A Study of Complication Identification Based on Weighted Association Rule Mining

Zhijun Yan, Kai Liu, Meiming Xing, Tianmei Wang, Baowen Sun

A Study of Complication Identification Based on Weighted Association Rule Mining

Zhijun Yan[1], Kai Liu[1], Meiming Xing[1], Tianmei Wang[2], and Baowen Sun[2]

[1] School of Management and Economics, Beijing Institute of Technology, Beijing, China
{yanzhijun, 2120141521, 2120131555}@bit.edu.cn
[2] School of Information, Central University of Finance and Economics, Beijing, China
{wangtianmei, sunbaowen}@cufe.edu.cn

Abstract. With the fast development of big data technology, data mining algorithms are widely used to process the medical data and support clinical decision-making. In this paper, a new method is proposed to mine the disease association rule and predict the possible complications. The concept of disease concurrent weight is proposed and Back Propagation (BP) neural network model is applied to calculate the disease concurrent weight. Adopting the weighted association rule mining algorithm, diseases complication association rule are derived, which can help to remind doctors about patients' potential complications. The empirical evaluation using hospital patients' medical information shows that the proposed method is more effective than two baseline methods.

Keywords: Weighted association rule, Complications, Data mining, Neural network

1 Introduction

Many hospitals use the Electronic Health Records system (EHRs) to integrate different kinds of medical data [1, 2]. The data mining technology is widely applied on the EHRs to support clinical decision, including drug supervision [3], complication identification [4], disease prediction [5], patient stratification [6] and other related directions [7]. Among them, medical complication identification gets special attention. Complications refer to diseases caused in the process of another disease. It is difficult to discover the potential relationship between disease and its complication based on a small dataset. The emerging of big data analytic technology makes it feasible to identify the possible complications by mining the massive data in EHRs.

Previous studies on complication identification normally focused on the specific disease and adopted the data mining algorithms to find some potential complications for the specific disease. However, these researches rarely consider the complications of different diseases together and ignore individual disease's features. Aiming to mine the potential complication relationship in the EHRs, this study firstly proposes the concurrent weight of diseases to depict the possibility that one disease will be a complication of other diseases. Based on the concurrent weight, we adopt the weighted association rule mining algorithm to mine potential complication relationship and predict the possible complications for clinical decision-making.

The rest of the paper is organized as follows. Section 2 introduces the related works on complication identification. Then, the proposed complication mining method is presented in Section 3, followed by the experiment results in Section 4. Finally, we discuss major research findings and future research of this study in Section 5.

2 Related Works

Complications represent the concurrence and association relationships among different diseases. They normally affect the patients seriously and cost much higher medical expenditure. In order to identify the possible complications for the disease, various data mining techniques are applied in complication identification.

Roque et al. [8] uses the text mining technology to analyze the electronic patient records of a Danish Psychiatric Hospital. By extracting phenotype information from electronic patient records, they analyze the disease co-occurrence relationship. Hanauer et al. [9] uses the Molecular Concept Map (MCM) algorithm to mine the electronic medical database and successfully find some interesting new diseases associations. Holmes et al. [10] use Application for Discovering Disease Associations using Multiple Sources (ADAMS) to identify the co-morbidities of the rare diseases Kaposi sarcoma, toxoplasmosis, and Kawasaki disease. By incorporating textual information from PubMed and Wikipedia, they find some rare or previously unreported associations.

As one of the most popular data mining algorithms, Association Rule Mining (ARM) is widely used in complication identification of a certain disease. Tai and Chiu [11] focuses on the complications identification of Attention Deficit/Hyperactivity Disorder (ADHD). By employing the association rule mining algorithm, they find that ADHD case group has apparently higher risk of comorbidity with psychiatric comorbidity than with other physical illnesses. Kim et al. [12] analyze the complications of type 2 diabetes mellitus. Based on the medical data of 411,414 patients from 1996 to 2007, they develop the Dx Analyze tool to clean data and reveal associations of comorbidity. The results show that association rule mining was practical for complication studies. Shin et al. [13] use Aprior algorithm and Clementine program to analyze the data of 5,022 patients with essential hypertension. The strong associations between hypertension, non-insulin-dependent diabetes mellitus and cerebral infarction are mined. Moreover, based on a large amount of data, Wright et al. [14] also use the association rule mining method to find out the association among different diseases and laboratory results. The results show that association rule mining is a useful tool for identifying clinically accurate associations and has a better performance over other knowledge-based methods.

In summary, some prior studies make great efforts to identify complication relationship by adopting association rule mining technique. However, they don't consider the different roles of different diseases on the complication identification, which can improve the mining performance. At the same time, patients' medical history is critical information for disease diagnosis and should be incorporated in mining complication relationship.

3 Methods

In this study, we firstly define the concurrent weight to evaluate the possibility that a disease develops as a complication of other diseases. Then we adopt the Back Propagation (BP) neural network to derive the concurrent weight of diseases and identify the disease complication relationship based on the weighted association rule mining algorithm. Finally, the list of possible complications is recommended to support clinical decision-making.

