

Received October 25, 2018, accepted December 6, 2018, date of publication December 14, 2018, date of current version February 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2886680

Patient Cluster Divergence Based Healthcare Insurance Fraudster Detection

CHENFEI SUN¹, QINGZHONG LI¹, HUI LI¹, YULIANG SHI^{1,2},
SHIDONG ZHANG¹, AND WEI GUO¹

¹Research Center of Software and Data Engineering, School of Computer Science and Technology, Shandong University, Jinan 250101, China

²Dareway Software Co., Ltd., Jinan 250101, China

Corresponding author: Qingzhong Li (lqz@sdu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1400100 and Grant 2018YFC0114709, in part by the Natural Science Foundation of Shandong Province for Major Basic Research Projects under Grant ZR2017ZB0419, in part by the Taishan Industrial Experts Program of Shandong Province under Grant tscy20150305, and in part by the Key Research and Development Program of Shandong Province under Grant 2016ZDJS01A09.

ABSTRACT Healthcare insurance frauds are causing millions of dollars in loss for public healthcare funds around the world. Healthcare fraud detection methods can help us to avoid the loss of medical healthcare insurance fund and to improve medical quality. The existing fraudster detection methods always consider people who violate normal behavior patterns as fraudsters. However, fraudsters can evade these monitors by camouflage, by adding normal behaviors so that they look “normal.” Our focus is to spot healthcare insurance patient fraudsters in the presence of camouflage. Although camouflage may hinder fraudster detection to some extent, we find that camouflage behaviors always sustain in a short period when the fraudster is conducting fraud. In other words, camouflage behaviors will not last long. Hence, if we can consider the cluster divergence of each patients’ hospital admission graph during a long time, we can detect healthcare insurance fraudsters free of the interference of fraudsters’ camouflage behaviors. In this paper, we propose the patient cluster divergence-based healthcare insurance fraudster detection (PCDHIFD), which can get rid of the disturbance of camouflage in fraud detection. Extensive experiment results show that our PCDHIFD outperforms the comparison approaches in terms of f -measure by over 15%.

INDEX TERMS Fraudster detection, camouflage, clustering, healthcare insurance.

I. INTRODUCTION

Healthcare insurance fraud is a serious threat to the proper use of public funds. An estimated 17 billion to 57 billion were lost due to fraud under the Healthcare scheme in 2012 [1]. This problem has motivated many researchers to develop fraud detection technologies in healthcare fraud detection. Traditional fraud detection techniques often rely on rules designed by experts which can be used as a basis for identifying behaviors violating some of these rules [2]. Besides, data-driven approaches for healthcare insurance fraud detection have now become popular recently. Most of the data-driven research in healthcare fraud is focused on statistical analysis and the use of machine learning algorithms like clustering, k-nearest neighbor, decision trees, neural networks, etc. However, these methods always have high false positive rate because normal patients may have some behaviors that violate behavior patterns while fraudsters may try their best to add normal behaviors so that they look “normal”.

Healthcare insurance patient fraudster detection in the face of camouflage is a non-trivial task. It has significant challenges as follows:

1) Camouflage: Smart patient fraudsters will also try to “look normal”, by mimicking normal patients as closely as possible - this kind of behavior is called “camouflage” in the recent literatures.

2) Heterogeneous: Healthcare insurance data are heterogeneous and longitudinal in nature. For example, a claim record consists of a series of physician orders, where each physician order usually consists of drug/treatment name, mode, dosage and time.

3) Time-evolving: Fraudsters will change their fraud behavior as time evolving. In other words, fraudsters’ behavior pattern is always changing, there is no obvious certain pattern.

To address the aforementioned challenges in healthcare insurance patient fraudster detection from healthcare insurance claim records, first, we need an effective method to

measure the similarity between claim records containing sequential and multifaceted information in physician orders. Second, base on the similarity measurements between claim records, we construct hospital admission graph G in patient-level. Third, we adopt graph based density peak clustering method in hospital admission graph G . Meanwhile, to make the results understandable, we need to extract semantic interpretation of each cluster. The semantic interpretation of a cluster indicates key components of the cluster and can help us make sense of the cluster. Finally, we compute Patient Cluster Divergence during the entire period and estimate the fraud probability of each patient.

Our main contribution in this paper are listed as follows:

1) Original density peak clustering-DPC [2] method need to select a cutoff distance which is a complex task in reality. So we improve DPC and adopt betweenness centrality of nodes in graph to measure the density of each object, which can avoid the complex selection of cutoff distance.

2) To get a better understanding of obtained cluster, we try to extract semantic interpretation of each cluster. For each cluster, we identify a dense core area around the density peak and extract a semantic interpretation for each cluster.

