

RESEARCH

Open Access



Health insurance fraud detection by using an attributed heterogeneous information network with a hierarchical attention mechanism

Jiangtao Lu¹, Kaibiao Lin^{1*}, Ruicong Chen¹, Min Lin², Xin Chen¹ and Ping Lu³

Abstract

Background With the rapid growth of healthcare services, health insurance fraud detection has become an important measure to ensure efficient use of public funds. Traditional fraud detection methods have tended to focus on the attributes of a single visit and have ignored the behavioural relationships of multiple visits by patients.

Methods We propose a health insurance fraud detection model based on a multilevel attention mechanism that we call MHAMFD. Specifically, we use an attributed heterogeneous information network (AHIN) to model different types of objects and their rich attributes and interactions in a healthcare scenario. MHAMFD selects appropriate neighbour nodes based on the behavioural relationships at different levels of a patient's visit. We also designed a hierarchical attention mechanism to aggregate complex semantic information from the interweaving of different levels of behavioural relationships of patients. This increases the feature representation of objects and makes the model interpretable by identifying the main factors of fraud.

Results Experimental results using real datasets showed that MHAMFD detected health insurance fraud with better accuracy than existing methods.

Conclusions Experiment suggests that the behavioral relationships between patients' multiple visits can also be of great help to detect health care fraud. Subsequent research fraud detection methods can also take into account the different behavioral relationships between patients.

Keywords Health insurance, Fraud detection, Heterogeneous graph, Graph neural network

Background

The popularisation of health insurance has provided many people with easy access to healthcare and medical protection for the public. In various countries, the rapid development of the medical services sector is inseparable from governmental support for healthcare. In the United States, healthcare generated about 3.5 trillion USD in 2017 [1], of which Medicare contributed about 20% or 702 billion USD [2]. However, health insurance fraud is also increasing, which poses a serious threat to the proper use of public funds. According to a survey by the

*Correspondence:

Kaibiao Lin
2010110706@txmut.edu.cn

¹ School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China

² School of Electromechanical and Information Engineering, Putian University, Putian 351100, China

³ School of Economic and Management, Xiamen University of Technology, Xiamen 361024, China



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Global Healthcare Anti-Fraud Network, approximately 260 billion USD is lost globally to health insurance fraud each year, which is equivalent to 6% of global healthcare spending [3]. According to the Coalition Against Insurance Fraud, the annual financial losses due to Medicare fraud in the United States is equivalent to 3–10% of healthcare expenditures [3], or 21–70 billion USD. A large number of medical claims are made every day and fraud can be committed in many ways, such as the illegal administration of drugs and consumables, purchase and resale of drugs and falsification of bills for medical services. Such fraud seriously harms the interests of users and service providers. In response, researchers have been developing methods to detect health insurance fraud. In China, 339 million USD was recovered from cases of health insurance fraud in 2020 alone [4]. In 2020, 1 USD = 6.58 RMB [5]. The main approach has been manual detection by government-organised experts, which is effective but requires a major expenditure of time and effort. It relies on sufficient a priori knowledge and it cannot detect incidents of health insurance fraud consistently and automatically. In addition, there are many types of health insurance fraud and methods relying on manual detection can have difficulty coping with complex and changing patterns. To address these limitations, another approach is to apply machine learning to finding incidents of health insurance fraud automatically. This approach finds outliers in the data, which represent incidents of fraud and uses data mining to identify special patterns of anomalous data that differ significantly from the rest of the data. Most methods based on machine learning extract statistical features from the user data such as the user's access trajectory, behaviour and cost. These methods then use such statistical features for prediction and classification by neural networks, random forests or other classifiers. Such methods rarely make full use of the interaction between users. However, many entities (i.e. objects) are involved in healthcare, such as patients, hospitals, departments and drugs. There are rich interactions between these objects that can be very important for detecting health insurance fraud. Some researchers have started to use graph embedding to incorporate user interactions. However, most graph embedding based models are black-box models and solving a specialised domain problem such as health insurance fraud detection often requires a certain degree of interpretability. In addition, previous approaches to health insurance fraud detection have tended to focus only on the characteristic attributes of a single visit. The behavioural attributes of multiple visits have been neglected. Such behavioural relationships can be modelled by a heterogeneous graph approach, which can also be used to find anomalies in the topology of the graph neural network. Graphs are

widely used to model complex relationships between objects in many fields [6], including computer vision [7], natural language processing, anomaly detection [8], academic network analysis and recommendation systems [9]. Figure 1(a) shows a general scenario of medical treatment, which has several important objects such as the patient, doctor, hospital, department and medicine. In addition to attribute information, such objects also possess rich interactive information, such as the patient going to the hospital on a certain date or drugs being prescribed to a patient by a particular department. Integrating the patient's attribute information requires expanding the traditional heterogeneous information network (HIN) into an attributed HIN (AHIN) so that the objects in the heterogeneous information network can possess characteristics. Figure 1(b) shows the architecture of the AHIN for the medical treatment scenario, which contains the objects and their interactions. However, there are still several challenges. First, if health insurance data are modelled by an AHIN, the complexity of the data will lead to a large number of useless nodes in the graph. A graph neural network works by representing a learning task as the propagation and aggregation of attributes among nodes. A large number of noisy nodes will inevitably affect the results. Second, fraudulent users need to be distinguished from innocent patients. The results of the detected need to be interpretable considering its potential application in a professional setting. Finally, distinguishing which behavioural relationship of a medical visit is important for analysis is another challenge. We propose a health insurance fraud detection model based on a multilevel attention mechanism that we call MHAMFD. The basic idea of MHAMFD is to enhance the user representation by considering interactions between objects through neighbour nodes obtained from multilevel behavioural relationships in an AHIN. Specifically, we build bridges between patient nodes through behavioural relationships for medical visit scenarios. The different behavioural relationships are used to construct a medical multi-relational graph network. Meanwhile, we propose a multi-level graph neural network which is used to model multiple behavioural relationships within the network for fraud detection. The model has several advantages: (1) Using different levels of behavioural relationships in a way to select appropriate neighbour nodes to form different isomorphic graphs. This can effectively reduce the graph structure and can integrate the semantic information interwoven with multiple behavioural relationships [10]. It is also possible to obtain a comprehensive fraud detection result. (2) Due to the effectiveness of attention mechanism in various machine learning tasks [11, 12], we design a hierarchical attention mechanism. The first level of intra-relationship aggregation (intra-relation AGG) is

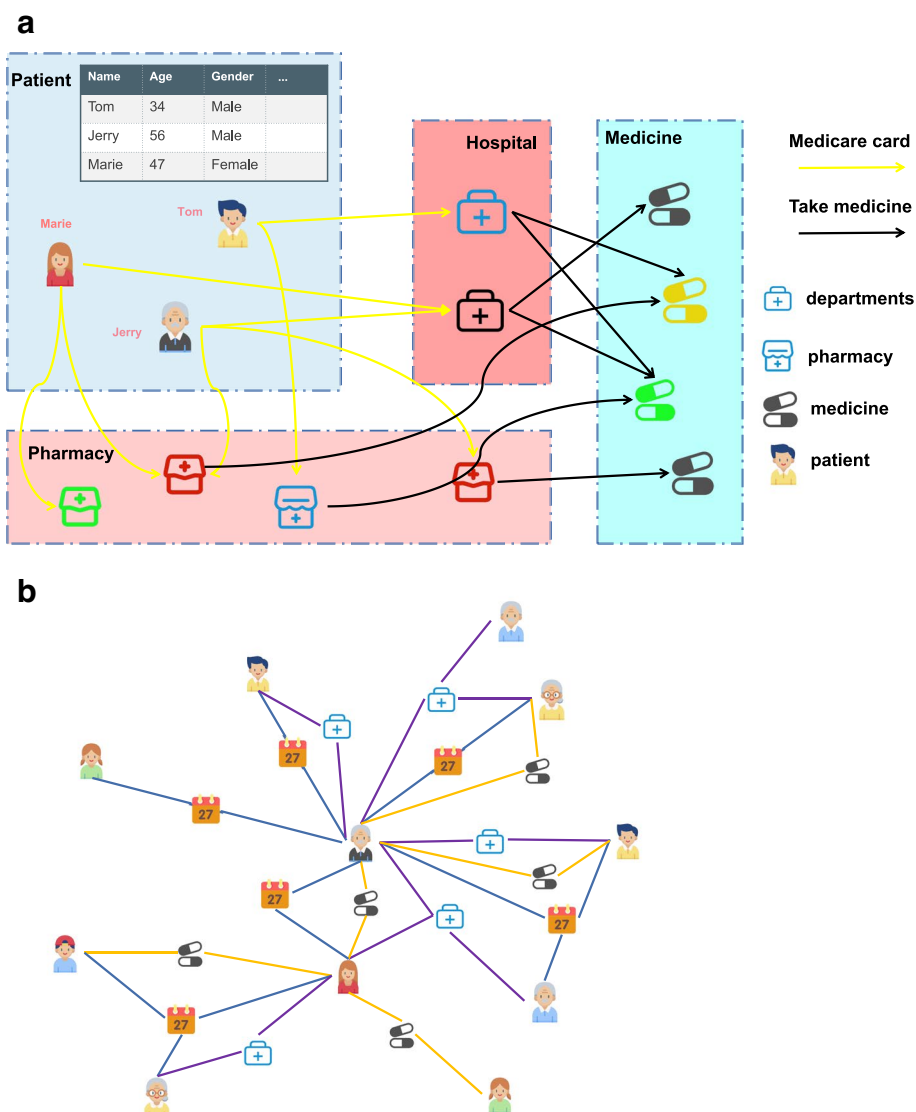


Fig. 1 AHIN of the scenario of medical payment service: **a** Scenario of medical payment service; **b** Heterogeneous Network

an attention mechanism that can effectively associate different attributes of different neighbours. The second level of inter-relationship aggregation (inter-relation AGG) can effectively correlate different behavioural relationships of patients. The third layer of hierarchical behavioural aggregation (hierar-relation AGG) can aggregate complex semantic information from the interweaving of different levels of behavioural relationships of patients. This effectively reflects which levels of behavioural relationships are more important for the final task. This attention-based model can improve interpretable results to provide more insights about the health care fraud detection task and outcomes. In summary, the contributions of our study can be summarised as follows:

