Hospitalization Cost Prediction for Cardiovascular Disease by Effective Feature Selection

Wei Dai[1], Mengxing Huang[1](✉), Qian Wu[1], Hanzhi Cai[2],
Ming Sheng[3], and Xin Li[4]

[1] Hainan University, Haikou 570228, China
{1848429933,854682796}@qq.com, huangmx09@163.com
[2] University of Sheffield, Sheffield, UK
caihanzhi1996@qq.com
[3] BNRist Tsinghua University, Beijing 100084, China
shengming@tsinghua.edu.cn
[4] Beijing Tsinghua Changgung Hospital, Beijing 102218, China
Horsebackdancing@sina.com

Abstract. The burden of cardiovascular diseases is increasing, and the annual growth rate of hospitalization expenses for cardiovascular diseases is much higher than that of GDP. Therefore, researchers have developed a number of intelligent systems to predict hospitalization costs for cardiovascular disease. However, there are some problems with these methods, such as the performance of real world data sets and the differences between the feature selection and the actual selection of doctors. This paper proposes a method to construct a Medical Concept Knowledge Graph (MCKG) by combining open source knowledge graphs such as Wikidata and OpenKG, open source knowledge bases such as UMLS, and doctors' prior medical knowledge. A Medical Instance Knowledge Graph (MIKG) is constructed based on MCKG and the data of cardiovascular disease related medical records from the cooperative hospital. We conduct feature selection according to MIKG, draw feature alternatives, and combine with doctor-defined rules to arrive at final feature selection. We predict hospitalization costs with random forest algorithm. Experimental results show that the average error rate of our method is lower than that of the baseline algorithms.

Keywords: Cardiovascular diseases · Concept knowledge graph · Instance knowledge graph · Feature selection · Machine learning

1 Introduction

Cardiovascular disease is a serious threat to human beings. In China, the mortality rate of cardiovascular disease is still the highest among all diseases. Cardiovascular disease is considered to be one of the major causes of death in the world. With the aging of society and the acceleration of urbanization, the prevalence of unhealthy lifestyles among Chinese resident, the risk factors of cardiovascular disease are generally exposed. At the same time, the national burden of cardiovascular disease is growing increasingly heavy.

© Springer Nature Switzerland AG 2020
G. Wang et al. (Eds.): WISA 2020, LNCS 12432, pp. 314–325, 2020.
https://doi.org/10.1007/978-3-030-60029-7_29

Since 2004, the average annual growth rate of hospitalization expenses for cardiovascular disease is much higher than the growth rate of gross domestic product (GDP) [1]. Therefore, being able to predict hospitalization expenses in advance is of great significance to both patients and hospitals [2], and how to select features according to sample data and doctors' needs is crucial. Feature selection to improve the accuracy of prediction and combined with doctors' prior knowledge can effectively reduce the error rate of prediction is a major research of machine learning [3]. Many researchers have created different algorithms to predict the hospitalization costs of cardiovascular diseases. However, these systems have the problems of unsatisfactory accuracy when facing real world data sets [4] and different requirements from actual doctors in feature selection.

The concept of knowledge graph is proposed by Google on May 17, 2012. Google will use this as a basis to build a next-generation intelligent search engine. In essence, knowledge graph is a semantic network that reveals the relationship between entities. Formal descriptions of real-world things and their relationships can be made. With theproliferation of semantic Web resources and the publication and sharing of vast amounts of RDF data, researchers in academia and industry have spent a great deal of effort building a variety of structured knowledge bases. These knowledge bases can be roughly divided into two categories: open link knowledge base and industry knowledge base. Typical examples of open linked knowledge base are Freebase, Wikidata, OpenKG, YAGO; Typical examples of vertical industry knowledge base are: IMDB (movie data), MusicBrainz (music data), MusicBrainz (semantic knowledge network).

We apply the knowledge graph to the medical field [5], and use the knowledge graph in combination with the interaction between doctors for feature selection, and use the selected data to predict the hospitalization cost of cardiovascular diseases.

The main contributions of this paper include:

a) We create the medical health concept knowledge graph (MCKG) using the open source knowledge graph such as Wikidata, OpenKG and the open source knowledge base such as the language specification defined by UMLS.
b) Based on MCKG, we build the medical instance knowledge graph (MIKG) with real data from cooperative hospitals.
c) Based on the constructed knowledge graph, we use it to conduct feature selection and obtain feature alternatives. Doctors define rules and requirements in the alternative and further obtain the final feature selection scheme.
d) We use the selected feature data to predict the hospitalization cost of cardiovascular disease, and the experiment reduces the average error rate of the prediction.