3) To weaken the hinder of camouflage, we considers health seeking behaviors of patients during a long period while camouflage can only last a short time. Besides, we adopt graph to represent hospital admissions of patients which can retain the complex relations between variety kinds of objects. Therefore, our method can get rid of the disturbance of camouflage in fraud detection.

In summary, we propose Patient Cluster Divergence based Healthcare Insurance Fraudster Detection-PCDHIFD in this paper. In particular, it performs well compared to existing methods in healthcare insurance fraud detection in the face of camouflage.

II. RELATED WORK

In this section, we review the existing work in the literature related to fraudster detection in healthcare insurance.

Fraudster detection in healthcare insurance has received significant focus in recent years. Data mining is a popular method for detecting fraud. The large volumes of behavior data are difficult for conventional methods to process. With the increasing availability of clinical and behavior data, data mining is becoming an important tool for healthcare insurance fraud detection as well. Information and analyses obtained through data mining can improve operating efficiency [3]. More recently, Musal [4] proposed the use of clustering for geographical analysis of potential frauds. A data-mining framework that utilizes the concept of clinical pathways to facilitate automatic and systematic construction of fraud detection model has also been proposed in [5]. A graph analysis based approach for healthcare insurance fraud detection has been proposed in [6]. The approach relies on the availability of knowledge on the relationships among the many stakeholders involved in healthcare insurance (e.g., patients, doctors, pharmacies and insurance

companies) to identify suspicious relationships, suspicious spatial-temporal changes, suspicious graph structures and suspicious individuals. In the Chinese healthcare insurance claim system environment on which our research is based, users file their claims on a common technology platform without directly interacting with other stakeholders. The relationship data among these stakeholders are unavailable. Nevertheless, these approaches are not able to accurately identify fraud when fraudsters have camouflage behaviors.

Temporal data mining plays an important role in healthcare fraudster detection. The key challenge of temporal data mining is how to represent the temporal data, which can simplify the similarity computation between temporal sequences. For continuous time series data, many representation methods and similarity measurement algorithms have been developed [7]–[10]. For discrete event sequence data, some recent works are presented for diverse applications [8], [11]–[13]. However, healthcare insurance claim record in this paper is a sequence of physician orders, which is much more complex than simple event sequence and time series. Therefore, existing methods are not directly applicable. In this paper, to discover fraudster from large-scale healthcare insurance claim records, we proposed novel similarity measurement and representation methods for multifaceted and sequential event sets.

Most of clustering algorithms mentioned in [10] focus on solely dividing the homogeneous objects into different groups, rare work focus on extract semantic interpretation of the identified clusters which is helpful to understand the cluster result. In this paper, a hospital admission is much more complex than the homogeneous objects studied in traditional clustering problems. To address this challenge, we construct a dense core in each exemplar-based cluster to extract the semantic interpretation.

III. PROBLEM DEFINITION

Healthcare insurance claim records mainly contain three categories of patient information. They are demographic information, diagnose information and physician orders. Our goal is to detect fraudsters from the massive healthcare insurance claim records in the face of camouflage.

Definition 1 (Demographic Information): Demographic information is recorded when a patient visits a hospital, which includes the gender, age, occupation and other related information of a patient. These information has a great influence on clinical decisions, e.g., treatment option design and dosage selection. The demographic information of a patient can be formalized as

$$P = \{P^{age}, P^{gender}, P^{race}, \dots\}.$$

Definition 2 (Diagnose Information): Diagnose information is given by a physician when a patient visits a hospital. It contains the name and severity of the diseases. In practice, a patient may suffer from multiple healthcare problems and can be diagnosed with more than one kind of disease.

So diagnose information can be represented by

$$D = \{\{D_1^{name}, D_1^{severity}\}, \{D_2^{name}, D_2^{severity}\}, \dots\}$$

Definition 3 (Physician Order): A physician order refers to a medical prescription, which is implemented by a physician in the form of instructions that manage the care plan for a patient. A physician order can be expressed as a tetrad

$$PO = \{PO^{name}, PO^{mode}, PO^{dose}, PO^{time}\}$$

where PO^{name} represents the name of the used drug or treatment, PO^{mode} is the usage mode, which can be classified as “Intravenous injection” (IV), “Intramuscular” (IM), “hypodermic injection” (IH), “Oral” and so on. PO^{dose} indicates the dose. PO^{time} denotes the active time point of the order. For example, a physician order {Aspirin, Oral, 1, 2017.10.1} means that the medicine Aspirin is delivered by oral route, the dose is 1(a packed box of Aspirin), the active time of this physician order is 2017.10.1.

