BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 03/06 pp25–34
Volume 2, Number 1, March 2019
DOI: 10.26599/BDMA.2018.9020009

Heterogeneous Network-Based Chronic Disease Progression Mining

Chenfei Sun, Qingzhong Li*, Lizhen Cui, Hui Li, and Yuliang Shi

Abstract: Healthcare insurance fraud has caused billions of dollars in losses in public healthcare funds around the world. In particular, healthcare insurance fraud in chronic diseases is especially rampant. Understanding disease progression can help investigators detect healthcare insurance frauds early on. Existing disease progression methods often ignore complex relations, such as the time-gap and pattern of disease occurrence. They also do not take into account the different medication stages of the same chronic disease, which is of great help when conducting healthcare insurance fraud detection and reducing healthcare costs. In this paper, we propose a heterogeneous network-based chronic disease progression mining method to improve the current understanding on the progression of chronic diseases, including orphan diseases. The method also considers the different medication stages of the same chronic disease. Extensive experiments show that our method can outperform the existing methods by 20% in terms of F-measure.

Key words: disease progression; heterogeneous network; healthcare insurance fraud

1 Introduction

Healthcare insurance fraud has caused billions of dollars in losses in public healthcare funds around the world. In particular, healthcare insurance fraud involving chronic diseases is especially rampant. Detecting chronic disease-related healthcare insurance fraud is an important and difficult task. With the rapid evolution of computer software and hardware technologies, various types of data, such as healthcare insurance records, research data of pharmaceutical companies, and information captured from wearable devices, are becoming increasingly available. Effective mining of these data can help us obtain actionable insights into chronic disease progression. If we can understand the process of chronic disease progression, we will be able to identify patients at risk of developing chronic disease based on their previous healthcare history. Preventive measures can then be taken, and chronic disease-related healthcare insurance fraud can be detected to increase the quality of healthcare and reduce costs.

Obtaining chronic disease-related healthcare insurance records, which are defined as a systematic collection of patient healthcare information because of chronic disease, are a major motivation for conducting data-driven healthcare research. However, various challenges, such as sparsity, heterogeneity, noise, and bias, have been encountered when working directly with these records. To understand chronic disease progression, a considerable amount of work has been done in assessing the risk of developing chronic disease. Most of these methods are based on rule-based scoring models, which assign scores to various physiological observable factors, such as demographic information or family history. For example, these methods provide an intuitive method to assess patients within a short time frame in a specific healthcare setting. However, these

• Chenfei Sun, Qingzhong Li, Lizhen Cui, Hui Li, and Yuliang Shi are with Research Center of Software and Data Engineering, Shandong University, Jinan 250101, China. E-mail: sunchenfei@mail.sdu.edu.cn; lqz@sdu.edu.cn; clz@sdu.edu.cn; lih@sdu.edu.cn; shiyuliang@sdu.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2017-10-26; revised: 2018-04-26; accepted: 2018-05-03

methods are inapplicable to disease progression over a long period of time.

Healthcare insurance records in chronic disease, defined as systematic collection of patient healthcare information because of chronic disease, are one of the major carries for conducting data driven healthcare research. However, there are various challenges if we work directly with chronic disease healthcare insurance records, such as sparsity, heterogeneity, noisiness, bias, etc. To understand chronic disease progression, a considerable amount of work has been done in assessing the risk of developing chronic disease. Most of these methods are based on rule-based scoring models which assign scores to various physiological observable factors such as demographic information or family history. For example, these methods provide an intuitive way to assess patients within the short time frame in a specific healthcare setting. However, these methods cannot be applicable when it comes to disease progression during a long period of time.

Various data-mining methods have recently and effectively been applied in the healthcare setting. These methods include a variety of supervised learning algorithms, such as decision tree and artificial neural network, to predict heart disease, cancer, and other diseases. Clustering and vector similarity-based collaborative filtering methods have also been proposed to predict individual disease risk. Although these data mining methods can capture the comorbidity of diseases, they often ignore complex relations, such as time-gap and pattern of disease occurrence.

Countries worldwide and China, in particular, are undergoing healthcare reforms that aim to improve the quality of healthcare in a cost-effective manner. One way to achieve this aim is to understand the process of chronic disease progression and take protective measures to minimize healthcare costs from both the patients' and providers' perspective. How can the progression of chronic diseases be understood? The answer may be found in analyzing the healthcare insurance records of chronic patients who provide detailed process information during each hospital admission. Records collected from each hospital admission include a list of drugs and treatments used during this admission. The term "admission" refers to a time interval that represents an episode of care in a hospital setting. Over a certain period of time, a patient can have several admissions, and the patient's history is the sequence of such admissions. Because of the complex and interrelated nature of admission data, we adopt a heterogeneous information network to model the admission data. Chronic progression mining can then be transformed into a hidden pattern mining problem in a heterogeneous information network.

In this paper, we propose a Heterogeneous Network-based Chronic Disease Progression Mining (HNCDPM) method to help us understand the progression of chronic disease, including orphan diseases, detect chronic disease fraud, and reduce healthcare costs. HNCDPM considers the different medication stages of the same disease and obtains two types of rules: the pattern between different periods of different chronic diseases, which indicates the relationship between different types of chronic disease, and the pattern between different stages of the same chronic disease, which shows the clinical path of the chronic disease. These two types of rules can be used to detect chronic disease fraud. Extensive experiments show that our method can outperform the existing methods by 20% in terms of precision.

