

АНОТАЦІЯ

Забелін С.І. Моделі і методи прогнозування вулканічної активності з використанням технології штучного інтелекту. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 122 “Комп’ютерні науки”. – Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, 2020.

В дисертації вперше отримані такі нові наукові результати:

1. Вперше розроблено метод селекції інформативних змінних, який відрізняється тим що використовує результати роботи системи прогнозування для ітеративного поліпшення змінних які вибираються.

2. Вдосконалено ітеративний метод зменшення розмірності на основі методу Isomap який відрізняється тим що використовує зворотній зв’язок з перевіркою системою, який дозволяє досягти високої інформативності змінних.

3. Удосконалена задача багатокритеріального вибору, яка відрізняється тим, що для вибору змінних використовуються рівні довіри, експертна оцінка значимості, для підвищення точності моделі.

4. Розроблено новий генетичний алгоритм для генерування альтернатив при формуванні прогнозу, який відрізняється від існуючих, використанням спеціального полінома, який зв’язує всі вхідні змінні та дозволяє розкрити приховані взаємозв’язки всередині змінних.

Розроблено новий метод селекції інформаційних змінних який відрізняється від існуючих можливістю ефективно обробляти велику кількість

вхідних даних різної сутності, завдяки наявності декількох незалежних частин відбору.

Метод полягає в наступному. Прогнозуємо за n вхідними змінними $\{x_1, \dots, x_n\}$. Для оцінки змінних ми рахуємо помилки множинної регресії без однієї змінною по черзі і порівнюємо з множинною регресією і формуємо рейтинг як $r_{i0} = \frac{\min_{j=1,n}(\varepsilon_j - \varepsilon_0)}{\varepsilon_i - \varepsilon_0}$. Обчислюємо рейтинг для n вхідних змінних використовуючи кореляційний вибір змінних. Третя група оцінок обчислюється за допомогою розробленого ітеративного метода зменшення розмірності, який відрізняється від існуючих наявністю системи оцінок декількох отриманих підмножин і їх ітеративною зміною на основі їх результатів отриманих у системі прогнозування для зміни масштабування окремих тисних точок при зменшенні рейтингу. Отримавши три групи рейтингів використаємо модифікований багатокритеріальний вибір, щоб отримати m змінних що можна використовувати для побудування моделі прогнозування.

Розроблено систему прогнозування. На вхід в систему подаються різні вулканічні показники, за якими буде відбуватися прогнозування шуканих величин. На етапі автоматичної корекції даних відбувається нормалізація дани, їх регуляризація і інтерполяція, якщо в цьому є необхідність. Система робить вибір змінних використовуючи кореляційний аналіз, регресійний аналіз і ітеративний метод зменшення розмірності. Які проходять рейтинговий багатокритерійний вибір. Після багатокритеріального вибору отримані змінні проходять фазифікацію, яка була описана в цьому розділі, для більш точного побудови моделі. Використовуючи нечіткі дані відбувається навчання LSTM нечіткої мережі на випадково обраному часовому відрізку вхідних даних, тоді як решта даних буде використовуватися як перевірна вибірка. Після навчання моделі, відбувається дефазифікація даних, і перевірка моделі на перевірочній вибірці. Якщо отриманий результат відповідає заданому

критерію точності, то отримана модель є оптимальною, інакше починається нова ітерація з використанням ітеративного методу зменшення розмірності.

Безпосереднє практичне значення результатів, отриманих у ході дисертаційного дослідження, полягає в тому, що:

1. Розроблена система прогнозування вулканічної активності яка використовує ітеративний вибір змінних, яка може використовуватись для прогнозування не тільки вулканічної активності, але для прогнозування інших природних явищ, зокрема землетрусів.

2. Розроблено генетичний алгоритм прогнозування вулканічної активності, що може прогнозувати декілька альтернативних прогнозів.

3. Досягнута висока точність прогнозування в порівнянні з існуючими методами зменшення розмірності при роботі з вулканічною активністю завдяки модифікованому метод зменшення розмірності Isomap.

4. Збільшена точність прогнозування завдяки покращеній селекції інформаційних змінних і модифікованій задачі багатокритеріального вибору.

Були проведені експериментальні дослідження системи прогнозування і порівняльний аналіз з існуючими аналогами, де було показано перевагу розробленої системи прогнозування над ними.

Проведені експериментальні дослідження системи прогнозування і порівняльний аналіз з існуючими аналогами, де було показано перевагу розробленої системи прогнозування над ними. Так система прогнозування виявилася кращою в середньому на 50% ніж множинна регресія, на 38% ніж мережа Байеса. Проведено аналіз системи прогнозування до зашумлення даних і його вплив на вибір змінних.

Проведено аналіз отриманих популяцій генетичного алгоритму і їх поліномів і доведена ефективність поліноміального полінома, яка збільшила точність прогнозування особин в популяціях в середньому на 5%. Проведено

порівняльний аналіз існуючих методів прогнозування часових рядів. Так генетичний алгоритм виявився в середньому краще на 6% ніж VAR, і на 20% ніж множинна регресія.