Definition 4 (Hospital Admission): A hospital admission is composed of all the physician orders given to the patient during this stay, which can be represented as

$$HA = \{diagnose, PO_1, PO_2, \dots, PO_m\}$$

where diagnose is the diagnose disease of this hospital admission and m is the number of physician order in this hospital admission. In order to be easily understood, we present an example of hospital admission by Table 1.

TABLE 1. An example of hospital admission.

diagnose	name	mode	dose	time
pneumonia	sodium chloride injection	IV	1	2016.3.19
	piperacillin sodium	Oral	4	2016.3.19
	vitamin c	Oral	3	2016.3.20
	piperacillin sodium	Oral	3	2016.3.20
	ESR	Test	1	2016.3.22

We aim to detect fraud claim records from massive healthcare insurance claim records, so we consider patient-level hospital admission and propose a Patient Cluster Divergence based clustering method which can cluster patients into different groups. In addition, for each cluster, we extract semantic interpretation of this group. Finally, we compare the Patient Cluster Divergence of each patient and compute fraud probability of each patient. Through the patient cluster divergence during a long period, our method is able to against “camouflage” of fraudster due to “camouflage” usually last for a short time.

IV. PATIENT CLUSTER DIVERGENCE BASED HEALTHCARE INSURANCE FRAUDSTER DETECTION

In this section, we introduce the detail of the proposed methods. Our work is composed of three steps:

- 1) Compute similarity between patient-level hospital admission and construct patient hospital admission graph G.
- 2) Cluster and extract semantic interpretation of each cluster through a graph based dense peak clustering algorithm GDPC in G.
- 3) Compute Patient Cluster Divergence of each patient during the entire period and obtain the fraud probability of each patient.

A hospital admission defined in this paper is much more complex than previously studies objects, which poses non-trivial challenges to similarity measurement and cluster semantic interpretation extraction. To weaken the hinder of camouflage, we consider patients’ hospital admissions during the entire period. Therefore, we propose novel methods in each step.

A. SIMILARITY MEASUREMENT AND PATIENT HOSPITAL ADMISSION GRAPH CONSTRUCT

To cluster patients according to their hospital admissions, we need to compute similarity between each hospital admission pair. However, hospital admission includes not only nominal information like drug/treatment name, usage mode, but also numeric information such as dose and so on. Hence the recorded information in a hospital admission is heterogeneous. In this case, computing similarity between two hospital admissions is challenging.

According to section 3, a hospital admission is composed of a set of physician orders. Therefore, we need to define the similarity measurement between physician orders firstly. For two physician orders

$$PO_r = \{PO_r^{name}, PO_r^{mode}, PO_r^{dose}, PO_r^{time}\}$$

and

$$PO_s = \{PO_s^{name}, PO_s^{mode}, PO_s^{dose}, PO_s^{time}\},$$

the similarity between PO_r and PO_s is defined in (1), as shown at the bottom of this page.

$\delta(x, y)$ function denotes that if x and y are same, the value of $\delta(x, y)$ is 1 and equals 0 otherwise. As shown in Equation 1, we firstly compare the used drug or treatment name of two physician orders, if PO_r^{name} and PO_s^{name} are the same, we further consider the usage mode and dose; Otherwise, the similarity between two physician orders is set to 0.

After obtaining the similarity between two physician orders, the similarity between two hospital admissions can be considered as a similarity between two complex physician

$$Sim(PO_r, PO_s) = \frac{\delta(PO_r^{name}, PO_s^{name}) * [\delta(PO_r^{mode}, PO_s^{mode}) + \frac{\min(PO_s^{dose}, PO_r^{dose})}{\max(PO_r^{dose}, PO_s^{dose})}]}{2} \tag{1}$$

order sets. As mentioned in section 3, a hospital admission can be represented as a table composed of physician orders. Then the similarity between two hospital admissions

$$HA_1 = \{diagnose_1, PO_{11}, PO_{12}, \dots, PO_{1m}\}$$

and

$$HA_2 = \{diagnose_2, PO_{21}, PO_{22}, \dots, PO_{2h}\}$$

m,h denoted the number of physician orders in HA_1 and HA_2 respectively.

can be defined as

$$Sim(HA_1, HA_2) = \frac{|Common(HA_1, HA_2)|}{|Union(HA_1, HA_2)|} \quad (2)$$

where $|Common(HA_1, HA_2)|$ is the number of common physician orders of HA_1 and HA_2 , while $|Union(HA_1, HA_2)|$ indicates the number of physician orders in HA_1 and HA_2 .

