A medical insurance fraud detection model with knowledge graph and machine learning

Jie Li, Jiaying Liu, Xin Liu, Fang Yang, Yong Xu

Jie Li, Jiaying Liu, Xin Liu, Fang Yang, Yong Xu, "A medical insurance fraud detection model with knowledge graph and machine learning," Proc. SPIE 12260, International Conference on Computer Application and Information Security (ICCAIS 2021), 1226023 (24 May 2022); doi: 10.1117/12.2637418

SPIE.

Event: International Conference on Computer Application and Information Security (ICCAIS 2021), 2021, Wuhan, China

A medical insurance fraud detection model with knowledge graph and machine learning

Jie Li[a], Jiaying Liu[a], Xin Liu[a,*], Fang Yang[a], Yong Xu[b]
[a] School of Economics and Management, Hebei University of Technology, Tianjin, China; [b] School of Science, Hebei University of Technology, Tianjin, China

ABSTRACT

In recent years, cases of medical insurance fraud emerge in endlessly. We urgently need to develop an effective way to detect fraud. However, efficiently mining the heterogeneous medical text data is a complicated and tough assignment in fraud detection. Therefore, a medical insurance fraud detection model with knowledge graph and machine learning is proposed in this paper. Firstly, a knowledge graph with 53,164 nodes and 1,209,847 edges is built based on the medical insurance text data of 20,000 insured members. Secondly, representation learning and improved Label Propagation Algorithm (LPA) are used for feature engineering based on the knowledge graph. On this basis, combined with the expense data, the medical insurance fraud detection model is constructed by using Easy Ensemble and XGBoost. The experimental results show that the model proposed in this paper greatly improves the effect of medical insurance fraud detection. In addition, it is proved that text data plays an important role in medical insurance fraud detection.

Keywords: Medical insurance fraud detection, machine learning, knowledge graph, XGBoost

1. INTRODUCTION

In recent years, people pay more and more attention to health care, the number of medical insurance insured members is increasing dramatically. However, medical insurance fraud has become increasingly prominent. In 2020, a special campaign was launched to crack down on fraud and fraud of medical security funds, and a total of 22.311 billion yuan was recovered[1].

In June 2020, a document issued by the Chinese government pointed out that "Accelerating the construction of medical insurance standardization and informatization, establishing and improving the intelligent monitoring system of medical insurance, and strengthening the application of big data, constantly completing the medical knowledge base". It is imperative to promote the application of big data and information technology in medical insurance, establish medical knowledge base and implement intelligent monitoring of medical insurance.

In medical insurance fraud detection, numerical features such as expense features are mainly used. Medical text data, such as disease diagnosis, examination items, and treatment plans, also contains important information[2]. However, because it is difficult to process, medical text data is seldom used for medical insurance fraud detection. Even when used, it is processed mainly by one-hot encoding[3-5]. However, medical text data involves a wide variety of diseases, examination items, and so on, but each insured member's involvement is relatively few. The one-hot encoding may cause excessive characteristic dimension and serious data sparsity problems. For example, Lasaga et al.[6] constructed a 400-dimensional disease matrix and an 800-dimensional treatment matrix, but each disease only involved about 2-20 treatments, which caused serious data sparsity. Herland et al.[5] processed the original data with one-hot encoding, and the features were increased from 212 to 570, which is 268.87% of the original number of features. In addition to data sparsity and excessively high feature dimensions, the one-hot encoding will also cause the loss of semantic information. All of these make it challenging to achieve the effective mining of text data, and seriously affect the speed of the model training and the effect of medical insurance fraud detection.

Knowledge graph has significant advantages in text data processing due to the powerful semantic processing ability. The knowledge graph has been widely used in the medical field to process structured and unstructured medical data and build the medical knowledge base[7-10]. Knowledge graph has achieved very excellent effects in the processing of medical data. In medical insurance fraud, there is also a large amount of medical text data.

