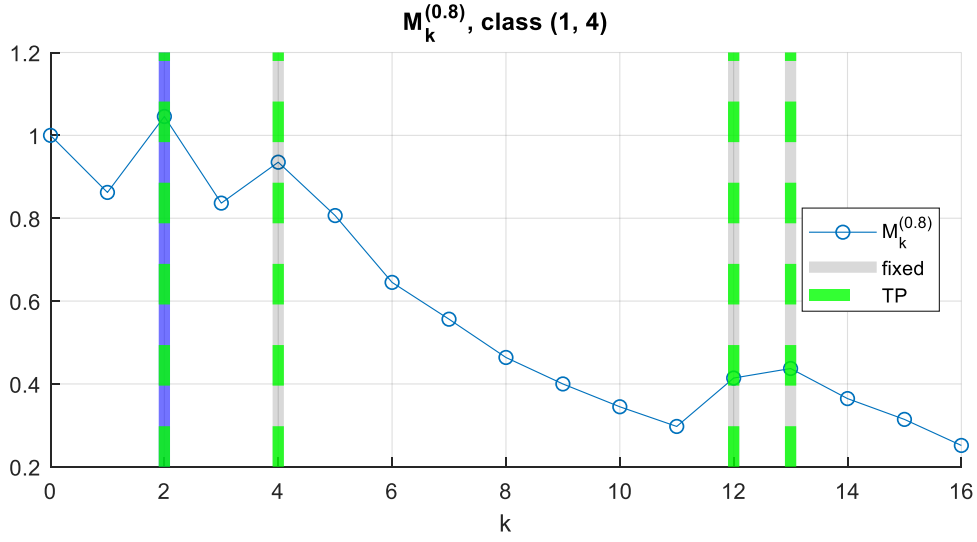


Рисунок 4.3 – Характеристика $M_k^{(\varepsilon)}$ для матчів класу (1, 4) і результати виявлення підозрілих матчів при $\eta = 0{,}8$ за правилом степеневого мартингала

На рис. 4.4 наведено результати обчислення ***інтегрального мартингалу*** (3.13) для класу матчів (1, 4), обчисленого за раніше розглянутими значеннями ступенів конформності. Далі ця характеристика аналізується правилом (3.22): точки, у яких значення інтегрального мартингалу зростає в порівнянні з

попередніми, відповідають матчам, які є підозрілими на фіксованість їх результату. На рис. 4.4 також штрихованими лініями виділено матчі, які за (3.19) є правильними спрацьовуваннями (true positives, TP, зелений колір). На основі цих виявлень обчислюються метрики точності (4.1), повноти (4.2) і міра F_1 (4.3). Отже, для класу матчів (1, 4) за методом виявлення на основі інтегрального мартингалу (3.22) за метрикою повноти (recall) маємо спрацьовування на 100 %: виявлено всі очікувані підозрілі матчі. За метрикою точності (precision) алгоритм спрацював на 100 %: всі очікувані матчі виявлено, і хибних виявлень немає. Міра F_1 для класу (1, 4) дорівнює 1, тобто правило (3.22) для класу (1, 4) також дало відмінні результати.

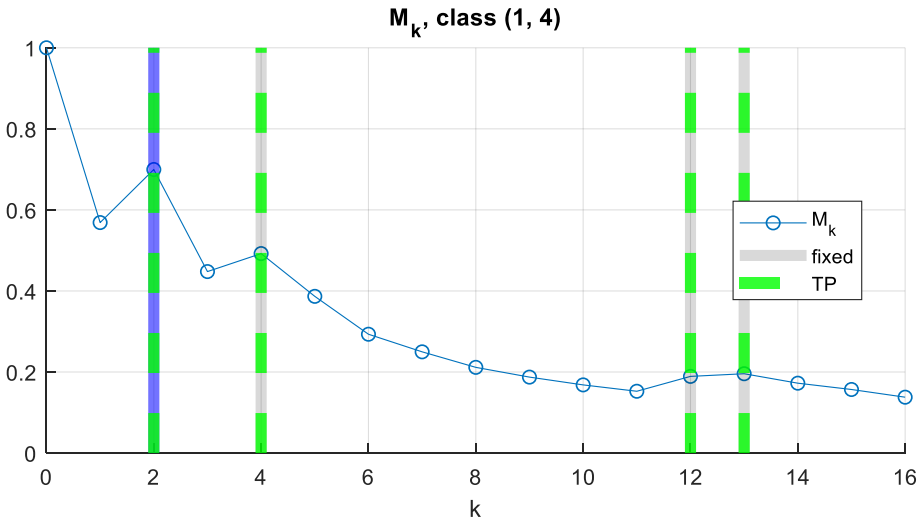


Рисунок 4.4 – Характеристика M_k для матчів класу (1, 4) і результати виявлення підозрілих матчів за правилом інтегрального мартингала

Розглянемо результати, отримані при виявленні аномальних матчів за ***гістограмою різниць м'ячів*** для поточного класу. На рис. 4.5 показано гістограму різниць м'ячів для класу матчів (1, 4), на якій за рівнем аномальності $p_A = 0{,}2$ визначено аномальні різниці м'ячів. В даному випадку, аномальними різницями м'ячів виявились 3, 4 і 5. Також на рис. 4.5 показано результати виявлення аномальних матчів за цією гістограмою: штрихованими лініями виділено матчі, які за гістограмою різниць м'ячів є правильними

спрацьовуваннями (true positives, TP, зелений колір). На основі цих виявлень, так само, як і для правила (3.22), обчислюються метрики точності (4.1), повноти (4.2) і міра F_1 (4.3). Отже, для класу матчів (1, 4) за методом виявлення на основі гістограми різниць м'ячів поточного класу матчів за метрикою повноти (recall) спрацював на 75 %: виявлено більшість очікуваних підозрілих матчів. За метрикою точності (precision) алгоритм спрацював на 100 %: всі очікувані матчі виявлено, і хибних виявлень немає. Міра F_1 для класу (1, 4) дорівнює 0,86, тобто виявлення за гістограмою різниць м'ячів для класу (1, 4) дало гарні результати, але є можливості для удосконалення.

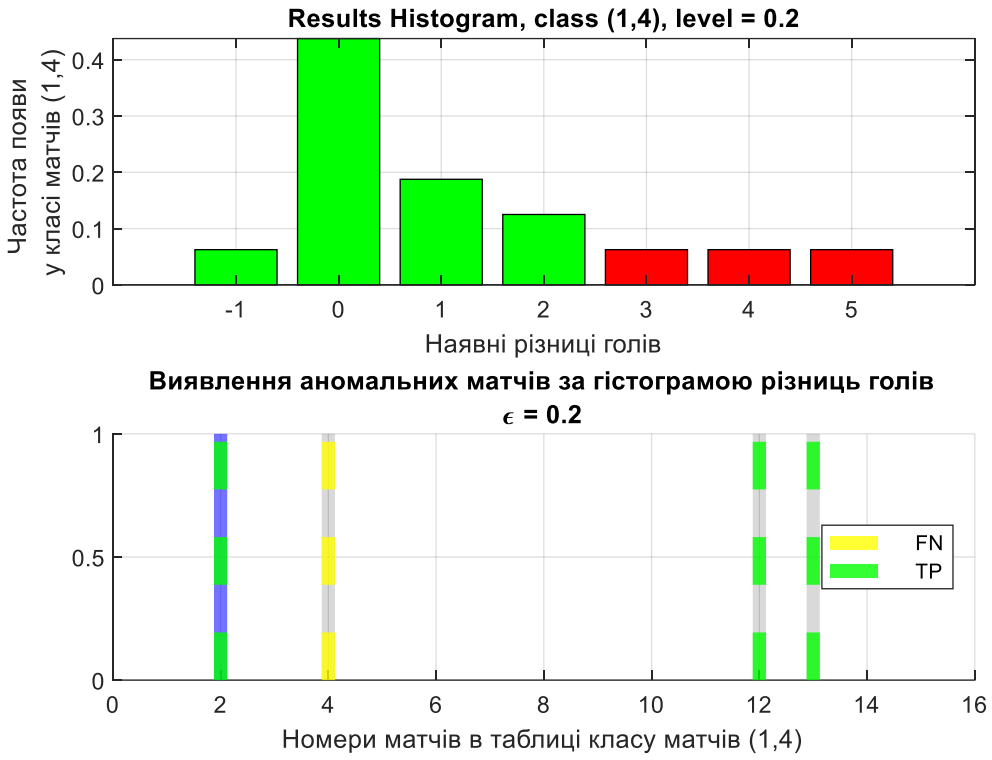


Рисунок 4.5 – Гістограма різниць м'ячів для матчів класу (1,4) і результати виявлення підозрілих матчів за цією гістограмою

Тепер аналогічно розглянемо результати, отримані при обчисленні міри неконформності, але вже за формулою (3.3) (рис. 4.6), тобто з округлення середнього результату по класу матчів. Значення цієї міри, так само як і міри (3.2), показують наскільки результат матчу відрізняється за значенням від

очікуваного результату, яким для цього методу є середній результат по класу матчів. Чим більшим є значення міри відмінності, тим більше цей матч виділяється з поміж інших за очікуваним результатом. Відмінність полягає лише в тому, що середній результат по класу матчів тепер є цілим числом. Середній результат по класу матчів (1, 4) з урахуванням округлення дорівнює 1. За отриманими значеннями міри неконформності на даному класі матчів чітко можна відділити матчі, які відносяться до нормальних, від тих, які є аномальними за своїм результатом. Таким чином, для цього класу можна було б застосувати спрощений принцип пошуку підозрілих матчів – за перевіркою чи є більшою міра неконформності за 1.

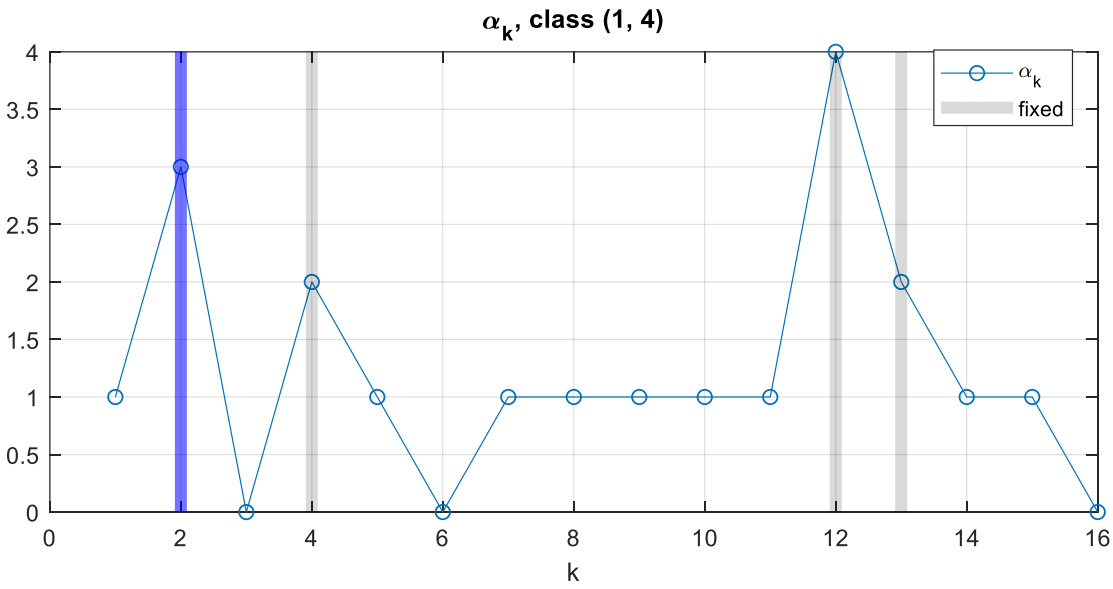


Рисунок 4.6 – Характеристика a_k, обчислена за формулою (3.3) для матчів класу (1, 4)

Далі за множиною матчів $\{z_1, z_2,\ldots, z_k,\ldots, z_N\}$ і отриманими значеннями міри неконфорності a_k для кожного спостереження z_k обчислюється ступінь конформності p_k (рис. 4.7) за (3.5). На рис. 4.7 продемонстровано результати виявлення підозрілих матчів для класу (1, 4) за (3.6) при $\varepsilon = 0{,}2$. При цьому порозі аномальності два з потрібних матчів не будуть виявлені. Хибних спрацьовувань при цьому також не відбувається.

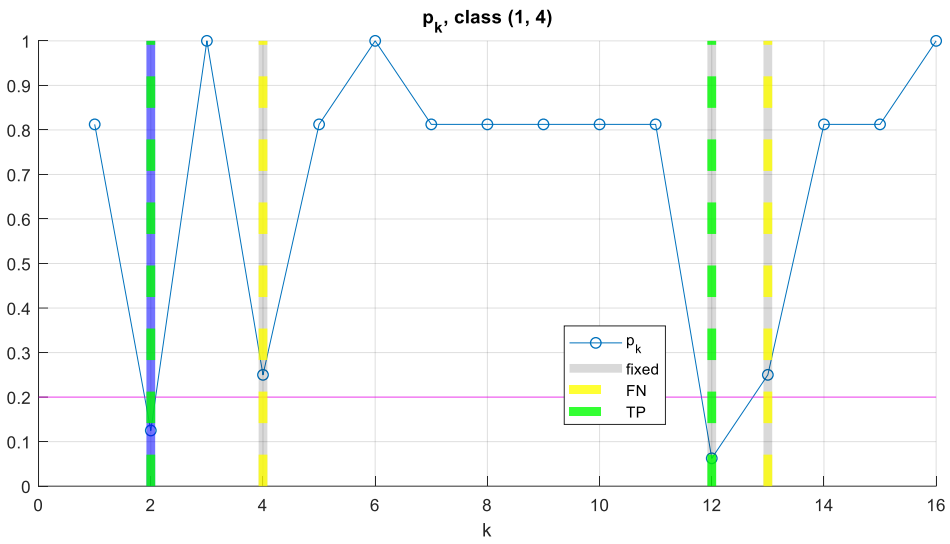


Рисунок 4.7 – Характеристика p_k для матчів класу (1, 4), обчислена за мірою неконформності (3.3) і результати виявлення підозрілих матчів при $\varepsilon = 0{,}2$ за правилом конформного аномального детектора (3.6)

Отже, для класу матчів $(1, 4)$ за (3.6) при $\varepsilon = 0{,}2$ конформний аномальний детектор за метрикою повноти (recall) спрацював на 50 %: виявлено половину очікуваних підозрілих матчів. За метрикою точності (precision) алгоритм спрацював на 100 %: серед виявлених матчів знаходяться лише очікувані матчі. Міра F_1 для класу (1, 4) в цьому випадку дорівнює 0,67, тобто алгоритм загалом для класу (1, 4) спрацював непогано, але є можливості для покращення. У порівнянні з результатами конформного аномального детектора, отриманими при використанні міри неконформності (3.2), поточні результати погіршились за мірою повноти: було виявлено на 1 очікуваний підозрілий матч менше, ніж при використанні міри неконформності (3.2).

Тепер обчислимо значення степеневого мартингала $M_k^{(\eta)}$ та інтегрального мартингалу M_k. На рис. 4.8 наведено результати обчислення степеневого мартингалу (3.12) з параметром чутливості $\eta = 0{,}8$, обчисленим за раніше розглянутими значеннями ступенів конформності для класу матчів (1, 4). Для класу матчів (1, 4) за методом виявлення на основі степеневого мартингалу (3.21) за метрикою повноти (recall) алгоритм спрацював на 100 %:

виявлено всі очікувані підозрілі матчі. За метрикою точності (precision) алгоритм спрацював на 100 %: всі очікувані матчі виявлено, і хибних виявлень немає. Міра F_1 для класу (1, 4) дорівнює 1, тобто правило (3.21) для класу (1, 4) дало відмінні результати. Результати, отримані в даному випадку є ідентичними до результатів, отриманих за допомогою степеневого мартингала з параметром чутливості $\eta = 0{,}8$, обчисленим при використанні міри неконформності (3.2).