However, different from previous similarity measurement problem, we need to consider the appearance times of elements and the similarities between elements. So we define $|Common(HA_1, HA_2)|$ in this paper as

$$|Common(HA_1, HA_2)| = \sum_{pq} Sim(PO_{1p}, PO_{2q}) * a_{pq} \quad (3)$$

where $Sim(PO_{1p}, PO_{2q})$ denotes similarity between p-th physician order PO_{1p} in HA_1 and q-th physician order PO_{2q} in HA_2 , $A = (a_{pq})_{|HA_1| * |HA_2|}$ is an allocation matrix. A is obtained by solving

$$\begin{aligned} \arg \max_A z = & \sum_{\substack{p,q \in \mathbb{Z} \\ p \in [1, |HA_1|] \\ q \in [1, |HA_2|]}} Sim(PO_{1p}, PO_{2q}) * a_{pq} \\ \text{s.t. } & \sum_q a_{pq} \leq freq(PO_{1p}) \\ & \sum_p a_{pq} \leq freq(PO_{2q}) \\ & a_{pq} \geq 0 \end{aligned} \quad (4)$$

where $freq(PO_{1p})$ and $freq(PO_{2q})$ represents the appearance times of p-th physician order in HA_1 and q-th physician order in HA_2 respectively. According to Equation 4, we can see that A is a allocation matrix which allocates the number of element occurrences in a two dimensional table with the goal of maximizing the same part of HA_1 and HA_2 .

Combining the definition from equation 2-4, similarity between HA_1 and HA_2 is finally defined as

$$Sim(HA_1, HA_2) = \frac{\sum_{pq} Sim(PO_{1p}, PO_{2q}) * a_{pq}}{|HA_1| + |HA_2| - \sum_{pq} Sim(PO_{1p}, PO_{2q}) * a_{pq}} \quad (5)$$

To cluster patient according to their hospital admission similarity, we construct hospital admission graph in patient level.

Definition 5 (Hospital Admission Graph): Hospital Admission Graph is a heterogeneous graph with two types of nodes and three kinds of edges. Two types of nodes are patient node and hospital admission node. We denote Hospital Admission

Graph as $G=(V,E,W)$, V is the vertex set and E is the edge set s.t $\forall(u, v) \in E, u \in V$ and $v \in V$. W is the weight of edge.

For each patient node p_i , demographic information in definition 1 is shown as properties of patient. Each hospital admission node HA_g is composed of a set of physician orders as defined in definition 4.

Three kinds of edges are defined as follows:

1) edge $e(p_i, HA_g)$ between a patient node and a hospital admission node indicates that patient p_i have an hospital admission. The property of this edge is the time of the hospital admission. We set the weight of such edges as -1 which indicates the patient conduct the hospital admission.

2) edge $e(HA_g, HA_h)$ between two hospital admission nodes HA_g and HA_h shows that they are similar and the similarity between them exceeds a threshold λ which is determined in experiments. The weight $w_{e(HA_g, HA_h)}$ of edge $e(HA_g, HA_h)$ shows the similarity between the two hospital admissions and is computed according to Equation 2. To be more specifically, there is an edge between two hospital admission nodes if and only if the similarity between them is bigger than threshold λ .

3) edge $e(p_i, p_j)$ between two patient nodes p_i and p_j represents the two patients are similar in demographic and diagnose information. The weight $w_{e(p_i, p_j)}$ is computed as

$$w_{e(p_i, p_j)} = \alpha * \frac{\sum_{f \in \{age, sex, \dots\}} \delta(P_i^f, P_j^f)}{N} + \beta * \frac{|D_i \cap D_j|}{|D_i \cup D_j|} \quad (6)$$

where $\{age, sex, \dots\}$ denotes the selected features and N is the number of selected relevant features in patient demographic information. And the larger the $w_{e(p_i, p_j)}$ is, the more similar between patient p_i and p_j . α and β indicate the weights of demographic and diagnose information to patient similarity calculation.

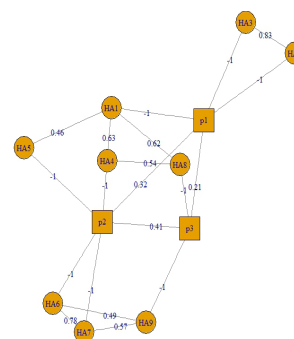


FIGURE 1. Example of Patient Hospital Admission Graph G. There are two kinds of vertices and three kinds of edges. Circle nodes indicate patients and square nodes present hospital admissions. The weight ‘-1’ between patient node and hospital admission node denotes the patient conducts the hospital admission. Other weight indicates the similarity between two connected nodes.

Figure 1 shows an example of Patient Hospital Admission Graph G. There are two kinds of vertices and three kinds of edges as mentioned above.