*ljrsch@126.com

International Conference on Computer Application and Information Security (ICCAIS 2021), edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12260, 1226023 · © The Authors. Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2637418

Efficiently mining the heterogeneous medical text data is a complicated and tough assignment in fraud detection. Aiming at that, we apply the knowledge graph to the processing of medical text data in medical insurance fraud detection, and propose a medical insurance fraud detection model with knowledge graph and machine learning. Firstly, a knowledge graph for medical insurance fraud detection is constructed based on medical text data of 20,000 insured members, which contains 53,164 nodes and 1,209,847 edges. Then, by using representation learning and improved LPA, feature engineering based on knowledge graph is carried out to make the network structure and semantic information in knowledge graph can be used for machine learning. Finally, combined with the expense data, medical insurance fraud detection model is constructed by using the EasyEnsemble and XGBoost. And the recall of 0.81, Accuracy of 0.83, and AUC of 0.82 are achieved. The model proposed in this paper can assist the medical insurance administration in making rapid and accurate judgments in the medical insurance reimbursement audit. And this paper can provide references for the processing of medical text data and the application of knowledge graph in medical insurance fraud detection.

2. MEDICAL INSURANCE FRAUD DETECTION MODEL FOR INSURED MEMBERS

This section proposes a medical insurance fraud detection model for insured members based on knowledge graph and machine learning. In section 2.1, a knowledge graph for medical insurance fraud detection is constructed based on the text data. In section 2.2, by using the representation learning and improved LPA, feature engineering is carried out based on the knowledge graph, which makes the information in it could be used in machine learning. In section 2.3, combined with the expense features, the EasyEnsemble and XGBoost are used to construct the medical insurance fraud detection model for insured members.

2.1. Construction of knowledge graph for medical insurance fraud detection

The knowledge graph for medical insurance fraud detection is constructed in this paper based on the actual medical insurance settlement data provided by the medical insurance department in a region of China. The data consists of three data sets. The first data set contains 1.83 million medical expense records. The second data set contains 6.53 million expense detail records. The third data set is the label of insured members reviewed and determined by medical insurance experts. This data covers 20,000 insured members, including 19,000 normal insured members and 1,000 fraudulent insured members. Each medical bill of the insured member is stored as a record in the medical expense data set, and expense details of the bills are stored in the expense detail data set. Two parts of the records can be connected by "SNID". There are 15 columns in medical expense records, including SNID, PID (ID of insured members), hospital, disease, declaration amount of drug expense, examination expense, treatment expense, operation expense, bed expense, medical material expense and other expense, approval amount of the bill, the amount paid by the individual, the amount paid by medical insurance fund and reimbursement time. There are 5 columns in expense detail records, including SNID, hospital, service (service items covered by medical insurance), unit price, and quantity. For privacy protection, the names of insured members and hospitals have been desensitized.

Knowledge graph has powerful capability of knowledge representation and reasoning[11]. At present, knowledge graph are actively used in the medical field to process medical data[12]. In this paper, the knowledge graph is applied to the processing of medical text data. The text data used in this paper mainly contain three types: hospital, disease, and service. Therefore, the knowledge graph for medical insurance fraud detection is built based on these three columns. The construction of the knowledge graph consists of three steps:

Firstly, entities, attributes, and relationships are extracted from medical text data using natural language processing technology. The knowledge graph is essentially the semantic network that reveals the relationship between entities, consisting of nodes and edges[13]. The entity, attribute, and relationship are the essential elements of the knowledge graph[14]. Nodes represent entities, and edges represent relationships between entities.

In this paper, four types of entities are extracted: insured member, hospital, disease, and service. Then, define the ID for each entity as the unique identifier. And the name of the disease and service entity is regarded as the attribute of the entity.

The knowledge graph constructed in this paper aims for the medical insurance fraud detection of insured members, so it is centered on the insured members. Therefore, three relationship types, visit, ill, and serve, are defined between the insured member and other entities, and the triples composed of these three relationships are insured member-visit-hospital, insured member-ill-disease, and insured member-serve-service.