Ключові слова: нейроні мережі, генетичний алгоритм, вибір змінних, зменшення розмірності, нечіткі числа, вулкани, вулканічна активність, прогнозування.

SUMMARY

Zabielin S. Models and methods of forecasting volcanic activity using artificial intelligence technology. - Qualifying scientific work, the manuscript.

Thesis for a PhD degree in specialty 122 "Computer Science". – National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, 2020.

In the dissertation the following new scientific results were first received:

1. A method of variable selection was developed for the first time, characterized in that it uses the results of the forecasting system to iteratively improve the selected variables.

2. Improved iterative method of reducing the dimension based on the method of Isomap, which differs in that it uses feedback from the test system, which allows to achieve high informativeness of variables.

3. Improved multi-criteria selection task, characterized in that the selection of variables uses levels of confidence, expert assessment of significance, to improve the accuracy of the model.

4. A new genetic algorithm has been developed to generate alternatives in the formation of a forecast that differs from the existing ones, using a special polynomial that connects all input variables and allows to reveal hidden relationships within variables.

A new method of variable selection has been developed that differs from the existing ability to efficiently process large amounts of input variable, due to the presence of several independent criteria of the selection. We predict n input variables $\{x_1, \dots, x_n\}$. To estimate the variables, we count the linear regression errors

without one variable each step and compare with the linear regression and rank as

$$r_{i0} = \frac{\min_{j=1,n} (\varepsilon_j - \varepsilon_0)}{\varepsilon_i - \varepsilon_0}.$$

We also calculate the rating for n input variables using the correlation selection of variables. The third group of estimates is calculated using the developed iterative method of dimensionality reduction, which differs from the existing methods by iterative change of created subsets based on their results obtained in the prediction system to change the scaling of individual narrow points with decreasing rating. Having received three sets of ratings, we will use modified multicriteria selection to obtain m variables that can be used to build a forecasting model.

A forecasting system was developed. Various volcanic indicators are put forward to the system, according to which the required values will be predicted. At the stage of automatic data correction, the data is normalized, regularized, and interpolated, if necessary. The system then selects variables using correlation analysis, regression analysis and an iterative dimensionality reduction method. Further passing the multi-criteria rating choice. After multicriteria selection, the resulting variables are subjected to the facetization described in this section for more accurate model construction. Further, using fuzzy data, the LSTM fuzzy network is trained on a randomly selected time slice of input, while the rest of the data will be used as a test sample. After training the model, the data defuzzied and the model is validated on a test sample. If the obtained result meets the specified accuracy criterion, then the obtained model is optimal, otherwise a new iteration begins using an iterative dimensionality reduction method.

The immediate practical significance of the results obtained during the dissertation research is that:

1. System was developed for predicting volcanic activity that uses iterative selection of variables, which can be used to predict not only volcanic activity, but to predict other natural phenomena, including earthquakes.

2. An evolutionary algorithm for predicting volcanic activity has been developed, which can create several alternative forecasts.

3. Achieved high forecasting accuracy compared to existing methods of dimension reduction when working with volcanic activity due to the modified method of dimension reduction of Isomap.

4. Increased forecasting accuracy due to improved selection of information variables and modified multicriteria selection problem.

Experimental studies of the forecasting system and comparative analysis with existing analogues, which showed the advantage of the developed forecasting system over them. Thus, the prediction system was better by an average of 50% than linear regression, 38% better than the Bayesian network. The analysis of the forecasting system before data noise and its influence on the choice of variables is carried out.

The analysis of the received populations of the evolutionary algorithm and their polynomials was carried out and efficiency of the polynomial which has increased accuracy of forecasting of individuals in populations on the average by 5%. A comparative analysis was done of existing methods of forecasting time series. Thus, the evolutionary algorithm turned out to be on average 6% better than VAR, and 20% better than linear regression.

Keywords: neural networks, genetic algorithm, variable selection, dimensionality reduction, fuzzy numbers, volcanoes, volcanic activity, forecasting.

List of main publications of the applicant:

1. Zabielin, S. I. (2018). Big Data analysis via model reduction methods. System Research and Information Technologies, 0(2), 35-41. doi:10.20535/srit.2308-8893.2018.2.04 (фактова)

2. Zabielin, S. I. (2018). Forecasting activity of the Kilauea volcano using intelligent methods of data analysis. *International Journal of Information Theories and Applications*, 25(4), 94-99. (scopus, іноземного видання)
3. Zabielin, S. (2019). Forecasting SO₂ emission of Kilauea volcano using intelligent method of data analysis. *System Research and Information Technologies*, 0(4), 30-38. doi:10.20535/srit.2308-8893.2019.4.03(фахова)
4. Забелин С. И. (2017). Геометрические методы анализа Big Data. 19-th International conference on System Analysis and Information Technology SAIT 2017, May 22–25 (конференція)
5. Забелин С. И. (2018). Интеллектуальный анализ вулканических газов с помощью интеллектуальных методов анализа данных. 20-th International conference on System Analysis and Information Technology SAIT 2018, May 21–24. (конференція)