A Hybrid Approach for Detecting Fraudulent Medical Insurance Claims

(Extended Abstract)

Chenfei Sun[1], Yuliang Shi[1], Qingzhong Li[1], Lizhen Cui[1], Han Yu[2], Chunyan Miao[2]
[1]School of Computer Science and Technology, Shandong University, Jinan, China
[2]Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798
sun.chenfei@163.com, {shiyuliang, lqz, clz}@sdu.edu.cn, {han.yu, ascymiao}@ntu.edu.sg

ABSTRACT

Medical insurance frauds are causing huge losses for public healthcare funds in many countries. Detecting medical insurance frauds is an important and difficult challenge. Because of the complex granularity of data, existing fraud detection approaches tend to be less effective in terms of recalling fraudulent claim behaviours. In this paper, we propose a Hybrid Fraud Detection Approach (HFDA) to address this problem, which is compared with four state-of-the-art approaches using a real-world dataset. Extensive experiment results show that the proposed approach is significantly more effective and efficient.

Keywords

Behaviour patterns; outliers; evidence theory; fraud

1. INTRODUCTION

In today's world, population aging has brought new challenges to the healthcare system [1]. As many people live longer, governments around the world are increasing their spending on healthcare. Medical insurance fraud is a serious threat to public funds and public trust [5]. This has motivated many researchers to develop fraud detection technologies.

As more and more healthcare insurance systems move online, many claim-related activities take place over the Internet and can thus be tracked by the claim system. Therefore, the combination of behaviour trajectory big data and machine learning techniques offer promising solutions to the medical insurance fraud problem. The granularity of behaviours, represented by mathematical symbols, can be very fine-grained. This can result in some behaviours belonging to the same category being represented by different sets of symbols. Due to the curse of cardinality, existing pattern recognition based fraud detection methods cannot accurately identity latent patterns which are more effectively identified at a coarser behaviour granularity [3].

In this paper, we propose a Hybrid Fraud Detection Approach (HFDA) which can detect both unusual categories of claim behaviours and unusual frequencies of claim behaviours simultaneously. Our proposed approach is compared against four state-of-the-art approaches using a large-scale real-world dataset containing more than 40 million medical claim activity records from over 40,000 users of the medical insurance claim system used in Zibo, Shandong, China. Extensive experiments show that the proposed hybrid approach is significantly more effective and efficient.

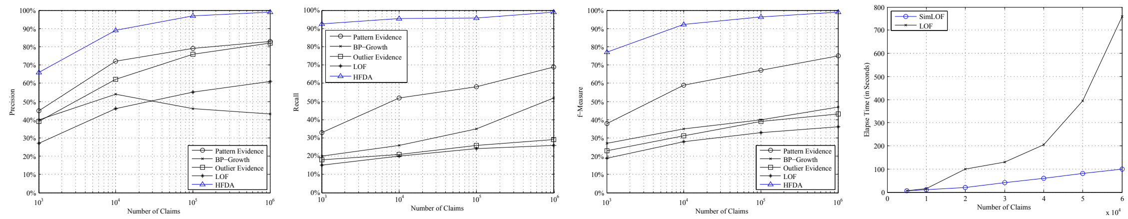
Appears in: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016), J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.), May 9–13, 2016, Singapore.*

2. THE PROPOSED APPROACH

In a medical insurance claim system, there can be millions of transactions from a large number of users. In order to make better sense of the users' actions, it is advantageous to organize the transactions into behaviour sequences.In this paper, we use the term *behaviour*, b, to refer to an activity, a, performed by a subject, s, at a given point in time, t. Thus, $b = \langle s, a, t\rangle$. An example behaviour b_1 can be ⟨"ID123456", "received asthma medications","2012/10/15 14:30:08"⟩. A *behaviour sequence* refers to a collection of behaviours carried out by the same subject in chronological order over a given period of time. It is can be represented conceptually as $BS = (b_1 \to b_2 \to ... \to b_m)$, where $b_1, b_2, ..., b_m$ represent different behaviours from the same subject.

We denote a behaviour sequence as a weighted behaviour graph, $G(V, E, W)$, where V, E and W represent the set of vertices, edges and weights of the graph, respectively. Every vertex represents a distinct behaviour. Vertex i corresponds to b_i and vertex j corresponds to b_j in a set of behaviours. $d(i, j)$ represents the distance between vertices i and j in a behaviour sequence. An edge e_{ij} between two vertices i and node j only exists if $d(i, j) < r$ more than once across multiple behaviour sequences in a behaviour sequence set $\mathbf{\Phi}$. In this case, r is a predefined threshold for neighbors to be regarded as relevant. The weight of the edge between vertices i and j is defined as $w_{ij} = \sum_{\forall BS_n, 1 \le n \le N} \frac{1_{[d(i,j)<r]}}{N}$ where N is the number of behaviour sequences in a behaviour sequence set $\mathbf{\Phi}$. $1_{[\text{condition}]}$ is a function which evaluates to 1 if [condition] is true. Otherwise, it evaluates to 0. The larger the value of w_{ij}, the more often b_i and b_j appear close to each other in the behaviour sequences. Thus, w_{ij} indicates the closeness between behaviours b_i and b_j across multiple behaviour sequences.