Secondly, the data are organized into the triple, and entity table and relationship table are established. Triple is the general representation form of the knowledge graph, and its forms mainly include entity1-relation-entity2 and entity-attribute-attribute value. The entity table stores entities and their attributes; namely, the triples in the form of entity-attribute-attribute value are stored in the entity table. The relationship table stores the relationship between entities; namely, the triples in the form of entity1-relationship-entity2 are stored in the relationship table. Examples of the entity table are shown in Table 1. Examples of the relationship table are shown in Table 2.

Table 1. Examples of the entity table.

Examples of entity table of the insured member		**Examples of entity table of disease**		
pid:ID	LABEL	did:ID	disease-name	LABEL
20001503353	insured member	i1	Renal Anemia	disease
20000627033	insured member	i77	Osteoporosis	disease
20003573959	insured member	i648	Colon cancer	disease
Examples of entity table of hospital		Examples of entity table of service		
hid:ID	LABEL	sid:ID	service-name	LABEL
2	hospital	s60	Bailing Capsule	service
180	hospital	s228	Disposable syringe	service
1260	hospital	S369	Routine urianlysis	service

Table 2. Examples of the relationship table.

START_ID	**END_ID**	**TYPE**
20000000231	186	visit
20001889611	i1	ill
20000000231	s5691	serve

Finally, entities are denoted as nodes for the knowledge graph, relationships are denoted as edges that connect nodes, and Neo4j is used to construct the knowledge graph. The knowledge graph constructed in this paper contains 53,164 nodes and 1209,847 edges, and its scale is shown in Table 3. The visualization function of the Neo4j graph database is used to show a part of the network in the knowledge graph, as shown in Figure 1.

Table 3. The scale of knowledge graph.

Type	**Name**	**Number**	**Total**
Entity	Insured member	20000	53164
	Hospital	456	
	Disease	4479	
	Service	28229	
Relationship	Visit	44050	1209847
	Ill	139033	
	Serve	1026764	

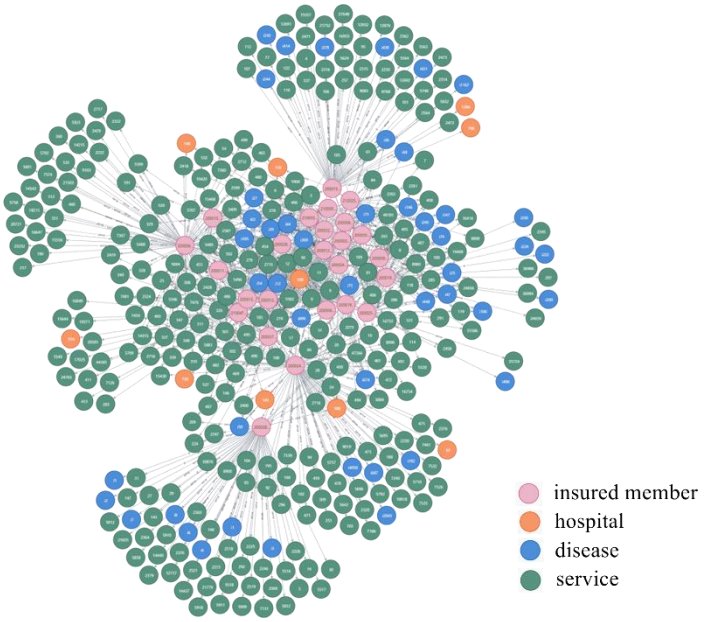


Figure 1. Part network in the knowledge graph.

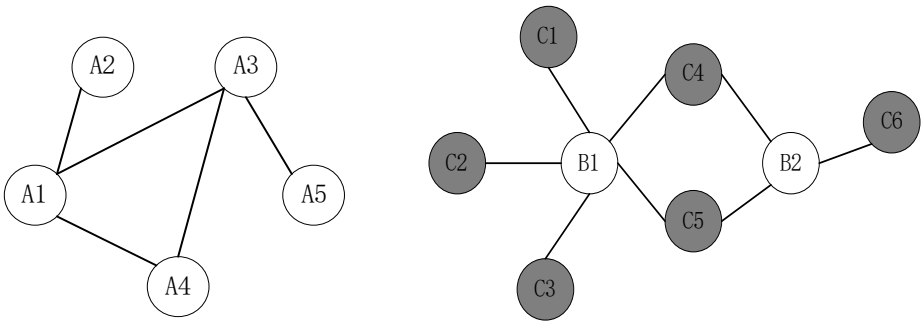
2.2. Feature engineering based on knowledge graph

The data form of knowledge graph is graph, which is generally difficult to be directly used in machine learning. In this paper, we use representation learning to vectorize the knowledge graph, so that the information in the knowledge graph can be used for machine learning.