The rest of the paper is organized as follows. Section 2 reviews the related work on the problem of chronic disease progression and Section 3 describes the problem of chronic disease progression. Section 4 introduces HNCDPM, and Section 5 provides an empirical study of our algorithm with real healthcare insurance data. Finally, Section 6 concludes our work and discusses several interesting research directions.

2 Related Work

This section reviews the existing work on disease progression mining and healthcare fraud detection, which is closely related to the research proposed in this paper.

Healthcare fraud has considerably inflated loss for individuals, entities, and governments. Combating healthcare fraud has turn out to be a vital concern. Hence, several researchers have developed healthcare fraud detection systems. The job of fraud detection systems is to find, detect, and report frauds as they appeared in the system[1–4]. Ko et al.[5] specifically considered only one field, Urology, while using 2012 Content Management System (CMS) data. The authors attempted to determine an estimated saving from a standardized service utilization by analyzing variability among Urologists within the field's service utilization and payment. A study, which uses 2013 CMS dataset, built a machine learning model to detect

when physicians exhibit anomalous behavior in their medical insurance claims[6]. It attempts to determine if, and when, physicians are acting outside the norm of their respective specialty, which could indicate misuse, fraud, or lack of knowledge around billing procedures.

Charlson Comorbidity Index[7] was proposed in 1987 to predict the 10-year mortality of patients by ranking a range of demographic and comorbid conditions, such as heart disease, cancer, and AIDS. The Elixhauser index[8] showed slightly better prediction performance than this index[9], especially when predicting mortality beyond 30 days. Similar models, such as APACHE-II[10] and Mortality Probability Models (MPM)[11], have also been used to assess the condition of ICU patients and determine the aggressiveness of treatment. A more recent approach introduced in healthcare informatics is derived from social network analysis methods[12]. Based on graph theory, these methods treat the healthcare data as complex relations between different entities, including physicians, diseases, and hospitals. The goal of this approach is to mainly understand the interactions between healthcare entities[13], improve collaboration efficiency among physicians[14], map the knowledge structure in healthcare research[15], and understand the progression of chronic comorbidity[16], among others. However, the actual context of healthcare settings, approaches, and entities considered in social network analysis varies widely[17]. Although these methods can help us understand chronic disease progression to some extent, they are often only applicable to a specific kind of chronic disease and lack generality. Besides, they consistently ignore complex relations, such as time-gap and pattern of disease occurrence.

Frequent Subgraph Mining (FSM) is the essence of graph mining. The objective of FSM is to extract all frequent subgraphs in a given data set with occurrence counts above a specified threshold. The straightforward idea behind FSM is to grow candidate subgraphs in either a breadth-first or depth-first manner (candidate generation) and then determine whether the identified candidate subgraphs occur frequently enough in the graph data set for them to be considered interesting (support counting). However, in existing FSM methods, such as gSpan[18], nodes that appear less often than the threshold are removed and do not appear in the obtained frequent subgraph. This issue may result in our inability to obtain information on orphan diseases. Thus, in this paper, we propose the application of Constrained Frequent Subgraph Mining (CFSM), which can maintain rare nodes and mine only subgraphs with a certain structure. Our methods can reduce the size of the candidate subgraph set and remarkably improve computation efficiency.

Complex network communities can be categorized as overlapping or non-overlapping community structures. Of particular interest to this work is the overlapping community, which is a notable feature in many networked systems. In healthcare networks, a disease can belong to multiple groups. Hence, the detection of overlapping communities in complex networks can be expressed as a disease progression mining problem. Among the existing community detection methods currently available, Infomap[19] shows the most rapid calculations and the highest accuracy. This method considers a random walk over the network. The more extensively the nodes are connected one with each other, the more likely the walker will stay within them and, thus, form a community. Analysis of the flows over the network gives access to the underlying community structure. However, Infomap considers only the topological structure of the network; it cannot obtain an optimal community detection result in healthcare networks. Thus, we improve Infomap by combining the semantic information of the nodes and make the community detection results more meaningful.

3 Problem Definition

We denote $P = (p_1, p_2, ..., p_n)$ as a set of patients who have chronic healthcare insurance records during period (T_s, T_e), where T_s indicates the start time of the time range and T_e is the end time of the time range. The purpose of our algorithm is to mine the progression of chronic disease based on these records. We consider the different stages of the same chronic disease and mine two types of progression rules of the disease. One type of rules describes the pattern between different stages of different chronic diseases, which indicates the relationship between different kinds of chronic diseases. The second type of rules describes the patterns between different stages of the same chronic disease, which shows the clinical path of the disease. These two types of rules can be used to help us detect chronic disease fraud.

We propose an HNCDPM method to mine the chronic disease progression and help us detect chronic disease-related healthcare insurance fraud. The

proposed method can be divided into five steps.

Step 1. Construct a health seeking temporal graph for each patient.

Step 2. Mine frequent disease-process subgraphs from the graph set in Step 1 using Constrained Frequent Subgraph Mining (CFSM), and recode the health-seeking temporal graph set with the mined frequent disease-process subgraphs.

Step 3. Construct the base disease progression network by statistical aggregation of the recoded graph set in Step 2.

Step 4. Conduct community detection on the base disease progression network and transform them into chronic disease progression rules.

Step 5. Conduct chronic disease-based healthcare insurance fraud detection according to the rules obtained in Step 4.

4 Heterogeneous Network-Based Chronic Disease Progression Mining

In this section, we introduce the details of our HNCDPM method. First, we present the basic definition of a health-seeking temporal graph and explain how it is constructed.

