Far Eastern Mathematical Journal. 2022. V. 22. No 2. P. 179–184

UDC 004.89
MSC2020 68T50

© L. S. Grishina[1], A. Yu. Zhigalov[1], I. P. Bolodurina[1,2], E. L. Borshhuk[2], D. N. Begun[2], Yu. V. Varennikova[3]

Investigation of the efficiency of graph data representation for a cardiovascular disease predictive model by deep learning methods

Currently, cardiovascular diseases (CVD) are the most common cause of death in the world. Artificial intelligence methods provide extensive opportunities for extracting new knowledge from the raw data of medical information systems (MIS). This study is aimed at building a model for predicting the diagnosis of CVD based on patient complaints at a doctor's appointment using natural language processing methods. The formation of the initial data set is based on a graph model of the patient's medical history with CVD according to the visit protocols. A comparative analysis of machine learning models such as the naive Bayesian classifier, the support vector machine and convolution neural networks is carried out. As a result of the experiments, the most effective model for predicting CVD has been selected.

Key words: *natural language processing, graph model, cardiovascular disease, convolutional neural networks, support vector machine, medical information systems, disease prediction model.*

DOI: https://doi.org/10.47910/FEMJ202222

Introduction

Cardiovascular diseases top the ranking of the most important causes of death in the world – every year 17 million people die due to CVD. This problem has many initiatives to reduce the mortality rate, the main idea of which is to develop screening and early diagnosis programs [1, 2]. The introduction of machine learning methods into medical and clinical processes will allow automating many tasks to ensure timely patient care. Within the framework of this study, a model for predicting the diagnosis of International Classification of Diseases (ICD) for CVD is being built based on textual information of patient complaints at a doctor's appointment using natural language processing (NLP)

[1] Orenburg State University, Russia, 460018, Orenburg, prosp. Pobedy, 13.
[2] Orenburg State Medical University, Russia, 460000, Orenburg, Sovetskaya street, 6.
[3] Medical Information and Analytical Center of the city of Orenburg, Russia, 460024, Orenburg, Marchal Zhukova street, 42.
E-mail: prmat@mail.osu.ru (I. P. Bolodurina).

methods. The formation of the initial data set is based on a graph model of the patient's medical history with CVD by converting xml files of the visit protocols and extracting the patient's complaints in chronological order. The data of the visit protocols are provided by the Medical Information and Analytical Center (MIAC) of the city of Orenburg. As a result of the experiments carried out, the effectiveness of predictive models – Multinomial NB, SVC and CNN model was investigated.

1 Related works

Scientists all over the world are engaged in the application of artificial intelligence methods and the development of decision support systems in the field of healthcare.

The authors of the study [3] review a meta-analytical methodology for assessing the predictive ability of machine learning algorithms and analyze the use of convolution neural networks (CNN), support vector machine (SVM) and extreme gradient boosting algorithm for predicting coronary heart disease and stroke risk. Experiments have confirmed the effectiveness of the introduction of models into medical practice.

The article [4] provides a comparative analysis of the accuracy of various machine learning methods, such as the SVM, the K-Nearest Neighbor (KNN) algorithm, decision trees and the simplest artificial neural network (ANN) for assessing the risk of coronary heart disease and building binary classification models. ANN shows the highest accuracy of diagnosis of CVD diseases (about 85%). The application of deep learning methods based on the UCI Machine Learning Heart Disease dataset for predicting the presence of CVD diseases is carried out in [5]. Experiments have shown a high accuracy of binary classification — 94%. Deep learning methods, such as CNN and recurrent neural networks (RNN), are analyzed in the study [6] in the task of predicting CVD in patients with COVID-19. The authors confirmed the effectiveness of the models for detecting the early stage of the disease.

Studies [7, 8] have demonstrated the ability of NLP to create a classifier based on discharge texts for four groups of critical illnesses, as well as to predict early psychiatric readmission. The article [9] presents models for assessing the presence of important concomitant CVD in medical records with arbitrary text. The predictive ability of the models is estimated at 85%, and the highest accuracy is obtained for conditions with greater diagnostic clarity (diabetes, hypertension, etc.).

Thus, at the moment it is an urgent task to build a model for predicting the diagnosis of CVD based on textual information of patient complaints at a doctor's appointment using NLP methods. At the same time, most of the works cover only the solution of individual problems of binary classification of diseases, or multiclass classification based on clearly structured data. This work is aimed at developing models for diagnosing CVD based on raw unstructured textual data of medical information systems collected in the original dataset.

2 Problem statement

To develop predictive models of CVD, modules of interaction with the regional MIS of the city of Orenburg are implemented: the module XMLParseModule loads impersonal protocols via the MIAC API in xml format for patients diagnosed with CVD, and the

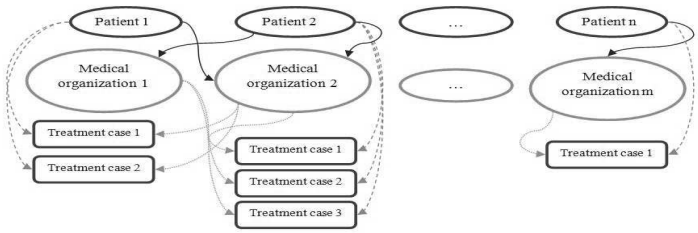

Fig. 1: Graph model of structural and semantic relations of MIS

module DictParseModule extracts information from heterogeneous xml documents by recursive traversal and analysis of all markup branches.