Representation learning can realize the mapping of high-dimensional graph data to low-dimensional vector, enabling the data to be directly inputted into machine learning while preserving the structure and semantic information in the knowledge graph. Node2vec is an effective and extensible representation learning algorithm, which can reflect network characteristics and node neighbor characteristics. This method integrates two random walk methods of the depth-first walk and breadth-first walk. The sequence obtained by random walk is processed by the method of processing the word vector, and the vector representation of nodes is obtained. In a nutshell, the graph is processed by Node2vec, and a unit vector with dimensions of n for each node in the knowledge graph is obtained. In this paper, the Node2vec is used to obtain the node vector features of insured members by taking the knowledge graph as input and setting n as 30.

To further explore the key information that can be used for medical insurance fraud detection in the knowledge graph, we construct the fraud risk feature for insured members based on the improved LPA.

LPA is a semi-supervised learning algorithm based on the graph, and the basic idea is to predict the label of the unlabeled node from the label of the labeled node. Each node propagates the label to its neighbor nodes according to the similarity. LPA is often used in fraud detection, but it has some limitations. LPA is mainly used in graphs containing only one entity type, as shown in Figure 2a. For example, in telephone fraud detection[15], only one kind of entity is included in the graph. Each entity represents a telephone number, and the connection between entities indicates the call record between two numbers. In this case, LPA can be directly used to judge the labels of unknown numbers. However, in graphs containing multiple entity types, as shown in Figure 2b, LPA cannot be applied directly. Such as in the medical insurance fraud, assuming that the white circle represents the entities of the insured members (B1, B2), the grey circle represents other types of entities (C1-C6) such as hospital, disease, et al. Only the entities of the insured members have labels, other types of entities do not have labels. Different entities of the insured members are not directly connected but indirectly connected by other types of nodes (C4, C5). In this type of graph, LPA cannot be directly applied.



(a) (b)

Figure 2. (a) Graph of single type entity; (b) Graph of multiple types entity.

This paper improves the LPA and applies it to the graph containing multiple entities, and then constructs the fraud risk feature of the insured members. The fraud risk feature S_i of the insured member i with the unknown label is calculated according to the information of the insured members with the known label. The specific methods are as follows:

(1) The similarity between the insured member node i with the unknown label and all other insured member nodes with the known label are calculated. Based on the knowledge graph, the Jaccard similarity coefficient is used to calculate the similarity between the insured member nodes. Assuming that N_i represents the set of nodes directly connected to the insured member node i and N_j represents the set of nodes directly connected to the insured member node j, then the node similarity between the insured member node i and the insured member node j is:

$$Jaccard(N_i, N_j) = \frac{|N_i \cap N_j|}{|N_i \cap N_j|} \tag{1}$$

For example, if the node similarity between node B1 and node B2 in Figure 2b is calculated, find the set of nodes directly connected to node B1 is $N_{B1} = \{C1, C2, C3, C4, C5\}$, and the set of nodes directly connected to node B2 is $N_{B2} = \{C4, C5, C6\}$, so the intersection of two sets $N_{B1} \cap N_{B2} = \{C4, C5\}$, the union of two sets $N_{B1} \cup N_{B2} = \{C1, C2, C3, C4, C5, C6\}$, the similarity of node B1 and B2 is $Jaccard(N_{B1}, N_{B2}) = |N_{B1} \cap N_{B2}| / |N_{B1} \cap N_{B2}| = |\{C4, C5\}| / |\{C1, C2, C3, C4, C5, C6\}| = 2/6 = 1/3$.