B. GRAPH BASED DENSITY PEAK CLUSTERING AND CLUSTER SEMANTIC INTERPRETATION EXTRACTION

After getting Patient Hospital Admission Graph G, we conduct clustering method on graph G. Clustering aims to partition a set of objects into multiple clusters so that objects in the same cluster are similar to each other as much as possible while objects in different clusters are not similar. However, in this case, except cluster objects into groups, we also need to obtain semantic interpretation of each cluster. So we propose a graph based density peak based clustering GDPC method to accomplish this task.

Our clustering method derives from Density Peak Clustering(DPC) [2], which is a recently proposed clustering algorithm. The idea of this method is that cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from points with higher densities. DPC can discover clusters with complex shapes while traditional clustering methods can only find spherical clusters. However, the selection of cutoff distance d_c in DPC is complex. Meanwhile, DPC didn't introduce the calculation of distance in detail. Hence we proposed a graph based density peak based clustering method GDPC to cluster objects in G. In GDPC, we adopt degree centrality of node in graph to measure the density of object. Furthermore, we introduce distance calculation between objects in detail.

Algorithm 1 Graph Based Density Peak Clustering and Cluster Semantic Extraction

Require: Patient Hospital Admission graph $G=(V,E,W)$, cutoff distance d_c

for each node u in G **do**

2: **for** any other node v in $G-\{u\}$ **do**

$d_{uv} = e^{-w_{uv}}$

4: **end for**

$\rho_u = (\sum_{s \neq u \neq t \in V} \frac{g_{st}(u)}{g_{st}})^{\frac{n(n-1)}{2}}$

6: $\gamma_u = \min_{v: \rho_v \geq \rho_u} (d_{uv})$

$\eta_u = \rho_u * \gamma_u$

8: **end for**

$E = \{j | \eta_j \geq \epsilon, \epsilon \text{ is the } k\text{-th largest } \eta \text{ value}\}$

10: **for** any other node i $\notin E$ **do**

$c(i) = c(\arg \min_{j \in E} d_{ij})$

12: **end for**

for each cluster c_i in C **do**

14: $DCore_i = \{j | d(j, ex_i) \leq d_i\}$

$Support_i(PO_l) = \frac{\sum_{j \in DCore_i} \lambda(PO_l, j)}{|DCore_i|}$

16: obtain semantic meaning of each cluster c_i

end for

In GDPC, for each object, we compute two indicators: 1) local density ρ and 2) the minimum distance between object and any other objects with higher local density γ , where ρ is defined as

$$\rho_u = (\sum_{s \neq u \neq t \in V} \frac{g_{st}(u)}{g_{st}})^{\frac{n(n-1)}{2}} \tag{7}$$

where g_{st} is the number of shortest paths from node s to node t, and $g_{st}(u)$ indicates the number of shortest paths from node s to node t through node u. n is the number of nodes in G.

The meaning of ρ is to measure the density of object.

The second indicator γ is computed by the minimum distance between object u and any other object v with higher density:

$$\gamma_u = \min_{v: \rho_v \geq \rho_u} (d_{uv}) \tag{8}$$

where d_{uv} is defined as

$$d_{uv} = e^{-w_{uv}} \tag{9}$$

$w_{u,v}$ is the weight of edge (u,v) in G. d_{uv} indicated the distance between object u and v, the larger the similarity between u and v, the smaller the distance d_{uv} is. Objects with larger ρ and γ are considered as the illustrations. The intuition is that illustrations are the points with highest density in a relative large range. Then clustering result can be obtained according to mined illustrations. For each non-illustration object, its cluster is the same as its nearest illustration node, which is

$$c(i) = c(\arg \min_{j \in E} d_{ij}) \tag{10}$$

$j \in E$ indicates j is a cluster exemplar.

After we obtain the clustering result, we want to extract the semantic meaning of each cluster. We adopt k-nearest neighbor of its illustration to define the dense core area of each cluster. A dense core can be represented as a set

$$DCore_i = \{j | d(j, ex_i) \leq d_i\} \tag{11}$$

ex_i is the exemplar of the i-th cluster, $d(j, ex_i)$ is the distance between objects j and exemplar ex_i , d_i is the distance between exemplar ex_i and its k-th nearest neighbor.

In order to extract semantic meaning of each cluster from its dense core, we define the support of a physician order occurred in the cluster as

$$Support_i(PO_l) = \frac{\sum_{j \in DCore_i} \lambda(PO_l, j)}{|DCore_i|} \tag{12}$$

where $\lambda(PO_l, j)=1$ if physician order PO_l appears in hospital admission j which is in dense core are $DCore_i$ of i-th cluster, $\lambda(PO_l, j)=0$ otherwise.