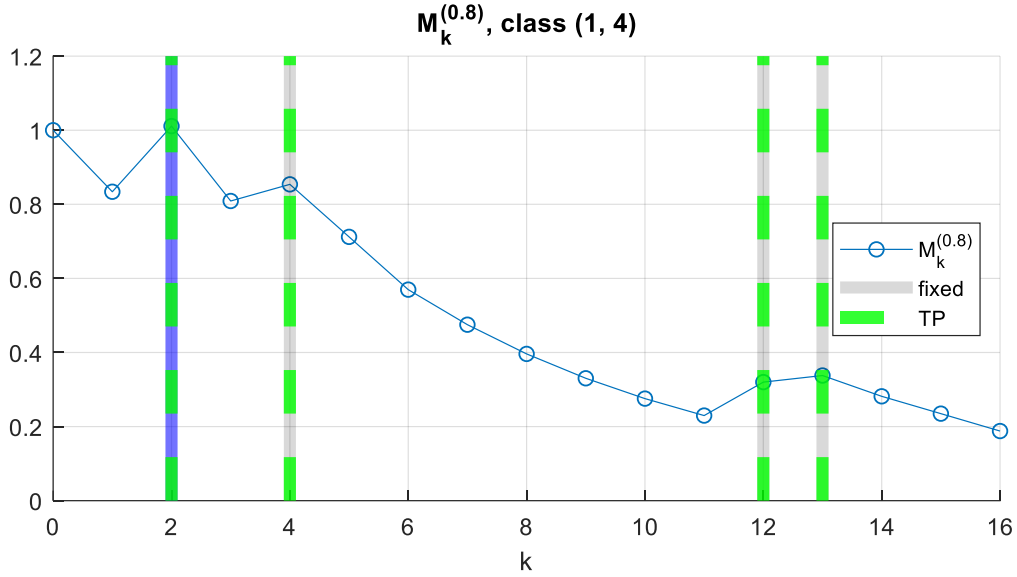


Рисунок 4.8 – Характеристика $M_k^{(\eta)}$ для матчів класу (1, 4) і результати виявлення підозрілих матчів при $\eta = 0{,}8$ за правилом степеневого мартингала (3.21)

На рис. 4.9 наведено результати обчислення інтегрального мартингалу (3.13) для класу матчів (1, 4), обчисленим за раніше розглянутими значеннями ступенів конформності. Для класу матчів (1, 4) за методом виявлення на основі інтегрального мартингалу (3.22) за метрикою повноти (recall) спрацював на 100 %: виявлено всі очікувані підозрілі матчі. За метрикою точності (precision) алгоритм спрацював на 100 %: всі очікувані матчі виявлено, і хибних виявлень немає. Міра F_1 для класу (1, 4) дорівнює 1, тобто правило (3.22) для класу (1, 4) дало відмінні результати.

Результати, отримані в даному випадку, є ідентичними до результатів, отриманих за допомогою інтегрального мартингала правила (3.22), обчисленим при використанні міри неконформності (3.2).

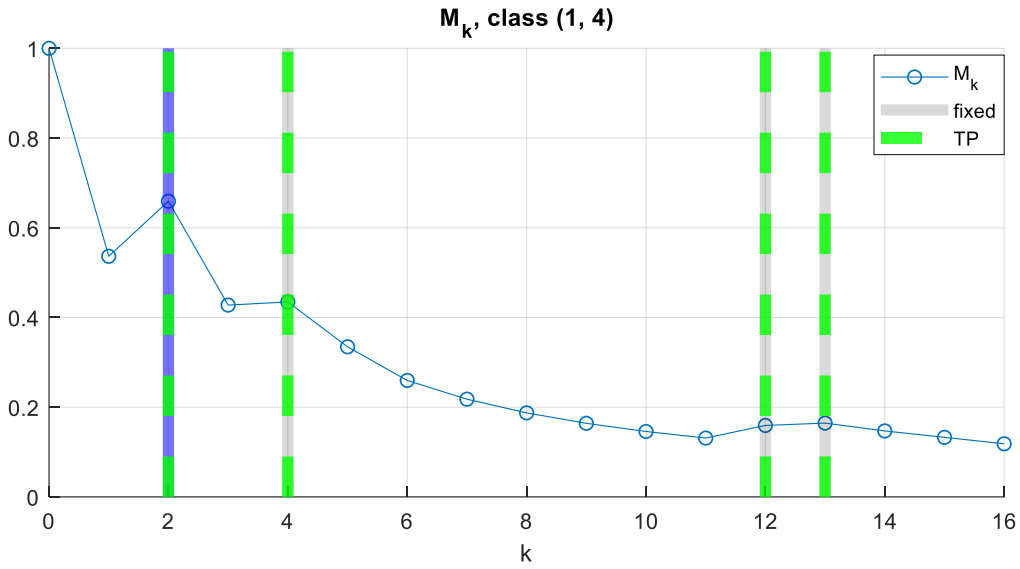


Рисунок 4.9 – Характеристика M_k для матчів класу (1, 4) і результати виявлення підозрілих матчів за правилом інтегрального мартингала (3.22)

Тепер розглянемо роботу методів на класі матчів (4, 1), тобто на класі, симетричному до попереднього. Середній результат по класу матчів $avg(i,j)$ дорівнює –0,875. Отже, очікуваним результатом матчу є нічия або виграш гостьової команди з різницею в один м’яч. Більшість підозрілих матчів у цьому класі мають велику різницю м’ячів — не менша за 3 м’ячі (табл. 4.2).

Графік міри відмінності a_k кожного матчу даного класу, обчислену за формулою (3.2), показано на рис. 4.10. На відміну від ситуації з класом (1, 4), у цьому класі існують матчі, міра відмінності яких знаходиться на одному рівні з потенційно підозрілими матчами. У даному класі матчів всього один такий матч (№ 7), але за певних умов такі матчі можуть бути причиною хибних виявлень.

Таблиця 4.2 – Матчі класу (4, 1) модельного сезону

№	Домашня команда	Виїзна команда	Результат	Потенційно підозрілий
1	team 15	team 1	0:0	0
2	team 15	team 2	3:3	0
3	team 15	team 3	3:3	0
4	team 15	team 4	3:2	1
5	team 17	team 1	1:1	0
6	team 17	team 2	4:1	2
7	team 17	team 3	0:3	0
8	team 17	team 4	1:2	0
9	team 18	team 1	1:1	0
10	team 18	team 2	1:1	0
11	team 18	team 3	0:4	1
12	team 18	team 4	1:2	0
13	team 20	team 1	1:2	0
14	team 20	team 2	0:2	0
15	team 20	team 3	0:5	1
16	team 20	team 4	1:2	0

На рис. 4.11 показано результати виявлення підозрілих матчів для класу (4, 1) за принципом, сформульованим для *конформного аномального детектора* (3.6) при $\varepsilon = 0{,}2$. На цьому рисунку штрихованими лініями виділено матчі, які за (3.6) є правильними спрацьовуваннями (true positives, TP, зелений колір) та хибними пропусками (false negatives, FN, жовтий колір). На основі цих виявлень обчислюються метрики точності (4.1), повноти (4.2) і міра F_1 (4.3). Отже, при $\varepsilon = 0{,}2$ для класу матчів (4, 1) конформний аномальний детектор (3.6) за метрикою повноти (recall) спрацював на 75 %: виявлено більшість очікуваних підозрілих матчів. За метрикою точності

(precision) алгоритм спрацював на 100 %. Міра F_1 для класу (4, 1) дорівнює 0,86, що є ознакою того, що на даному класі матчів детектор спрацював гарно, але все одно є можливості для покращення результату.

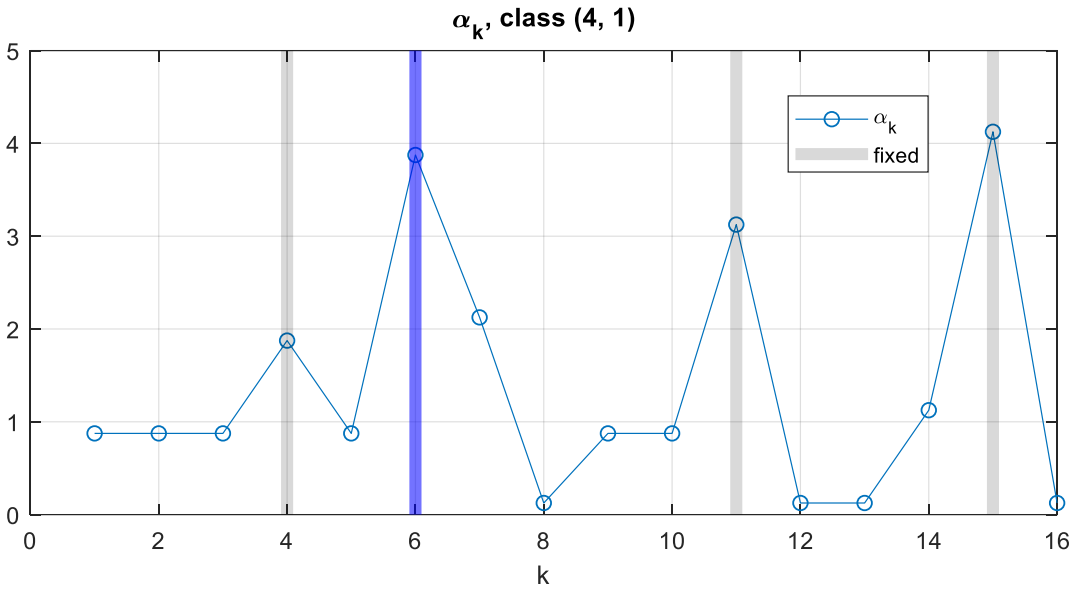


Рисунок 4.10 – Характеристика a_k, обчислена за формулою (3.2) для матчів класу (4, 1)

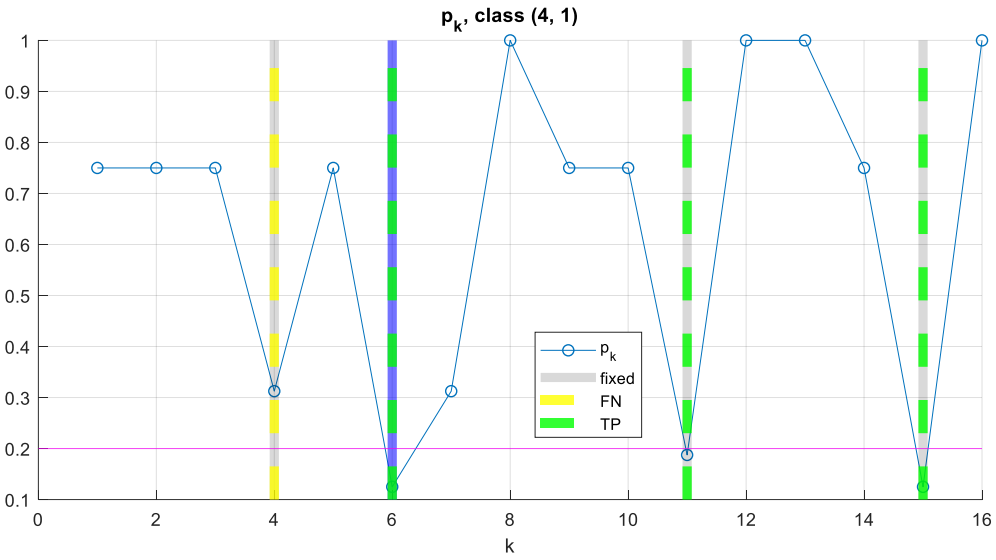


Рисунок 4.11 – Характеристика p_k для матчів класу (4, 1), обчислена за мірою неконформності (3.2) і результати виявлення підозрілих матчів при $\varepsilon = 0{,}2$ за правилом конформного аномального детектора (3.6)

На рис. 4.12 показано результати виявлення підозрілих матчів для класу (4, 1) за принципом, сформульованим для *степеневого мартингалу* (3.12) з параметром чутливості $\eta = 0{,}8$, обчисленим за раніше розглянутими значеннями ступенів конформності. Додатково до позначень, прийнятих на рис. 4.11, на рис. 4.12 штрихованими лініями червоного кольору виділено матчі, які є хибними спрацьовуваннями (false positives, FP). За метрикою повноти (recall) алгоритм спрацював на 100%: виявлено всі очікувані підозрілі матчі. За метрикою точності (precision) алгоритм спрацював на 80%: всі очікувані матчі виявлено, але присутні й хибні виявлення. Міра F_1 для класу (4, 1) дорівнює 0,89, тобто правило (3.12) для класу (4, 1) дало результати, які є кращими, ніж у конформного аномального детектора, але все одно є можливості для покращення. У порівнянні з результатами конформного аномального детектора, отриманими при використанні міри неконформності (3.2), поточні результати погіршились за метрикою точності (зменшення до 80%), але покращились за метрикою повноти (зростання до 100%). Водночас, за метрикою F_1 поточні результати є трохи кращими, ніж результати конформного аномального детектора.

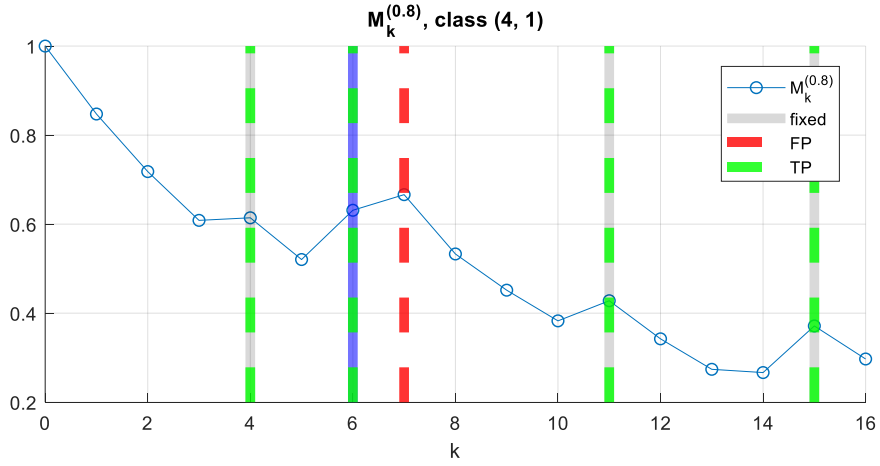


Рисунок 4.12 – Характеристика $M_k^{(\eta)}$ для матчів класу (4, 1) і результати виявлення підозрілих матчів при $\eta = 0{,}8$ за правилом степеневого мартингала (3.21)

На рис. 4.13 наведено результати обчислення інтегрального мартингалу (3.13) для класу матчів (4, 1), обчисленим за раніше розглянутими значеннями ступенів конформності. Для класу матчів (4, 1) за методом виявлення на основі інтегрального мартингалу (3.22) за метрикою повноти (recall) алгоритм спрацював на 75 %: виявлено більшість очікуваних підозрілих матчів. За метрикою точності (precision) алгоритм спрацював на 75 %: всі очікувані матчі виявлено, але присутні хибні спрацьовування. Міра F_1 для класу (4, 1) дорівнює 0,75, що є ознакою того, що метод виявлення на основі інтегрального мартингалу (3.22) спрацював нормально, але є можливості для покращення результату.

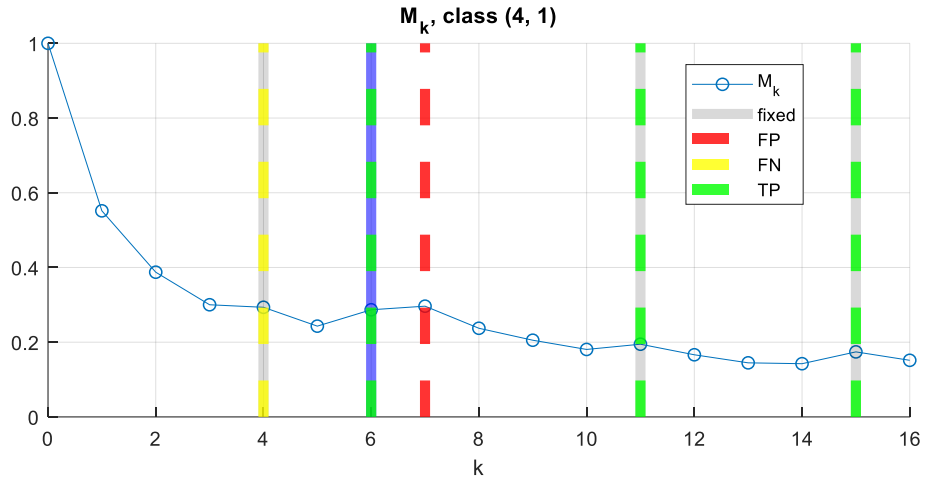


Рисунок 4.13 – Характеристика M_k для матчів класу (4, 1) і результати виявлення підозрілих матчів за правилом інтегрального мартингала (3.22)

Розглянемо результати, отримані при виявленні аномальних матчів за гістограмою різниць м’ячів для поточного класу. На рис. 4.14 показано гістограму різниць м’ячів для класу матчів (4, 1), на якій за рівнем аномальності $p_A = 0{,}2$ визначено аномальні різниці м’ячів. В даному випадку, аномальними різницями м’ячів виявились –2, 1 і 3. Також на рис. 4.14 показано результати виявлення аномальних матчів за цією гістограмою: зеленими штрихованими лініями виділено матчі, які мають результат, що відповідає

червоним ствопцям на гістограмі і при цьому відмічені як дійсно аномальні матчі (true positives, TP, зелений колір), червоними штрихованими лініями виділено матчі, які мають результат, що відповідає червоним ствопцям на гістограмі і при цьому не відмічені як аномальні матчі (false positives, FP, червоний колір). На основі цих виявлень, так само, як і для правила (3.22), обчислюються метрики точності (4.1), повноти (4.2) і міра F_1 (4.3). Отже, для класу матчів (4, 1) за методом виявлення на основі гістограми різниць м'ячів поточного класу матчів за метрикою повноти (recall) спрацював на 67 %: з виявлених матчів більша частина є очікуваними підозрілими матчами. За метрикою точності (precision) алгоритм спрацював на 50 %: половину очікуваних матчів виявлено. Міра F_1 для класу (4, 1) дорівнює 0.571, тобто виявлення за гістограмою різниць м'ячів для класу (4, 1) дало результати, які потребують удосконалення.