(2) Node similarity is used as the weight in the information propagation of insured members with the known label to insured members with the unknown label, and the fraud risk feature of insured members with the unknown label is calculated. For the insured member node i with the unknown label, finding out k insured member nodes with the known label with the highest similarity is $j_m (m = 1, 2, ...k)$. The similarity between the node i and the node j_m is taken as the weight of the influence of the node j_m on the label of the node i. The larger the similarity with the node i, the larger the weight on its label.

Based on this, the fraud risk feature S_i of the insured members is constructed for the insured member i, as shown in equation (2). In this formula, y_{j_m} represents the label (0 or 1) of the insured member j_m. The higher the value of S_i, the larger the risk that the insured member i is a fraudulent insured member.

$$S_i = \frac{\sum_{m=1}^{k} Jaccard(N_i, N_{jm}) \cdot y_{jm}}{k} \tag{2}$$

In this part, representation learning is firstly used to transform the knowledge graph for medical insurance fraud detection into the 30-dimensional node vector features of the insured members, which can be directly used in the machine learning model while retaining the structure and semantic relations in the graph. Then, based on the improved LPA, the fraud risk feature of the insured members S_i is constructed, and the key information that can be used to detect the medical insurance fraud in the knowledge graph is further mined.

2.3. Construction of medical insurance fraud detection model

Aiming to detect more accurately and effectively in medical insurance fraud detection, we not only use the node vector features and the fraud risk feature S_i of insured members constructed based on the knowledge graph, but also construct the expense features which can greatly reflect the behavior of the insured members when building the medical insurance fraud detection model of insured members.

The data collected in this paper contains ten columns of expense data, which can be roughly divided into three categories. The first category is the declared amount of 7 types of expenses, including drug, examination, treatment, operation, bed, medical material, and other expenses. The second category is the approval amount of the bill. The third category is the payment amount of different payment subjects, including the amount paid by the individual and the amount paid by the medical insurance fund.

Each bill is a line of the record in the collected data, and an insured member may have many records. Therefore, the number of bills is counted for each insured member, and the sum, the mean, and standard deviation of each bill are respectively counted for ten columns of expense data. In addition, the proportions of each expense in the sum of this category are calculated respectively for the expenses in the first and third categories. All of these are expense features.

The node vector features of the insured members, the fraud risk feature S_i, and the expense features are integrated, and a total of 71 features of insured members are obtained. These features are integrated with the PID and label of the insured members to obtain the data set D, which is used for medical insurance fraud detection of insured members.

The stratified sampling method is adopted to divide the data set D into train set D_r and test set D_e in the ratio of 8:2 to ensure that the proportion of normal insured members and fraudulent insured members in the train set D_r and test set D_e are consistent with the data set D. Data set D contains 20,000 insured members, including 19,000 normal insured members and 1000 fraudulent insured members. Train set D_r contains 16,000 insured members, including 15,200 normal insured members and 800 fraudulent insured members. The test set D_e contains 4,000 insured members, including 3,800 normal insured members and 200 fraudulent insured members. The train set D_r is used to train the model, and the test set D_e is used to test the effect of the model.

In particular, when constructing the fraud risk feature S_i, the parameter k is set to 64. In the train set, the insured members' label is known. When constructing the fraud risk for the insured member i in the train set, the node i of the insured member is regarded as the node of the unknown label. And all other nodes of insured member in the train set are regarded as the node of the known label to calculate the fraud risk for the insured member i. In the test set, the label of the insured members is unknown. When constructing the fraud risk for the insured member i in the test set, the node i of the insured member is taken as the node of the unknown label. And all the nodes of insured member in the train set are taken as the nodes of the known label to calculate the fraud risk for the insured member i.

In model training, firstly, the data should be balanced. Medical insurance fraud detection is a typical data imbalance problem. In this paper, the ratio of normal insured members to fraudulent insured members is 19:1. The number of fraudulent insured members is far lower than the number of normal insured members. If the data is not balanced, the effect of the medical insurance fraud detection model will be seriously affected. This paper uses the idea of the EasyEnsemble method to balance the data. Assuming that the majority class samples set in the train set is D_+, the minority class samples set in the train set is D_- ($|D_-| \ll |D_+|$). The D_+ is randomly sampled, which a subset D_{+k} of the majority class samples with the same number as the minority class samples set D_- is sampled, and the balanced subset D_k is obtained by combining D_{+k} and D_-. The process is repeated nine times to obtain nine balanced subsets.