The rest of the paper is organized as follows: In Sect. 2 we discuss the related work. Section 3 introduces the methodology about how to construct MCKG and MIKG. Section 4 shows the experiment and prediction results. At last, we conclude the paper in Sect. 5.

2 Related Work

Knowledge graph is an important part of artificial intelligence technology [6]. It has been a hot trend in the field of artificial intelligence to make use of core technologies such as knowledge extraction and knowledge representation [7] of knowledge graph to carryout relevant research. Knowledge graph has a very broad application prospect in the medical services, the technology can solve the problems of strong data professionalism and complex structure in the medical field, improve medical and health services [8] and plays an important role in clinical decision support system [9].

At present, most of the studies related to cardiovascular diseases use data sets of UCI CLEVELAND [10]. Aiming at feature selection, Senthilkumarmohan et al. proposed a method about Hybrid Random Forest with Linear Model, which uses artificial neural network model with feedback for feature selection [11]. FajrI et al. used discrete minimum wavelet method for feature selection [12]. AliL et al. explained method of exhaustion to search the best configuration of the network to select relevant features from the feature space [13], and Fatih et al. selected features based on the simplified rule library [14]. Jesmin et al. combined with medical knowledge, computing intelligently to delete clinical features [15]. Prakash, S et al. used optimality criterion feature selection method for feature selection [16]. Chandra Babu Gokulnath et al. combining genetic algorithm with support vector machine used to select features in feature space [17]. Ting-Ting Zhao et al. used discriminant minimum class locality preserving canonical correlation analysis to extract features from two data sets based on gain and entropy of motion vector [18]. Sarah P et al. used convolutional neural network to make sense of feature selection [19]. Ashirjaveed et al. employed random searching algorithm to select relevant features [20].

These feature selection methods are not combined with knowledge graph. In this paper, we used a different feature selection method. We first construct MCKG based on doctors' prior knowledge, open source knowledge base and open source knowledge graph, and then integrate the structured data of hospital database and case data to obtain MIKG. We use MIKG for feature selection and get the alternative scheme of features. Then, we further screen the alternative scheme according to the rules defined by doctors and the actual needs of doctors to get the final feature selection scheme.

3 Methodology

Most of the existing medical knowledge graphs are constructed based on medical literature published on the Internet as well as various public data sets and electronic medical records. Although such data are easy to obtain, there are some problems such as limited knowledge sources, low data purity and data redundancy. The existing feature selection methods are rarely combined with knowledge graph. Using more efficient data storage method of medical knowledge graph and combining with more authoritative medical knowledge of doctors to screen the hospitalization features of cardiovascular diseases can effectively reduce the average error of prediction costs.

To deal with these problems, this article proposes such a method: Open source knowledge graphs, such as Wikidata, OpenKG, etc. and open source knowledge base, such as medical language specifications defined by UMLS and doctors' prior medical knowledge are used to construct the medical concept knowledge graph (MCKG), the medical instance knowledge graph(MIKG) is completed using data of cardiovascular disease related cases from cooperative hospitals. Based on the constructed MIKG, we obtain a feature alternative scheme, and then combine with the actual needs of doctors and rules to generate the final feature selection scheme in the feature alternative scheme.

The knowledge graph data combined with doctor's interaction, the all features data, and the feature data selected by random search algorithm [20] are compared in three dimensions by combining the machine learning algorithm of the three schools, random forest [21], support vector machine [22], and line regression [23], the training set and the test set use a ratio of 70%: 30%, the evaluation standard is the average hospitalization cost error. The average error rate of the selected feature data combined with the random forest algorithm is reduced to 11.86%. This is a significant improvement over the feature data selected by other methods. Figure 1 is the core process of this paper:

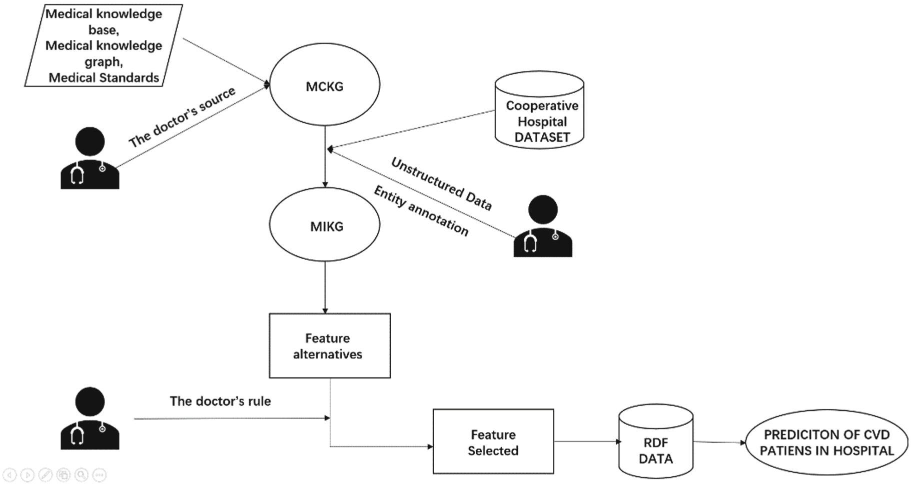

Fig. 1. Hospital cost prediction flow chart

MCKG's data sources mainly include public medical knowledge base, medical knowledge graph, and unified medical standards and specifications, which further guide the construction of MIKG. The data source of MIKG is mainly the structured data of the cooperative hospitals and the unstructured data entities marked by the doctors. The detailed process of MCKG and MIKG construction will be introduced in Sect. 3.1 and Sect. 3.2.

3.1 The Construction of MCKG

As we all know, natural language has the characteristics of polysemy and multiple synonyms, so there is a problem of concept confusion in the traditional medical knowledge graph. In this paper, open source knowledge graph such as Wikidata and OpenKG published on the Internet are combined with the prior medical knowledge of doctors in cooperative hospitals. The knowledge of doctors' dictionaries in unified standardized language provided by UMLS is imported into the conceptual knowledge graph. The knowledge graph is defined with entities as nodes and relationships and attributes as edges. Using ontology notation, that is a triplet(entity-relationship-entity) represents two associated nodes.

The MCKG constructed include the medical knowledge of Chinese and English knowledge as well as the medical specifications defined by UMLS, the main sources of data are from medical knowledge base, medical knowledge graph and doctor. MCKG includes 8,298,580 medical concepts from 116 word-lists and 51 entity words from cooperative hospital. The part of the MCKG constructed in this article is shown in Fig. 2, strictly in accordance with the UMLS definition specification, which is helpful to accurately understand the concepts and relationships between entities (only a part of the concept graph is intercepted in the figure).

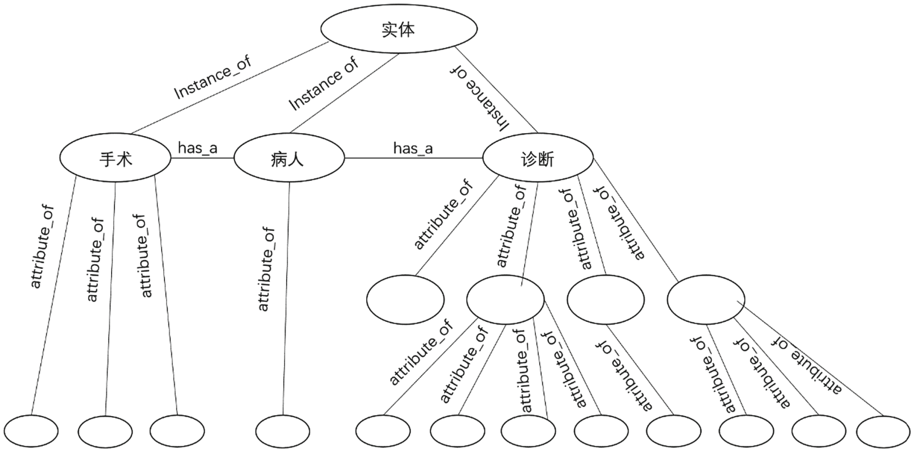

Fig. 2. Medical concept knowledge graph

It can be seen from the figure that the entity part has a surgery part, a patient part, and a diagnosis part. The attribute value part will be completed by the MIKG mentioned in the next section. The main role of the MCKG is mainly two points. First, it clarifies the relationship between the various parts of the graph, and second it guides the construction of MIKG.