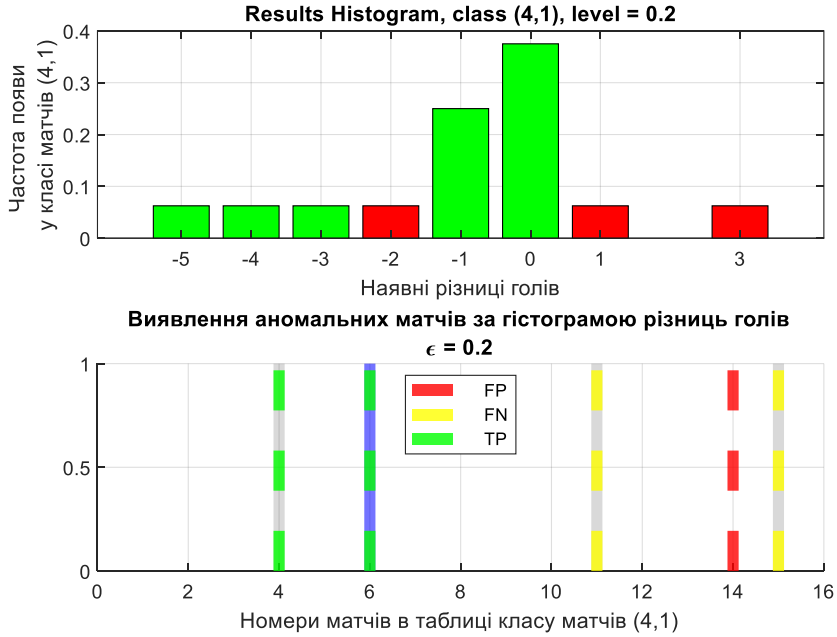


Рисунок 4.14 – Гістограма різниць м'ячів для матчів класу (4, 1) і результати виявлення підозрілих матчів за цією гістограмою

Гістограми різниць м'ячів для матчів класів (1,4) і (4,1) максимально відрізняються між собою у порівнянні з іншими «симетричними» класами,

тому вони й були обрані для демонстрації особливостей роботи розроблених методів.

4.2 Порівняльний аналіз розроблених методів за даними модельного сезону

Оцінки ефективності результатів виявлення матчів, підозрілих на фіксований результат, за гістограмою різниць м'ячів класу матчів поточного модельного сезону за рівнями аномальності даних $p_A = 0{,}2$ та $p_A = 0{,}3$ наведено в табл. 4.3. Розглянуті модельні сезони сформовано за алгоритмом, наведеним у розділі 2, використовуючи дані реального сезону 2013-2014 років Ліги II Франції. Розмітка даних виконується за гістограмами різниць м'ячів класів матчів, отриманими за 100 модельних сезонів (рис. 2.12).

Таблиця 4.3 – Оцінки ефективності результатів виявлення матчів, підозрілих на фіксований результат, за гістограмами різниць м'ячів класів матчів поточного сезону

Клас	Обсяг вибірки, N	$p_A = 0{,}2$						$p_A = 0{,}3$					
		TP	FN	FP	P	R	F_1	TP	FN	FP	P	R	F_1
(1, 1)	12	0	2	1	0%	0%	0	2	0	1	67%	100%	0,8
(1, 2)	16	0	0	3	Немає аномальних матчів			0	0	3	Немає аномальних матчів		
(1, 3)	35	2	6	2	50%	25%	0,33	2	6	2	50%	25%	0,33
(1, 4)	16	3	1	0	100%	75%	0,86	4	0	0	100%	100%	1
(2, 1)	16	2	0	1	67%	100%	0,8	2	3	1	67%	40%	0,5
(2, 2)	12	0	0	1	Немає аномальних матчів			0	3	3	0%	0%	0
(2, 3)	32	5	0	0	100%	100%	1	8	0	0	100%	100%	1
(2, 4)	16	3	0	0	100%	100%	1	3	3	0	100%	50%	0,67
(3, 1)	32	6	0	0	100%	100%	1	6	4	0	100%	60%	0,75
(3, 2)	32	1	5	4	20%	17%	0,18	5	5	0	100%	50%	0,67
(3, 3)	56	5	0	4	56%	100%	0,71	9	0	0	100%	100%	1
(3, 4)	32	2	5	3	40%	29%	0,33	2	5	3	40%	29%	0,33
(4, 1)	16	2	2	1	67%	50%	0,57	3	2	1	75%	60%	0,67
(4, 2)	16	3	0	0	100%	100%	1	3	3	0	100%	50%	0,67
(4, 3)	32	4	2	1	80%	67%	0,73	7	0	0	100%	100%	1
(4, 4)	12	1	2	1	50%	33%	0,4	2	2	1	67%	50%	0,57
Середні характеристики					66%	64%	0,64				78%	61%	0,66

Виявлення матчів, підозрілих на фіксований результат, в свою чергу, за цим методом відбувається за гістограмами різниць м'ячів класів матчів,

побудованими лише за поточним сезоном. Комірки у стовпцях метрик точності P, R і F_1 мають діапазонне чотирьох-кольорове зафарбування. Червоним кольором зафарбовані комірки зі значеннями з діапазону [0,4; 0,6) або [40 %; 60 %). Помаранчевим кольором зафарбовані комірки зі значеннями з діапазону [0,6; 0,75) або [60 %; 75 %). Жовтим кольором зафарбовані комірки зі значеннями з діапазону [0,75; 0,9) або [75 %; 90 %). Зеленим кольором зафарбовані комірки зі значеннями з діапазону [0,9; 1] або [90 %; 100 %].

Як випливає з табл. 4.3 за рівнями аномальності даних $p_A = 0{,}2$ та $p_A = 0{,}3$ гістограмний метод виявлення аномалій продемонстрував низьку якість роботи, причому за метрикою F_1 якість роботи в обох випадках є в середньому майже однаковою.

Зокрема, у результатах зустрічається така унікальна ситуація: у класі (1, 1) алгоритм не виявив жодного справжнього аномального матчу. З цього можна зробити висновок, що гістограма різниць м’ячів по класу матчів, утворена лише за поточним сезоном, може суттєво відрізнятись від такої гістограми, побудованої за багатьма сезонами. Це, в свою чергу, призводить до того, що аномальні матчі, визначені за гістограмою різниць м’ячів багатьох сезонів, за певного рівня аномальності можуть вважатись неаномальними за гістограмою різниць м’ячів поточного сезону.

Необхідно також відмітити, що підвищення рівня аномальності призвело в середньому до підвищення якості виявлення за метрикою точності, а також зменшенню якості за метрикою повноти. Це обумовлено тим, що при збільшенні рівня аномальності p_A збільшується кількість аномальних даних у вибірці і, відповідно, зменшується кількість неаномальних результатів. Це призводить до збільшення правильних виявлень аномалій TP та зменшення хибних виявлень аномалій FP у виразі (4.1), що дає збільшення метрики точності. І навпаки, зі збільшенням у вибірці аномальних даних збільшується кількість хибних виявлень нормальних даних FN у виразі (4.2), що призводить до зменшення метрики повноти. Ці закономірності в подальшому

простежуються в усіх алгоритмах виявлення матчів, підозрілих на фіксований результат.

Оцінки ефективності застосування на різних класах матчів методу виявлення матчів підозрілих на фіксований результат на основі конформного аномального детектора (3.6) при мірі неконформності (3.2), наведено в табл. 4.4. Результати наведені для двох випадків розмітки даних: при розмітці на рівнях аномальності $p_A = 0{,}2$ та $p_A = 0{,}3$. Значення порогу аномальності ε обрано за правилом $\varepsilon = p_A$ у відповідності до рекомендацій щодо порога аномальності з розділу 3.2. Комірки у стовпцях метрик точності P, R, F_1 мають таке ж кольорове оформлення, що й в табл. 4.3.

Таблиця 4.4 – Оцінки ефективності методу виявлення матчів, підозрілих на фіксований результат, на основі конформного аномального детектора (3.6)

$p_k < \varepsilon, \varepsilon = p_A$													
Клас	Обсяг вибірки, N	$p_A = 0{,}2$						$p_A = 0{,}3$					
		TP	FN	FP	P	R	F_1	TP	FN	FP	P	R	F_1
(1, 1)	12	2	0	0	100%	100%	1	2	0	1	67%	100%	0,8
(1, 2)	16	0	0	0	Немає аномальних матчів			0	0	0	Немає аномальних матчів		
(1, 3)	35	1	7	0	13%	100%	0,22	8	0	0	100%	100%	1
(1, 4)	16	3	1	0	75%	100%	0,86	4	0	0	100%	100%	1
(2, 1)	16	2	0	0	100%	100%	1	2	3	0	100%	40%	0,57
(2, 2)	12	0	0	2	Немає аномальних матчів			0	3	2	0%	0%	0
(2, 3)	32	5	0	0	100%	100%	1	8	0	0	100%	100%	1
(2, 4)	16	3	0	0	100%	100%	1	3	3	0	100%	50%	0,67
(3, 1)	32	6	0	0	100%	100%	1	6	4	0	100%	60%	0,75
(3, 2)	32	1	5	4	20%	17%	0,18	5	5	0	100%	50%	0,67
(3, 3)	56	5	0	4	56%	100%	0,71	9	0	0	100%	100%	1
(3, 4)	32	2	5	0	100%	29%	0,44	7	0	0	100%	100%	1
(4, 1)	16	3	1	0	100%	75%	0,86	4	1	0	100%	80%	0,89
(4, 2)	16	3	0	0	100%	100%	1	3	3	0	100%	50%	0,67
(4, 3)	32	6	0	0	100%	100%	1	7	0	0	100%	100%	1
(4, 4)	12	2	1	0	100%	67%	0,8	3	1	0	100%	75%	0,86
Середні характеристики					83%	85%	0,79				91%	74%	0,79

Підвищення рівня аномальності призвело в середньому до підвищення якості виявлення за метрикою точності на 8%, а також зменшенню якості за метрикою повноти на 11% у порівнянні з випадком $p_A = 0{,}2$. Показник метрики F_1 у середньому не змінився.

У таблиці 4.5 наведено оцінки ефективності застосування на різних класах матчів методу виявлення матчів, підозрілих на фіксований результат, на основі степеневих мартингалів з $\eta = 0{,}2$ та $\eta = 0{,}5$ за описаним правилом (3.21) при мірі неконформності (3.2). Розмітка матчів відбувалась на рівні аномальності $p_A = 0{,}2$. Комірки у стовпцях метрик точності P, R, F_1 мають таке ж кольорове оформлення, що й в табл. 4.3.

Таблиця 4.5 – Оцінки ефективності методу виявлення матчів, підозрілих на фіксований результат, на основі степеневих мартингалів за (3.21) при $\eta = 0{,}2$ та $\eta = 0{,}5$ при рівні аномальності $p_A = 0{,}2$

Клас	Обсяг вибірки, N	$M_k^{(0,2)} > M_{k-1}^{(0,2)}$						$M_k^{(0,5)} > M_{k-1}^{(0,5)}$					
		TP	FN	FP	P	R	F_1	TP	FN	FP	P	R	F_1
(1,1)	12	0	2	0	0%	0%	0	2	0	0	100%	100%	1,00
(1,2)	16	0	0	0	Немає аномальних матчів			0	0	16	Немає аномальних матчів		
(1,3)	32	1	7	0	100%	13%	0,22	7	1	0	100%	88%	0,93
(1,4)	16	2	2	0	100%	50%	0,67	3	1	0	100%	75%	0,86
(2,1)	16	2	0	0	100%	100%	1	2	0	0	100%	100%	1,00
(2,2)	12	0	0	0	Немає аномальних матчів			0	0	2	Немає аномальних матчів		
(2,3)	32	3	2	0	100%	60%	0,75	5	0	0	100%	100%	1,00
(2,4)	16	1	2	0	100%	33%	0,5	3	0	0	100%	100%	1,00
(3,1)	32	2	4	0	100%	33%	0,5	6	0	0	100%	100%	1,00
(3,2)	32	1	5	0	100%	17%	0,29	1	5	4	20%	17%	0,18
(3,3)	56	5	0	0	100%	100%	1	5	0	4	56%	100%	0,71
(3,4)	32	2	5	0	100%	29%	0,44	7	0	0	100%	100%	1,00
(4,1)	16	2	2	0	100%	50%	0,67	3	1	0	100%	75%	0,86
(4,2)	16	2	1	0	100%	67%	0,8	3	0	0	100%	100%	1,00
(4,3)	32	3	3	0	100%	50%	0,67	6	0	1	86%	100%	0,92
(4,4)	12	1	2	0	100%	33%	0,5	2	1	0	100%	67%	0,80
Середні характеристики					93%	45%	57%				90%	87%	88%

У результатах, отриманих за правилом (3.21) метрика точності (P) в переважній більшості ситуацій сягає 100 %, тобто при значеннях параметра $\eta = 0{,}2$ та $\eta = 0{,}5$ чутливості правило степеневого мартингала дозволяє виявити саме аномальні матчі у більшості класів матчів.

При цьому, при $\eta = 0{,}2$ середнє значення метрика повноти R дорівнює 45%, тобто є низьким і не є чутливим до аномальних результатів матчів. При $\eta = 0{,}5$ метрика повноти підвищується на 40%, тобто з підвищенням значення η метод стає значно більш чутливим до аномальних матчів.

За мірою F_1 спостерігається така ж тенденція. При $\eta = 0{,}2$ значення міри F_1 дорівнює 57%, що обумовлено низьким значенням метрики *R*. При $\eta = 0{,}5$ значення міри F_1 підвищується до 88%, що викликано різким зростанням метрики *R*. Наведені результати демонструють вплив параметра чутливості степеневого мартингала η на метрику *R,* яка характеризує чутливість алгоритму виявлення на основі степеневого мартингала.

Також у результатах присутня унікальна ситуація: у класі (1,1) алгоритм при значенні $\eta = 0{,}2$ не виявив жодного аномального матчу. З цього можна зробити висновок, що при значенні $\eta = 0{,}2$ метод на основі степеневого мартингала (3.21) не здатний виявляти аномальні матчі, *p*-value яких $p_k \leq \frac{2}{12} \sim 0{,}167$. З іншого боку, у випадку з класами матчів (2,1) та (3,3), чутливість алгоритму виявилась достатньою для того, що виявити всі необхідні матчі і не зробити хибних виявлень: значення метрик точності (*P*) та повноти (*R*) сягають 100%, і відповідно міра $F_1 = 1$.

У табл. 4.6 наведено оцінки ефективності застосування на різних класах матчів методу виявлення матчів, підозрілих на фіксований результат, на основі степеневого мартингалу з $\eta = 0{,}8$ (3.21), обчисленого при мірі неконформності (3.2). Розмітка матчів відбувалась на рівнях аномальності $p_A = 0{,}2$ та $p_A = 0{,}3$. Комірки у стовпцях табл. 4.6 метрик точності *P*, *R* і F_1 мають таке ж кольорове оформлення, що й в табл. 4.3.

При використанні степеневого мартингалу з $\eta = 0{,}8$ і $p_A = 0{,}2$ у порівнянні з результатами, наведеними в таблиці 4.5, відбулося збільшення показника метрики повноти до 100% по всім класам матчів, зменшення показника метрики точності на 19%, але показник метрики F_1 підвищився на 0,27. Отже, при збільшенні параметра чутливості η відбувається збільшення показника метрики повноти. При значенні параметра чутливості $\eta = 0{,}8$ і $p_A = 0{,}3$ також відбувається збільшення показника метрики точності і зниження метрики повноти, але вони приймають близькі високі значення (86% та 85% відповідно), а значення параметра метрики F_1 не змінюється.