- To the best of our knowledge, MHAMFD is the first to sample multiple behavioural relationships among neighbours to obtain a heterogeneous network of healthcare attributes. The interactions between different types of objects are learned by capturing structural information between objects in real healthcare scenarios.
- We designed a hierarchical attention mechanism for automatically learning the importance of different patient access behaviours in the healthcare domain for fraud detection.
- Extensive experiments with two real datasets showed that MHAMFD performed better at fraud detection than existing graph representation learning methods.

Related work

Health insurance fraud detection

Health insurance fraud has a significant impact on healthcare and daily life. Early research on fraud detection focused on a rule-based approach. Bauder and Khoshgoftaar [13] concluded that health insurance fraud has some obvious behavioural signals, which can be detected by defining some combination of rules. Because of the simplicity and interpretability of rule-based approaches, they have long been used for fraud detection [14]. However, designing such rules depends on a priori knowledge. In addition, there are so many ways to commit health insurance fraud that this approach has difficulty in dealing with complex and changing patterns. Recent methods have started using machine learning to find the intrinsic patterns in data indicating fraud automatically. Current detection methods can be classified into two general categories [15, 16]: unsupervised and supervised learning. Methods based on unsupervised learning depend on the data distribution. Machine learning is used to find outliers in the data, which indicate fraud and data mining is used to find abnormal data that differ significantly from general data for a specific mode. However, such methods are limited because they examine fraudulent behaviour according to a specific model. For example, Zhang and He [17] examined outliers in the treatment of diseases and the costs incurred to identify fraud related to falsified expenses. However, health insurance-related businesses are becoming increasingly detailed and fraudulent behaviour is becoming increasingly complex, changeable and concealed. New patterns of fraud continue to appear and anomaly detection algorithms based on fixed patterns lack immunity to these new patterns. Thus, fraud detection methods based on fixed patterns have difficulty with meeting current needs [18]. Predictive models based on supervised learning require large amounts of labelled data, but the data of actual scenarios are generally unlabelled and less data are available to protect patient privacy. Bauder and Khoshgoftaar [19] used random forest to classify unbalanced data. Pandey et al. [20] proposed using rule-based scoring systems, logistic regression models and decision trees, all of which strongly rely on a large amount of training data.

Graph-based learning

Our proposed model is related to graph-based methods. Network embedding is an effective method for modelling graph structures. Representation learning is performed by mapping the network nodes into a low-dimensional vector space. Then, low-dimensional dense vectors are used to represent any node in the network, which can be flexibly applied to different data mining tasks. Many early studies focused on representation learning with

homogeneous networks, where existing deep models were combined with network features to learn feature representations of nodes or edges. DeepWalk [21] combines random walk and the skip-gram model to learn network node representation. LINE [22] adds a second-order similarity based on first-order neighbour similarity to learn strongly distinguished node representation for large scale sparse networks. SDNE [23] uses a deep autoencoder to extract nonlinear characteristics of a network structure. However, homogeneous network modelling often extracts only part of the information of an actual interactive system or does not distinguish the heterogeneity of objects and their relationships, which results in irreversible information loss [24, 25]. In contrast, heterogeneous network modelling provides two benefits. First, a heterogeneous network is an effective tool for fusing information: not only different types of objects and their interactions but also information from heterogeneous data sources [26]. Second, the coexistence of multiple types of objects and relationships in heterogeneous networks contains rich structural and semantic information, which provides a precise and interpretable new way of discovering hidden patterns. Further methods have been proposed that use meta-paths [27] to explore the network topology and node characteristics. Some models guide the selection of neighbours by defining multiple meta-paths in a heterogeneous graph. However, such models only consider the importance of nodes and meta-paths and ignore the structural information of different levels of relationships intertwined in heterogeneous graphs. This can lead to underutilisation of structural information in the graph. In our study, one of our aims was to use the structural information obtained from the hierarchy of different behavioural relationships to learn the implicit interactions among objects in the healthcare services domain to improve the fraud detection accuracy.

Preliminaries

A HIN is an information network that contains multiple types of objects and multiple types of links [6]. Because actual patient data have information on certain attributes, the HIN can be extended into an AHIN to integrate these attributes. Here we define several terms necessary for understanding the explanation of our proposed model.

Heterogeneous graph

A heterogeneous graph is a special information network that can be denoted as a graph $G = \{V, \varepsilon, X\}$ comprising the set of nodes V , set of links and attribute information matrix $X \in \mathbb{R}^{|V| \times k}$. A heterogeneous graph is associated with a node type mapping function $\emptyset : V \rightarrow A$ and link

type mapping function $\varphi : \varepsilon \rightarrow R$, where A and R denote the sets of predefined object types and link types, respectively. Normally, a network can be considered a heterogeneous graph when the number of node types $|A|$ and link types $|R|$ meet the condition $|A| + |R| > 2$. Otherwise, it is a homogeneous graph.

Multiple behavioural relationship paths

The behavioural relationship paths in the medical domain can be used to capture rich semantic information for the heterogeneous graph. A behavioural relationship path is an abstract sequence of node types connected by link types that is denoted in the form of $A_1 \xrightarrow{M_1 R_1} A_2 \xrightarrow{M_2 R_2} A_3 \xrightarrow{M_3 R_3} \dots \xrightarrow{M_l R_l} A_{l+1}$, which can be abbreviated as $A_1 A_2 A_3 \dots A_{l+1}$. It describes a composite relationship $MR = M_1 R_1 \circ M_2 R_2 \circ \dots \circ M_l R_l$ between node types A_1 and A_l , where \circ denotes the composition operator of relationships. Depending on the number of edges connecting two adjacent patient nodes, the behavioural relationship can be classified as single-level or multilevel.

Neighbours according to multiple behavioural relationships

For a given user u in a HIN with a given attribute, the neighbours of the behavioural relationship paths are defined as the aggregate neighbour set for a given behavioural relationship path of the user u in the AHIN.

Multi-relational medical graph

A heterogeneous medical graph is a HIN extracted from health insurance data. To prevent the loss of heterogeneous information between different types of nodes and to reduce the size of the graph, we can map this to multi-relationship medical graph that only keeps the patient nodes. This converts the heterogeneous graph to an isomorphic graph while preserving the rich interaction information.

Node selection

The different levels of impurities in different behavioural relationships in a multi-relationship graph will affect the embedding results. There may be cases where the neighbours of a patient node are connected by different numbers of edges. To analyse which behavioural relationship or composite behavioural relationship has a greater impact on identifying health insurance fraud, we can first sample different behavioural relationships and then find neighbour nodes connected with the same behavioural relationship. Figure 2 shows the process of decomposing and reconstructing a heterogeneous graph. As an example, Phase b in Fig. 2 shows that the p2, p4 and p5 nodes are connected to the patient node p1 through the single-level behavioural relationship PDP. Only the p2 node is connected by the multi-level behavioural relationship PDTMP.