Thus, within the framework of this work, a graph model of structural and semantic relations between entities of various weakly structured documents of information systems necessary for intelligent processing of big data is constructed using the example of electronic medical records of patients (see figure 1).

With the help of this software and hardware complex, 364020 protocols of patient visits processed from October 1 to December 31, 2021. The volume of xml files ranged from 3 Kb to 1008 Kb (depending on the degree of fullness). The combined data on all available protocols formed into a single database. Thus, structured information has been obtained that can be analyzed by machine learning methods.

Formal mathematical formulation of the CVD classification problem: Let us X is the information about the patient's complaints at the doctor's appointment, and the set Y consists of the labels corresponding to the patient's disease according to the ICD classifier. Let us consider the problem of constructing a machine learning model $a : X \to Y$ that allows us to classify incoming complaints of new patients according to the corresponding ICD codes in some way with a given accuracy.

3 Data preprocessing

Since doctors describe the section "Complaints of the patient" inside the protocol in a free form, natural language processing methods can be used to predict diseases.

Let us do some preliminary data processing to build predictive models. Firstly, we exclude all ECG protocols, blood tests, etc. (we will leave only patient admission protocols that include complaints). Secondly, we delete records with missing values and evaluate the distribution of patient complaint records by diagnosis. Due to the rather short time period for uploading data (October-December), the number of patients who came for examination again is a small number (less than 6%), however, in order to fully take into account the patient's complaints, additional aggregation of the protocols of visits within a single diagnosis was carried out.

As a result of the analysis of the distribution of CVD, we note a strong imbalance in the data set for 43 classes. In this regard, we reduce classes with fewer than 1000 entries (see figure 2). The final data set included 41329 protocols for 9 classes of diseases. For each complaint text, 18096 features were generated (TF-IDF score for unigrams, bigrams, etc.). Random oversampling on the vector representation of complaint texts was used to balance classes.

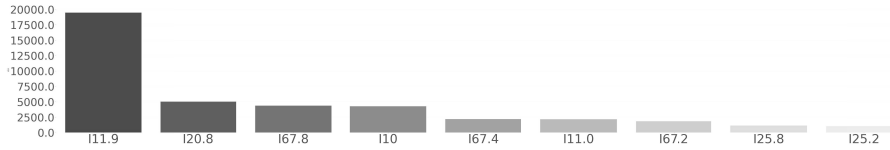

Fig. 2: Distribution of records by disease diagnoses

4 Models for predicting cardiovascular diseases

As part of this study, the application of 3 machine learning models was considered:

1) Naive Bayes Classifier. The polynomial naive Bayesian classifier is suitable for classification with discrete features. The model with the most effective parameters showed an accuracy of 66.85%. The classification was obtained unevenly, classes I11.9 and I67.8 are better determined by the model. However, the model determines most diseases less effectively in comparison with others (less than 60%).

2) Linear SVM. The linear SVM has more flexibility in choosing the penalty function and the loss function and scales better on large datasets. The Linear SVM model showed an accuracy of 68.84%. Classes I11.9, I67.8, and I67.2 are better defined by the model.

3) CNN model. The structure of the CNN is constructed as follows: after the input layer comes the embedding layer (output size is 128) with dropout (0.2), then Convolution operator for filtering neighborhoods of 1-D inputs (kernel size is 3, activation function is 'relu') and the Fully Connected classifier (250 $\times$ 9) are implemented. The training uses weighted cross-entropy loss, the Adam optimization method. The CNN model showed an accuracy of 70.17%, which is the best result among other models. The CNN confusion matrix is shown in figure 3.

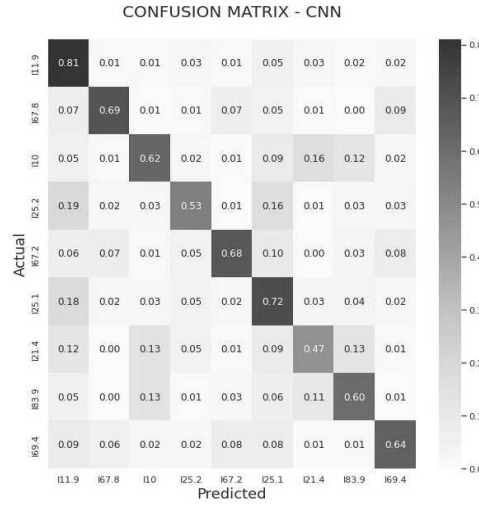

Fig. 3: Confusion matrix of CNN

A comparative analysis of methods with 5-cross-validation showed that the CNN model is the most effective method (the mean accuracy is equal to 69.96%). At the same time, this model has the standard deviation of ± 0.57% during the cross-validation, which indicates the stability of the result (see Table 1).