Таблиця 4.6 – Оцінки ефективності методу виявлення матчів, підозрілих на фіксований результат, на основі степеневого мартингалу (3.18) при $\eta = 0{,}8$, обчислених для рівнів аномальності $p_A = 0{,}2$ та $p_A = 0{,}3$

$M_k^{(0,8)} > M_{k-1}^{(0,8)}$													
Клас	Обсяг вибірки N	$p_A = 0{,}2$						$p_A = 0{,}3$					
		TP	FN	FP	P	R	F_1	TP	FN	FP	P	R	F_1
(1, 1)	12	2	0	1	67%	100%	0,8	2	0	1	67%	100%	0,8
(1, 2)	16	0	0	0	Немає аномальних матчів			0	0	0	Немає аномальних матчів		
(1, 3)	35	8	0	2	80%	100%	0,89	8	0	2	80%	100%	0,89
(1, 4)	16	4	0	0	100%	100%	1	4	0	0	100%	100%	1
(2, 1)	16	2	0	3	40%	100%	0,57	5	0	0	100%	100%	1
(2, 2)	12	0	0	2	Немає аномальних матчів			0	3	2	0%	0%	0
(2, 3)	32	5	0	3	63%	100%	0,77	8	0	0	100%	100%	1
(2, 4)	16	3	0	0	100%	100%	1	3	3	0	100%	50%	0,67
(3, 1)	32	6	0	4	60%	100%	0,75	10	0	0	100%	100%	1
(3, 2)	32	6	0	4	60%	100%	0,75	10	0	0	100%	100%	1
(3, 3)	56	5	0	4	56%	100%	0,71	9	0	0	100%	100%	1
(3, 4)	32	7	0	3	70%	100%	0,82	7	0	3	70%	100%	0,82
(4, 1)	16	4	0	1	80%	100%	0,89	5	0	0	100%	100%	1
(4, 2)	16	3	0	0	100%	100%	1	3	3	0	100%	50%	0,67
(4, 3)	32	6	0	4	60%	100%	0,75	7	0	3	70%	100%	0,82
(4, 4)	12	3	0	0	100%	100%	1	3	1	0	100%	75%	0,86
Середні характеристики					74%	100%	0,84				86%	85%	0,84

Водночас, була перевірена умова еквівалентності методу на основі конформного аномального детектора і степеневого мартингала, що доведена у рамках твердження в підрозділі 3.3 і має вигляд:

$$\varepsilon = \eta^{\frac{1}{1-\eta}}.$$

При $\eta = 0{,}8$ маємо $\varepsilon = 0{,}32768$. Розраховані показники метода виявлення на основі конформного аномального детектора повністю співпали з показниками метода виявлення на основі степеневого мартингала.

У табл. 4.7 наведено оцінки ефективності застосування на різних класах матчів методу виявлення матчів підозрілих на фіксований результат на основі інтегрального мартингалу (3.22), обчисленого при мірі неконформності (3.2). Розмітка матчів відбувалась на рівнях аномальності $p_A = 0{,}2$ та $p_A = 0{,}3$. Комірки у стовпцях метрик точності P, R і F_1 мають таке ж кольорове оформлення, що й в табл. 4.3.

При використанні інтегрального мартингалу при рівні аномальності $p_A = 0{,}2$ досягаються високі показники всіх розглянутих метрик. При

збільшенні рівня аномальності відбувається збільшення показника метрики точності на 9% і зниження показників метрик повноти на 17% та F_1 на 0,05.

Таблиця 4.7 – Оцінки ефективності методу виявлення матчів, підозрілих на фіксований результат, на основі інтегрального мартингалу (3.22), обчислених для рівнів аномальності $p_A = 0{,}2$ та $p_A = 0{,}3$

$M_k > M_{k-1}$													
Клас	Обсяг вибірки N	$p_A = 0{,}2$						$p_A = 0{,}3$					
		TP	FN	FP	P	R	F_1	TP	FN	FP	P	R	F_1
(1, 1)	12	1	1	0	100%	50%	0,67	1	1	0	100%	50%	0,67
(1, 2)	16	0	0	0	Немає аномальних матчів			0	0	0	Немає аномальних матчів		
(1, 3)	35	7	1	2	78%	88%	0,82	7	1	2	78%	88%	0,82
(1, 4)	16	4	0	0	100%	100%	1	4	0	0	100%	100%	1
(2, 1)	16	2	0	1	67%	100%	0,8	3	2	0	100%	60%	0,75
(2, 2)	12	0	0	2	Немає аномальних матчів			0	3	2	0%	0%	0
(2, 3)	32	5	0	2	71%	100%	0,83	7	1	0	100%	88%	0,93
(2, 4)	16	3	0	0	100%	100%	1	3	3	0	100%	50%	0,67
(3, 1)	32	5	1	2	71%	83%	0,77	7	3	0	100%	70%	0,82
(3, 2)	32	5	1	4	56%	83%	0,67	9	1	0	100%	90%	0,95
(3, 3)	56	5	0	4	56%	100%	0,71	9	0	0	100%	100%	1
(3, 4)	32	7	0	2	78%	100%	0,88	7	0	2	78%	100%	0,88
(4, 1)	16	3	1	1	75%	75%	0,75	4	1	0	100%	80%	0,89
(4, 2)	16	3	0	0	100%	100%	1	3	3	0	100%	50%	0,67
(4, 3)	32	6	0	4	60%	100%	0,75	7	0	3	70%	100%	0,82
(4, 4)	12	2	1	0	100%	67%	0,8	2	2	0	100%	50%	0,67
Середні характеристики					79%	89%	0,82				88%	72%	0,77

У табл. 4.8 наведено середні показники метрик точності, повноти і F_1 для розглянутих методів виявлення підозрілих щодо фіксованого результату матчів при використанні міри неконформності (3.2) та рівнях аномальності $p_A = 0{,}2$ та $p_A = 0{,}3$.

При використанні простої міри неконформності на даних модельного сезону запропоновані методи виявлення на основі конформного аномального детектора, степеневого мартингалу і інтегрального мартингалу забезпечують виграші щодо виявлення потенційно підозрілих матчів з фіксованим результатом у порівнянні з відомим гістограмним методом на 8%-17% за метрикою точності, 11%-36% за метрикою повноти і 0,11-0,20 за метрикою F_1.

При використанні простої міри неконформностні на даних модельного сезону запропоновані методи виявлення на основі конформного аномального

детектора й інтегрального мартингалу мають близькі показники (відрізняються на 2%-4% за метриками точності і повноти та 0,02-0,03 за метрикою F_1) для розглянутих рівнів аномальностей 0,2, 0,3 і забезпечують точність виявлення на 2%-9% вище у порівнянні із методом на основі ***степеневого мартингалу*** при значенні параметра чутливості 0,8. Разом з тим останній метод забезпечує виграш на 11-15% за метрикою повноти і на 0,02-0,07 за метрикою F_1.

Таблиця 4.8 – Середні показники метрик точності, повноти і F_1 розглянутих методів виявлення матчів, підозрілих на фіксований результат, при використанні міри неконформності (3.2)

Метод	$p_A = 0{,}2$			$p_A = 0{,}3$		
	P	R	F_1	P	R	F_1
Гістограмний метод пошуку аномалій	66%	64%	0,64	78%	61%	0,66
Конформний аномальний детектор при $p_k < p_A$	83%	85%	0,79	91%	74%	0,79
Степеневий мартингал при $\eta = 0{,}8$	74%	100%	0,84	86%	85%	0,84
Інтегральний мартингал	79%	89%	0,82	88%	72%	0,77

У табл. 4.9 наведено середні показники метрик точності, повноти і F_1 для розглянутих методів виявлення підозрілих щодо фіксованого результату матчів при використанні міри неконформності (3.3) та рівні аномальності $p_A = 0{,}3$.

При використанні міри неконформності з округленням середньої різниці м’ячів запропоновані методи виявлення на основі конформного аномального детектора, степеневого мартингалу і інтегрального мартингалу забезпечують гірші показники, ніж при використанні міри неконформності (3.2). Більш чутливими до округлення середньої різниці м’ячів виявилися методи на основі конформного аномального детектора і інтегрального мартингалу.

Таблиця 4.9 – Середні показники метрик точності, повноти і F_1 розглянутих методів виявлення матчів підозрілих на фіксований результат при використанні міри неконформності (3.3)

Метод	$p_A = 0{,}3$		
	P	R	F_1
Конформний аномальний детектор при $p_k < p_A$	91%	62%	0,67
Степеневий мартингал при $\eta = 0{,}8$	91%	81%	0,83
Інтегральний мартингал	87%	62%	0,69

Степеневий мартингал при параметрі чутливості $\eta = 0{,}8$ забезпечив найкращі значення показників метрик точності, повноти та F_1 при виявленні матчів, потенційно підозрілих на фіксованість результату. Таким чином, для вирішення практичних задач доцільно використовувати міру неконформності (3.2).

4.3 Аналіз методів виявлення підозрілих щодо фіксованості результату матчів за даними реального сезону

Тестування розглянутих методів виявлення підозрілих щодо фіксованості результату матчів було також проведено на реальних даних. Для цього було обрано сезон 2014-2015 років Серії *B* Італії, оскільки правоохоронні органи встановили, що керівництво команди «Катанія» (Catania) організувало для своєї команди ряд договірних матчів саме у цьому сезоні[12]. Матчі, які юридично визнані договірними, наведені у табл. 4.10.

Розглянуті матчі були організовані наприкінці сезону, що дозволяє говорити про те, що команда «Катанія» мала на меті турнірні цілі: як мінімум,

[12] Press, A. (2015, June 30). Catania's owner admits to match fixing in five Serie B games. The Guardian. Retrieved 3 September 2023, from https://www.theguardian.com/football/2015/jun/30/catania-match-fixing-serie-b

команда намагалась зберегти себе від можливої участі у плей-оф матчах і переведення у нижчу лігу. Також, деякі результати матчів є достатньо великими (з різницею у 3 м’ячі), що може говорити про можливий заробіток на цих результатах.

Таблиця 4.10 – Договірні матчі команди «Катанія» у сезоні 2014-2015 років Серії *B* Італії

Домашня команда	Виїзна команда	Результат	Дата проведення
Varese	Catania	0:3	02-Apr-2015
Catania	Trapani	4:1	11-Apr-2015
Latina	Catania	1:2	19-Apr-2015
Catania	Ternana	2:0	24-Apr-2015
Catania	Livorno	1:1	02-May-2015

Остаточна турнірна таблиця сезону 2014-2015 років Серії *B* Італії наведена у таблиці 4.11. Однією з особливостей результатів розглянутого сезону є те, що багато команд завершили основний залік турніру з однаковими загальними очками.

Як наслідок, після офіційного завершення сезону відбулось 10 додаткових футбольних матчів у серіях плей-оф та плей-аут. В основному заліку сезону відбулося 462 матчі. Для подальшого групування команд було використано лише дані з таблиці 4.11, не враховуючи результати додаткових матчів. Також через вказану особливість сезону для групування його команд, крім загальних очок $p(t)$ команд, було використано також ознаку загальної різниці м’ячів $gd(t)$.

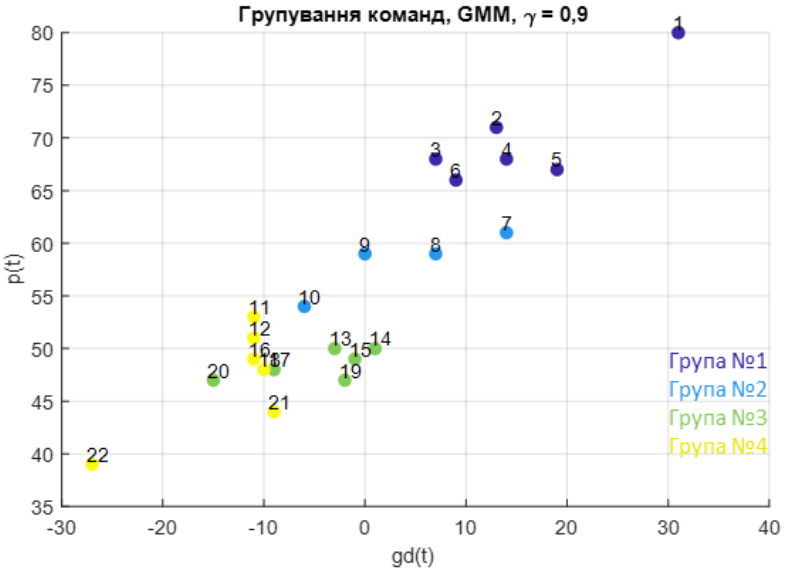
Для групування команд було здійснено їх розбиття на 4 кластери методом на основі Гаусівських сумішей з коефіцієнтом регуляризації $\gamma = 0{,}9$ (рис. 4.15). При цьому під час кластеризації було використано дані всіх команд, крім команд №1 та №22, які за свої показниками явно виділяються з поміж усіх команд: точки цих команд знаходяться на великій відстані від основної маси точок.

Таблиця 4.11 – Показники успішності команд у сезоні 2014-2015 років Серії *B* Італії

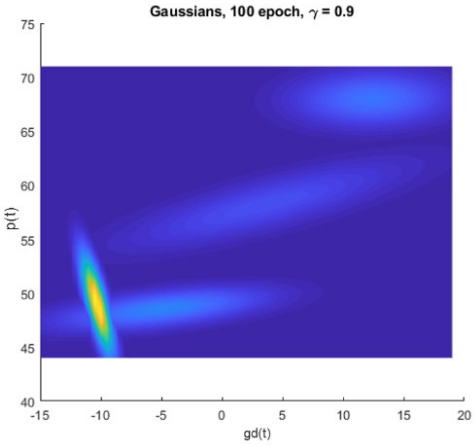
i	t	$s(t)$	$w(t)$	$d(t)$	$gf(t)$	$ga(t)$	$gd(t)$
1	'Carpi'	80	22	14	59	28	31
2	'Frosinone'	71	20	11	62	49	13
3	'Vicenza'	68	18	14	44	37	7
4	'Bologna'	68	17	17	49	35	14
5	'Spezia'	67	18	13	59	40	19
6	'Perugia'	66	16	18	49	40	9
7	'Pescara'	61	16	13	69	55	14
8	'Avellino'	59	15	14	57	50	7
9	'Livorno'	59	15	14	42	42	0
10	'Bari'	54	14	12	43	49	-6
11	'Trapani'	53	13	14	56	67	-11
12	'Ternana'	51	13	12	36	47	-11
13	'Latina'	50	11	17	38	41	-3
14	'Lanciano'	50	10	20	49	48	1
15	'Catania'	49	12	13	59	60	-1
16	'Pro Vercelli'	49	12	13	46	57	-11
17	'Brescia'	48	12	12	54	63	-9
18	'Crotone'	48	12	12	42	52	-10
19	'Entella'	47	10	17	37	39	-2
20	'Modena'	47	10	17	37	52	-15
21	'Cittadella'	44	9	17	47	56	-9
22	'Varese'	39	9	12	40	67	-27

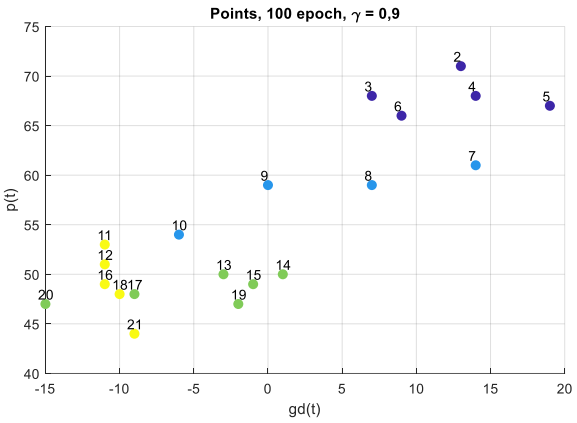
Загалом, утворені групи виокремились за близькістю по загальних очках $p(t)$ команд сезону: у групу №1 потрапили команди з очками з діапазону [66; 71], у групу №2 — відповідно, з діапазону [54; 61], а у групу №3 — відповідно, з діапазону [47; 50]. Водночас, при формуванні групи №4 основним фактором виявилася близькість за різницею м'ячів у сезоні. При цьому, різниця м'ячів кожної команди з групи № 4 є меншою, ніж у більшості команд інших груп. Цей фактор також вплинув на вибір кластеру, до якого потрібно було приєднати команду №22: ця команда має найменшу різницю м'ячів. Також команда №22 за дистанцією до кластерного центроїду є

ближчою до групи №4, ніж до групи №3. Найбільш близькою групою до команди №1 за дистанцією до кластерного центроїду є група №1.