3.2 The Construction of MIKG

MIKG is constructed under the guidance of the MCKG described in Sect. 3.1. The cardiovascular diseases data sets used in this paper are all from cooperative hospital. The data is divided into structured data and unstructured data.

The instantiation of the KG is mainly the process of knowledge extraction. The main process of this experiment generates a medical dictionary based on the latest cardiovascular disease diagnosis rules defined by experts, unstructured data (the medical record data of some patients) mainly adopts the method of entity annotation, defines relevant rules, extracts features related to hospitalization costs, and imports them into the MIKG. Here is an example of entity annotation of unstructured data in Table 1(Only part of a patient's case data is intercepted):

Table 1. Entity annotation sample

Annotation text	Entity tags			
Case history	Symptom	Examination	Disease	Medication
Patients were due to eight years ago, no significant incentives in paroxysmal retrosternal pain, for the stuffy pain, no radiation, for 20 to 30 min each time, postoperative oral "aspirin, wave force d" antiplatelet therapy In July 2009, coronary angiography showed that after the anterior descending branch and the first diagonal branch stenting, the intima of the stent was slightly enlarged, the wall of the anterior descending branch was irregular, and the distal segment was 90% narrow. After the operation, angina pectoris occurred several times, and the hospital treatment improved and discharged	paroxysmal retrosternal pain	irregular branch wall; coronary angiography;	angina pectoris;	stenting; aspirin;

The table shows that we divide the types of entity annotation into four categories: symptom entity, examination entity, disease entity and medication entity. The entity tags serve as the entity node of MIKG and they are imported into MIKG in the form of RDF triples.

The structured data of cardiovascular disease comes from the hospital database, including basic patient information, surgical information, diagnosis information and other information. The structured data is mapped according to the rules of relational data (ER)-mapping-RDF data. For example, If the table contains "cardiovascular diseases" and related hospital information, we can map it to an RDF triple. The goal of instantiation of MCKG is to extract the entities and relationships of cardiovascular diseases from textual data and structured data, then realize the visualization of MIKG and select features through interaction with doctors.

First, under the guidance of doctors, seven tables related to the prediction of hospitalization costs for cardiovascular diseases were extracted, as shown in Table 2:

Table 2. Related ER data

No	Table_Name	Description
1	PATIENT_VISIT	Patient's operation information
2	OPERATION	Patient's operation information
3	LAB_MASTER	Patient's test information
4	LAB_RESULT	Patient's test result
5	DIAG_TYPE	Diagnosis type
6	MASTER	Patient master index
7	ORDER	Doctor's advice information

We extract the entities of all patient records in these tables, namely patient entity, surgery entity, diagnosis entity, diagnosis result entity, diagnosis type entity, main index entity, medical order entity (the above-mentioned information related to patient's privacy has been desensitized):

1. Patient entity extraction: extract the ID and admission ID of each patient with cardiovascular disease from the patient ID (PATIENT_ID) and the patient's admission ID (VISIT_ID) as the attribute value of the patient's entity.
2. Surgery entity extraction: due to the different conditions of each patient and the different operations performed, different types of operations such as vascular exploration, coronary angiography, and coronary artery bypass grafting are extracted from the surgical entities of the patient as a subclass of surgical entities entity.
3. Diagnosis entity extraction: Each patient's examination number, examination date, and patient's basic information such as gender and age were extracted as the subclass entities of the diagnostic entity.
4. Diagnosis result entity extraction: The diagnosis results of each patient are necessarily different, and indicators such as WBC, NEUT%, RBC, etc. as well as the diagnosis result time are extracted as the subclass entities of the diagnosis result entity.
5. Diagnosis type entity extraction: Different patients have different types of diagnosis according to the needs of different types of cardiovascular diseases. Different diagnosis types such as vascular headache, carotid atherosclerosis, coronary atherosclerotic heart disease are extracted from the diagnosis entities as diagnosis type entity.
6. Main index entity extraction: The payment types of each patient, such as out-of-pocket, public expense, medical insurance, as well as entities such as place of birth and date of birth, are extracted as the subclass entities of the main index entity.
7. medical order entity extraction: Entities such as the medical examination performed by each patient, the drugs related to cardiovascular disease used, the corresponding dose, the starting time and the end time of the medication are extracted as the subclass entities of the medical order entity.