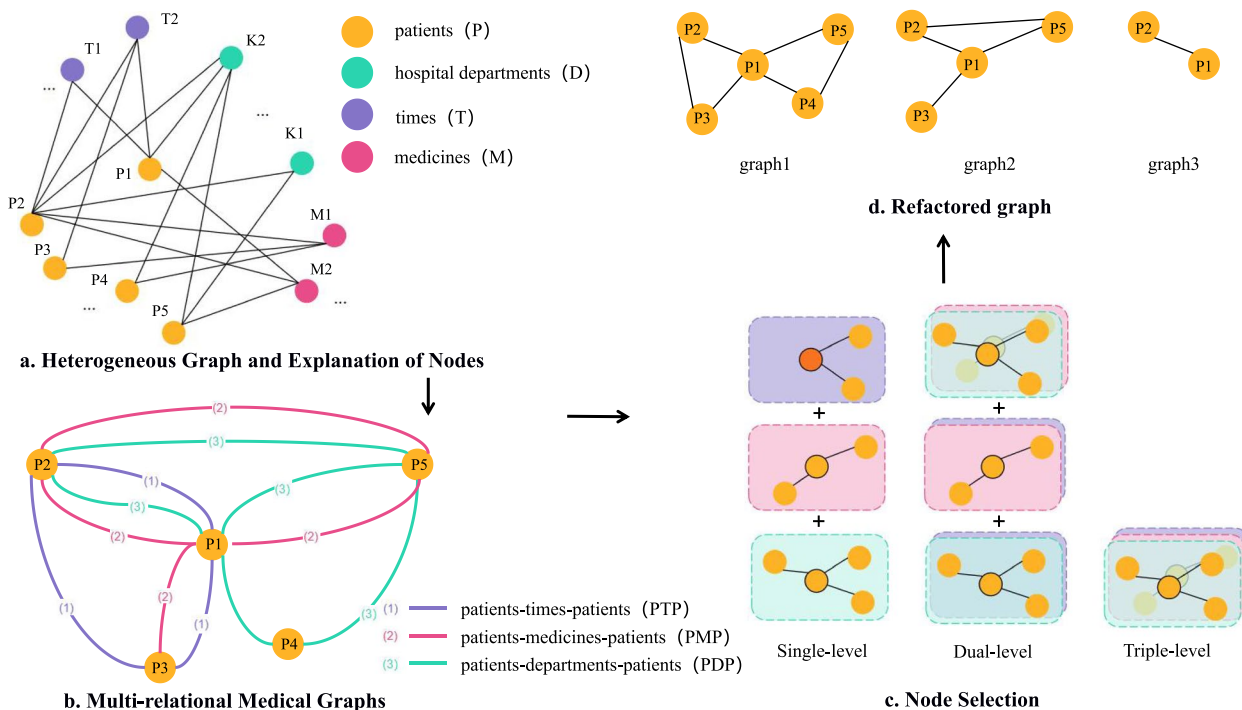


Fig. 2 Process of decomposing and reconstructing heterogeneous graphs

Health insurance fraud detection problem with an AHIN

The health insurance fraud detection problem can be defined by modelling different objects and their interactions in a real medical treatment scenario as the AHIN $G = \{V, \varepsilon, X\}$. In our experiments, the patient node set was a subset of the node set denoted as $U \subset V$. In the dataset, each patient $u \in U$ had the label $Y_u \in \{0, 1\}$ to indicate whether the patient was fraudulent. For a given AHIN $G = \{V, \varepsilon, X\}$ and training set $D = \{(u, Y_u)\}$, the ultimate goal is to predict the probability that a patient in the test set is fraudulent.

Methods

Theory

Here, we present the proposed model MHAMFD, which uses a hierarchical attention mechanism for fraudulent user detection. Figure 3 shows the general architecture of MHAMFD. The basic concept is to enhance the user’s representation by making full use of interactions with neighbour nodes based on multilevel behavioural relationships in an AHIN.

Observations of real data

Intuitively, fraudulent users of health insurance tend to cluster closely through different types of interactions. For example, in the AHIN shown in Fig. 1(a), fraudulent users tend to make medical visits to specific departments or use health insurance cards on a large scale at a uniform point in time. To aggregate the diverse behavioural

relationships of fraudulent users, we conducted experiments on a real dataset. First, we collected path-based neighbours for each user according to two different paths: PDP (i.e. the user visited the same department) and PDTMP (i.e. the user visited the same department on the same day, and both users picked up the same medication). For each path, we counted the number of neighbours who are users and divided all users into groups based on the number of fraudulent neighbours. The probability of health insurance fraud was calculated for each group. Figure 4 shows the percentage boost in fraud rate for the user groups with and without fraudulent neighbours for both paths. The different paths produced different percentage boosts. This confirmed to us that different behavioural relationships have different levels of importance to users, which can be captured by an attention mechanism.

Intra-relational aggregation

To better guide the embedding of a multi-relational graph neural network, we propose aggregation based on different levels of behavioural relationships. The overall aggregation process of MHAMFD is divided into three parts depending on the type of relationship: intra, inter and hierarchical. For the intra-relational aggregation process, the selected neighbours of a node play different roles in node embedding based on the behavioural relationship. In other words, each neighbour node has a different importance. For example, in the neighbour