(а)





(б) (в)

Рисунок 4.15 – Результати групування команд сезону 2014-2015 років Серії *B* Італії

(а) остаточне групування команд сезону, (б) проєкція результуючих Гаусіан на координати кластеризації, (в) групування команд №№ 2-21 методом на основі Гаусівських сумішей

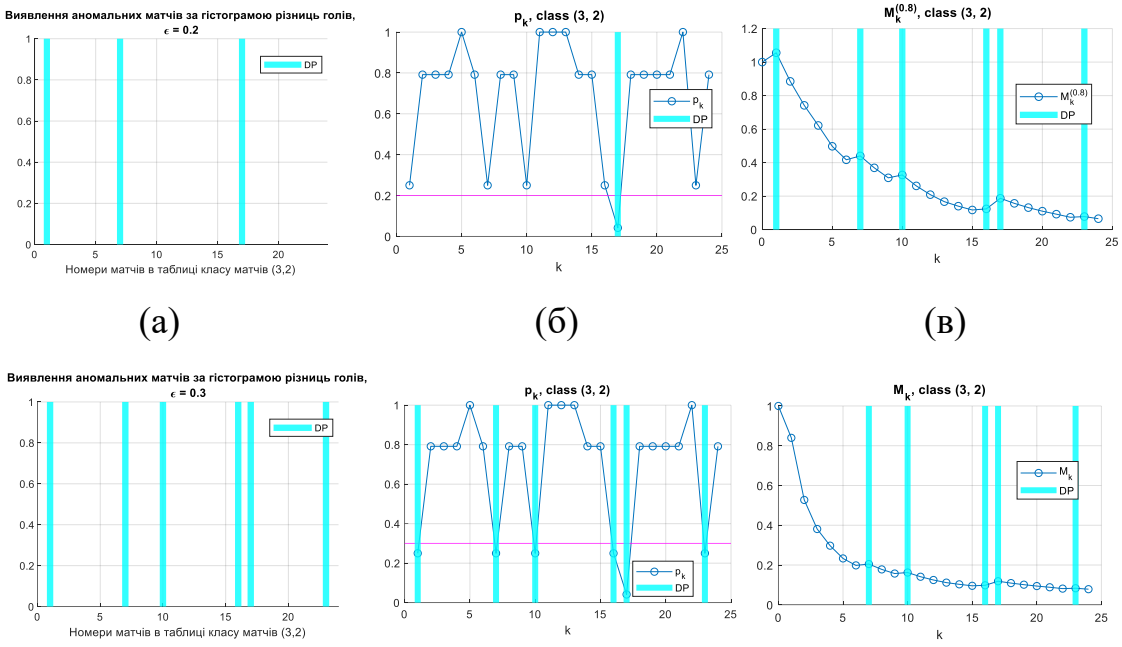
У відповідності до отриманого групування команд утворено класи матчів і на отриманих класах матчів протестовано роботи розглянутих і

запропонованих методів виявлення підозрілих на фіксований результат футбольних матчів. Результати запропонованих методів виявлення отримано при використанні простої міри неконформності (3.2). Інформацію про розподіл матчів з таблиці 4.10 за класами матчів наведено у таблиці 4.12. У класі матчів (3, 4) опинилось два договірних матчі, у класах матчів (4, 3), (3, 3) та (3, 2) опинилось по одному договірному матчу.

Таблиця 4.12 – Розподіл договірних матчів команди «Катанія» за класами матчів

Домашня команда	Виїзна команда	Результат	Дата проведення	Клас матчів	Номер у класі матчів
Varese	Catania	0:3	02-Apr-2015	(4, 3)	30
Catania	Trapani	4:1	11-Apr-2015	(3, 4)	28
Latina	Catania	1:2	19-Apr-2015	(3, 3)	27
Catania	Ternana	2:0	24-Apr-2015	(3, 4)	30
Catania	Livorno	1:1	02-May-2015	(3, 2)	22

Розглянемо результати виявлення аномальних матчів у класі матчів (3, 2) (рис. 4.16). За гістограмним методом пошуку аномалій при порозі аномальності 0,2 було виявлено 3 аномальних матчі (рис. 4.16, а), а при порозі аномальності 0,3 – 5 аномальних матчів (рис. 4.16, г). За конформним аномальним детектором при порозі аномальності 0,2 (рис. 4.16, б) у цьому класі матчів було виявлено 1 аномальний матч, а при порозі аномальності 0,3 — 6 аномальних матчів (рис. 4.16, д). Такі ж результати отримано за допомогою степеневого мартингала з параметром чутливості 0,8 (рис. 4.16, в). За інтегральним мартингалом було виявлено 5 аномальних матчів (рис. 4.16, е). Договірний матч №22 не було виявлено за жодним з цих методів. Ця ситуація пов’язана з тим, що матч завершився нічиєю, у той час як у цьому класі матчів очікуваною різницею м’ячів є 0, тобто теж нічия. Загалом, нічия є одним з частих результатів матчу, тому такий матч важно визначити як аномальний на основі лише різниці м’ячів.

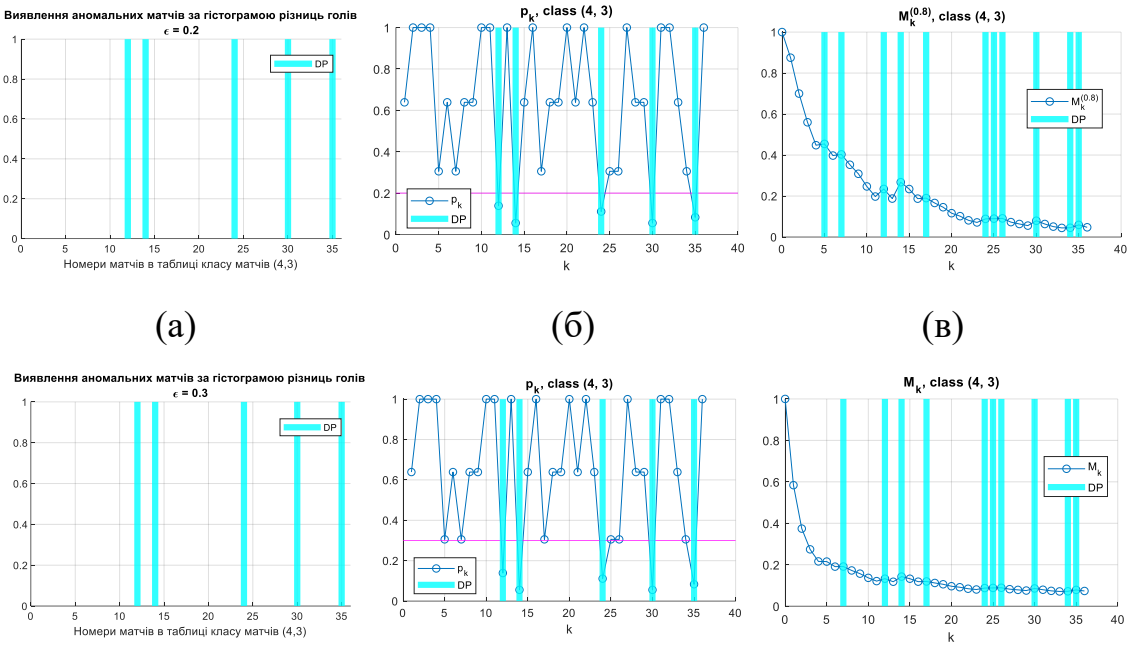


(а) (б) (в)

(г) (д) (е)

Рисунок 4.16 – Результати виявлення договірних матчів у класі матчів (3, 2) (а, г) гістограмним методом при $p_A = 0{,}2$ та $p_A = 0{,}3$, (б, д) конформним аномальним детектором при $\varepsilon = 0{,}2$ та $\varepsilon = 0{,}3$, (в) методом степеневого мартингалу з $\eta = 0{,}8$, (е) методом інтегрального мартингалу

Розглянемо результати виявлення аномальних матчів у класі матчів (4, 3) (рис. 4.17). За гістограмним методом пошуку аномалій з порогами аномальності 0,2 та 0,3 було виявлено 5 аномальний матчів (рис. 4.17, а, г). Такий самий результат отримано за конформним аномальним детектором (рис. 4.17, б, д). За степеневим мартингалом з параметром чутливості 0,8 було виявлено 11 аномальних матчів (рис. 4.17, в). За інтегральним мартингалом було виявлено 10 аномальних матчів. Договірний матч №30 було виявлено за усіма розглянутими методами. Ця ситуація пов’язана з тим, що матч завершився перемогою «Катанії» з перевагою у 3 м’ячі при тому, що середня різниця м’ячів в класі матчів дорівнює 0,08, тобто матч за своїм результатом сильно відхиляється від очікуваного результату матчу.



(а) (б) (в)

(г) (д) (е)

Рисунок 4.17 – Результати виявлення договірних матчів у класі матчів (4, 3) (а, г) гістограмним методом при $p_A = 0{,}2$ та $p_A = 0{,}3$, (б, д) конформним аномальним детектором при $\varepsilon = 0{,}2$ та $\varepsilon = 0{,}3$, (в) методом степеневого мартингалу з $\eta = 0{,}8$, (е) методом інтегрального мартингалу

Розглянемо результати виявлення аномальних матчів у класі матчів (3, 3) (рис. 4.18). За гістограмним методом пошуку аномалій при порогах аномальності 0,2 та 0,3 було виявлено 6 аномальних матчів (рис. 4.18, а, г). За конформним аномальним детектором при порозі аномальності 0,2 (рис. 4.18, б) у цьому класі матчів було виявлено 3 аномальних матча, а при порозі аномальності 0,3 — 7 аномальних матчів (рис.4.18, д). За інтегральним мартингалом було виявлено 7 аномальних матчів (рис.4.18, е). Такі ж результати отримано за допомогою степеневого мартингала з параметром чутливості 0,8 (рис.4.18, в). Договірний матч №27 було виявлено за допомогою конформного аномального детектора при порозі аномальності 0,3, степеневого мартингалі з параметром чутливості 0,8 та інтегрального мартингала. Середня різниця м’ячів у цьому класі матчів дорівнює 0,67, а різниця договірного матчу №27 дорівнює –1. Даний договірний матч відрізняється від очікуваного

результату, але у цьому класі матчів існують матчі, у яких p-value ще менше, ніж у поточного — через це матч не розпізнається за гістограмним методом та конформним аномальним детектором при порозі аномальності 0,2.

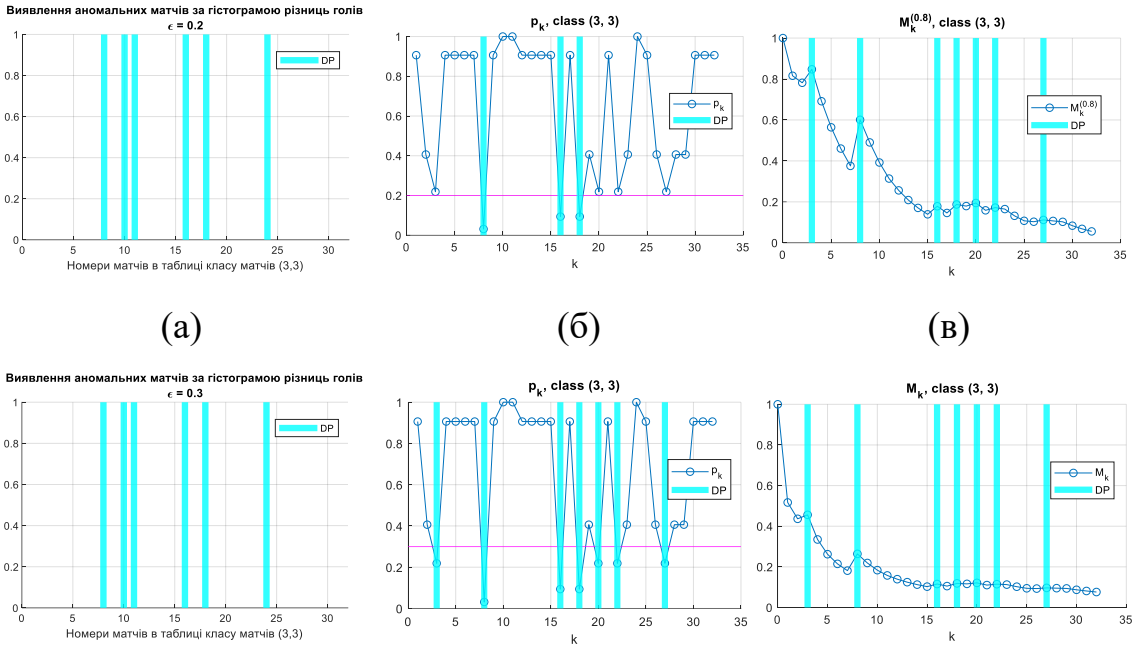


Рисунок 4.18 – Результати виявлення договірних матчів у класі матчів (3, 3): (а, г) гістограмним методом при $p_A = 0{,}2$ та $p_A = 0{,}3$, (б, д) конформним аномальним детектором при $\varepsilon = 0{,}2$ та $\varepsilon = 0{,}3$, (в) методом степеневого мартингалу з $\eta = 0{,}8$, (е) методом інтегрального мартингалу

Розглянемо результати виявлення аномальних матчів у класі матчів (3, 4) (рис. 4.19). За гістограмним методом пошуку аномалій при порозі аномальності 0,2 було виявлено 5 аномальних матчів (рис. 4.19, а), а при порозі аномальності 0,3 – 9 аномальних матчів (рис. 4.19, г). За конформним аномальним детектором при порозі аномальності 0,2 (рис. 4.19, б) у цьому класі матчів було виявлено 7 аномальних матчів, а при порозі аномальності 0,3 — 9 аномальних матчів (рис. 4.19, д). За інтегральним мартингалом було виявлено 9 аномальних матчів (рис. 4.19, е). Такі ж результати отримано за допомогою степеневого мартингала з параметром чутливості 0,8 (рис. 4.19, в).

Договірний матч №28 виявлено за допомогою конформного аномального детектора при порозі аномальності 0,3, степеневого мартингалі з параметром чутливості 0,8 та інтегрального мартингала. Договірний матч №30 виявлено за усіма розглянутими методами.

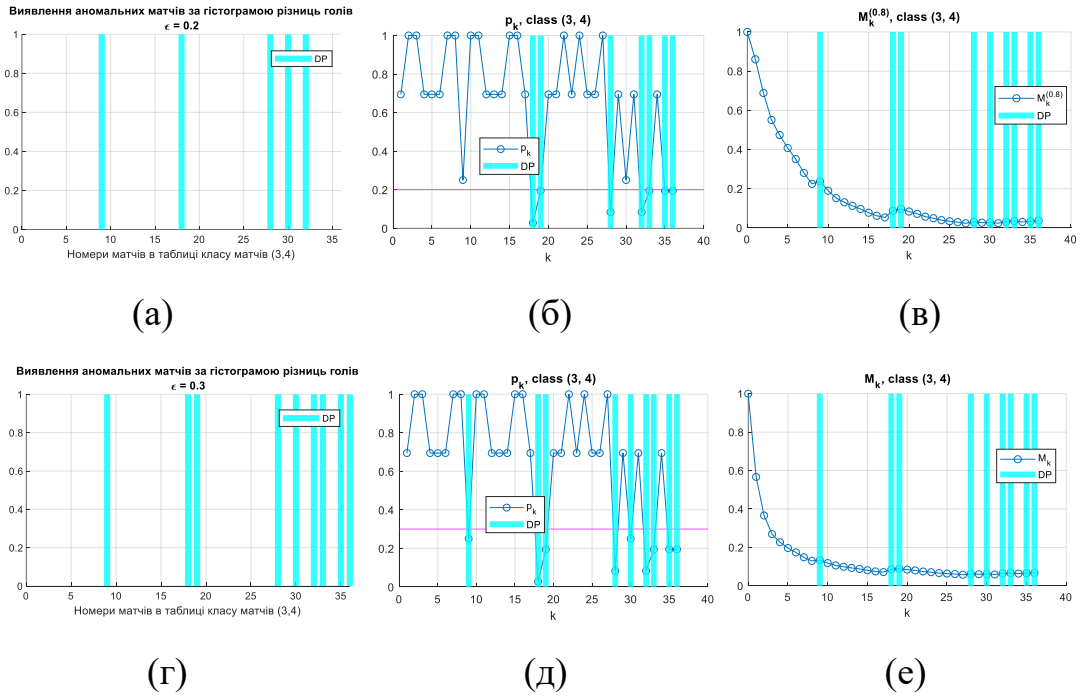


Рисунок 4.19 – Результати виявлення договірних матчів у класі матчів (3, 4): (а, г) гістограмним методом при $p_A = 0{,}2$ та $p_A = 0{,}3$, (б, д) конформним аномальним детектором при $\varepsilon = 0{,}2$ та $\varepsilon = 0{,}3$, (в) методом степеневого мартингалу з $\eta = 0{,}8$, (е) методом інтегрального мартингалу

Отже, за допомогою гістограмного методу виявлено 3 з 5 договірних матчів. За допомогою конформного аномального детектора з порогом аномальності 0,2 виявлено лише 2 з 5 договірних матча, а з порогом аномальності 0,3 вдалося виявити вже 4 з 5 договірних матчів. За допомогою степеневого мартингала з параметром чутливості 0,8 та інтегрального мартингала виявлено також 4 з 5 договірних матчів. Загалом у розглянутих класах матчів виявлено до 33 аномальних матчів, що становить до 26% усієї кількості матчів цих класів. Це значення є близьким до порогу конформного

аномального детектора, який співпадає із заданою ймовірністю появи аномальних даних $p_A = 0{,}3$.