Table 1: The average accuracy and the standard deviation for different models.

Model name	**Average accuracy**	**Standart deviation**
Linear SVC	68.42%	±0.15%
Multinomial NB	66.4%	±0.67%
CNN model	69.96%	±0.57%

Note that the Linear SVC model provides the smallest standard deviation in the cross-validation (± 0.15%), on the other hand, the Multinomial NB model demonstrates the lowest classification accuracy (66.4%) with a standard deviation of ±0.67%.

Conclusion

A model for predicting the diagnosis of CVD has been developed that analyzes the complaints of patients at a doctor's appointment using NLP methods. As a result of the experiments carried out, the effectiveness of predictive models (Multinomial NB, Linear SVC, and CNN model) has been investigated. The highest accuracy of classification of diseases (by 1.7%) was demonstrated by the CNN method – 70.17%. At the same time, this model has a small standard deviation during the cross-validation (±0.57%), which demonstrates the stability of the prediction result.

Acknowledgements

The research was carried out within the framework of the Priority 2030 program (Agreement No. 075-15-2021-1171/2 dated May 11, 2022), with the financial support of the RFBR (project No. 20-07-01065), as well as scholarships of the President of the Russian Federation to young scientists and graduate students (No. SP-919.2022.5).

References

[1] D. K. Arnett, R. S. Blumenthal, M. A. Albert, A. B. Buroker, Z. D. Goldberger, E. J. Hahn, C. D. Himmelfarb, "2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines", *JACC*, **140**, (2019), 596–646.

[2] S. Xu, T. Zhu, Zh. Zang, D. Wang, J. Hu, X. Duan, "Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework", *Ann. of Math.*, **1**, (2017), 228–232.

[3] C. Krittanawong, H. Virk, S. Bangalore, Z. Wang, K. Johnson, R. Pinotti, H. Zhang, S. Kaplin, B. Narasimhan, T. Kitai, U. Baber, J. Halperin, W. Tang, "Machine learning prediction in cardiovascular diseases: a meta-analysis", *Sci Rep.*, **10**, (2020), 16057.

[4] S. N. Pasha, D. Ramesh, S. Mohmmad, A. Harshavardhan, "Cardiovascular disease prediction using deep learning techniques and Diophantine approximation", *IOP Conf. Ser.*, **981**, (2020), 022006.

[5] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, P. Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning", *Computational Intelligence and Neuroscience*, **1**, (2021), 8387080.

[6] S. Malathi, "Prediction of cardiovascular disease using deep learning algorithms to prevent COVID-19", *Journal of Experimental & Theoretical Artificial Intelligence*, **100**, (2021), 1–15.

[7] G. E. Weissman, M. O. Harhay, R. M. Lugo, B. D. Fuchs, S. D. Halpern, M. E. Mikkelsen, "Natural Language Processing to Assess Documentation of Features of Critical Illness in Discharge Documents of Acute Respiratory Distress Syndrome Survivors", *Ann. Am. Thorac. Soc.*, **13**, (2016), 1538–1545.

[8] A. A. Rumshisky, M. Ghassemi, T. Naumann, P. Szolovits, V. M. Castro, T. McCoy, R. H. Perlis, "Predicting early psychiatric readmission with natural language processing of narrative discharge summaries", *Transl. Psychiatry*, **6**:10, (2016), e921.

[9] A. N. Berman, "Natural language processing for the assessment of cardiovascular disease comorbidities: The cardio-Canary comorbidity project", *Clin Cardiol*, **44**:9, (2021), 1296–1304.

Received by the editors
June 15, 2022

Гришина Л. С., Жигалов А. Ю., Болодурина И. П., Борщук Е. Л., Бегун Д. Н., Варенникова Ю. В. Исследование эффективности графового представления данных для модели прогнозирования сердечно-сосудистых заболеваний методами глубокого обучения. *Дальневосточный математический журнал.* 2022. Т. 22. № 2. С. 179–184.

АННОТАЦИЯ

В настоящий момент сердечно-сосудистые заболевания (ССЗ) являются наиболее распространенной причиной смертности в мире. Методы искусственного интеллекта дают обширные возможности для извлечения новых знаний из необработанных данных медицинских информационных систем (МИС). Настоящее исследование направлено на построение модели прогнозирования диагноза ССЗ на основе жалоб пациента на приеме у врача с применением методов NLP. Формирование исходного набора данных основано на графовой модели истории болезни пациента с ССЗ по протоколам посещения. Проведен сравнительный анализ моделей машинного обучения, таких как наивный байесовский классификатор, метод опорных векторов и сверточная нейронная сеть. В результате проведенных экспериментов выбрана наиболее эффективная модель прогнозирования ССЗ.

Ключевые слова: *обработка естественного языка, графовая модель, сердечно-сосудистые заболевания, сверточные нейронные сети, метод опорных векторов, медицинские информационные системы, модель прогнозирования заболеваний.*