To sum up, the relationship between different entities is extracted by applying the MCKG to MIKG, for example: the relationship bet ween the patient entity and the patient entity is has_a, The relationship between the type of surgery and the surgical entity is attribute_of, The relationship between the type of diagnosis and diagnosis entity attribute_of and so on. Converting the cardiovascular disease data from the cooperative hospital into RDF data, the construction of the medical instance knowledge graph is shown in Fig. 3:

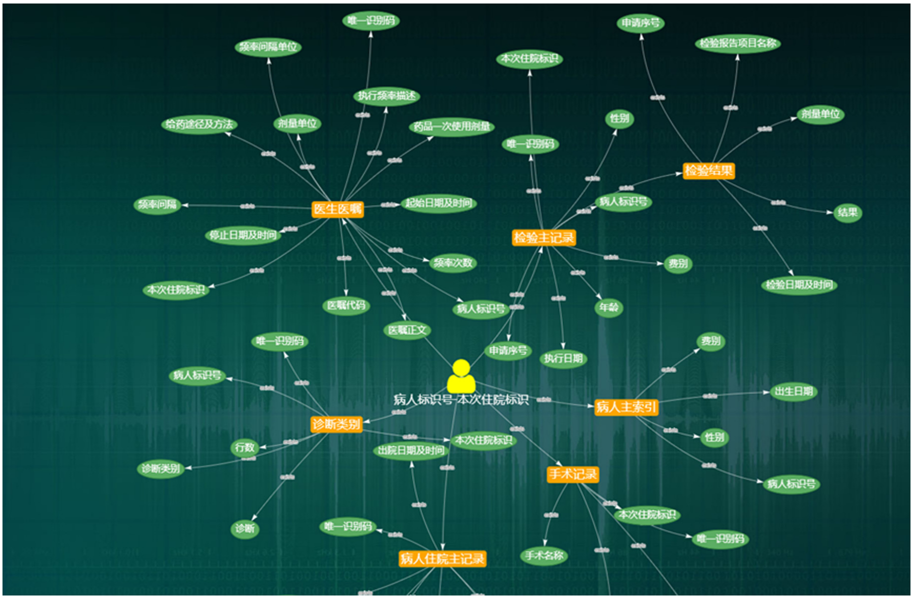

Fig. 3. Medical instance knowledge graph

The size of the constructed medical instance knowledge graph is 83.6 GB which contains 698946023 triples. Taking the patient entity as the center, different entity nodes (rectangular nodes in the figure) are connected and different nodes are connected to the corresponding instance nodes (elliptical nodes in the figure).

3.3 KG Feature Selection

Based on the already generated MCKG and MIKG, we have screened up to 189 features provided by the original database into 47 features as shown in Fig. 3. Combing the selected KG with the doctor's needs and regulations, according to the 1–2 steps closest to the patient's hospitalization information, the nine feature KG with the highest correlation with hospitalization costs are finally extracted as shown in Fig. 4:

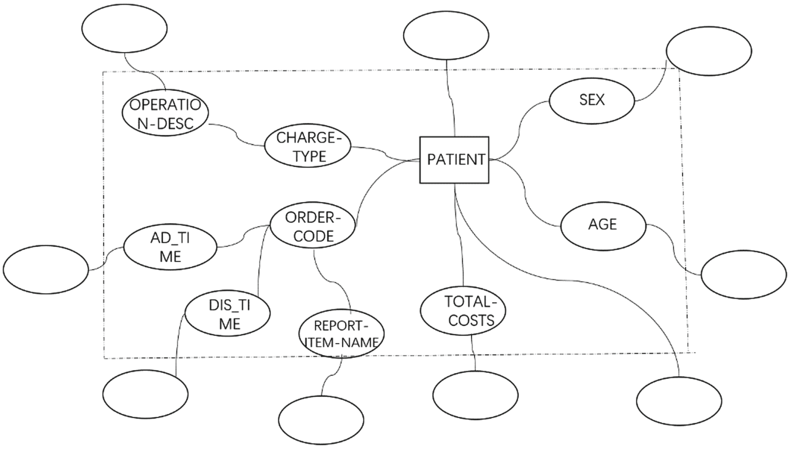

Fig. 4. Knowledge graph feature selection

The dotted frame in the figure is divided into alternative plans submitted to the doctor, who selects features based on his/her prior medical knowledge and clinical needs. ORDER_CODE is the doctor's advice code, AD_TIME is patient's admission time, DISCHARGE_TIME is patient's discharge time, CHARGE_TYPE is patient's type of payment, REPORT_ITEM_NAME is patient' examination items, SEX is patient's sex, Age is patient's age, OPERATION_DESC is patient's type of operation, TOTAL_COSTS is Total cost of patient hospitalization.