Then the semantic meaning of each cluster can be represented as physician orders and their supports.

C. PATIENT CLUSTER DIVERGENCE BASED HEALTHCARE INSURANCE FRAUD DETECTION

For each patient p, we can estimate the fraud probability according to the consistency between patient similarity and their hospital admission similarity. Under normal circumstances, the more similar two patients are, the more similar their hospital admissions are. So healthcare insurance fraud detection can be transformed to a Patient Cluster

TABLE 2. Three extracted typical cluster semantic meaning of hospital admission.

cluster	physician order	support
Cluster 1	cardiac color ultrasound,test,1	0.16
	venous blood,test,1	0.29
	Bayaspirin,Oral,6	0.28
	sodium chloride injection,IV,1	0.23
	liver function,test,1	0.14
	urine sediment and microscopic examination,test,1	0.12
	myocardial enzyme,test,1	0.18
routine electrocardiogram examination,test,1	0.11	
Cluster 2	shexiangbaoxin pills,Oral,8	0.31
	tongxinluo capsule,Oral,4	0.33
	metoprolol tartrate,Oral,2	0.27
	rosuvastatin,Oral,5	0.19
	aspirin,Oral,3	0.25
	Clopidogrel,Oral,2	0.19
	isosorbide mononitrate,Oral,8	0.16
	nitroglycerin,Oral,6	0.21
	computer glucose monitoring,test,1	0.33
Cluster 3	sodium chloride injection,IV,1	0.24
	repaglinide tablets,Oral,1	0.29
	continue subcutaneous insulin injection,IH,1	0.23
	diformin tablets,Oral,2	0.31
	adenosine deaminase measurement,test,1	0.17
	plasma viscosity measurement,test,1	0.19
	long-lasting human insulin analogue,IH,1	0.21

the severity of CHD patient, while the semantic meaning of cluster 2 is some medicines which can help CHD patients to maintain normal condition. Cluster 3 indicates the main treatments and drugs used for diabetes.

As discussion in section 3,the semantic meaning of cluster can help us calculated the fraud probability of each patient. The method is then compared to competing methods which can be grouped into the following categories:

Classification methods (CM). The classification methods, such as decision trees (DT) [14] and support vector machines (SVM) [15], are straight-forwardly fitted with the training set and evaluated with the test set.

Anomaly detection (AD). Anomaly detection methods aims to identify objects which do not conform to an expected pattern or other objects in a dataset. GdiLOF which improved LOF Outlier Detection Algorithm [16], seem more appropriate for our healthcare insurance fraud detection.

Pattern Mining(PM). Pattern mining methods tend to mine the behavior patterns of the whole crowd. BP-Growth [3] proposed optimizing strategies for association rule mining for behavior pattern analysis.

Neural Network(NN). A three layer MLP architecture was selected as is was the common choice in the previously mentioned research and so is representative of the range of classifiers that could be used [17]. All neuron values were in the range [0,1] using a standard sigmoidal activation function. Besides, we adopt LSTM [18] which considers temporal information in the hospital admissions.

For all the methods with parameters, we optimize the parameters with 10-fold cross-validation by further dividing the training set into 80% for model fitting and 20% for parameter validation.

Evaluation metrics. We use precision, recall, and F-score computed with test set to evaluate the performances of different methods. Precision = $\frac{t_p}{t_p+f_p}$ is the fraction of patients

identified as fraudulent which are indeed fraudulent. Recall = $\frac{t_p}{t_p+f_n}$ is the fraction of all fraudulent patients that have been correctly identified. f-measure = $\frac{(2*Precision*Recall)}{Precision+Recall}$ is the weighted harmonic mean of precision and recall. Here, t_p (true positive) is the number of patients correctly classified as fraudulent, f_p (false positive) is the number of persons incorrectly classified as fraudulent, and f_n (false negative) is the number of patients incorrectly classified as non-fraudulent.

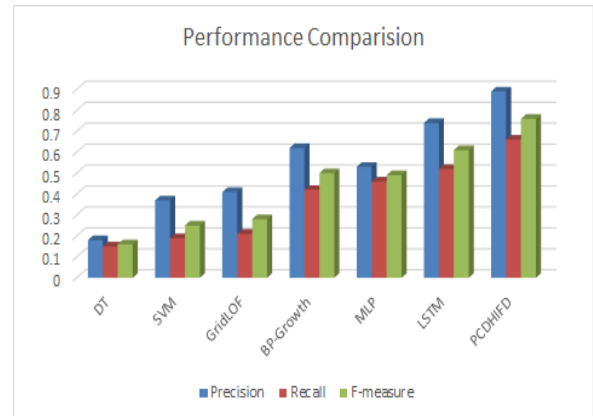


FIGURE 5. Performance of PCDHIFD against other approaches. Our PCDHIFD outperforms the comparison approaches in terms of f-measure by over 15%.