Based on original behaviour sequences alone, it is hard to define the distances between the sequences or extract useful features. When the behaviour sequences are composed of a large number of behaviours, it is difficult to cluster the behaviour sequences. One way to tackle this problem is to



(a) precision (b) recall (c) f-measure (d) time cost

Figure 1: Experiment Results.

represent the behaviour sequences by behaviour classes. By denoting the behaviour sequences in the form of behaviour classes, repeated subsequences can be observed more clearly and sequential features can be more meaningful.

In order to reduce the dimensionality of the myriad of behaviours in a typical medical insurance claim system, we need to embed the behaviours into a lower dimension. Manifold learning [2] is the process of estimating a low-dimensional structure which underlies a collection of high-dimensional data. To carry out the embedding, we take advantage of the manifold embedding of a graph.

As medical insurance claim data for each individual user tend to be sparse, peer group comparison is favoured over self-comparison when it comes to outlier-based fraud detection. The outlier based fraud probability, p', for an incoming claim, c, can be obtained by computing the density of known fraudulent claims which share some features with c. The top K nearest neighbors of c in the historical claims data (i.e., $x_1, x_2, ..., x_K$) are selected for the calculation as $density(c) = \frac{1}{K}\sum_{i=1}^{k} p(x_i)$ where $p(x_i)$ is the probability density function of x_i.

Through this process, we obtain two sources of evidence to determine if an incoming claim is fraudulent: 1) *behaviour pattern-based evidence*, which represents unusual categories of behaviours; and 2) *outlier-based evidence*, which represents unusual frequencies of behaviours. In order to combine these two sources of evidence, we leverage on the Dempster-Shafer Evidence Theory [4]. Since the two sources of evidence are independent from each other, we base our proposed approach on Dempster's Rule of Combination [7].

3. EXPERIMENTAL EVALUATIONS

In this section, we evaluate the performance of the proposed approach using a real-world dataset. The dataset used in this experiment is collected from the Dareway Medical Insurance Claim System which is currently used in Zibo City, Shandong Province, China. The dataset contains medical insurance claim activities from over 40,000 users.

We adopt *precision*, *recall*, *f-measure* and *time cost* to evaluate how effectively each approach identifies fraudulent medical insurance claims. It can be observed from Figures 1(a), 1(b) and 1(c) that the proposed approach, which combines pattern-based evidence and outlier-based evidence, can achieve significantly better performance than other comparison approaches. BP-Growth proposed optimizing strategies for association rule mining for behaviour pattern analysis. However, because of the curse of cardinality, BP-Growth is not effective in mining large-scale frequent itemsets consisting of more than two types of behaviours. The proposed approach is significantly more efficient, too (Figure 1(d)).

4. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a hybrid approach to detect both unusual categories of claim behaviours and unusual frequencies of claim behaviours simultaneously. The proposed approach is evaluated against four state-of-the-art approaches using a large-scale real-world dataset. Extensive experiments show that the proposed approach is significantly more effective and efficient than existing approaches. In future, we will infuse human factors such as computational curiosity [6, 8] into the approach to build decision support agents.

Acknowledgements

This research is supported, in part, by the National Natural Science Foundation of China under Grant No. 61572295, 61573212, 61272241; the Natural Science Foundation of Shandong Province of China under Grant No. ZR2014FM031, ZR2013FQ014; the Shandong Province Independent Innovation Major Special Project No. 2015ZDXX0201B03; the Shandong Province key research and development plan No. 2015GGX101015, 2015GGX101007; the Fundamental Research Funds of Shandong University No. 2015JC031; the National Research Foundation Singapore under its Interactive Digital Media (IDM) Strategic Research Programme; and the Lee Kuan Yew Post-Doctoral Fellowship Grant.

REFERENCES

[1] Y. Cai, Z. Shen, S. Liu, H. Yu, *et. al.* An agent-based game for the predictive diagnosis of Parkinson's disease. In *AAMAS*, pages 1663–1664, 2014.

[2] C.-C. Chiu and S. Marsella. Gesture generation with low-dimensional embeddings. In *AAMAS*, pages 781–788, 2014.

[3] J. Liu, E. Bier, A. Wilson, T. Honda, *et. al.* Graph analysis for detecting fraud, waste, and abuse in healthcare data. In *IAAI*, pages 3912–3919, 2015.

[4] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, 1976.

[5] Z. Shen, H. Yu, C. Miao, and J. Weng. Trust-based web service selection in virtual communities. *Web Intelligence and Agent Systems*, 9(3):227–238, 2011.

[6] Q. Wu, X. Han, H. Yu, Z. Shen, and C. Miao. The innovative application of learning companions in virtual singapura. In *AAMAS*, pages 1171–1172, 2013.

[7] R. R. Yager. On the dempster-shafer framework and new combination rules. *Information Sciences*, 41(2):93–137, 1987.

[8] H. Yu, Z. Shen, C. Miao, and A.-H. Tan. A simple curious agent to help people be curious. In *AAMAS*, pages 1159–1160, 2011.