4.1 Health-seeking temporal graph construction

Suppose the time range of healthcare insurance records is (T_s, T_e), where T_s indicates the start time of the time range and T_e is the end time of the time range.

Definition 1. Health-seeking Behavior. Each health-seeking behavior b_i can be denoted as $b_i = (p, d, t)$, where p is the patient, d denotes the diagnose, and t is the health-seeking time of the health-seeking behavior. As mentioned above, a health-seeking behavior may contain multiple kinds of drugs/treatments. For example, $b_1 = (p_1, \text{dA}, \text{“2016.12.11”})$ indicates that patient p_1 conducted health-seeking behavior b_i on December 11, 2016 and was diagnosed with dA.

Definition 2. Health-seeking Temporal Graph. Health-seeking temporal graph G is a heterogeneous information network with three types of nodes: patient, health seeking-behavior, and process. Three types of edges are observed in G.

The edge between patient node p_i and health-seeking behavior node b_j shows that patient p_i conducts the health-seeking behavior b_j.

The edge between health-seeking behavior node b_u and health-seeking behavior node b_v indicates that b_v occurs after b_u, and the weight of edge e_{uv} is defined as

$$W_{e_{uv}} = \frac{1}{|t_u - t_v| + 1} \tag{1}$$

where t_u and t_v refer to the time of b_u and b_v. The shorter the time interval between b_u and b_j, the less the weight of edge e_{uv}.

The edge between health-seeking behavior node b_j and process m_r indicates that process m_r is used in b_j, and the weight of edge e_{jr} shows the dose of m_r in b_j. Each patient has a health seeking temporal graph G, which can be donated as (p_i, G_i). Figure 1 shows an example of the health-seeking temporal graph of patient p_1.

From Fig. 1, we can see the different health-seeking behaviors of the same patient. Although the diagnoses are the same, they may indicate different periods of the same disease. Thus, we can identify the period of disease according to the process information. Conventional disease progression mining methods consider the same diagnoses as the same disease.

However, they may actually refer to different periods of the same disease. For instance, the first and last health-seeking behaviors in Fig. 1 are both diagnosed as disease A, but they are, in fact, different periods of A. To be more specific, using one kind of hypotensive drug and using two kinds of hypotensive drugs may refer to different periods of hypertension.

To identify whether the two health-seeking behaviors with the same diagnoses refer to the same disease period, we must mine the frequent disease-process patterns.

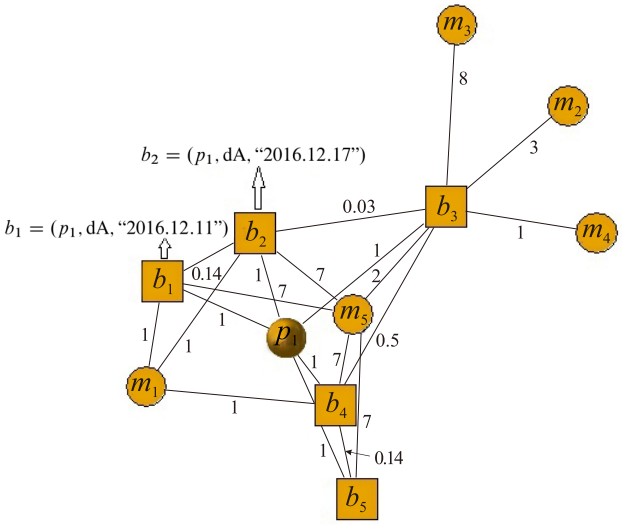


Fig. 1 Example of health-seeking temporal graph of patient p_1. The spherical node indicates a patient with a chronic disease, each square node represents a health-seeking behavior, and each circular node is a process.

4.2 Constrained frequent disease-process subgraph mining

In this section, we consider the diagnosis of each health-seeking behavior. Frequent disease-process pattern mining can be considered as a CFSM problem in the graph set $G^S = \langle G_i : i = 1, 2, \ldots, N_p \rangle$, where N_p is the number of patients.

Definition 3. Frequent subgraph. For a given graph set G^S, if a subgraph g occurs more frequently than the specified threshold, then g is a frequent subgraph of G^S.

Definition 4. Frequent subgraph mining. The objective of FSM is to extract all of the frequent subgraphs in a given data set with occurrence counts above a specified threshold.

The general process of subgraph mining algorithms, such as gSpan, is as follows:

(1) Traverse all graphs and calculate the frequency of all nodes and edges.

(2) Compare the frequency and threshold of all nodes and edges, and remove those with frequencies smaller than the threshold.

(3) Recalculate the frequency of the remaining nodes and edges, order them by frequency, and then recode them by their order.

(4) Conduct submining for each edge in the remaining edge set.

We can infer that, according to the existing frequent subgraph mining method, nodes that show up less frequently than the threshold specified will be removed and do not appear in the obtained frequent subgraph. Unfortunately, this treatment may result in an inability to obtain information on orphan diseases. Orphan diseases also play an important role in our healthcare insurance. To address this challenge, we propose a new CFSM method.

Definition 5. Constrained frequent subgraph mining. A constrained frequent subgraph means the structure of the subgraph is predefined, and the support calculation is redefined as

$$\text{support}(g + v) = \frac{\text{count}((g + v), G^S)}{\min(\text{count}(g, G^S), \text{count}(v, G^S))} \tag{2}$$

We adopt the vertex-increased iteration mode. Here, g indicates the original subgraph, and v indicates the added vertex. If support$(g + v)$ is greater than the predefined threshold, then subgraph $g + v$ is added to the frequent subgraph set.