Висновки до розділу 4

1. Якість роботи ***гістограмного методу*** виявлення аномалій на даних модельного сезону в середньому дорівнює 66 % за метрикою точності, 64 % за метрикою повноти і 0,64 за метрикою F_1 при використанні рівня аномальності $p_A = 0{,}2$, та 78 % за метрикою точності, 61 % за метрикою повноти і 0,66 за метрикою F_1 при використанні порогу аномальності $p_A = 0{,}3$. В обох випадках, гістограмний метод виявлення аномалій продемонстрував низьку якість роботи, причому за метрикою F_1 якість роботи в обох випадках була в середньому майже однаковою.

2. Експериментально було підтверджене доведене теоретично в дисертації твердження про те, що умова еквівалентності методів виявлення на основі конформного аномального детектора і степеневого мартингалу зв'язує їх параметри за формулою $\varepsilon = \eta^{\frac{1}{1-\eta}}$ (розглядався випадок з $\eta = 0{,}8$ і $\varepsilon = 0{,}32768$).

3. При використанні простої міри неконформності на даних модельного сезону запропоновані методи виявлення на основі ***конформного аномального детектора***, ***степеневого мартингалу*** і ***інтегрального мартингалу*** забезпечують перевагу щодо виявлення потенційно підозрілих матчів з фіксованим результатом у порівнянні з відомим гістограмним методом на 8%-17% за метрикою точності, 11%-36% — за метрикою повноти і 0,11-0,20 — за метрикою F_1.

4. Підвищення рівня аномальності призводить в середньому до підвищення якості виявлення розглянутими алгоритмами за метрикою точності, а також зменшенню якості за метрикою повноти. Для забезпечення необхідного рівня метрики повноти доцільно використовувати алгоритм на

основі степеневого мартингалу із відповідним значенням параметра чутливості.

5. При використанні міри неконформності з округленням середньої різниці м’ячів запропоновані методи виявлення на основі ***конформного аномального детектора***, ***степеневого мартингалу*** і ***інтегрального мартингалу*** забезпечують гірші показники, ніж при використанні міри неконформності (3.2). Більш чутливими до округлення середньої різниці м’ячів виявилися методи на основі ***конформного аномального детектора*** і ***інтегрального мартингалу***. Степеневий мартингал при параметрі чутливості $\eta = 0{,}8$ забезпечив найкращі значення показників метрик точності, повноти та F_1 щодо виявлення потенційно підозрілих на фіксований результат матчів.

6. Запропоновані методи на основі ***конформного аномального детектора***, ***степеневого мартингалу*** і ***інтегрального мартингалу*** виявили 4 з 5 матчів сезону 2014–2015 років Серії *B* Італії, які вважаються договірними за інформацією від офіційних правоохоронних органів Італії.

ОСНОВНІ РЕЗУЛЬТАТИ І ВИСНОВКИ

1. При групуванні команд сезону запропонована процедура регуляризації недіагональних елементів коваріаційних матриць в методі Гаусівських сумішей, яка забезпечує зменшення чутливості до початкових умов, а також формування кластерів еліпсоподібної форми, що дозволяє врахувати неочевидні кореляційні зв'язки між точками набору даних.

2. Розроблена імітаційна модель футбольного сезону з матчами з фіксованим результатом, особливістю якої є те, що матчі поділяються на класи за контекстуальними атрибутами «сила команди» і «тип гри» – домашня або виїзна, і відповідно ймовірності забиття голів командою під час матчу розраховуються для класів, а не по усьому сезону. Моделювання матчів з фіксованим результатом відбувається шляхом випадкового вибору і заміни результатів матчів з нормальним результатом на аномальний з врахуванням розподілів різниць м'ячів класу матчів. Отримані з використанням імітаційної моделі гістограми різниць м'ячів по кожному класу матчів мають очікувані закономірності в результатах матчів: сильніші команди мають кращі результати в грі зі слабшими командами; вдома команди грають краще, ніж на виїзді. Розподіли типів результатів за всіма класами матчів реального і модельних сезонів є подібними за критерієм Колмогорова-Смірнова на рівні значущості 0,001.

3. На основі аномального конформного детектору з використанням запропонованих мір неконформності розроблено метод виявлення підозрілих щодо фіксованого результату футбольних матчів, причому спрацьовування детектору відбувається, коли ступень конформності (p-value) матчу досягає значення певного порогу. Цей метод відноситься до класу методів виявлення без вчителя і дозволяє вводити оцінки гарантованої точності для отриманих рішень. Для забезпечення балансу між чутливістю та точністю виявлення, значення порогу слід встановлювати близьким до апріорної ймовірності появи аномальних об'єктів.

4. На основі степеневого мартингалу розроблено метод виявлення підозрілих щодо фіксованості результату футбольних матчів, причому прийняття рішення відбувається при зростанні значення степеневого мартингалу для поточного спостереження по відношенню до значення цього ж мартингала для попереднього спостереження. Зміна параметра чутливості дозволяє налаштовувати степеневий мартингал на виявлення аномалій відповідного рівня.

5. На основі інтегрального мартингалу розроблено метод виявлення підозрілих щодо фіксованості результату футбольних матчів, причому прийняття рішення відбувається при зростанні значення інтегрального мартингалу для поточного спостереження по відношенню до значення цього ж мартингала для попереднього спостереження. Використання інтегрального мартингала не вимагає налаштування параметрів.

6. При використанні простої міри неконформності на даних модельного сезону запропоновані методи виявлення на основі конформного аномального детектора, степеневого мартингалу й інтегрального мартингалу забезпечують виграші щодо виявлення потенційно підозрілих матчів з фіксованим результатом у порівнянні з відомим гістограмним методом на 3%-13% за метрикою точності, 11%-30% за метрикою повноти і 10%-18% за метрикою F_1. Підвищення рівня аномальності призводить запропонованими методами в середньому до підвищення якості виявлення аномалій за метрикою точності і зменшенню якості за метрикою повноти. Для забезпечення необхідного рівня метрики повноти доцільно використовувати метод на основі степеневого мартингалу із відповідним значенням параметра чутливості.

7. Запропоновані методи на основі конформного аномального детектора, степеневого мартингалу й інтегрального мартингалу дозволили виявити 4 з 5 матчів сезону 2014–2015 років Серії *B* Італії, які вважаються договірними за інформацією від офіційних правоохоронних органів Італії.

8. Розроблені методи можуть бути використані для виявлення підозрілих на фіксований результат матчів у інших спортивних турнірах,

таких як: хокей, волейбол, бейсбол, баскетбол, кіберспорт тощо. За відповідного переформулювання і підбору адекватної міри неконформності запропоновані в дисертаційному дослідженні методи можуть бути використані для пошуку широкого кола контекстних аномалій (нетипові транзакції по банківському рахунку, проникнення до закритої мережі, аномальна кількість повідомлень в соціальних мережах на певну тематику тощо).

СПИСОК ЛІТЕРАТУРИ

1. Zhuk, I., & Chertov, O. (2023). Framework based on conformal predictors and power martingales for detection of fixed football matches. *Eastern-European Journal of Enterprise Technologies*, *2*(4 (122), 6–15. https://doi.org/10.15587/1729-4061.2023.276977

2. Chertov, O., & Zhuk, I. (2023). Detection of fixed football matches based on the theory of conformal predictors using the modified Stepanets indicator function. *Eastern-European Journal of Enterprise Technologies*, *3*(4 (123), 22–32. https://doi.org/10.15587/1729-4061.2023.282645

3. Чертов, О. Р., & Жук, І. С. (2022). Імітаційна модель футбольного сезону з матчами з фіксованим результатом, *Наукові вісті КПІ, 1–2*, с. 82–94, 2022. https://doi.org/10.20535/kpisn.2022.1-2.287916

4. Chertov, O. R., & Zhuk, I. S. (2023). Clusterization of soccer season teams using K-means and GMM methods. *Intelligent Solutions-S: Proceedings of the International Symposium, September 28, 2023, Kyiv-Uzhorod, Ukraine / Ministry of Education and Science of Ukraine, Taras Shevchenko National University of Kyiv and* [etc]; Vitaliy Ye. Snytyuk (Editor) (pp. 14-15).

5. Жук, І. (2020). Застосування конформних предикторів і степеревих мартингалів для виявлення підозрілих матчів футбольних турнірів. *Прикладна математика та комп'ютинг. ПМК, 2022* : п’ятнадцята наук. конф. магістрантів та аспірантів, Київ, 16-18 лист. 2022 р. : зб. тез доп. / [редкол.: Дичка І. А. та ін.]. — К. : Просвіта, 2022. — 368 с. ISBN 978-617-7010-23-3 С. 24 – 29.

6. Chertov, O., Zhuk, I., & Serdyuk, A. (2021). Search of the Deviation from the Natural Process Using Stepanets Approach for Classification of Functions. *2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)* (pp. 720-724), doi: 10.1109/IDAACS53288.2021.9660997.

7. Жук, І. (2020). Виявлення зовнішнього впливу в інформаційних потоках мережі Internet як проблема ідентифікації образу. *Філософія і науково-технічна творчість у хронотопі технічного університету: Матеріали III Міжнародної науково-практичної конференції.* – 410 с. С. 144 – 147.

8. Жук, І. С., & Чертов, О. Р. (2020). Використання математичного апарату наближень Степанця для виявлення штучних втручань у сигналах різної природи. *Інтегровані інтелектуальні робототехнічні комплекси (ІІРТК-2020). Тринадцята міжнародна науково-практична конференція 19-20 травня 2020 р.*, Київ, Україна. – К.: НАУ, 2020. – 305 с. С. 276 – 278.

9. Lilley, E. (2015). A Review of the recommendations of the 'Report of the Sports Betting Integrity Panel' in assessing the progress towards tackling Match-fixing in Sport. *Laws of the Game, 1*(1), Article 1.

10. Huggins, M. (2018). Match-Fixing: A Historical Perspective. *The International Journal of the History of Sport, 35*(2–3), 123–140. https://doi.org/10.1080/09523367.2018.1476341

11. Match-fixing in sport a mapping of criminal law provisions in EU 27. (2015). *Trends in Organized Crime*, *18*(3), 251–260. https://doi.org/10.1007/s12117-015-9241-4

12. Haberfeld, M. R., & Sheehan, D. (Eds.). (2013). *Match-Fixing in International Sports: Existing Processes, Law Enforcement, and Prevention Strategies*. Springer International Publishing. https://doi.org/10.1007/978-3-319-02582-7

13. Sportradar. (2023). (rep.). *Betting corruption and match-fixing in 2022*. Retrieved September 3, 2023, from https://sportradar.com/wp-content/uploads/2023/03/Betting-Corruption-And-Match-Fixing-In-2022.pdf

14. 2022 Integrity Report. (2023). *IBIA*. https://ibia.bet/2022-annual-report/

15. Закон України. Про запобігання впливу корупційних правопорушень на результати офіційних спортивних змагань від 03.11.2015 № 743-VIII. Відомості Верховної Ради, 2015, № 51, ст. 472.

16. Park, J.-H., Choi, C.-H., Yoon, J. & Girginov, V. (2019). How should sports match fixing be classified? *Cogent Social Sciences*, *5*(1), https://doi.org/10.1080/23311886.2019.1573595

17. Constandt, B., & Manoli, E. (2022). *Understanding match-fixing in sport: Theory and practice.*

18. Hill, D. (2009). How Gambling Corruptors Fix Football Matches. *European Sport Management Quarterly*, *9*(4), 411–432. https://doi.org/10.1080/16184740903332018

19. Chertov, O., & Tavrov, D. (2015). Microfiles as a Potential Source of Confidential Information Leakage. *Intelligent Methods for Cyber Warfare*, 87–114. Springer International Publishing. https://doi.org/10.1007/978-3-319-08624-8_4

20. Tzeng, C.-C., & Lee, P.-C. (2021). Understanding match-fixing from the perspective of social capital: A case study of Taiwan's professional baseball system. *International Review for the Sociology of Sport*, *56*(4), 558–577. https://doi.org/10.1177/1012690220917060

21. Tzeng, C.-C., & Ohl, F. (2023). Examining the fabrics of match-fixing: The underground sport betting system. *International Review for the Sociology of Sport*, *58*(1), 188–207. https://doi.org/10.1177/10126902221095688

22. Forrest, D., & McHale, I. G. (2019). Using statistics to detect match fixing in sport. *IMA Journal of Management Mathematics*, 30(4), 431–449. https://doi.org/10.1093/imaman/dpz008

23. Manoli, A. E., & Antonopoulos, G. A. (2015). 'The only game in town?': Football match-fixing in Greece. *Trends in Organized Crime*, *18*(3), 196–211. https://doi.org/10.1007/s12117-014-9239-3

24. Yilmaz, S., Manoli, A. E., & Antonopoulos, G. A. (2019). An anatomy of Turkish football match-fixing. *Trends in Organized Crime*, *22*(4), 375–393. https://doi.org/10.1007/s12117-018-9345-8

25. Razali, N., Mustapha, A., Yatim, F. A., & Aziz, R. A. (2017). Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL). *IOP Conference Series: Materials Science and Engineering*, *226*(1), 012099.

https://doi.org/10.1088/1757-899X/226/1/012099

26. Anfilets, S., Bezobrazov, S., Golovko, V., Sachenko, A., Komar, M., Dolny, R., Kasyanik, V., Bykovyy, P., Mikhno, E., & Osolinskyi, O. (2020). Deep multilayer neural network for predicting the winner of football matches. *International Journal of Computing*, 70–77. https://doi.org/10.47839/ijc.19.1.1695

27. Narizuka, T., Yamazaki, Y., & Takizawa, K. (2021). Space evaluation in football games via field weighting based on tracking data. *Scientific Reports*, *11*(1), 5509. https://doi.org/10.1038/s41598-021-84939-7

28. Badur, B., & Topaçan, U. (2011). Forecasting Football Game Outcomes: A Data Mining Approach. *16th International Business Information Management Association Conference* (pp. 1905–1914), Kuala Lumpur, Malaysia, 29-30 June 2011, ISBN: 978-0-9821489-5-2.

29. Eryarsoy, E., & Delen, D. (2019). Predicting the Outcome of a Football Game: A Comparative Analysis of Single and Ensemble Analytics Methods. *Proceedings of the 52nd Annual Hawaii International Conference on System Sciences* (pp. 1107-1115), January 8-11, 2019, Maui, Hawaii, ISBN: 978-0-9981331-2-6, http://dx.doi.org/10.24251/HICSS.2019.136

30. Azeman, A. (2020). Football Match Outcome Prediction by Applying Three Machine Learning Algorithms. *International Journal of Emerging Trends in Engineering Research*, *8*(1.1), 73–77.
https://doi.org/10.30534/ijeter/2020/1181.12020

31. Бутина, Д. В. (2021). Нейросетевая система прогнозирования результатов сезона немецкой футбольной лиги «БУНДЕСЛИГИ». *Искусственный интеллект в решении актуальных социальных и экономических проблем XXI века*, 373–380.

32. Бутина, Д. В., & Ясницкий, Л. Н. (2023). Нейросетевая система прогнозирования результатов сезона итальянской футбольной лиги ‘Серия А’. *Вестник Пермского университета. Серия: Математика. Механика. Информатика*, *1*(60). https://doi.org/10.17072/1993-0550-2023-1-84-92

33. Jia, R., Wong, C., & Zeng, D. (2013). *Final Project*: *Predicting the Major League Baseball Season*. CS 229.
https://cs229.stanford.edu/proj2013/JiaWongZeng-PredictingTheMajorLeagueBaseballSeason.pdf

34. Valero, C. S. (2016). Predicting Win-Loss outcomes in MLB regular season games – A comparative study using data mining methods. *International Journal of Computer Science in Sport*, *15*(2), 91–112. https://doi.org/10.1515/ijcss-2016-0007

35. Ясницкий, Л. Н., Киросова, А. В., Ратегова, А. В., & Черепанов, Ф.М. (2014). Методика нейросетевого прогнозирования результатов спортивных состязаний на примере чемпионата мира-2015 по легкой атлетике. *Вестник Пермского университета. Серия: Математика. Механика. Информатика*, *3*(26), 90-97.

36. Jiang, W., Zhao, K., & Jin, X. (2021). Diagnosis Model of Volleyball Skills and Tactics Based on Artificial Neural Network. *Mobile Information Systems*, *2021*, e7908897. https://doi.org/10.1155/2021/7908897

37. Titman, A. C., Costain, D. A., Ridall, P. G., & Gregory, K. (2015). Joint modelling of goals and bookings in association football. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(3), 659–683. https://doi.org/10.1111/rssa.12075

38. Gudmundsson, J., & Wolle, T. (2014). Football analysis using spatio-temporal tools. *Computers, Environment and Urban Systems*, *47*, 16–27. https://doi.org/10.1016/j.compenvurbsys.2013.09.004

39. Janetzko, H., Sacha, D., Stein, M., Schreck, T., Keim, D. A., & Deussen, O. (2014). Feature-driven visual analytics of soccer data. *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 13–22. https://doi.org/10.1109/VAST.2014.7042477

40. Decroos, T. (2020). *Soccer Analytics Meets Artificial Intelligence: Learning Value and Style from Soccer Event Stream Data*.

https://lirias.kuleuven.be/retrieve/587585

41. Spearman, W. (2018). *Beyond Expected Goals*.

42. Rios-Neto, H. M. R., Jr, W. M., & Vaz-de-Melo, P. O. S. (2020). A new look into Off-ball Scoring Opportunity: Taking into account the continuous nature of the game. *Conference: Barcelona Analytics in Sports Tomorrow*.