4 Experiments

In order to verify the effectiveness of the MIKG combined with the feature selection of doctors' interactive, we divide the experimental data into three groups:

The first group is DAF (data of all features), the second group is the data filtered by Random Searching Algorithm (RSA), and the third group is the data filtered by MIKG mentioned in Sect. 3 combined with the knowledge of doctors (KG-D).

The data set used in the experiment is RDF triplet data set, 10,000 of these triples are randomly selected, the training set and the test set use a ratio of 70%: 30%.

Experimental environment for this experiment: Processor: Inter® Core ™ i5-8265U CPU @ 1.8 Hz; RAM: 8.0 GB, operating system: WIN 10. This experiment uses Python3.7 software package.

The experiment uses machine learning algorithms: SVM, RF and LR. The average prediction error of hospitalization cost (Averr) is used as the evaluation index.

$$Averr = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\% \tag{1}$$

$\hat{y}_i$ represents the predicted value of hospitalization costs,y_i represents the actual value of hospitalization expenses, n is the total number of samples. The process of feature selection of RSA-RF [17] is shown in Fig. 5:

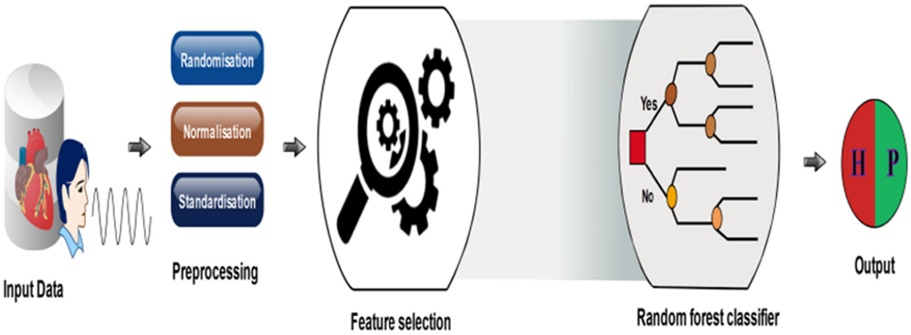

Fig. 5. RSA feature selection

The average prediction error rate of this experiment is shown in Table 3:

Table 3. Average prediction error for different feature selection methods

Feature selection method	SVM	LR	RF
DAF	36.34%	39.69%	35.12%
RSA	17.42%	21.17%	16.48%
KG-D	12.74%	14.55%	11.86%

It can be seen from the table that when all the features related to cardiovascular disease of patients are used to predict hospitalization cost, no matter which classifier is used, SVM, LR or RF, there is a high prediction error. When we use the feature data selected by the random search algorithm to predict the hospitalization cost, we can see that the prediction error is reduced. When we used MIKG in combination with the feature data of doctors' interactive selection for prediction, the prediction error of the classifier was significantly reduced, among which the best effect was achieved when RF was used, and the prediction cost error was reduced to 11.86%. This experiment proves the proposed the effectiveness of this method.

5 Conclusion and Future Work

We build MCKG using open source knowledge graphs such as Wikadata, OpenKG, etc. and open source knowledge base such as the medical language specifications defined by UMLS and the doctor's prior medical knowledge. Then we integrate the structured data in the database of the cooperative hospital and the unstructured data processed by doctors through entity annotation. We select features through the above-mentioned knowledge graph to get a feature alternative, and then combine the doctor's clinical needs and definition rules to get the final feature selection. Based on the feature selected by the above-mentioned method, we compare the corresponding RDF data with the data obtained by the features selected by the random search algorithm and the data corresponding to all the features related to hospitalization costs using SVM, RF, LR three different genres of machine learning algorithms to perform the hospitalization cost error prediction, experimental results prove that our feature selection method combined with random forest algorithm effectively reduces the prediction error of cardiovascular disease hospitalization costs.

Future work will focus on the application of MIKG to other data sets, as well as the selection of different deep learning models, and apply MIKG to a broader field of artificial intelligence.