In this paper, we define the structure of a frequent subgraph as a disease node and multiple process nodes. Then, the process of CFSM can be divided into four steps.

(1) Remove the two-node subgraphs that do not match the predefined structure of the frequent subgraph.

(2) Calculate the support of the filtered two-node subgraphs using the new support definition and remove those subgraphs with supports lower than the threshold.

(3) Conduct submining on the obtained two-node subgraph. In contrast to existing submining processes, our method maintains all of the mined frequent subgraphs instead of the maximal frequent subgraphs.

(4) Return the mined frequent subgraphs.

Figure 2 shows examples of mined constrained frequent disease-process subgraphs. We denote these frequent disease-process subgraphs as $\text{FS}_z, z = 1, 2, \ldots, Z$, where Z is the number of frequent subgraphs mined from CFSM mentioned above.

Based on the mined frequent disease-process subgraph, we can recode the health-seeking temporal graph of patients. Algorithm 1 gives a detail description of our CFSM method. We replace the health-seeking behavior and process nodes with the most similar frequent disease-process subgraph. Figure 3 shows the recoded health-seeking temporal graph of patient p_i in Fig. 1.

To calculate the similarity between health-seeking

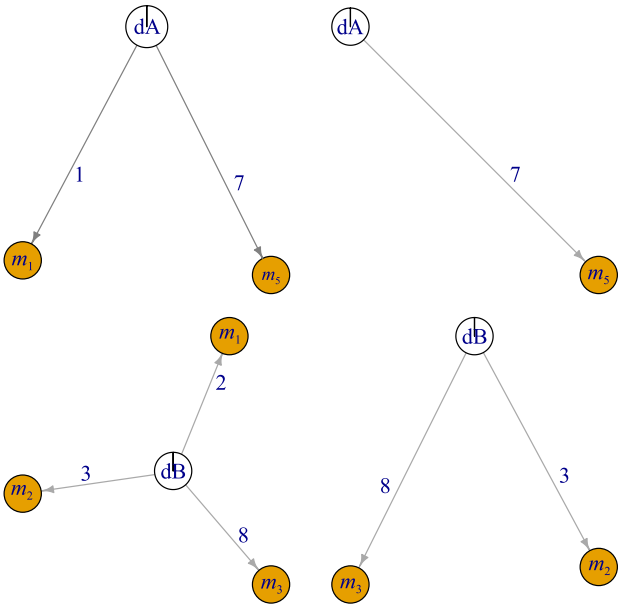


Fig. 2 Examples of mined constrained frequent disease-process subgraphs. Four frequent disease-process subgraphs are shown. The first two are used to diagnose dA, while the last two are used to diagnose dB.

Algorithm 1 Constrained frequent disease-process subgraph mining

Require: Patient set $P = \{p_1, \ldots, p_n\}$, behavior record of each patient during a time period $T_s - T_e$, FS $= \varnothing$

1: **for** p_i in P **do**
2: Construct temporal graph G_i
3: **end for**
4: Filter two-node subgraphs matching predefined structure
5: **for** g_j in filtered two-node subgraphs **do**
6: Calculate support (g_j)
7: **if** support$(g_j) \geqslant$ threshold **then**
8: FS $=$ FS $\cup\, g_j$
9: subMining (g_j) – same as the node-increased iteration in gSpan
10: **end if**
11: **end for**
12: **return** FS

Fig. 3 Recoded health-seeking temporal graph of patient p_i. Here, p_1 indicates patient p_i and the rest of the nodes FS1, FS2, and FS3 represent different frequent disease-process subgraphs.

temporal graphs and frequent subgraphs, we adapt the cosine similarity method. After recoding with the mined frequent disease-process subgraph, different periods of the same disease can be clearly denoted by different subgraphs, and the dimensionality of the graph is reduced to a great extent. As a result, we can obtain more meaningful insights from analyzing the obtained graph set denoted by the mined frequent disease-process subgraph.

4.3 Base disease progression network construction

The base disease progression network is constructed from a recoded graph set using statistical aggregation. In the base network Gbase, each node indicates a mined frequent disease-process subgraph, and the edge between nodes refers to the frequency with which nodes tend to occur sequentially. A node attribute called frequentness, which calculates the number of times the mined frequent disease-process subgraph has occurred for all chronic patients over all admissions, is also obtained. Each edge has two attributes: strength and standard delay. The strength attribute denotes the probability of two nodes occurring in consecutive time, and the standard delay attribute denotes the time span after which the later node occurs after the previous one. This attribute is actually calculated by generating a frequency distribution graph of time intervals and then taking the most significant three-time duration from this distribution.

Figure 4 shows an example of a base disease progression network. The thickness of the edges denotes the attribute strength, while the weight of the edge shows the standard delay.

4.4 Heterogeneous network-based chronic disease progression mining

To understand the process of chronic disease progression, we adopt community detection methods based on the obtained base disease progression network. Nodes that have a close connection with each other tend to be divided into the same community. In contrast to traditional community detection, multiple edges may be found between two nodes in Gbase. Because of the rapid and high accuracy of Infomap[19], in this paper, we improve this algorithm to conduct community detection in Gbase.

Infomap considers only the topological structure of the network; it cannot obtain an optimal community detection result in Gbase. We improve Infomap by combining the semantic information of the nodes to make the community detection result more meaningful.