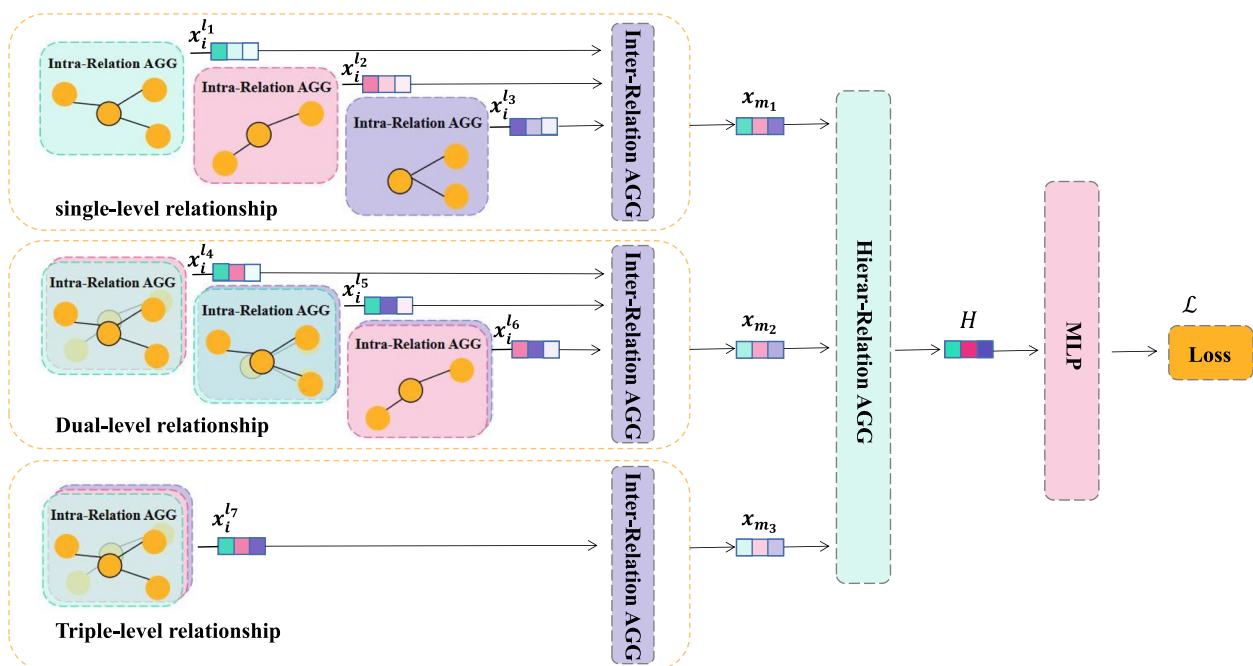


Fig. 3 Architecture of the MHAMFD model

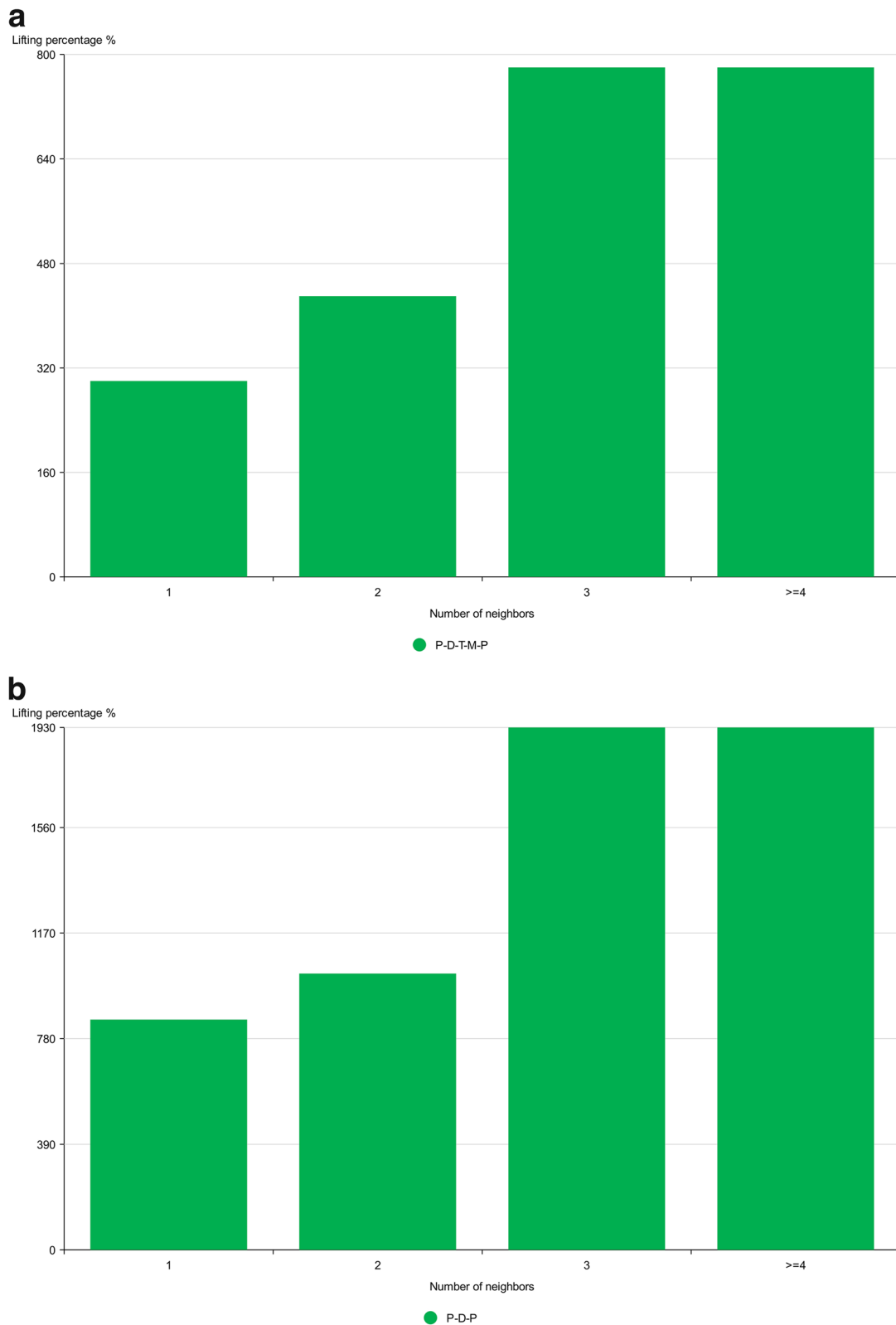


Fig. 4 Lifting percentages of fraud rate in users with different amount of Medicare fraudster neighbours against users without any Medicare fraudster neighbour in two behaviour relationships: **a** Behavioural relationship PDTMP; **b** Behavioural relationship PDP

set based on the behavioural relationship PDP, the feature similarity varies between the target node and each neighbour node. This affects the feature representation of the target node that can be learned from the neighbour nodes. Therefore, we use the self-attention mechanism in the intra-relational aggregation process. We aggregate the embedding representations of meaningful neighbours, which forms node embedding representations that are specific to a certain behavioural relationship. Concretely, we use the self-attention mechanism to learn the importance of each neighbouring node to the target node. Corresponding weights are assigned to different neighbour nodes. Given a pair of nodes (i, j) connected by the behavioural relationship l , we can define the importance of the neighbour node j to the target node i as E_{ij}^l . The importance of the node pair (i, j) based on the relationship l can be expressed as follows,

$$E_{ij}^l = Att_{intra}[x_i, x_j; l] \tag{1}$$

where x_i and x_j respectively represent the embeddings of nodes i and j , and l represents the behavioural relationship connecting the two nodes. Att_{intra} represents the deep neural network that implements the attention mechanism for intra-relational aggregation. For a given behaviour relationship l , all node pairs share Att_{intra} . This is because the neighbours of all target nodes are selected by the same behavioural relationship. However, because of the asymmetry of the heterogeneous graph, E_{ij}^l is asymmetric. This means that the importance of node i to node j differs from the importance of node j to node i . The importance of a node depends on its characteristics. When the target node changes, E_{ij}^l also changes because the neighbour nodes of the target node based on the behavioural relationship l have changed. This can be expressed as $j \in N_i^l$ and $i \in N_j^l$, where N_i^l represents the neighbour of node i based on the behavioural relationship l . Afterwards, normalisation is performed with the *Softmax* function. This facilitates the embedding of aggregated neighbours, which is expressed by,

$$\alpha_{ij}^l = softmax_j(E_{ij}^l) = \frac{\exp(LeakyReLU(E_{ij}^l))}{\sum_{p \in N_i^l} \exp(LeakyReLU(E_{ip}^l))} \tag{2}$$

The attention mechanism in intra-relation aggregation is a single-layer feedforward neural network. Here, we need to add LeakyReLU function for nonlinear activation and we set the slope to 0.2. Calculating the importance of neighboring nodes using LeakyReLU, we can pay more attention to neighboring nodes that are more positively related to the target node. The weight coefficients of (i, j) depend on their characteristics. The weight coefficient α_{ij}^l

is also asymmetric, which means that the contributions of nodes i and j to each other also differ. This is because the target node changes, which changes the neighbour domain. Nodes i and j have different neighbours, so the normalisation term (denominator) differs. After the importance of all neighbouring nodes for the target node i is obtained, we can aggregate the attribute features of all neighbours connected by the behavioural relationship l with the corresponding coefficients,

$$x_i^l = \sigma \left(\sum_{j \in N_i^l} \alpha_{ij}^l \cdot x_j \right) \tag{3}$$

where x_i^l is the embedding representation learned by node i for the behavioural relationship l and σ is the activation function. Because heterogeneous graphs have scale-free characteristics, the variance of the graph data is relatively large. To reduce the variance, we can use multiple attention heads to make the training process more stable. Specifically, we repeat the attention mechanism k times and we concatenate the embeddings learned each time,

$$x_i^l = \parallel_{k=1}^K \sigma \left(\sum_{j \in N_i^l} \alpha_{ij}^{lk} \cdot x_j^k \right) \tag{4}$$

Given the feature vector x and the set of behavioral relationships $\{l_1, l_2, \dots, l_m\}$, the Intra-relational attention of MHAMFD generates m behavior-specific vector representations of the target node, which are denoted as $\{x_{l_1}, x_{l_2}, \dots, x_{l_m}\}$.

Inter-relational aggregation

Node embeddings specific to a certain behavioural relationship can only reflect node information one way. For a more comprehensive representation of node embeddings, information about these behavioural relationships needs to be fused. For health insurance fraud detection, each behavioural relationship has a different impact on node embedding. The behavioural relationships connecting target nodes in the heterogeneous graph have different meanings. For example, the set of patient nodes selected by the behavioural relationship PDP differs from the set of patient nodes selected by PTP and the embeddings learned by both are not the same. These have different degrees of importance for determining whether the target patient is fraudulent. Therefore, we use an attention mechanism in the inter-relational aggregation process to automatically learn the weights of different behavioural relationships and aggregate them. This is done by nonlinear transformation of the node vectors that have undergone intra-relational aggregation and then averaging them. The importance of each behavioural relationship is given by,

$$\omega_{l_i} = \frac{1}{|V|} \sum_{i \in V} q_0^T \cdot \tanh(W_0 \cdot x_i^l + b_0) \tag{5}$$

where ω_{l_i} is the importance of a behavioural relationship, q_0 is the attention vector, W_0 is the weight matrix and b_0 is the bias vector. For a meaningful comparison of the importance of different behavioural relationships, all of the above parameters are shared. Then, normalisation is performed by the *Softmax* function. The weight coefficient can be obtained as follows,

$$\beta_{l_i} = \text{softmax}(\omega_{l_i}) = \frac{\exp(\omega_{l_i})}{\sum_{i=1}^l \exp(\omega_{l_i})} \tag{6}$$

After the attention weights are obtained for each behavioural relationship, the attention weights of each behavioural relationship and the embedding representations between relationships are weighted and summed,

$$x_m = \sum_{i=1}^l \beta_{l_i} \cdot x_{l_i} \tag{7}$$

In summary, for a given embedding of nodes specific to a certain behavioural relationship, the contribution of each behavioural relationship can be expressed by Eq. (7). Att_{inter} represents the deep neural network that performs the attention mechanism. In summary, given a group behavior-specific vector embedding as input, the contribution of each behavioral relationship can be denoted as follows,

$$(\beta_{l_1}, \beta_{l_2}, \dots, \beta_{l_m}) = Att_{inter}(x_{l_1}, x_{l_2}, \dots, x_{l_m}) \tag{8}$$

After getting the importance of behavioral relationships, the behavioral relationships of each level are aggregated and generated n level-specific vector embedding, denoted as $\{x_{m_1}, x_{m_2}, \dots, x_{m_n}\}$.