Secondly, nine sample subsets are taken as input respectively, and the machine learning algorithm is used to train the base model $M_k (k = 1, 2, ..., 9)$. XGBoost algorithm is developed based on GBDT, which adds a regularization item to the error function of GBDT to control the complexity of the model, and uses the Random Forest algorithm for reference to support column sampling of data. Compared with other algorithms, XGBoost has better Accuracy and faster running speed. Therefore, we select XGBoost to train the base model.

Finally, according to the voting mechanism (equation (3)), the nine basic models are integrated to obtain the medical insurance fraud detection model of insured members. The process of model construction is shown in Figure 3.

$$y=\begin{cases}0, \ if \ \ count(M_{k(y_k=0)})>count(M_{k(y_k=1)})\\ 1, \ if \ \ count(M_{k(y_k=0)})<count(M_{k(y_k=1)})\end{cases} \tag{3}$$

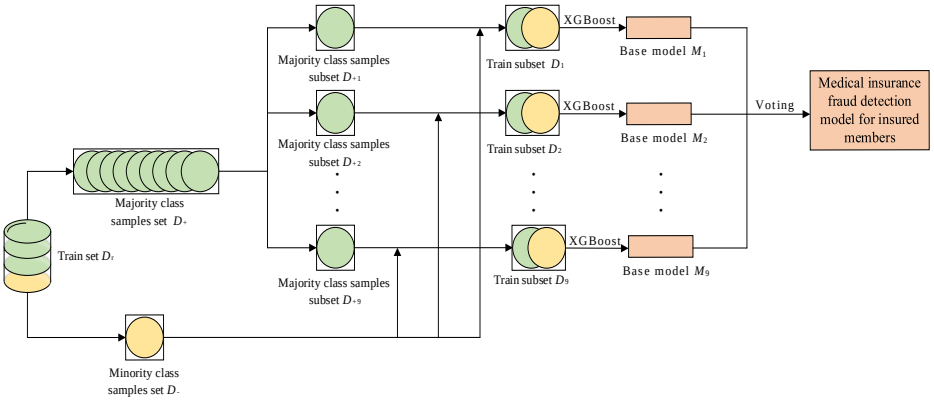


Figure 3. Flowchart of medical insurance fraud detection model.

3. MODEL EXPERIMENT VERIFICATION

Medical insurance fraud detection is a typical dichotomy problem, and its performance measurement standard mainly relies on the confusion matrix, as shown in Table 4. In the evaluation index of the model, three commonly used indexes, Recall, Accuracy, and AUC are selected. Recall is the recognition rate of fraudulent samples, Accuracy is the rate of correctly classified samples, and AUC is the area under the ROC curve. Recall and Accuracy are calculated as equations (4) and (5) based on the confusion matrix.

Table 4. Confusion matrix.

Real value	**Predicted value**	
	Fraud (1)	**Normal (0)**
Fraud (1)	TP	FN
Normal (0)	FP	TN

$$Recall=\frac{TP}{TP+FN} \tag{4}$$

$$Accuracy=\frac{TP+FN}{TP+FN+TN+FP} \tag{5}$$

The test set D_e is input into the medical insurance fraud detection model to evaluate the effect of the model. In order to ensure the stability of the model effect, ten experiments are conducted during the model test, and the average value of the ten experiments' results is taken as the final medical fraud detection results of the model. The results of ten experiments, the mean, and the standard deviation are shown in Table 5. The standard deviation of ten experiments is very small, which indicates that the model has a stable effect. The medical insurance fraud detection model has achieved the Recall of 0.8101, Accuracy of 0.8283, and AUC of 0.8197, and achieves a great detection effect.

In order to verify the application effect of the knowledge graph, the three kinds of features in the test set are taken as the model input, including the features constructed based on the knowledge graph, expense features, and all features. Ten experiments are conducted for each features set, and the average results of the ten experiments are shown in Table 6.