Infomap uses the probability flow of random walks on a network as a proxy for information flows in a real system and decomposes the network into modules by compressing a description of the probability flow. The result is a map that both simplifies and highlights

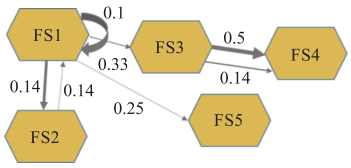


Fig. 4 An example of base disease progression network. Each node represents a different frequent disease-process subgraph. The weight of the edge indicates the probability of two nodes occurring in consecutive time, and the thickness of the edge is the time span after which the later node occurs after the previous one.

the regularities in the structure and their relationships. Based on the community detection result of the original Infomap, we redivide the community combining the semantic information of the nodes. To be more specific, if the nodes in two communities are similar in semantics, the distance between these nodes is reduced and the two communities can be merged into one community.

According to the community detection results of improved Infomap, we can deduce a rule set R with two types of rules: (1) The pattern between different stages of different chronic diseases, which indicates the relationships between different chronic diseases. (2) The pattern between different stages of the same chronic disease, which shows the clinical path of the chronic disease. These two types of rules can be used to detect chronic disease fraud.

4.5 Chronic disease fraud detection

This part of the framework essentially determines the similarity between the base chronic disease network and the healthcare history of a new patient (to check if the patients' records indicate existing chronic disease fraud). This method is called longitudinal node matching, which combines the sequential phases of rule-based and graph theory. The basic principles of this matching consider that the patients' risk for chronic disease fraud decreases if the patient has a high matching rate with the mine rules set R.

For a new patient p_{new}, suppose his health seeking temporal graph is G_{new}. We recode G_{new} with the mined disease-process subgraph and obtain G_{renew}. Then, the probability of fraud of patient p_{new} can be calculated as

$$\text{Pro}(p_{\text{new}}) = 1 - \text{Sim}(G_{\text{renew}}, R) = \frac{\sum_{\forall \text{edge} \in G_{\text{new}}} f(\text{edge}, R)}{|G_{\text{new}}|} \tag{3}$$

$$f(\text{edge}, R) = \begin{cases} 0, & \text{if edge} \in R; \\ 1, & \text{else} \end{cases} \tag{4}$$

$f(\text{edge}, R)$ indicates whether a rule in R can match the edge. If such a rule exists, $f(\text{edge}, R)$ is 1. Otherwise, $f(\text{edge}, R)$ is 0. The less the similarity between G_{new} with G_{base}, the larger probability that patient p_{new} will conduct chronic disease-related health insurance fraud.

5 Experiments

In this section, we apply our framework to actual healthcare insurance records, which include over 40 million admission records of 10 000 patients during the last five years. The dataset used in this experiment was collected from the Dareway Healthcare Insurance Claim System, which is currently used in a certain city of China. Figure 5 shows an example of the original healthcare insurance records. The records indicate the types and amount of process (drug/treatment) prescribed during each admission.

First, we construct a health-seeking behavior temporal graph for each patient. Then, we construct a constrained frequent disease-process subgraph from the graph set. Figure 6 shows some examples of mined frequent disease-process subgraphs. We can see that orphan diseases appear in the mined frequent disease-process subgraphs because our method redefines the support. Then, the orphan disease nodes can be maintained in the candidate frequent set instead of being deleted in the first step.

We recode the temporal graph with the mined frequent disease-process subgraph and construct a

SXH	YLXMBM	YYXMBM	JSXMBH	SXZFBL	XJ	DJ	SL	ZJE	QETC
112	yb1300193	yb1300193_SI	130	0.050000	0.000000	4.600000	2.0000	9.200000	0.000000
113	yb1001033	yb1001033_SI	100	0.000000	0.000000	5.400000	1.0000	5.400000	5.400000
114	yb1001039	yb1001039_SI	100	0.000000	0.000000	6.100000	1.0000	6.100000	6.100000
115	yb1001039	yb1001039_SI	100	0.000000	0.000000	6.100000	2.0000	12.200000	12.200000
116	yb1001039	yb1001039_SI	100	0.000000	0.000000	7.200000	1.0000	7.200000	7.200000
117	yb1001036	yb1001036_SI	100	0.000000	0.000000	7.000000	1.0000	7.000000	7.000000
118	yb1001033	yb1001033_SI	100	0.000000	0.000000	7.400000	1.0000	7.400000	7.400000
119	yb1001033	yb1001033_SI	100	0.000000	0.000000	5.400000	1.0000	5.400000	5.400000
120	yb1001033	yb1001033_SI	100	0.000000	0.000000	6.500000	2.0000	13.000000	13.000000
121	yb1300193	yb1300193_SI	130	0.050000	0.000000	4.600000	4.0000	18.400000	0.000000
122	yb1001033	yb1001033_SI	100	0.000000	0.000000	7.400000	1.0000	7.400000	7.400000
123	yb1001039	yb1001039_SI	100	0.000000	0.000000	6.400000	1.0000	6.400000	6.400000
124	yb1001039	yb1001039_SI	100	0.000000	0.000000	0.450000	1.0000	0.450000	0.450000
125	yb1000977	yb1000977_SI	100	0.000000	0.000000	12.200000	2.0000	24.400000	24.400000
126	yb1001038C	yb1001038C_SI	100	0.000000	0.000000	1.700000	1.0000	1.700000	1.700000
127	yb1000724C	yb1000724C_SI	100	0.100000	0.000000	14.700000	1.0000	14.700000	0.000000
128	yb1000697B	yb1000697B_SI	100	0.000000	0.000000	7.240000	2.0000	14.480000	14.480000
129	yb1001039	yb1001039_SI	100	0.000000	0.000000	0.450000	1.0000	0.450000	0.450000
130	yb1000699	yb1000699_SI	100	0.050000	0.000000	3.600000	1.0000	3.600000	0.000000
131	yb1001038C	yb1001038C_SI	100	0.000000	0.000000	1.700000	1.0000	1.700000	1.700000
132	yb1000699	yb1000699_SI	100	0.050000	0.000000	3.600000	2.0000	7.200000	0.000000

Fig. 5 Example of original healthcare insurance records. The process information of a patient is recorded during every admission.