Hierarchical relational aggregation

We perform a clipped reconstruction of the nodes and edges in the multi-relationship graph for health insurance fraud detection. Finally, isomorphic graphs are obtained with only single rows representing single-level, two-level and three-level behavioural relationships. After the intra-relational and inter-relational aggregations, we have obtained graph embedding representations of the target nodes based on three levels of behavioural relationships. Our observations of a real health insurance dataset indicated that the different levels of behavioural relationships are relevant for fraud detection. For more comprehensive node embedding representation, the information of the

different levels of behavioural relationships needed to be fused. Therefore, we designed a relational aggregation mechanism that captures the semantic information from the hierarchy of the behavioural relationships. For a given set of embeddings, if the degree of influence of different levels of behavioural relationships on the final task is added, then this can be represented as follows,

$$(\gamma_{m_1}, \gamma_{m_2}, \dots, \gamma_{m_n}) = Att_{hierar}(x_{m_1}, x_{m_2}, \dots, x_{m_n}) \tag{9}$$

$$\lambda_{m_i} = \frac{1}{|N|} \sum_{i \in N} q_1^T \cdot \tanh(W_1 \cdot x_{m_i} + b_1) \tag{10}$$

$$\gamma_{m_i} = \text{softmax}(\lambda_{m_i}) = \frac{\exp(\lambda_{m_i})}{\sum_{i=1}^n \exp(\lambda_{m_i})} \tag{11}$$

where λ_{m_i} is the level of importance of a behaviour, q_1 is the attention vector, W_1 is the weight matrix and b_1 is the bias vector. Specifically, a larger λ_{m_i} means that the behaviour at level m_i is more important for the final task. Finally, we need to aggregate the information contained at each level to obtain the final node embedding H ,

$$H = \sum_{i=1}^n \gamma_{m_i} \cdot x_{m_i} \tag{12}$$

We can apply the final embedding H to different downstream tasks. We can use cross-entropy as the loss function and optimise the model weights by backpropagation of the minimisation function. The cross-entropy is expressed as,

$$\mathcal{L} = - \sum_{s \in Y_S} Y_S \ln(W \cdot H^s) \tag{13}$$

where y_S is the set of node indices with labels, H^s and Y_S are the embeddings of labelled nodes and the corresponding labels and W is the classifier. We can use the hierarchical attention mechanism to aggregate information at each level and obtain meaningful node embeddings.

Experimental

We conducted experiments to evaluate the proposed model with real datasets, compare it against some baseline methods and use visualisation methods for a more intuitive presentation of the results. Here, we briefly describe the basic experimental setup. We performed ablation experiments to compare the effects of different modules. Finally, we analysed the effect of the hierarchical attention mechanism on the final task.

Datasets

We used two real datasets from a certain municipal health insurance bureau in China in 2018: Medical-1 and Medical-2. The fraud samples in the datasets were the same: patients detected because of abnormal kidney disease, repeated prescriptions, prescription of senile dementia drugs in their 80s and simultaneous outpatient hospitalisation. The main difference was that Medical-1 contained balanced samples with a ratio of positive to negative samples of 1:2. Medical-2 contained unbalanced samples with a ratio of positive to negative samples of about 1:70. Table 1 presents the detailed information on the two datasets.

Baselines

We compared the performance of our proposed MHAMFD against multiple advanced methods, including those based on graph structures and graph neural networks. The following baseline models were considered:

- Metapath2vec [28]: A heterogeneous graph embedding method that combines random walk guided by a meta-path with a skip-gram model. We tested different meta-paths to obtain the best performance.
- GCN [29]: A variant of a graph neural network designed for isomorphic graphs. It is a semi-supervised graph convolutional network. We tested all paths to obtain the best performance.
- GAT [30]: A variant of a graph neural network designed for isomorphic graphs. It is a semi-supervised neural network with an attention mechanism. We tested all paths to obtain the best performance.

- HAN [31]: A heterogeneous graph neural network for graph embedding. This model learns node embeddings specific to a meta-path from different isomorphic graphs and uses a two-level attention mechanism to aggregate them into a vector embedding that represents each node in the network.
- MHAMFD : The proposed model. It is a semi-supervised graph neural network that uses a nested aggregation mechanism to simultaneously learn the importance of each object in a heterogeneous graph.
- MHAMFD_{hierar}: A variant of MHAMFD that removes the hierarchical attention mechanism and assigns the same weight to each level of behavioural relationships.
- MHAMFD_{single}: A variant of MHAMFD that only uses single-level behavioural relationships for node selection.
- MHAMFD_{multi}: A variant of MHAMFD that use two- and three-level behavioural relationships for node selection.

Parameter settings

For Metapath2vec, we conducted a random walk for the patient nodes in the order of patient → medicine → patient → date → patient → hospital department → patient. Each patient node was randomly sampled 20 times and the number of negative samples was set to 5. For the semi-supervised graph neural networks (i.e. GCN, GAT and HAN), we used the same training set, validation set and test set for each to ensure fairness. For a fair comparison, we set the embedding dimension

Table 1 Dataset used in the experiment

Dataset	Node	Positive samples	Negative samples	Behavior-Relationship
Medical-1	Patient(P):440	152	288	PDP
	Time(T):351			PTP
	Medicine(M):2328			PMP
	Department(D):708			PDTP
				PDMP
				PTMP
				PDTMP
Medical-2	Patient(P):10647	152	10495	PDP
	Time(T):364			PTP
	Medicine(M):4718			PMP
	Department(D):2751			PDTP
				PDMP
				PTMP
				PDTMP

to 64 for all of the above algorithms. For MHAMFD, we used the Adam optimisation algorithm and set the learning rate to 0.005, number of attention heads to 8. We set the dimensions of the attention vectors and to 128. In order to cope with the small training set, regularization is widely used in the model to prevent over-fitting. In training, we adopt L2 regularization and set the regularization parameter λ to 0.001. We introduce L2 regularization to make some restrictions. When the weight parameters are updated, the absolute values of the weight parameters are continuously reduced to prevent over-fitting. Meanwhile, we set the dropout parameter to 0.6, which makes the data in each batch inconsistent, so it can be simply regarded as many different models for training, so as to get more robust weights, achieve multi-model fusion, improve the generalization of models and reduce the over-fitting rate of models. Also, we set an early stop parameter patience to 50. If the loss of the model does not decrease and the accuracy does not improve in 50 consecutive training cycles, the training will be terminated early. In this way, the problems of too long training and over-fitting can be prevented.

Results and Discussion

Node classification

To evaluate the effectiveness of MHAMFD, we started with node classification. First, we conducted experiments with Medical-1 to compare the performances of different models. We divide the dataset into three subsets: training, validation and test. The F-score (f1) and accuracy (acc) were used as validation metrics. We used end-to-end training for all models. A multilayer perceptron was connected to the end of a model, which was then optimised by using the cross-entropy loss function. Table 2 presents the node classification results. We can see that the performance of our proposed MHAMFD model is better than the baseline. Our MHAMFD model has statistical significance in all three evaluation indexes (two-tailed t test, $\alpha = 0.01, P < 0.01$). The overall effectiveness of Metapath2vec was quite different from that of the other models. This means that only using structural information for a small number of samples did not result

in learning better embedding representations. GAT, GCN and HAN all used meta-paths to find groups of identical behavioural trajectories. In contrast, MHAMFD mined the same behavioural trajectories to obtain different levels of behavioural relationships. Differences between patient nodes were detected by mining the deep semantic information of heterogeneous graphs through multilevel behavioural relationships. On average, f1 was 6% higher for MHAMFD than for the other models. This indicates that single-level relational paths are not enough to explore the structure of heterogeneous graphs and that the structural information of different levels of relationships needs to be mined.

Anomaly detection

Next, the above models were evaluated in terms of anomaly detection. We used Medical-2 for the experiment. We used the weighted cross-entropy loss function for training and obtained the embedding representation of each model with different training sets. Table 3 presents the anomaly detection results. Unlike for Medical-1, which had balanced samples, in this experiment the effectiveness of GCN was greatly reduced. This is because GCN treats all neighbours equally during aggregation, so the large proportion of non-fraudulent patients introduced a large amount of noise. The representations learned by GCN did not cope well with the task of anomaly detection. Overall, Metapath2vec was more effective with a large number of samples with rich structural information than with a small number of samples. Metapath2vec includes some structural information of the network but ignores the characteristic information of nodes, so the performance was generally mediocre. Both MHAMFD and HAN have an attention mechanism that allows nodes to distinguish the importance of neighbours. However, HAN cannot extract different levels of complex behavioural relationships, at least within the medical domain. These different levels of behavioural relationships have a significant impact on node embedding representation. MHAMFD considers more complex semantic information that it obtains from different levels of behavioural relationships and it aggregates more complex

Table 2 Classification effectiveness with balanced sample nodes

Train:Val:Test	Metrics	Metapath2Vec	GAT	GCN	HAN	MHAMFD _{hierar}	MHAMFD
1:1:3	f1	0.5595	0.7979	0.7914	0.791	0.8197	0.8694
	acc	0.7196	0.8523	0.8523	0.8598	0.8611	0.8961
2:1:2	f1	0.5964	0.7273	0.7547	0.7455	0.8354	0.8566
	acc	0.7386	0.8295	0.8523	0.8409	0.8678	0.8905
3:1:1	f1	0.6349	0.72	0.7772	0.7407	0.8123	0.8439
	acc	0.7386	0.8409	0.8371	0.8409	0.8423	0.8757

Table 3 Anomaly detection effectiveness with balanced sample nodes

Train:Val:Test	1:1:3			2:1:2			3:1:1		
	F1	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision
Metapath2vec	0.795	0.7191	0.8888	0.8037	0.7166	0.9148	0.7384	0.6857	0.8
GCN	0.6067	0.6175	0.6296	0.5737	0.5908	0.5894	0.6036	0.6304	0.6121
HAN	0.7658	0.7762	0.8051	0.7543	0.7695	0.796	0.7896	0.8025	0.8138
MHAMFD _{hierar}	0.8014	0.8131	0.9022	0.8095	0.8658	0.9194	0.8662	0.8346	0.9143
MHAMFD _{single}	0.7611	0.7917	0.8908	0.7979	0.8391	0.8948	0.7942	0.8417	0.9021
MHAMFD _{multi}	0.7981	0.8014	0.8989	0.8048	0.8490	0.8748	0.7816	0.8392	0.8846
MHAMFD	0.8361	0.8764	0.9194	0.8679	0.8813	0.9386	0.8806	0.8692	0.9435

information from neighbours. MHAMFD considers compound semantic relationships from interweaving different levels of behavioural relationships, which improves the quality of the selected neighbours of the target node. The results showed that MHAMFD performed the best with the different training set regardless of the metric (i.e. F1, Recall and Precision).