References

1. Chinese cardiovascular disease report compilation group: Summary of Chinese cardiovascular disease report 2016. China Circul. J. **032**, 521–530 (2017)
2. Zhang, Y., Wang, S.N., Liu, Y.: Application of ARIMA model on predicting monthly hospital admissions and hospitalization expenses for respiratory diseases. China Health statistics **032**, 197–200 (2015)
3. Guyon, I.: An introduction to variable and feature selection. JMLR.org (2003)
4. Guo, K.W., Pan, H.L., Hou, A.: Classification algorithm based on feature selection and clustering. J. Jilin Univ. (Science Ed.) **056**, 395–398 (2018)
5. Ansong, S., Eteffa, Kalkidan F., Li, C., Sheng, M., Zhang, Y., Xing, C.: How to empower disease diagnosis in a medical education system using knowledge graph. In: Ni, W., Wang, X., Song, W., Li, Y. (eds.) WISA 2019. LNCS, vol. 11817, pp. 518–523. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30952-7_52
6. Sheng, M., Hu, Q., Zhang, Y., Xing, C., Zhang, T.: A data-intensive CDSS platform based on knowledge graph. In: Siuly, S., Lee, I., Huang, Z., Zhou, R., Wang, H., Xiang, Wei (eds.) HIS 2018. LNCS, vol. 11148, pp. 146–155. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01078-2_13
7. Xu, Z.L., He, L.R, Wang, Y.F.: Overview of knowledge graph technology. J. Electr. Sci. Technol. 589–606
8. Research on current situation and strategy of artificial intelligence-assisted diagnosis and treatment. Chinese Eng. Sci. 20, 1–128 (2018)
9. Sheng, M., et al.: CLMed: a cross-lingual knowledge graph framework for cardiovascular diseases. Web Inf. Syst. Appl. 512–517 (2019)
10. Uyar, K., lhan, A.: Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. Procedia Comput. Sci. **120**, 588–593 (2017)
11. Mohan, S., Thirumalai, C., Srivastava, G.: Effective heart disease prediction using hybrid machine learning techniques. IEEE Access 1 (2019)
12. Alarsan, F.I., Younes, M.: Analysis and classification of heart diseases using heartbeat features and machine learning algorithms (2019)
13. Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., Khan, J.A.: An automated diagnostic system for heart disease prediction based on $\chi 2$ statistical model and optimally configured deep neural network. IEEE Access **7**, 34938–34945 (2019)

14. Basciftci, F., Eldem, A.: Using reduced rule base with Expert System for the diagnosis of disease in hypertension. Med. Biol. Eng. Comput. **51**, 1287–1293 (2013)
15. Nahar, J., Imam, T., Tickle, K.S., Chen, Y.-P.P.: Computational intelligence for heart disease diagnosis: a medical knowledge driven approach. Expert Syst. Appl. **40**, 96–104 (2013)
16. Prakash, S., Sangeetha, K., Ramkumar, N.: An optimal criterion feature selection method for prediction and effective analysis of heart disease. Cluster Comput. **22**, 11957–11963 (2019)
17. Gokulnath, C.B., Shantharajah, S.P.: An optimized feature selection based on genetic approach and support vector machine for heart disease. Cluster Comput. **22**, 1–11 (2019)
18. Zhao, T.T., Yuan, Y.B., Wang, Y.J., Gao, J., He, P.: Heart disease classification based on feature fusion. In: 2017 International Conference on Machine Learning and Cybernetics (2017)
19. Sarah, P., Ira, K.S., Enzo, F., Matthew, L., Ricardo, G., Ben, G., Daniel, R.: Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease. medical image analysis S1361841518303554 (2018)
20. Javeed, A., Zhou, S., Yongjian, L., Qasim, I., Noor, A., Nour, R.: An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection. IEEE Access **7**, 180235–180243 (2019)
21. Singh, Y.K., Sinha, N., Singh, S.K.: Heart disease prediction system using random forest. In: International Conference on Advances in Computing and Data Sciences (2017)
22. Saunders, C., et al.: Support vector machine. Comput. Sci. **1**, 1–28 (2002)
23. Allison, L.: Coding Ockham's Razor. Linear Regression, pp. 103–111. Springer, Heidelberg (2018). https://doi.org/10.1007/978-3-319-76433-7