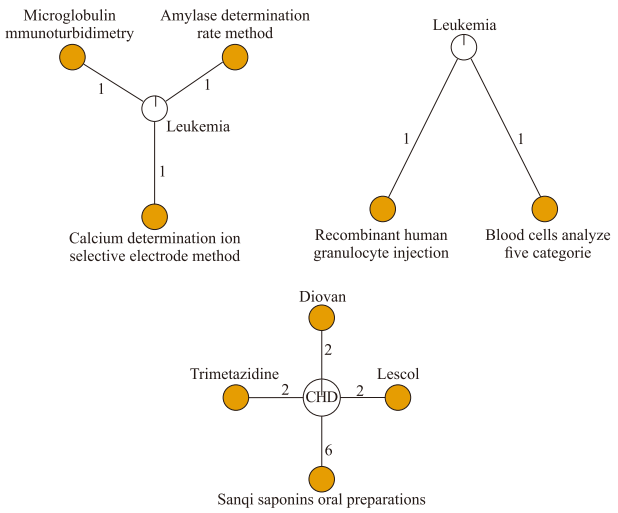


Fig. 6 Mined frequent disease-process subgraph. Although leukemia is a kind of orphan disease, it still appears in the mined frequent disease-process subgraphs.

base chronic disease network G_{base}. Next, we conduct community detection using the improved Infomap on G_{base}. Finally, we deduce rules from the mined community detection results. Table 1 shows an example of rules deduced from the community detection results. Figure 7 shows some rules mined from the conventional Apriori algorithm. All of the orphan disease nodes are deleted in the first step because their support is very low, and they do not appear in the rules.

Table 1 reveals that our method is able to mine the clinical paths of orphan diseases, such as leukemia, which is impossible in traditional subgraph mining methods because orphan disease nodes are removed in the first step in such methods. Our method also considers the different medication stages of the same disease, which is meaningful in efforts to understand the process chronic disease progression. Existing disease progression mining methods consider the same diagnosis as the same disease and ignore the different medication stages of the same disease.

Table 1 Example of rules deduced from the community detection result.

Rule type	Rule
Disease relation	Hypertension — 6 months ⟶ Hypertension phase two
Disease relation	Hypertension — 3 months ⟶ Coronary heart disease
Clinical path	Uremia:renal dialysis — 3 days ⟶ Local renal dialysis — 4 days ⟶ Hemodiafiltration
Clinical path	Leukemia: Index check — 4 days ⟶ Recombinant human granulocyte injection (once a day) — 3 days ⟶ Quantity of leukocyte

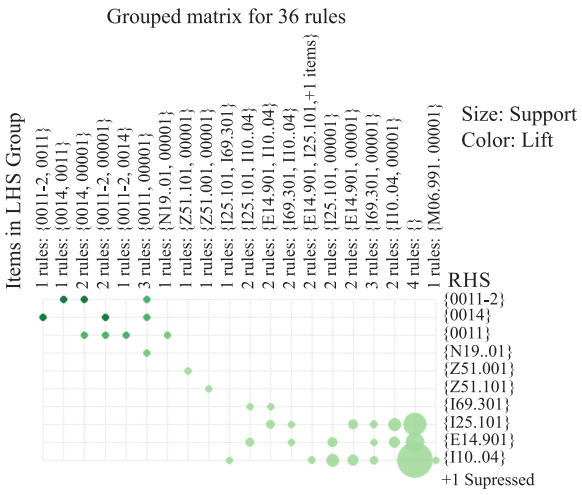
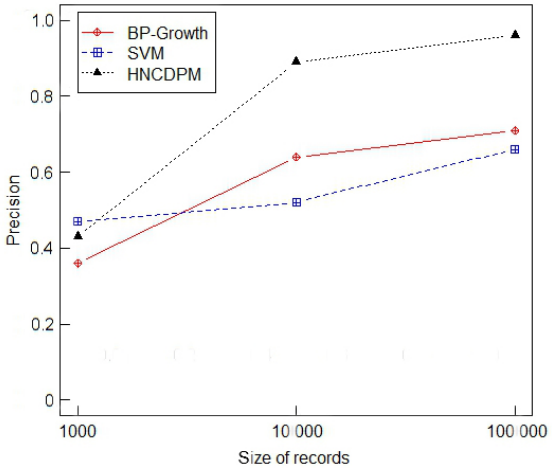


Fig. 7 Rules mined from the conventional Apriori algorithm. All of the orphan diseases are deleted in the first step and do not appear in the mined rules.