43. Langan, S. D. (2013). *Predict Football Matches: Using Spreadsheet Models to Become a Winning Sports Bettor* (Kindle Edition), 379 p.

44. Phua, C., Alahakoon, D., & Lee, V. (2004). Minority report in fraud detection: classification of skewed data. *SIGKDD Explorer Newsletter 6,* 1, 50–59. https://doi.org/10.1145/1007730.1007738

45. Chan, P., Fan, W., Prodromidis, A.L., & Stolfo, S.J. (1999). Distributed Data Mining in Credit Card Fraud Detection. *IEEE Intelligent Systems*, 67–74. https://doi.org/10.1109/5254.809570

46. Bolton, R.J., & Hand, D.J. (2001). Unsupervised Profiling Methods for Fraud Detection. *Conference of Credit Scoring and Credit Control VII* (pp. 5–7).

47. Kou, Y., C.-T. Lu, Sirwongwattana, S. & Huang, Y-P. (2004). Survey of Fraud Detection Techniques. *IEEE International Conference on Sensing and Control*, *2*, 749–754. https://doi.org/10.1109/ICNSC.2004.1297040

48. Basu, S., & Meckesheimer, M. (2007). Automatic outlier detection for time series: an application to sensor data. *Knowledge and Information Systems 11,* 2, 137–154. https://dl.acm.org/doi/10.5555/3225643.3225840

49. Zhang, K., Shi, S., Gao, H., & Li, J. (2007). Unsupervised outlier detection in sensor networks using aggregation tree. *Advanced Data Mining and Applications 4632*, 158–169. https://doi.org/10.1007/978-3-540-73871-8_16

50. He, H., Wang, J., Graco, W., and Hawkins, S. (1997). Application of neural networks to detection of medical fraud. *Expert Systems with Applications 13*(4), 329–336. https://doi.org/10.1016/S0957-4174(97)00045-6

51. Laurikkala, J., Juhola, M., & Kentala., E. (2000). Informal identification of outliers in medical data. *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology* (pp. 20–24).

52. Lin, J., Keogh, E., Fu, A., & Herle, H. V. (2005). Approximations to magic: Finding unusual medical time series. *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems. IEEE Computer Society, Washington, DC, USA* (pp. 329–334). https://doi.org/10.1109/CBMS.2005.34

53. Liu, J. P. and Weng, C. S. (1991). Detection of outlying data in bioavailability/bioequivalence studies. *Statistics Medicine 10*(9), 1375–89. https://doi.org/10.1002/sim.4780100906

54. Roberts, S. (2002). Extreme value statistics for novelty detection in biomedical signal processing. *Proceedings of the 1st International Conference on Advances in Medical Signal and Information Processing* (pp. 166–172). https://doi.org/10.1049/cp:20000333

55. Kazachuk, M., Petrovskiy, M., Mashechkin, I., & Gorohov, O. (2018). Novelty Detection Using Elliptical Fuzzy Clustering in a Reproducing Kernel Hilbert Space. In H. Yin, D. Camacho, P. Novais, & A. J. Tallón-Ballesteros (Eds.), *Intelligent Data Engineering and Automated Learning – IDEAL 2018* (pp. 221–232). Springer International Publishing. https://doi.org/10.1007/978-3-030-03496-2_25

56. Suzuki, E., Watanabe, T., Yokoi, H., & Takabayashi, K. (2003). Detecting interesting exceptions from medical test data with visual summarization. *Proceedings of the 3rd IEEE International Conference on Data Mining* (pp. 315–322). https://doi.org/10.1109/ICDM.2003.1250935

57. Agarwal, D. (2005). An empirical bayes approach to detect anomalies in dynamic multidimensional arrays. *Proceedings of the 5th IEEE International Conference on Data Mining. IEEE Computer Society, Washington, DC, USA*, (pp. 26–33). https://doi.org/10.1109/ICDM.2005.22

58. Agarwal, D. (2006). Detecting anomalies in cross-classified streams: a bayesian approach. *Knowledge and Information Systems 11*(1), 29–44. https://doi.org/10.1007/s10115-006-0036-4

59. Aggarwal, C. (2005). On abnormality detection in spuriously populated Data Streams. *Proceedings of 5th SIAM Data Mining* (pp. 80–91).

https://doi.org/10.1137/1.9781611972757.8

60. Savage, D., Zhang, X., Yu, X., Chou, P., & Wang, Q. (2014). Anomaly detection in online social networks. *Social Networks*, *39*, 62–70. https://doi.org/10.1016/j.socnet.2014.05.002

61. Kaur, R., & Singh, S. (2016). A survey of data mining and social network analysis based anomaly detection techniques. *Egyptian Informatics Journal 17*(2), 199–216. https://doi.org/10.1016/j.eij.2015.11.004

62. Pedersen, B. S., & Quinlan, A. R. (2017). Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. *The American Journal of Human Genetics*, *100*(3), 406–413. https://doi.org/10.1145/2808769.2808779

63. Xiao, C., Freeman, D. M., & Hwa, T. (2015). Detecting clusters of fake accounts in online social networks. *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security* (pp. 91–101).
https://doi.org/10.1145/2808769.2808779

64. Patcha, A., & Park, J.-M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, *51*(12), 3448–3470. https://doi.org/10.1016/j.comnet.2007.02.001

65. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), 15:1-15:58. https://doi.org/10.1145/1541880.1541882

66. Markou, M., & Singh, S. (2003). Novelty detection: A review—part 2:: neural network based approaches. *Signal Processing*, *83*(12), 2499–2521. https://doi.org/10.1016/j.sigpro.2003.07.019

67. Amer, M., Goldstein, M., & Abdennadher, S. (2013). Enhancing one-class support vector machines for unsupervised anomaly detection. *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description* (pp. 8–15). https://doi.org/10.1145/2500853.2500857

68. Зубков, Е. В., & Белов, В. М. (2016). Методы интеллектуального анализа данных и обнаружение вторжений. *Вестник СибГУТИ*, (1 (33)), 118.

69. Markou, M., & Singh, S. (2003). Novelty detection: A review—part 1: statistical approaches. *Signal Processing*, *83*(12), 2481–2497. https://doi.org/10.1016/j.sigpro.2003.07.018

70. Akpinar, M., Adak, M. F., & Guvenc, G. (2021). SVM-based anomaly detection in remote working: Intelligent software SmartRadar. *Applied Soft Computing*, *109*(C), 107457. https://doi.org/10.1016/j.asoc.2021.107457

71. Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, *13*(7), 1443–1471. https://doi.org/10.1162/089976601750264965

72. Schölkopf, B., Williamson, R. C., Smola, A., Shawe-Taylor, J., & Platt, J. (1999). Support Vector Method for Novelty Detection. *Advances in Neural NIPS'99: Proceedings of the 12th International Conference on Neural Information Processing Systems* (pp. 582–588). https://dl.acm.org/doi/10.5555/3009657.3009740

73. Hawkins, D. M. (1980). *Identification of Outliers*. Springer Netherlands. https://doi.org/10.1007/978-94-015-3994-4

74. Aggarwal, C. C., & Yu, P. S. (2008). Outlier detection with uncertain data. *Proceedings of the 2008 SIAM International Conference on Data Mining* (pp. 483–493). https://doi.org/10.1137/1.9781611972788.44

75. Sheikholeslami, G., Chatterjee, S., & Zhang, A. (1998). WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Database. *Proceedings of the International Conference on Very Large Databases.* New York, USA, (pp. 428–439). https://dl.acm.org/doi/10.5555/645924.671342

76. Adam, A., Rivlin, E., & Shimshoni, I. (2001). ROR: Rejection of Outliers by Rotation. *IEEE Transaction on Pattern Analysis and Machine Intelligence 23*(1), 78–84. https://doi.org/10.1109/34.899948

77. Knorr, E. M., & Ng, R.T. (1998). Algorithms for Mining Distance-Based Outliers in Large Dataset. *Proceedings of the 24th VLDB International Conference*, New York, USA, (pp. 392–403).

https://dl.acm.org/doi/10.5555/645924.671334

78. Hill, D. J., & Minsker, B. S. (2010). Anomaly Detection in Streaming Environmental Sensor Data: A Data-driven Modeling Approach. *Environmental Modelling and Software*, *25*(9), 1014–1022. https://doi.org/10.1016/j.envsoft.2009.08.010

79. Muthukrishnan, S., Shah, R., & Vitter, J. (2004). Mining Deviants in Time Series Data Streams. Proceedings. *16th International Conference on Scientific and Statistical Database Management* (pp. 41–50). https://doi.org/10.1109/SSDM.2004.1311192

80. Jerez, C. I., Zhang, J., & Silva, M. R. (2023). On Equivalence of Anomaly Detection Algorithms. *ACM Transactions on Knowledge Discovery from Data*, *17*(2), 21:1-21:26. https://doi.org/10.1145/3536428

81. Solberg, H. E. & Lahti, A. (2005). Detection of outliers in reference distributions: Performance of horn's algorithm. *Clinical Chemistry 51*(12), 2326–2332.

82. Han, J., Kamber, M., & Pei, J. (2012). 12—Outlier Detection. In J. Han, M. Kamber, & J. Pei (Eds.), *Data Mining (Third Edition)* 543–584. Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00012-5

83. Weigend, A. S., Mangeas, M., & Srivastava, A. N. (1995). Nonlinear gated experts for time-series: discovering regimes and avoiding overfitting. *International Journal of Neural Systems 6*(4), 373–399. https://doi.org/10.1142/s0129065795000251

84. Salvador, S. & Chan, P. (2003). Learning states and rules for time-series anomaly detection. *Tech. Rep. CS–2003–05, Department of Computer Science, Florida Institute of Technology Melbourne FL 32901.* https://cs.fit.edu/media/TechnicalReports/cs-2003-05.pdf

85. Kou, Y., Lu, C.-T., & Chen, D. (2006). Spatial weighted outlier detection. *Proceedings of SIAM Conference on Data Mining* (pp. 614–618). https://doi.org/10.1137/1.9781611972764.71

86. Shekhar, S., Lu, C.-T., & Zhang, P. (2001). Detecting graph-based spatial outliers: algorithms and applications (a summary of results). *Proceedings of the 7th ACM SIGKDD International conference on Knowledge discovery and data mining*. ACM Press, New York, NY, USA, (pp. 371–376). https://doi.org/10.1145/502512.502567

87. Lindemann, B., Maschler, B., Sahlab, N., & Weyrich, M. (2021). A survey on anomaly detection for technical systems using LSTM networks. *Computers in Industry*, *131*, 103498. https://doi.org/10.1016/j.compind.2021.103498

88. Hayes, M. A., & Capretz, M. A. (2015). Contextual anomaly detection framework for big sensor data. *Journal of Big Data*, *2*(1), 2. https://doi.org/10.1186/s40537-014-0011-y

89. Hojjati, H., Ho, T. K. K., & Armanfard, N. (2023). *Self-Supervised Anomaly Detection: A Survey and Outlook* (arXiv:2205.05173). arXiv. https://doi.org/10.48550/arXiv.2205.05173

90. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *ICML'20: Proceedings of the 37th International Conference on Machine Learning* (pp. 1597–1607). https://dl.acm.org/doi/abs/10.5555/3524938.3525087

91. Salehi, M., Eftekhar, A., Sadjadi, N., Rohban, M.H., & Rabiee, H.R. (2020). Puzzle-ae: Novelty detection in images through solving puzzles. https://doi.org/10.48550/arXiv.2008.12959

92. Tack, J., Mo, S., Jeong, J., & Shin, J. (2020). CSI: Novelty detection via contrastive learning on distributionally shifted instances. *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems* (pp. 11839–11852). https://dl.acm.org/doi/10.5555/3495724.3496717

93. Xu, J., & Stirenko, S. (2023). Mixup Feature: A Pretext Task Self-Supervised Learning Method for Enhanced Visual Feature Learning. *IEEE Access*, *11*, 82400–82409. https://doi.org/10.1109/ACCESS.2023.3301561

94. Xu, J., & Stirenko, S. (2022). Self-supervised Model Based on Masked Autoencoders Advance CT Scans Classification. *International Journal of Image, Graphics and Signal Processing*, *14*(5), 1–9. https://doi.org/10.5815/ijigsp.2022.05.01

95. Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explorations 6*(1), 1–6. https://doi.org/10.1145/1007730.1007733

96. Phua, C., Alahakoon, D., & Lee, V. (2004). Minority report in fraud detection: classification of skewed data. *SIGKDD Explorer Newsletter 6,* 1, 50–59. https://doi.org/10.1145/1007730.1007738

97. Steinwart, I., Hush, D., & Scovel, C. (2005). A classification framework for anomaly detection. *Journal of Machine Learning Research 6*, 211–232. https://dl.acm.org/doi/10.5555/1046920.1058109

98. Lee, K., Lee, K., Lee, H., & Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 7167–7177). https://dl.acm.org/doi/10.5555/3327757.3327819

99. Kim, S., Choi, Y., Lee, M. (2015). Deep learning with support vector data description. *Neurocomputing 165*, 111–117. https://doi.org/10.1016/j.neucom.2014.09.086

100. Doms, V., Gordienko, Y., Kochura, Y., Rokovyi, O., Alienin, O., & Stirenko, S. (2021). Deep Learning for Melanoma Detection with Testing Time Data Augmentation. In Z. Hu, Q. Zhang, S. Petoukhov, & M. He (Eds.), *Advances in Artificial Systems for Logistics Engineering* (pp. 131–140). Springer International Publishing. https://doi.org/10.1007/978-3-030-80475-6_13

101. Ахієзер, О., Грінберг, Г., Любчик, Л., & Ямковий, К. (2023). Побудова регресійної моделі інтенсивності відмов за агрегованими даними з використанням методів ядерного машинного навчання. *Вісник Національного технічного університету «ХПІ». Серія: Системний аналіз, управління та*

інформаційні технології, (2 (8), 51–56. https://doi.org/10.20998/2079-0023.2022.02.08

102. Malinovic-Milicevic, S., Radovanovi, M.M., Radenkovic, S.D., Vyklyuk, Y., Milovanovic, B., Milanovic Pešic, A., Milenkovic, M., Popovic, V., Petrovic, M., & Sydor, P. et al (2023). Application of Solar Activity Time Series in Machine Learning Predictive Modeling of Precipitation-Induced Floods. *Mathematics*, *11*(4), 795. https://doi.org/10.3390/math11040795

103. Белас, А. О., & Бідюк, П. І. (2021). Вибір критерію якості для оцінювання прогнозів нелінійних нестаціонарних процесів. *Наукові вісті КПІ*. (2 (2021), 38–45. https://doi.org/10.20535/kpisn.2021.2.236936

104. Левенчук, Л. Б., & Бідюк, П. І. (2020). Байєсівський аналіз даних у моделюванні та прогнозуванні нелінійних нестаціонарних процесів. *Наукові вісті КПІ*. (3 (2020), 14–23. https://doi.org/10.20535/kpi-sn.2020.3.209877

105. Bidyuk , P. I., Korshevnyuk, L. O., Gozhyj, O. P., Kalinina, I. O., Prosyankina-Zharova, T. I., & Terentiev O. M. (2019). Modeling and forecasting financial and economic processes with decision support system. *Наукові вісті КПІ*. (5-6 (2019), 7–17. https://doi.org/10.20535/kpi-sn.2019.5-6.176835

106. Левенчук, Л. Б., Гуськова, В. Г., & Бідюк, П. І. (2021). Ймовірнісне моделювання операційних ризиків. *Наукові вісті КПІ*, (3 (2021), 26–37. https://doi.org/10.20535/kpisn.2021.3.251681

107. Hosseini, M., McNairn, H., Mitchell, S., Robertson, L. D., Davidson, A., Ahmadian, N., Bhattacharya, A., Borg, E., Conrad, C., Dabrowska-Zielinska, K., de Abelleyra, D., Gurdak, R., Kumar, V., Kussul, N., Mandal, D., Rao, Y. S., Saliendra, N., Shelestov, A., Spengler, D., … Becker-Reshef, I. (2021). A Comparison between Support Vector Machine and Water Cloud Model for Estimating Crop Leaf Area Index. *Remote Sensing*, *13*(7), 1348. https://doi.org/10.3390/rs13071348

108. Яйлимова, Г. О., Яйлимов, Б. Я., Шелестов, А. Ю., & Красільнікова Т. М. (2022). Інтелектуальні методи та моделі обробки супутникових даних у задачі моніторингу звалищ. *Проблеми керування та*

інформатики, *2*, 128–140.