Figure 5 shows the performance of PCDHIFD against other approaches. We have several interesting observations which confirm our research motivation from Figure 5. First, the precisions of all pure classification methods are very low. Since the proportion of positive instances are extremely low, the classification problem is unbalanced. The AD methods perform somehow better, but it have low recall because most fraudsters will try their best to avoid to bypass regular detection rules. The PM method has low recall because there is few behavior pattern in the crowd. In other words,because of the curse of cardinality, BP-Growth [3] can hardly find meaningful frequent itemsets from the whole crowd. LSTM behaves better because it considers time information of the hospital admissions but the precision is not high enough to be applied in practical fraud detection systems.In contrast, our PCDHIFD method significantly improve the precision by more than 15%. This observation shows that our approach can effectively reduce the false positives. Moreover, our method also performs better in terms of other metrics. For example, the recall of our method is ten percent more than existing methods. As a result of high precision and high recall, when these two metrics are combined together to form the f-measure shown in Figure 3, PCDHIFD consistently beats the comparison approaches in the experiments. On average, PCDHIFD outperforms the comparison approaches in terms of f-measure by over 15%.

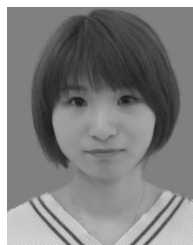
VI. CONCLUSION

In this paper, we propose a Patient Cluster Divergence based healthcare insurance fraudster detection method-PCDHIFD.

We consider hospital admissions of each patient during the entire period, however, fraudsters' camouflage behaviors last for a short time. Therefore, our PCDHIFD method can detect healthcare insurance fraudsters free of the interference of fraudsters' camouflage behaviors. Specifically, we compute similarity between patient-level hospital admission and construct similarity graph G . Then we cluster and extract semantic meaning of each cluster through a graph based density peak clustering algorithm GDPC in G . Finally we compute Patient Cluster Divergence of each patient during the entire period and obtain the fraud probability of each patient. Experimental results show that our method can significantly improve the fraud detection accuracy in the face of camouflage. To be more specifically, our PCDHIFD outperforms the comparison approaches in terms of f -measure by over 15%.

REFERENCES

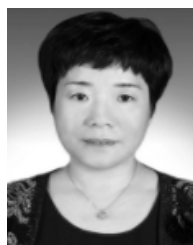
- [1] N. Aldrich, J. Crowder, and B. Benson, "How much does medicare lose due to fraud and improper payments each year," *Sentinel*, 2014.
- [2] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.
- [3] X. Li, H. Cao, E. Chen, H. Xiong, and J. Tian, "Bp-growth: Searching strategies for efficient behavior pattern mining," in *Proc. IEEE 13th Int. Conf. Mobile Data Manage. (MDM)*, Jul. 2012, pp. 238–247.
- [4] R. M. Musal, "Two models to investigate medicare fraud within unsupervised databases," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8628–8633, 2010.
- [5] W.-S. Yang and S.-Y. Hwang, "A process-mining framework for the detection of healthcare fraud and abuse," *Expert Syst. Appl.*, vol. 31, no. 1, pp. 56–68, 2006.
- [6] J. Liu et al., "Graph analysis for detecting fraud, waste, and abuse in healthcare data," *AI Mag.*, vol. 37, no. 2, pp. 33–46, 2016.
- [7] I. Batal, D. Fradkin, J. Harrison, F. Moerchen, and M. Hauskrecht, "Mining recent temporal patterns for event detection in multivariate time series data," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 280–288.
- [8] C. Liu, F. Wang, J. Hu, and H. Xiong, "Temporal phenotyping from longitudinal electronic health records: A graph based framework," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 705–714.
- [9] G. E. Batista, X. Wang, and E. J. Keogh, "A complexity-invariant distance measure for time series," in *Proc. SIAM Int. Conf. Data Mining*, 2011, pp. 699–710.
- [10] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.
- [11] L. Sun, C. Liu, C. Guo, H. Xiong, and Y. Xie, "Data-driven automatic treatment regimen development and recommendation," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1865–1874.
- [12] C. Liu, K. Zhang, H. Xiong, G. Jiang, and Q. Yang, "Temporal skeletonization on sequential data: Patterns, categorization, and visualization," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 1, pp. 211–223, Jan. 2016.
- [13] J. Yang, C. Liu, M. Teng, H. Xiong, M. Liao, and V. Zhu, "Exploiting temporal and social factors for B2B marketing campaign recommendations," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2015, pp. 499–508.
- [14] W. Dai and W. Ji, "A mapreduce implementation of C4. 5 decision tree algorithm," *Int. J. Database Theory Appl.*, vol. 7, no. 1, pp. 49–60, 2014.
- [15] R. A. Berk, "Support vector machines," in *Statistical Learning From a Regression Perspective*. Cham, Switzerland: Springer, 2016, pp. 291–310.
- [16] Z. Xie et al., "An improved outlier detection algorithm to medical insurance," in *Proc. Int. Conf. Intell. Data Eng. Autom. Learn.* Cham, Switzerland: Springer, 2016, pp. 436–445.
- [17] G. Hinton, O. Vinyals, and J. Dean. (2015). "Distilling the knowledge in a neural network." [Online]. Available: <https://arxiv.org/abs/1503.02531>
- [18] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient subtyping via time-aware LSTM networks," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 65–74.