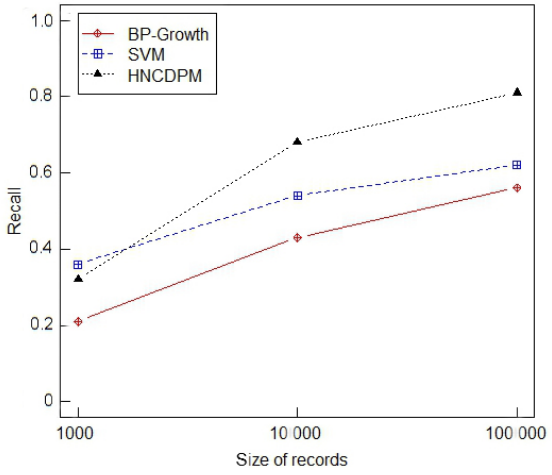
Then, we detect chronic disease-related healthcare insurance fraud with the obtained rules. We adopt commonly used metrics including precision, recall, and F-measure to evaluate how effectively each approach identifies fraudulent medical insurance claims. Time cost refers to how much time an approach takes on a standard hardware/software system to produce the fraud detection results. Precision $= \dfrac{t_p}{t_p + f_p}$ is the fraction of claims identified as fraudulent which are indeed fraudulent. Recall $= \dfrac{t_p}{t_p + f_n}$ is the fraction of all fraudulent claims that have been correctly identified. F-measure $= \dfrac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ is the weighted harmonic mean of precision and recall. Here, t_p (true positive) is the number of claims correctly classified as fraudulent, f_p (false positive) is the number of claims incorrectly classified as fraudulent, and f_n (false negative) is the number of claims incorrectly classified as non-fraudulent. Figure 8 shows the performance of our developed HNCDPM versus those of several existing insurance fraud detection methods (BP-Growth[20] and Support Vector Machine (SVM)[21]). HNCDPM, which considers different medication stages of the same chronic disease, demonstrates significantly better performance than the other comparison approaches. The recall values of the studied methods increase with increasing record size. HNCDPM is able to correctly identify about 80% of the fraudulent claim records from the datasets. Hence, our method can outperform the existing methods by 20% in terms of F-measure.

6 Conclusion

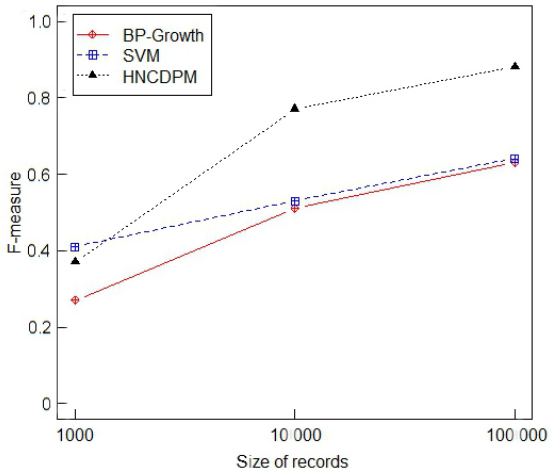
This paper proposes HNCDPM to help detect health insurance fraud. The developed method helps us understand the progression of chronic disease, including orphan diseases, and is helpful in detecting chronic disease-related fraud and reducing healthcare costs. HNCDPM considers different medication periods of the same disease and produces two types of rules: the pattern between different stages of different chronic diseases, which indicates the relationship between different types of chronic disease, and the pattern between different stages of the same chronic disease, which shows the clinical path of the disease. These two types of rules can be used to help detect chronic disease fraud. Extensive experiments show that our method can



(a) Precision



(b) Recall



(c) F-measure

Fig. 8 Performance of HNCDPM compared with those of existing fraud detection methods. Our method shows an improved F-measure of over 20% compared with existing methods.

outperform the existing methods by over 20% in terms of F-measure.

Acknowledgment

This work was partially supported by the National Key Research and Development Plan (No. 2016YFB-1000602), Science and Technology Development Plan Project of Shandong Province (No. 2016GGX101034), Shandong Province Independent Innovation Major Special Project (No. 2016ZDJS01A09), and Taishan Industrial Experts Programme of Shandong Province (Nos. tscy20150305 and tscy20160404).

References

[1] S. S. Waghade and A. M. Karandikar, A comprehensive study of healthcare fraud detection based on machine learning, *Int. J. Appl. Eng. Res.*, vol. 13, no. 6, pp. 4175–4178, 2018.

[2] H. Joudaki, A. Rashidian, B. Minaei-Bidgoli, M. Mahmoodi, B. Geraili, M. Nasiri, and M. Arab, Using data mining to detect health care fraud and abuse: A review of literature, *Glob. J. Health Sci.*, vol. 7, no. 1, pp. 194–202, 2015.

[3] R. A. Bauder and T. M. Khoshgoftaar, A novel method for fraudulent Medicare claims detection from expected payment deviations (application paper), in *Proc. 17th Int. Conf. Information Reuse and Integration* (*IRI*), Pittsburgh, PA, USA, 2016, pp. 11–19.

[4] H. Joudaki, A. Rashidian, B. Minaei-Bidgoli, M. Mahmoodi, B. Geraili, M. Nasiri, and M. Arab, Improving fraud and abuse detection in general physician claims: A data mining study, *Int. J. Health Policy Manag.*, vol. 5, no. 3, pp. 165–172, 2016.

[5] J. S. Ko, H. Chalfin, B. J. Trock, Z. Y. Feng, E. Humphreys, S. W. Park, H. B. Carter, K. D. Frick, and M. Han, Variability in Medicare utilization and payment among urologists, *Urology*, vol. 85, no. 5, pp. 1045–1051, 2015.