Model analysis

Some previous graph embedding models have been based on the assumption that the learning performance increases with the contribution of the path to node embedding. In a graph constructed for health insurance fraud detection, two patient nodes may be connected by more than a single-level behavioural relationship. For nodes that are intertwined and connected by multilevel behavioural relationships, the semantic relationships are often more complex. We argue that different levels of behavioural relationships have different degrees of influence on the final embedding of patient nodes. Figure 5 shows the results with the hierarchical attention mechanism, the dual-level behavior relationship has the greatest contribution, while the triple-level behavior relationship has the worst contribution. This shows that the neighbor nodes obtained from the dual-level behavior relationship are more similar to the target nodes and make greater contributions to the final task. The triple-level behavior relationship has the least influence, because the number of neighbor nodes connected by the target node through the triple-level relationship is far less than that of the single-level relationship and the dual-level relationship. The target node gets less information from neighboring nodes, which leads to the influence of the triple-level behavior relationship is not as good as that of the other two levels. The specific attention results are shown in Table 4. In the results of the attention value for behavior relationship, we can see that behavioral relationship PDTP has the greatest influence on node embedding in the dual-level behavior relationship. This means that

the characteristics of patients who have seen a doctor in the same department on the same day are similar to each other and highly correlated. The table shows that the degree of attention may reflect the interpretability of the model. This can be used to select neighbour nodes according to the appropriate hierarchy of behavioural relationships.

Ablation experiment

To validate each component of our proposed model, we performed further experiments with different variants of MHAMFD (see the 'Baselines' section). Figure 6 compares the performances of the different variants. The horizontal axis shows the variants and the vertical axis represents the performance. The performance was evaluated according to the following metrics: f1, recall and precision. The results are presented for 20%, 40% and 60% of the training set of Medical-2. Between MHAMFD_{single} and MHAMFD_{multi}, the latter performed slightly better. This indicates that the multilevel behavioural relationship path contains more complex and accurate information about the nodes and confirms the accuracy of selecting neighbour nodes with different levels of behavioural relationships. The results for MHAMFD_{hierar}, MHAMFD_{single} and MHAMFD_{multi} showed that the model performance was significantly improved by aggregating different levels of behavioural relationships simultaneously. Finally, the results for MHAMFD and MHAMFD_{hierar} showed that different levels of behavioural relationships had different impacts on the final task and that an attention mechanism was needed to learn these impacts and obtain better results.

Visualisation

For a more intuitive comparison, we visualised the learned embeddings in two-dimensional space. This allows the node distribution to be represented in a low-dimensional space. We used t-distributed stochastic neighbour embedding (t-SNE) to visualise patient nodes

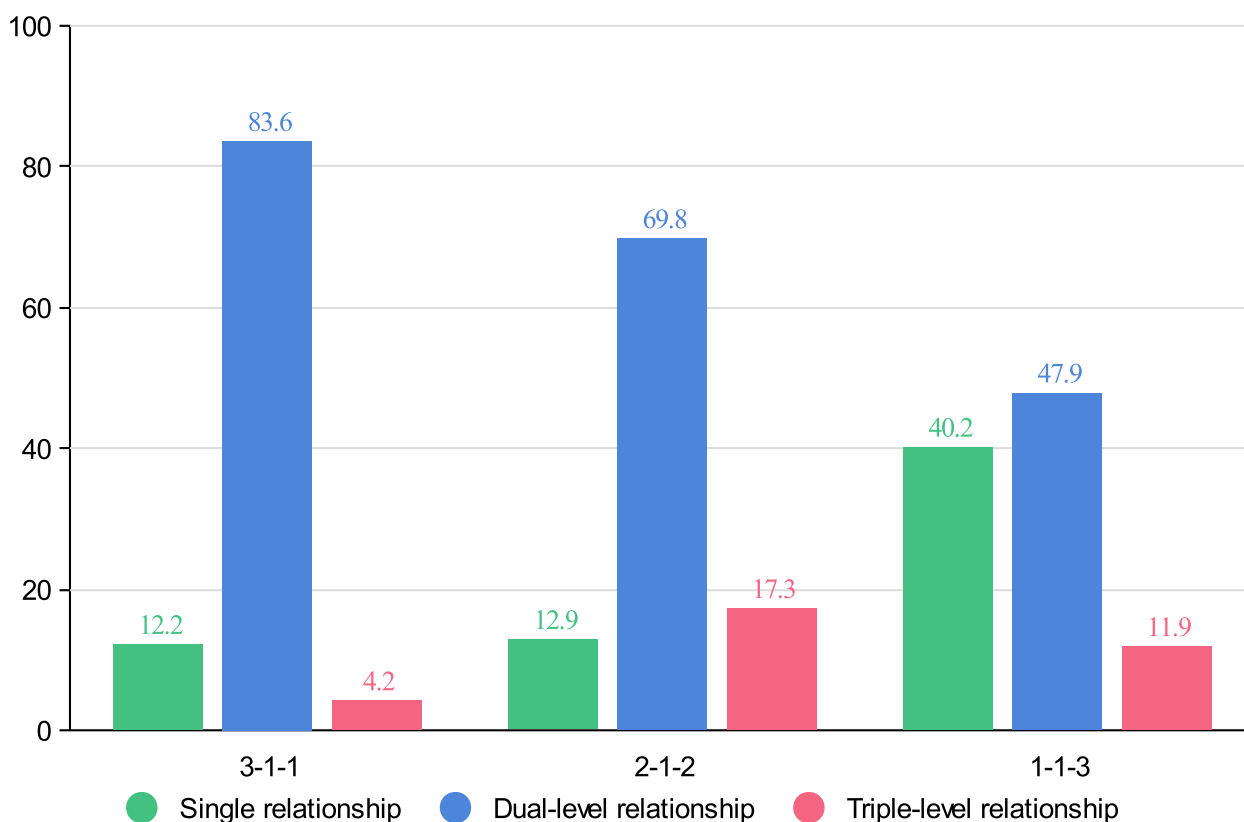


Fig. 5 Attention values (%) for different levels of behavioural relationships

Table 4 Attention result in Medical-1 datasets

Levels of behavioural relationship	Single relationship			Dual-level relationship			Triple-level relationship
Attention value for levels of behavioural relationship	0.129			0.698			0.173
Behavioural relationship	PDP	PTP	PMP	PDTP	PDMP	PTMP	PDTMP
Attention value for behavioural relationship	0.475	0.346	0.179	0.619	0.302	0.079	1

embedded in Medical-2 for each model. For convenience of observation, we took all of the positive samples and randomly selected 500 nodes from the negative samples. Figure 7 shows the results. Metapath2vec performed poorly and could not distinguish the node distribution. GAT and GCN could mostly distinguish between positive and negative samples, but there was no sharp boundary and a considerable number of positive and negative samples were mixed together. HAN and MHAMFD_{hierar} perform better than GAT and GCN. This is because HAN used multiple meta-paths and thus could obtain more semantic information. Meanwhile, MHAMFD_{hierar} not only used single-level behavioural relationships to sample neighbour nodes but also added multilevel behavioural relationships to select appropriate neighbour

nodes. Overall, MHAMFD performed the best. It clearly distinguished between different classes and there was a clear boundary between positive and negative samples. These results further illustrate the effectiveness of adding different levels of aggregation modules.

Hyper-parameter analysis

We further studied the sensitivities of several key hyper-parameters by varying them in different scales. We first tested the effect of the dimension of the final embedding Z, and the result is shown in Fig. 8(a). We set the dimensional range of the final embedding Z to $Z \in (16, 32, 64, 128, 256, 512)$. From Fig. 8(a), we observed that the performance of the model improves when the value of Z is gradually increased from 16 to 128.