CHENFEI SUN received the bachelor's and master's degrees from the Computer Science and Technology College, Shandong University, in 2012, where she is currently pursuing the Ph.D. degree. Her research interests include data mining and fraud detection.



QINGZHONG LI is currently a Professor and a Ph.D. Supervisor with Shandong University. He is also the Academic Leader of the Jinan Big Data Integration and Intelligent Analysis Outstanding Innovation Team. He has presided over more than 20 national, provincial, and ministerial level projects. He has published more than 80 papers in important journals at home and abroad. His main research interests include data science and data analysis. He is a member of the China Computer Society Database Committee, the China Computer Society System Software Committee, and the China Computer Society Electronic Government and Office Automation Committee.



HUI LI was born in 1967. She received the bachelor's degree from the Computer Science Department, Hohai University, in 1989, the master's degree from the Computer Science and Technology College, Shandong University, in 2001, and the Ph.D. degree in engineering from Shandong University, in 2010. She mainly teaches the database system, the database design and realization, and other undergraduate and graduate courses. She presided over and participated in more than ten national and provincial projects and horizontal projects, such as the National Natural Science Foundation of China, the National Development and Reform Commission of China's Industrialization, the Shandong Province Science and Technology Development Plan, and the Shandong Province Natural Science Foundation. She has published in the *Journal of Communication* and other important journals and conferences in China and other countries; more than ten papers and many articles were included in EI. She was a recipient of the Shandong Provincial Science and Technology Progress Award and the Outstanding Achievement Award of Computer Application in Shandong Province.



YULIANG SHI was born in 1978. He received the bachelor's degree from the Computer Science Department, Shandong University, in 2000, the master's degree from the Computer Science and Technology College, Shandong University, in 2003, and the Ph.D. degree in engineering from Fudan University, in 2006. He is currently the Deputy Director of the Computer Software and Data Engineering Research Center, Shandong University, where he is involved in the calculation of collaborative computing for the China Computer Society Service Branch Committee Member Shandong Taishan Industry Leading Talent. He is also an Associate Professor with Shandong University. He is also a Master Mentor. He teaches software engineering, the object-oriented development technology, the database design and implementation, and many undergraduate and postgraduate courses. He hosted and participated in the National Science and Technology Support Plan, the National Natural Science Foundation of China, the Shandong Province Science and Technology Development Plan, the Shandong Province Natural Science Foundation of China, and other countries and more than ten provincial projects. He has published more than 20 papers in the important domestic or foreign journals.



SHIDONG ZHANG was born in 1969. He received the degree and the master's degree from the Computer Science Department, Shandong University, in 1990 and 1993, respectively, and the D.Sc. degree from Shandong University, in 2003, where he is currently a Professor and a Ph.D. Supervisor. He is also the Academic Leader of the Jinan Big Data Integration and Intelligent Analysis of Outstanding Innovation Team. He taught many main specialized courses, such as the database system and the database design and implementation. He presided over and participated in the National Science and Technology Support Plan, the National Natural Science Fund, the National Development and Reform Commission Industrialization Projects, the Development of Science and Technology Plan of Shandong Province, the Shandong Province Natural Science Foundation of China, and other countries and more than 20 provincial projects. He has published more than 40 papers in the *Journal of Computers*, and meeting important journals, such as the *Journal of System Simulation*. He was a recipient of the Progress Prize in Science and Technology in Shandong Province repeatedly.



WEI GUO received the bachelor's, master's, and Ph.D. degrees from the Computer Science and Technology College, Shandong University, in 2001, 2005, and 2015, respectively. He is a Master Mentor.

...