There is no doubt that the expense features that are widely used in the detection of medical insurance fraud have a good

effect. However, the features constructed based on the knowledge graph obviously also have an excellent effect on medical insurance fraud detection. The difference between these two results is slight, indicating that the features constructed based on the knowledge graph in this paper have a good effect on medical insurance fraud detection. In addition, it is proved that text data plays an important role in medical insurance fraud detection.

Table 5. Results of ten experiments.

	Recall	Accuracy	AUC
Exp.1	0.8068	0.8240	0.8159
Exp.2	0.8309	0.8243	0.8274
Exp.3	0.8116	0.8358	0.8243
Exp.4	0.8116	0.8315	0.8221
Exp.5	0.8068	0.8238	0.8157
Exp.6	0.7826	0.8353	0.8104
Exp.7	0.7826	0.8373	0.8114
Exp.8	0.8357	0.8183	0.8265
Exp.9	0.8019	0.8348	0.8192
Exp.10	0.8309	0.8185	0.8244
Average	0.8101	0.8283	0.8197
Std.	0.0169	0.0067	0.0056

The recall of 0.8101, Accuracy of 0.8283, and AUC of 0.8197 are achieved after the combination of the two parts of features, and the fraud detection effect is significantly improved. In summary, the application of the knowledge graph can effectively extract the key information that can be used for medical insurance fraud detection in the text data, realizing the deep mining of the text data. Besides, by combining text data with expense data, the medical insurance fraud detection model with knowledge graph and machine learning proposed in this paper achieves a better fraud detection effect.

Table 6. Effect comparison of different features.

Feature	Recall	Accuracy	AUC
Features based on knowledge graph	0.6309	0.7095	0.6724
Expense features	0.7145	0.7985	0.7588
All features	0.8101	0.8283	0.8197

Medical insurance fraud frequently occurs, which causes enormous losses to the medical insurance fund. Based on the research in this paper and the reality of medical insurance fraud, the following anti-fraud suggestions are proposed:

First, the medical insurance management departments should consider various aspects when auditing medical insurance reimbursement for insured members. In addition to the audit of the declaration amount and other information of the insured members, it should also carefully distinguish whether the diseases of the insured members are consistent with the treatment methods. And guarding against medical insurance fraud in the form of hanging bed in hospital, false treatment, and false invoice, so as to maintain the security of the medical insurance fund.

Second, increasing penalties to increase the cost of medical insurance fraud. Fundamentally, medical insurance fraud is to defraud medical insurance funds and obtain illegal benefits. However, the punishment for fraud is relatively low at the present stage, and the punishment methods such as refusing to pay the violation fee and fine are mostly adopted, which makes the cost of fraud relatively low. Those are not enough to deter criminals and keep medical insurance fraud cases

continue despite repeated prohibition. It is necessary to strengthen the punishment of fraud to effectively deter criminals and make them not dare to breed fraud.

Third, strengthen the publicity of medical insurance policies and regulations. Among the exposed cases of medical insurance fraud in China, there are many cases in which hospitals employ healthy older adults to cheat medical insurance funds. There are also cases of using other people's medical insurance cards to get medical insurance funds. Many people have a weak sense of law and fail to realize that such behavior has violated the law and may face criminal punishment. With the gradual expansion of the coverage of medical insurance, it is necessary to strengthen the publicity and education of medical insurance policies and regulations for citizens to make the public aware of the common forms of medical insurance fraud, the severe losses caused by medical insurance fraud and the possible punishment of medical insurance fraud, to eliminate the occurrence of medical insurance fraud fundamentally.

In order to ensure the safety of medical insurance funds, the audit system of medical insurance reimbursement should be improved. It is vital to effectively detect medical insurance fraud, recover the defrauded medical insurance fund, strengthen the punishment of medical insurance fraud to form a deterrent to criminals. There is great significance to establish and improve the supervision and management system of medical insurance from multiple perspectives to make it impossible for lawbreakers to take advantage, ensure the safety of the medical insurance system and promote the effective use of medical insurance funds.