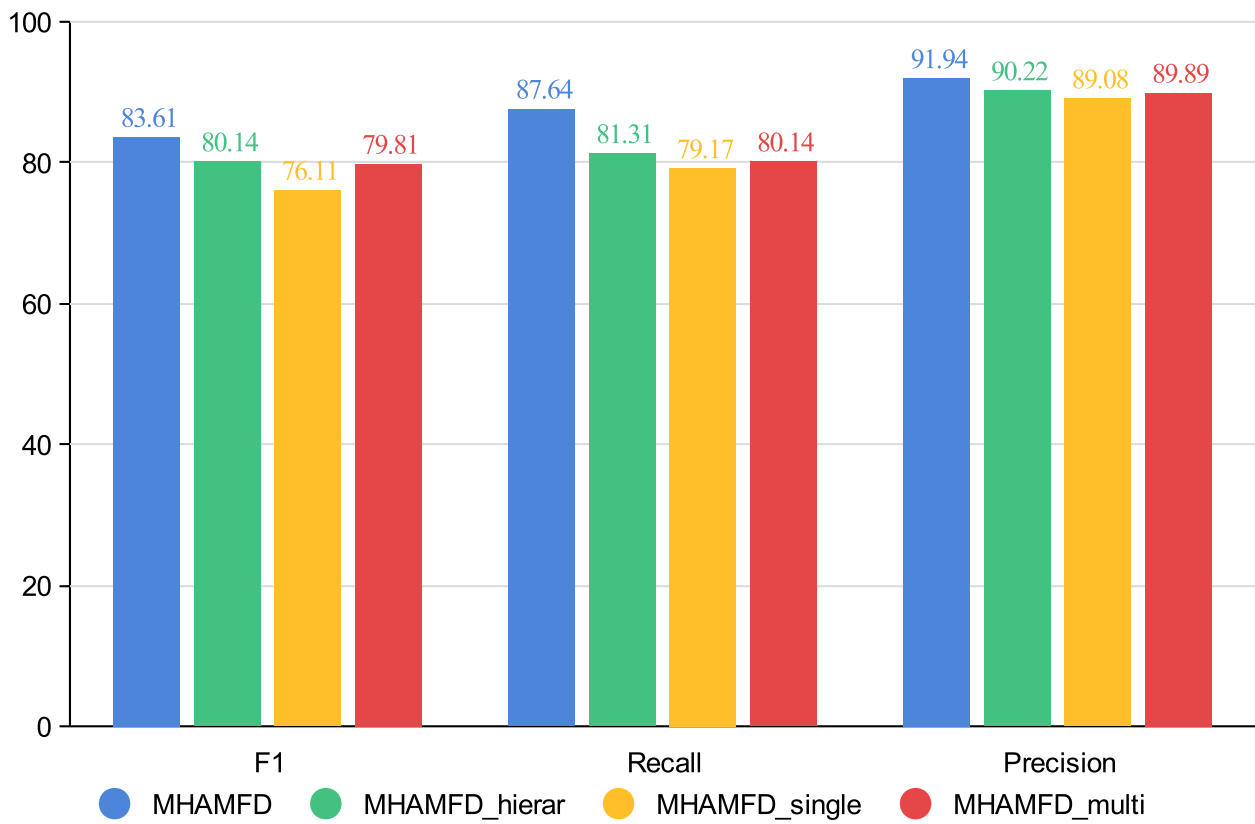


Fig. 6 Experimental results (%) with different variants of MHAMFD

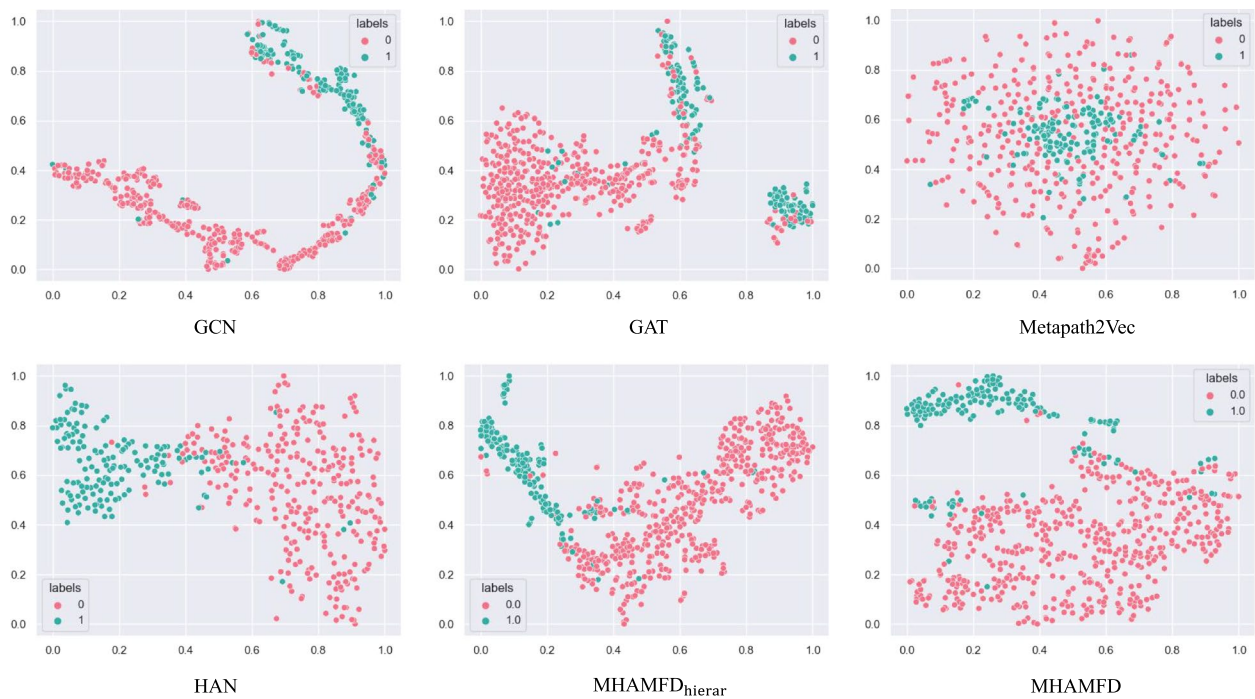


Fig. 7 Visualisation of the learned node embeddings with the Medical-2 dataset

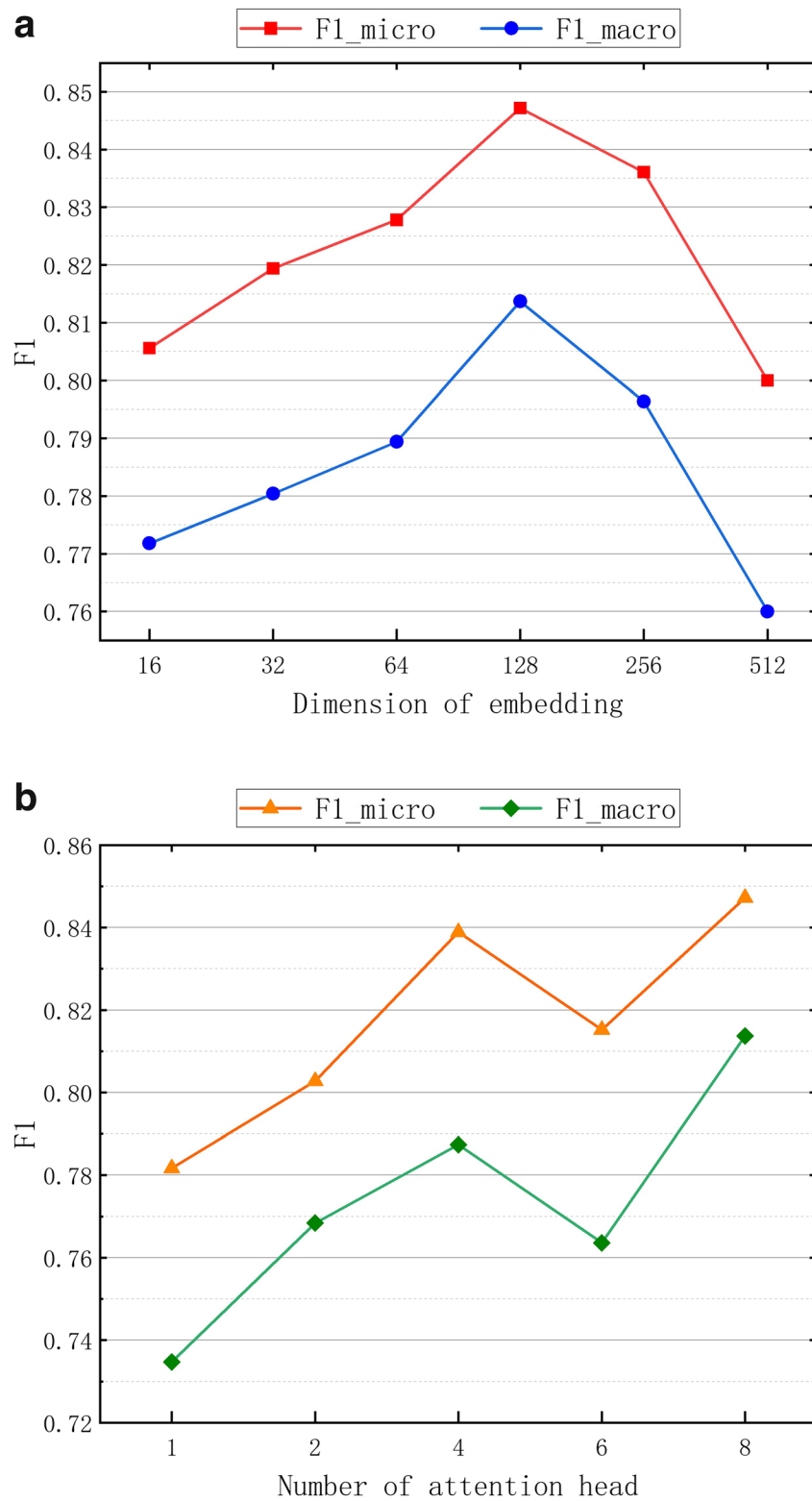


Fig. 8 Hyper-parameter sensitivity analysis of MHAMFD: **a** Dimension of the final embedding Z ; **b** Number of attention head K

However, when the value of Z is increased further from 128, the performance of the model starts to decrease. This is because MHAMFD needs a suitable dimension to encode the behavioral relationship information. And as the dimension increases further, it may introduce additional redundancy. We explored and analyzed the performance of MHAMFD with various number of attention head. We set the value of K to $K \in (1, 2, 4, 6, 8)$ and the result is shown in Fig. 8(b). We observed that the performance of MHAMFD is basically improved as the attention head increases. Meanwhile, we also find that multihead attention can make the training process more stable.