109. Fujimaki, R., Yairi, T., & Machida, K. (2005). An approach to spacecraft anomaly detection problem using kernel feature space. *Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM Press, New York, NY, USA, (pp. 401–410). https://doi.org/10.1145/1081870.1081917

110. Dasgupta, D. & Majumdar, N. (2002). Anomaly detection in multidimensional data using negative selection algorithm. *Proceedings of the IEEE Conference on Evolutionary Computation*. Hawaii, 1039–1044. https://doi.org/10.1109/CEC.2002.1004386

111. Ruff, L., Vandermeulen, R.A., Gornitz, N., Binder, A., Muller, E., Muller, K.R., & Kloft, M. (2019). Deep semi-supervised anomaly detection. *2020 International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1906.02694

112. Kiran, B.R., Thomas, D.M., & Parakkal, R. (2018). An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging*, *4*, 36. https://doi.org/10.3390/jimaging4020036

113. Lyubchyk, L., & Yamkovyi, K. (2022). Comparative Analysis of Modified Semi-Supervised Learning Algorithms on a Small Amount of Labelled Data. *System Research & Information Technologies*, *4*, 34–43. https://doi.org/10.20535/SRIT.2308-8893.2022.4.03

114. Lyubchyk, L., Galuza, A., & Grinberg, G. (2020). Semi-supervised Learning to Rank with Nonlinear Preference Model. *Studies in Fuzziness and Soft Computing*, *391*, 81–103. ISSN 1434-9922. https://doi.org/10.1007/978-3-030-38893-5_5

115. Ruff, L., Kauffmann, J.R., Vandermeulen, R.A., Montavon, G., Samek, W., Kloft, M., Dietterich, T.G., & Muller, K.R. (2021). A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*. *109*(5), 756–795 https://doi.org/10.1109/JPROC.2021.3052449

116. Hodge, V., & Austin, J., (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*, 85–126. https://doi:10.1023/B:AIRE.0000045502.10941.a9

117. Lande, D., Soboliev, A., & Dmytrenko, O. (2022). Intelligent technologies in information retrieval systems. *Artificial Intelligence*, *27*(1), 260–268. https://doi.org/10.15407/jai2022.01.260

118. Ланде, Д. В., Страшной, Л., & Балагура, І. В. (2021). Метод формування та кластеризації кореляційних мереж понять. *Реєстрація, зберігання і обробка даних*, *23*(2), 27–36. https://doi.org/10.35681/1560-9189.2021.23.2.239209

119. Ланде, Д. В., & Снарський, А. О. (2020). Мережі, що визначаються динамікою тематичних інформаційних потоків. *Реєстрація, зберігання і обробка даних*, *22*(1), 56–61. doi.org/10.35681/1560-9189.2020.1.1.207784

120. Parmar, P., Gharat, A., & Rhodin, H. (2022). Domain Knowledge-Informed Self-Supervised Representations for Workout Form Assessment. *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022*, *Proceedings, Part XXXVIII* (pp. 105–123). https://doi.org/10.1007/978-3-031-19839-7_7

121. Koshkina, M., Pidaparthy, H., & Elder, J. H. (2021). Contrastive Learning for Sports Video: Unsupervised Player Classification. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (pp. 4523–4531). https://doi.org/10.1109/CVPRW53098.2021.00510

122. Han, J., Kamber, M., & Pei, J. (2012). 9 - Classification: Advanced Methods. In J. Han, M. Kamber, & J. Pei (Eds.), *Data Mining (Third Edition)* (pp. 393–442). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00009-5

123. Li, G., & Jung, J. J. (2023). Deep learning for anomaly detection in multivariate time series: Approaches, applications, and challenges. *Information Fusion*, *91*(C), 93–102. https://doi.org/10.1016/j.inffus.2022.10.008

124. Chalapathy, R., & Chawla, S. (2019). *Deep Learning for Anomaly Detection: A Survey* (arXiv:1901.03407). arXiv. https://doi.org/10.48550/arXiv.1901.03407

125. Pang, G., Shen, C., Cao, L., & Hengel, A. V. D. (2021). Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*, *54*(2), 38:1-38:38. https://doi.org/10.1145/3439950

126. Petrović, Marko D.; Radovanović, Milan M.; Vyklyuk, Yaroslav; Milenković, Milan, & Tretiakova, Tatiana N. (2021). The Conditionality of Outdoor Sports Events on Weather-Induced Impacts and Possible Solution. *Journal of hospitality & tourism research (Washington, D.C.)*, 2021, *45*(7), 1303-1323. https://doi.org/10.1177/1096348020971028

127. Fu, M., Le, C., Fan, T., Prakapovich, R., Manko, D., Dmytrenko, O., Lande, D., Shahid, S., & Yaseen, Z. M. (2021). Integration of complete ensemble empirical mode decomposition with deep long short-term memory model for particulate matter concentration prediction. *Environmental Science and Pollution Research*, *28*(45), 64818–64829. https://doi.org/10.1007/s11356-021-15574-y

128. Oh, M., & Iyengar, G. (2020). Sequential Anomaly Detection using Inverse Reinforcement Learning. *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1480–1490. https://doi.org/10.1145/3292500.3330932

129. Pang, G., Hengel, A. van den, Shen, C., & Cao, L. (2021). Toward Deep Supervised Anomaly Detection: Reinforcement Learning from Partially Labeled Anomaly Data. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 1298–1308). https://doi.org/10.1145/3447548.3467417

130. Novotarskyi, M. A., Stirenko, . S. G., Gordienko, Y. G., & Kuzmych, V. A. (2021). Deep Reinforcement Learning with Sparse Distributed Memory for "Water World" problem solving. *Radio Electronics, Computer Science, Control*, (1), 136–143. https://doi.org/10.15588/1607-3274-2021-1-14

131. Belhadi, A., Djenouri, Y., Srivastava, G., & Lin, J. C.-W. (2021). Reinforcement learning multi-agent system for faults diagnosis of mircoservices in industrial settings. *Computer Communications*, *177*(C), 213–219. https://doi.org/10.1016/j.comcom.2021.07.010

132. Garcia, N. M. (2019). Multi-agent system for anomaly detection in Industry 4.0 using Machine Learning techniques. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, (4), 33-40. https://doi.org/10.14201/adcaij2019843340

133. Dorofieiev, Y.I., & Lyubchyk, L. M. (2019). Consensus control of multi-agent systems with input delays: A descriptor model approach. *Mathematical Modeling and Computing*, *6*(2), 333–343. https://doi.org/10.23939/mmc2019.02.333

134. Han, J., Kamber, M., & Pei, J. (2012). 10 - Cluster Analysis: Basic Concepts and Methods. In J. Han, M. Kamber, & J. Pei (Eds.), *Data Mining (Third Edition)* (pp. 443–495). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00010-1

135. Raschka, S. (2015). *Python Machine Learning.* Packt Publishing.

136. Burkov, A. (2019) *The Hundred-Page Machine Learning Book*. 159.

137. Giordani, P., Ferraro, M. B., & Martella, F. (2020). *An Introduction to Clustering with R* (Vol. 1). Springer. https://doi.org/10.1007/978-981-13-0553-5

138. Кокорева, Я. В., & Макаров, А. А. (2015). Поэтапный процесс кластерного анализа данных на основе алгоритма кластеризации k-means. *Молодой ученый*. *13*(93). 126–128.

139. Han, J., Kamber, M., & Pei, J. (2012). 8 - Classification: Basic Concepts. In J. Han, M. Kamber, & J. Pei (Eds.), *Data Mining (Third Edition)* (pp. 327–391). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00008-3

140. Balasubramanian, V., Ho, S.-S., & Vovk, V. (2014). *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications*

(1st ed.). Morgan Kaufmann Publishers Inc. https://dl.acm.org/doi/10.5555/2671155

141. Ho, S.-S., & Wechsler, H. (2010). A Martingale Framework for Detecting Changes in Data Streams by Testing Exchangeability. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(12), 2113–2127. https://doi.org/10.1109/TPAMI.2010.48

142. Laxhammar, R. (2014). Conformal anomaly detection : Detecting abnormal trajectories in surveillance applications [PhD dissertation, University of Skövde]. Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-8762

143. Матвеев, А., Бурнаев, Е., & Тетко, И. (2017). Конформная классификация в задачах прогнозирования физико-химических свойств молекул. *ИТuС 2017*, 155–163.

144. Gammerman, A., & Vovk, V. (2007). Hedging Predictions in Machine Learning: The Second Computer Journal Lecture. *The Computer Journal*, *50*(2), 151–163. https://doi.org/10.1093/comjnl/bxl065

145. Laxhammar, R., & Falkman, G. (2011). Sequential Conformal Anomaly Detection in trajectories based on Hausdorff distance. *14th International Conference on Information Fusion*, (pp. 1–8).

146. Shafer, G., & Vovk, V. (2007). A tutorial on conformal prediction. *Journal of Machine Learning Research*, *9*, 371–421. https://dl.acm.org/doi/10.5555/1390681.1390693

147. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

148. Vovk, V., Nouretdinov, I., & Gammerman, A. (2003). Testing Exchangeability On-Line. *ICML'03: Proceedings of the Twentieth International Conference on International Conference on Machine Learning* (pp. 768–775). https://dl.acm.org/doi/abs/10.5555/3041838.3041935

149. Vovk, V. G. (1993). A Logic of Probability, with Application to the Foundations of Statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, *55*(2), 317–351. https://doi.org/10.1111/j.2517-6161.1993.tb01904.x

150. Dai, L. & Bouguelia, M. (2021). Testing Exchangeability With Martingale for Change-Point Detection. *International Journal of Ambient Computing and Intelligence (IJACI), 12*(2), 1-20. http://doi.org/10.4018/IJACI.2021040101

151. Хмаладзе, Э. В. (1982). Некоторые применения теории мартингалов в статистике. *УМН*, *37*(6), 193–212

152. Doob, J. L. (1953). *Stochastic processes*. New York: Wiley.

153. Ho, S.-S., Schofield, M., Sun, B., Snouffer, J., & Kirschner, J. (2019). A Martingale-Based Approach for Flight Behavior Anomaly Detection. *2019 20th IEEE International Conference on Mobile Data Management (MDM)* (pp. 43–52). https://doi.org/10.1109/MDM.2019.00-75

154. MacQueen, J. B. (1967). Some Methods for Classification and Analysis of MultiVariate Observations. In L. M. L. Cam & J. Neyman (Eds.). *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281-297), : University of California Press.

155. Li, R.-P., & Mukaidono, M. (1995). A maximum-entropy approach to fuzzy clustering. *Proceedings of 1995 IEEE International Conference on Fuzzy Systems.*, *4*, 2227–2232. https://doi.org/10.1109/FUZZY.1995.409989

156. Li, R.-P., & Mukaidono, M. (1999). Gaussian clustering method based on maximum-fuzzy-entropy interpretation. *Fuzzy Sets and Systems*, *102*(2), 253–258. https://doi.org/10.1016/S0165-0114(97)00126-7

157. Krishnapuram, R., Joshi, A., Nasraoui, O., & Yi, L. (2001). Low-complexity fuzzy relational clustering algorithms for Web mining. *IEEE Transactions on Fuzzy Systems*, *9*(4), 595–607. https://doi.org/10.1109/91.940971

158. Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer US. https://doi.org/10.1007/978-1-4757-0450-1

159. Ju, Z., & Liu, H. (2012). Fuzzy Gaussian Mixture Models. *Pattern Recognition*, *45*(3), 1146–1158. https://doi.org/10.1016/j.patcog.2011.08.028

160. Baid, U., & Talbar, S. (2016). Comparative Study of K-means,

Gaussian Mixture Model, Fuzzy C-means algorithms for Brain Tumor Segmentation. *Proceedings of the International Conference on Communication and Signal Processing 2016 (ICCASP 2016)* (pp. 583–588). https://doi.org/10.2991/iccasp-16.2017.85

161. Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D., & Reynolds, D. A. (2004). A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Advances in Signal Processing*, *2004*(4), 101962. https://doi.org/10.1155/S1110865704310024

162. Lehmann, E. L. (1999) *Elements of Large Sample Theory*. Springer, New York. https://doi.org/10.1007/b98855

163. Ross, S. M. (2017). *Introductory Statistics*. Academic Press.

164. Kremer, N. Sh. (2001). *Theory of Probability and Mathematical Statistics*. UNITY-DANA.

165. Taha, H. A. (2005). *Operations Research: An Introduction.* (7th ed.). Pearson Education, Singapore Pte. Ltd., Indian Branch, Delhi.

166. Axelsson, S. (2000). The base-rate fallacy and the difficulty of intrusion detection. *ACM Transactions on Information and System Security*, *3*(3), 186–205. https://doi.org/10.1145/357830.357849

ДОДАТОК А. СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Наукові праці, в яких опубліковані основні наукові результати дисертації:

1. Zhuk, I., & Chertov, O. (2023). Framework based on conformal predictors and power martingales for detection of fixed football matches. *Eastern-European Journal of Enterprise Technologies*, *2*(4 (122), 6–15. https://doi.org/10.15587/1729-4061.2023.276977 [Scopus Q3].

У роботі здобувачем розроблені методи виявлення підозрілих на фіксований результат футбольних матчів на основі конформного предиктора і степеневого мартингала, які застосовано до даних реального футбольного сезону 2013-2014 років Ліги 2 Франції.

2. Chertov, O., & Zhuk, I. (2023). Detection of fixed football matches based on the theory of conformal predictors using the modified Stepanets indicator function. *Eastern-European Journal of Enterprise Technologies*, *3*(4 (123), 22–32. https://doi.org/10.15587/1729-4061.2023.282645 [Scopus Q3].

У роботі здобувачем розроблено удосконалений метод виявлення підозрілих на фіксований результат футбольних матчів та застосовано запропоновані методи до даних реального футбольного сезону 2014-2015 років Серії В Італії.

3. Чертов, О. Р., & Жук, І. С. (2022). Імітаційна модель футбольного сезону з матчами з фіксованим результатом, *Наукові вісті КПІ, 1–2*, с. 82–94, 2022. https://doi.org/10.20535/kpisn.2022.1-2.287916

У роботі здобувачем розроблено алгоритм формування імітаційної моделі футбольного сезону з договірними матчами.

2. Наукові праці, які засвідчують апробацію матеріалів дисертації:

4. Chertov, O. R., & Zhuk, I. S. (2023). Clusterization of soccer season teams using K-means and GMM methods. *Intelligent Solutions-S: Proceedings of the International Symposium, September 28, 2023, Kyiv-Uzhorod, Ukraine / Ministry of Education and Science of Ukraine, Taras Shevchenko National University of Kyiv and* [etc]; Vitaliy Ye. Snytyuk (Editor) (pp. 14-15).

Здобувачем запропоновано і реалізовано процедуру регуляризації коваріаційної матриці для методу кластеризації на основі Гаусівських сумішей.

5. Жук, І. (2020). Застосування конформних предикторів і степеневих мартингалів для виявлення підозрілих матчів футбольних турнірів. *Прикладна математика та комп'ютинг. ПМК, 2022* : п’ятнадцята наук. конф. магістрантів та аспірантів, Київ, 16-18 лист. 2022 р. : зб. тез доп. / [редкол.: Дичка І. А. та ін.]. — К. : Просвіта, 2022. — 368 с. ISBN 978-617-7010-23-3 С. 24 – 29.

6. Chertov, O., Zhuk, I., & Serdyuk, A. (2021). Search of the Deviation from the Natural Process Using Stepanets Approach for Classification of Functions. *2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)* (pp. 720-724), doi: 10.1109/IDAACS53288.2021.9660997. [Scopus].

Здобувачем запропоновано алгоритм застосування математичного апарату класифікації функцій для розв’язання задачі пошуку матчів підозрілих на фіксований результат, проведено моделювання з використанням даних футбольного сезону 2013-2014 років Ліги 2 Франції.

7. Жук, І. (2020). Виявлення зовнішнього впливу в інформаційних потоках мережі Internet як проблема ідентифікації образу. *Філософія і науково-технічна творчість у хронотопі технічного університету: Матеріали ІІІ Міжнародної науково-практичної конференції.* – 410 с. С.144 – 147.

8. Жук, І. С., & Чертов, О. Р. (2020). Використання математичного

апарату наближень Степанця для виявлення штучних втручань у сигналах різної природи. *Інтегровані інтелектуальні робототехнічні комплекси (ІІРТК-2020). Тринадцята міжнародна науково-практична конференція 19-20 травня 2020 р.*, Київ, Україна. – К.: НАУ, 2020. – 305 с. С. 276 – 278.

Здобувачем проаналізовано можливості застосування індикаторних функцій Степанця для пошуку аномалій у характеристиках, які описуються періодичними функціями.