[6] R. A. Bauder, T. M. Khoshgoftaar, A. Richter, and M. Herland, Predicting medical provider specialties to detect anomalous insurance claims, in *Proc. 28th Int. Conf. Tools with Artificial Intelligence (ICTAI)*, San Jose, CA, USA, 2016, pp. 784–790.

[7] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation, *J. Chron. Dis.*, vol. 40, no. 5, pp. 373–383, 1987.

[8] A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey, Comorbidity measures for use with administrative data, *Med. Care*, vol. 36, no. 1, pp. 8–27, 1998.

[9] M. T. A. Sharabiani, P. Aylin, and A. Bottle, Systematic review of comorbidity indices for administrative data, *Med. Care*, vol. 50, no. 12, pp. 1109–1118, 2012.

[10] D. T. Wong and W. A. Knaus, Predicting outcome in critical care: The current status of the APACHE prognostic scoring system, *Can. J. Anaesth.*, vol. 38, no. 3, pp. 374–383, 1991.

[11] M. J. Breslow and O. Badawi, Severity scoring in the critically ill: Part 1—Interpretation and accuracy of outcome prediction scoring systems, *Chest*, vol. 141, no. 1, pp. 245–252, 2012.

[12] M. Baglioni, S. Pieroni, F. Geraci, F. Mariani, S. Molinaro, M. Pellegrini, and E. Lastres, A new framework for distilling higher quality information from health data via social network analysis, in *Proc. 13th Int. Conf. Data Mining Workshops*, Dallas, TX, USA, 2013, pp. 48–55.

[13] J. G. Anderson, Evaluation in health informatics: Social network analysis, *Comput. Biol. Med.*, vol. 32, no. 3, pp. 179–193, 2002.

[14] S. Uddin, A. Khan, and M. Piraveenan, Administrative claim data to learn about effective healthcare collaboration and coordination through social network, in *Proc. 48th Hawaii Int. Conf. System Sciences*, Kauai, HI, USA, 2015, pp. 3105–3114.

[15] S. Uddin, A. Khan, and L. A. Baur, A framework to explore the knowledge structure of multidisciplinary research fields, *PLoS One*, vol. 10, no. 4, p. e0123537, 2015.

[16] H. Luijks, T. Schermer, H. Bor, C. Van Weel, T. Lagro-Janssen, M. Biermans, and W. De Grauw, Prevalence and incidence density rates of chronic comorbidity in type 2 diabetes patients: An exploratory cohort study, *BMC Med.*, vol. 10, p. 128, 2012.

[17] D. Chambers, P. Wilson, C. Thompson, and M. Harden, Social network analysis in healthcare settings: A systematic scoping review, *PLoS One*, vol. 7, no. 8, p. e41911, 2012.

[18] X. F. Yan and J. W. Han, gSpan: Graph-based substructure pattern mining, in *Proc. 2002 IEEE Int. Conf. Data Mining*, Maebashi, Japan, 2002, pp. 721–724.

[19] M. Rosvall and C. T. Bergstrom, Maps of random walks on complex networks reveal community structure, *Proc. Natl. Acad. Sci. USA*, vol. 105, no. 4, pp. 1118–1123, 2008.

[20] X. Y. Li, H. H. Cao, E. H. Chen, H. Xiong, and J. L. Tian, BP-growth: Searching strategies for efficient behavior pattern mining, in *Proc. 13th Int. Conf. Mobile Data Management*, Bengaluru, India, 2012, pp. 238–247.

[21] J. A. K. Suykens, Support vector machines: A nonlinear modelling and control perspective, *Eur. J. Control*, vol. 7, nos. 2&3, pp. 311–327, 2001.

Qingzhong Li received the PhD degree from Chinese Academy of Sciences in 2000. He is a professor and PhD supervisor in Shandong University. His main research interests include data science and data analysis. He has presided over more than 20 national, provincial, and ministerial level projects, published more than 80 papers in important journals. He is a member of China Computer Society Database Committee, China Computer Society System Software Committee, and China Computer Society Electronic Government and Office Automation Committee, and he is the academic leader of Jinan Big Data Integration and Intelligent Analysis Outstanding Innovation Team.

Lizhen Cui received the BS, MS, and PhD degrees from Shandong University in 1999, 2002, 2005, respectively. He was a visiting scholar in Georgia Institute of Technology in 2013. He is a professor in the School of Computer Science and Technology at Shandong University. He published over 20 papers in journal and conference proceedings. His research interests include workflow and distributed data management for cloud computing and service computing. He is a member of China Computer Society Professional Committee of the Standing Committee of Service Computing, Database Committee, Collaborative Computing Committee, and China Computer Federation Young Computer Scientists & Engineers Forum (CCF YOCSEF).

Chenfei Sun is a PhD candidate in Shandong University from 2014. She received the BS and MS degrees from Shandong University in 2012 and 2015, respectively. Her research interests include data mining and fraud detection.

Hui Li received the BS degree from Hohai University in 1989, and MS and PhD degrees from Shandong University in 2001 and 2010, respectively. She is an associate professor in the School of Computer Science and Technology at Shandong University. Her research interests include business process management and distributed data management for cloud computing.

Yuliang Shi received the PhD degree from Fudan University in 2006, and obtained BS and MS degrees from Shandong University in 2000 and 2003, respectively. He is an associate professor in the School of Computer Science and Technology at Shandong University. He published over 20 papers in journal and conference proceedings. His research interests include cloud computing, large data management, and business process management. He is the deputy director of Software and Data Engineering Research Center in Shandong University.