Conclusion

In this study, we considered the problem of health insurance fraud detection. We used a real healthcare dataset to explore different behavioural relationships of visits by patients. The impact of interactions between different objects in a healthcare scenario on fraud detection was analysed. These interactions were captured by an AHIN in a model called MHAMFD. Different levels of behavioural relationships are used to select appropriate neighbours, which considers the composite semantic information from the interweaving of different relationships and improves the quality of neighbour nodes. The embedding representation of target nodes is comprehensively learned by the aggregation of behavioural relationships within, between and at different levels. The effectiveness of MHAMFD at health insurance fraud detection was verified through experiments using real medical data. In future work, we will conduct explanatory studies on applying MHAMFD to health insurance fraud detection in different scenarios.

Abbreviations

HIN	Heterogeneous information network
AHIN	Attributed heterogeneous information network
AGG	Aggregation

Acknowledgements

The authors would like to thank Jinpo Chen for his contribution to the feature engineering part of the research and Harrison O for his help with polishing.

Authors' contributions

JTL and KBL conceived the study and led the research. JTL and RCC developed the source codes. JTL implemented and coordinated the analyses. JTL, XC, PL and ML performed the analysis. JTL wrote the paper. All authors contributed to revising the manuscript. All authors read and approved the final manuscript.

Funding

This research was supported by the Science Foundation of Fujian Province (No. 2021J011188, 2019Y0057), the Open Fund of Engineering Research Center of Big Data Application in Private Health Medicine, Fujian Provincial University (KF2020003), and the XMUT Scientific Research Project (No. YKJCX2020103). These funds were used in the design of the study and in the

collection, analysis, and interpretation of data, as well as in the writing of the manuscript.

Availability of data and materials

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Declarations

Ethics approval and consent to participate

This study was carried out in accordance with the tenets of the Helsinki Declaration and was approved by the China Medical Security Bureau. All methods were performed in accordance with the relevant guidelines and regulations. To ensure the protection of patients' personal information, all datasets underwent anonymization procedures prior to analysis. A city of medical security bureau in China has allowed us to use the Medical-1 and Medical-2 datasets. The informed consent of the experimental subjects were obtained before this study.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 7 October 2022 Accepted: 15 March 2023

Published online: 06 April 2023

References

- Sisko AM, Keehan SP, Poisal JA, Cuckler GA, Smith SD, Madison AJ, et al. National health expenditure projections, 2018–27: economic and demographic trends drive spending and enrollment growth. *Health Aff.* 2019;38(3):491–501.
- Cubanski J, Neuman T, Freed M. The facts on Medicare spending and financing. Washington, DC: Kaiser Family Foundation; 2018.
- Morris L. Combating fraud in health care: an essential component of any cost containment strategy. *Health Aff.* 2009;28(5):1351–56.
- China Medical Security Bureau. Statistical bulletin on the development of China's medical security business in 2020. 2020. http://www.nhsa.gov.cn/art/2021/6/8/art_7_5232.html. Accessed 20 Feb 2021.
- BANK OF CHINA. Bank of China Foreign Exchange Rates. 2020. <https://www.boc.cn/sourcedb/whpj/>. Accessed 8 Feb 2020.
- Xu D, Ruan C, Korpeoglu E, Kumar S, Achan K. Inductive representation learning on temporal graphs. 2020. arXiv preprint [arXiv:2002.07962](https://arxiv.org/abs/2002.07962).
- Felzenszwalb PF, Huttenlocher DP. Efficient graph-based image segmentation. *Int J Comput Vis.* 2004;59(2):167–81.
- Akoglu L, Tong H, Koutra D. Graph based anomaly detection and description: a survey. *Data Min Knowl Disc.* 2015;29(3):626–88.
- Chang J, Gao C, Zheng Y, Hui Y, Niu Y, Song Y, et al. Sequential recommendation with graph neural networks. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: Association for Computing Machinery; 2021. p. 378–87.
- Sun Y, Han J. Mining heterogeneous information networks: principles and methodologies. *Synth Lect Data Min Knowl Discov.* 2012;3(2):1–159.
- Cheng Z, Ding Y, He X, Zhu L, Song X, Kankanhalli MS. A[^] 3NCF: An Adaptive Aspect Attention Model for Rating Prediction. In: Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI). Stockholm: AAAI Press; 2018. p. 3748–54.
- You Q, Jin H, Wang Z, Fang C, Luo J. Image captioning with semantic attention. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE Computer Society; 2016. p. 4651–59.
- Bauder RA, Khoshgoftaar TM. The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data. *Health Inf Sci Syst.* 2018;6(1):1–14.

14. Sadiq S, Tao Y, Yan Y, Shyu ML. Mining anomalies in medicare big data using patient rule induction method. In: 2017 IEEE third international conference on multimedia Big Data (BigMM). Laguna Hills: IEEE; 2017. p. 185–92.
15. Viegas JL, Cepeda NM, Vieira SM. Electricity fraud detection using committee semi-supervised learning. In: 2018 International Joint Conference on Neural Networks (IJCNN). Rio de Janeiro: IEEE; 2018. p. 1–6.
16. Joudaki H, Rashidian A, Minaei-Bidgolli B, Mahmoodi M, Geraili B, Nasiri M, et al. Using data mining to detect health care fraud and abuse: a review of literature. *Glob J Health Sci.* 2015;7(1):194.
17. Zhang W, He X. An anomaly detection method for medicare fraud detection. In: 2017 IEEE International Conference on Big Knowledge (ICBK). Hefei: IEEE; 2017. p. 309–14.
18. Seo J, Mendelevitch O. Identifying frauds and anomalies in medicare-B dataset. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Jeju: IEEE; 2017. p. 3664–67.
19. Bauder R, Khoshgoftaar T. Medicare fraud detection using random forest with class imbalanced big data. In: 2018 IEEE international conference on information reuse and integration (IRI). Salt Lake City: IEEE; 2018. p. 80–7.
20. Pandey P, Saroliya A, Kumar R. Analyses and detection of health insurance fraud using data mining and predictive modeling techniques. In: *Soft Computing: Theories and Applications*. Singapore: Springer; 2018. p. 41–9.
21. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York: Association for Computing Machinery; 2014. p. 701–10.
22. Tang J, Qu M, Wang M, Zhang M, Yan J, Mei Q. Line: Large-scale information network embedding. In: *Proceedings of the 24th international conference on world wide web*. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee; 2015. p. 1067–77.
23. Wang D, Cui P, Zhu W. Structural deep network embedding. In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. New York: Association for Computing Machinery; 2016. p. 1225–34.
24. Lichtenwalter RN, Lussier JT, Chawla NV. New perspectives and methods in link prediction. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York: Association for Computing Machinery; 2010. p. 243–52.
25. Leroy V, Cambazoglu BB, Bonchi F. Cold start link prediction. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York: Association for Computing Machinery; 2010. p. 393–402.
26. Shi C, Li Y, Zhang J, Sun Y, Philip SY. A survey of heterogeneous information network analysis. *IEEE Trans Knowl Data Eng.* 2016;29(1):17–37.
27. Sun Y, Han J, Yan X, Yu PS, Wu T. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proc VLDB Endowment.* 2011;4(11):992–1003.
28. Dong Y, Chawla NV, Swami A. metapath2vec: Scalable representation learning for heterogeneous networks. In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. New York: Association for Computing Machinery; 2017. p. 135–44.
29. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. 2016. arXiv preprint [arXiv:1609.02907](https://arxiv.org/abs/1609.02907).
30. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. 2017. arXiv preprint [arXiv:1710.10903](https://arxiv.org/abs/1710.10903).
31. Wang X, Ji H, Shi C, Wang B, Ye Y, Cui P, et al. Heterogeneous graph attention network. In: *The world wide web conference*. New York: Heterogeneous graph attention network. ; 2019. p. 2022–32.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