4. CONCLUSIONS

Efficiently mining the heterogeneous medical text data is a complicated and tough assignment in fraud detection. Aiming at that, this paper applies the knowledge graph to medical text data and proposes a medical insurance fraud detection model with knowledge graph and machine learning. The experimental results show that the application of knowledge graph in medical insurance fraud has achieved a good effect. The medical insurance fraud detection model with knowledge graph and machine learning proposed in this paper can effectively judge the label of insured members. The experimental results prove the importance of text data in medical insurance fraud detection. This paper can provide the reference for the application of knowledge graph and the processing of text data in the medical insurance fraud, and the model constructed in this paper can assist the audit of medical insurance reimbursement.

Facing the future, the knowledge graph for medical insurance fraud detection needs to be further expanded and improved to ensure the comprehensiveness of the insured members' information. The mining of the knowledge graph needs to be further studied and can combine with graph theory and other related methods. All of these can lay the foundation for the widespread application of knowledge graph in this field, and it is helpful to further improve the effectiveness of medical insurance fraud detection.

ACKNOWLEDGMENTS

This research was funded by the National Natural Science Foundation of Hebei Province (G2019202350).

REFERENCES

[1] National Healthcare Security Administration, [Statistics on the Development of Medical Securityin 2020], http://www.nhsa.gov.cn/art/2021/3/8/art_7_4590.html, (2021).
[2] Chandola, V., Sukumar, S. R. and Schryver, J. C., "Knowledge discovery from massive healthcare claims data," Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, 1312-1320 (2013).
[3] Bauder, R. and Khoshgoftaar, T., "Medicare fraud detection using random forest with class imbalanced big data," 2018 IEEE International Conference on Information Reuse and Integration (IRI), IEEE, Salt Lake City, 80-87 (2018).
[4] Musal, R. M., "Two models to investigate Medicare fraud within unsupervised databases," Expert Systems with Applications, 37(12), 8628-8633 (2010).
[5] Herland, M., Khoshgoftaar, T. M. and Bauder, R. A., "Big data fraud detection using multiple medicare data sources," Journal of Big Data, 5(1), 1-21 (2018).

[6] Lasaga, D. and Santhana, P., “Deep learning to detect medical treatment fraud,” KDD 2017 Workshop on Anomaly Detection in Finance, PMLR Halifax, 114-120 (2017).
[7] Miao, F., Liu, H., Huang, Y., et al., “Construction of semantic-based traditional Chinese medicine prescription knowledge graph,” 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), IEEE, Chongqing, 1194-1198 (2018).
[8] Yu, T., Li, J., Yu, Q., et al.,” Knowledge graph for TCM health preservation: Design, construction, and applications,” Artificial Intelligence in Medicine, 77(3), 48-52 (2017).
[9] Shao, L. I. and Bo, Z., “Traditional Chinese medicine network pharmacology: Theory, methodology and application,” Chinese Journal of Natural Medicines, 11(2), 110-120 (2013).
[10] Zou, X., “A survey on application of knowledge graph,” Journal of Physics Conference Series, 1487(1), 012016 (2020).
[11] Shi, L., Li, S., Yang, X., et al., “Semantic health knowledge graph: Semantic integration of heterogeneous medical knowledge and services,” BioMed Research International, 1-12 (2017).
[12] Wu, T., Qi, G., Li, C., et al., “A survey of techniques for constructing Chinese knowledge graphs and their applications,” Sustainability, 10(9), 1-26 (2018).
[13] Nickel, M., Murphy, K., Tresp, V., et al., “A review of relational machine learning for knowledge graphs,” Proceedings of the IEEE, 104(1), 11-33 (2015).
[14] Wang, Y., Xiao, W., Tan, Z., et al., “Caps-OWKG: A capsule network model for open-world knowledge graph,” International Journal of Machine Learning and Cybernetics, 1-11 (2021).
[15] Peng, H. and You, M., “The health care fraud detection using the pharmacopoeia spectrum treeand neural network analytic contribution hierarchy process,” 2016 IEEE Trustcom/BigDataSE/ISPA, IEEE, Tianjin, 2006-2011 (2016).