3.1 The Concurrent Weight of Diseases

Prior medical experience is very helpful to predict the future health status and possible diseases. We assume that some diseases will appear as complication more frequent than others. A concurrent weight is proposed to represent the possibility that a disease becomes the complication of other diseases. The higher the disease's concurrent weight, the more likely it is the complication of certain diseases.

Assume that the disease set is *D*, and every item in *D* is one disease. Each disease d_i also has a corresponding set C_i to describe its known complications. So the concurrent weight of the disease d_i can be defined as follows:

$$W(d_i) = \frac{\sum_{j=1}^{n} m_{ij}}{n} \tag{1}$$

where $w(d_i)$ is the concurrent weight of the disease d_i, n is the total number of diseases in the set D, and m_{ij} describes whether the disease d_i appears in the complication set of d_j or not. When the disease d_i appears in the C_j, then m_{ij} is equal to 1, otherwise it is zero.

However, it is impossible to generate a complete complication list for each disease, and we can't directly obtain concurrent weights for all diseases by prior knowledge. Thus we firstly collect the complications of some traditional diseases, then apply some artificial intelligent algorithms to train the collected data and predict the whole set of concurrent weight. For the initial set of traditional diseases and its complications, we used the web crawler to download disease information from domestic professional medical website "Clove Garden" (http://www.dxy.cn/). Then 596 kinds of diseases and their occurrence frequency as complications are collected as the known complication knowledge.

3.2 BP Neural Network Model

The Back Propagation (BP) neural network is one of the most used forward neural networks [15]. We selected the three layer BP neural network to predict the concurrent weight of the diseases. It includes input, output and hidden layer.

Input Layer of the Model. First of all, we describe the disease as a three-dimension vector to characterize the properties of comorbid diseases. The three dimensions

include the position of the disease in the International Classification of Disease (ICD) coding schema, the importance of the disease, and the appearing order in diagnosis.

The position of the disease in the International Classification of Disease (ICD) coding schema represents the location of the disease in the ICD list. The importance of the disease represents the impact of the disease on the patient's recovery process. For inpatients, the hospital will normally record all diseases they have and each disease will be assigned an importance value, which describes how it is important for patients' treatment. During the hospital stay, patients may have several diseases. These diseases will be recorded in the appearing order. By converting each disease to a vector, we used the 596 kinds of diseases and their weights as the input of the BP neural network model. At the same time, each dimension of the disease vector is normalized to makes sure that each value is ranged from 0 to 1.

Output Layer of the Model. The output layer represents the learning result of the BP neural network model. In this study, BP model is adopted to predict the disease concurrent weight. Thus the output layer is the disease concurrent weight.

We firstly discretize the concurrent weights of diseases. The discretization process can effectively avoid the hidden defects in the training dataset and make the model more stable. Moreover, the values of the concurrent weights of diseases are commonly very small considering the large number of diseases, thus we amplify the concurrent weights to get significant results. The discretization calculation formula of the concurrent weights is given as follows:

$$f(w) = \begin{cases} 0.2, & w \leq 0.022 \\ 0.4, & 0.022 < w \leq 0.044 \\ 0.6, & 0.044 < w \leq 0.066 \\ 0.8, & w > 0.066 \end{cases} \tag{2}$$

where *f(w)* is the discretization result of the concurrent weight *w* of a disease.

Then the number of neurons in output layer is set to 2, and the output value of each neuron cell is 0 or 1. Four output values of two neurons: (0,0), (0,1), (1,0) and (1,1) are mapped to four kinds of weights: 0.2, 0.4, 0.6 and 0.8 respectively. Thus we establish the direct connection between input layer and output layer.

Hidden Layer of the Model. The hidden layer of BP model is responsible for the information transformation. It can have one or several layers. As a single hidden layer BP neural network can approximate any nonlinear function with high precision [16], only one hidden layer is set in this study. The number of the neuron in the input and output layer is determined according to the input data and output data. For the hidden nodes number in the hidden layer, although many approaches have been proposed, no one works efficiently for all problems. The most common method is to determine the appropriate number of hidden nodes by experiments performance comparison. Thus we do experiments on a set of values as the number of the neuron in the hidden layer. The value that brings the least training time is the final number of hidden nodes.

3.3 The Weighted Association Rule Mining

Differences among diseases are significant and different disease will have different roles in identifying complications. Thus the weighted association rule mining method [17] is adopted in this study. It attempts to provide a weight to individual items that are not based solely on item support. And thresholds of weighted support and confidence are also defined to measure the significance of the association rules mined.

Similar with the traditional association rule mining algorithm, the support of the item set X is denoted as support(X), if the number of items in X is n, the weighted support of X is:

$$\text{Wsupport(X)} = \text{support(X)} \times \left(\frac{1}{n} \times \sum_{j=1}^{n} w_j\right) \tag{3}$$

The item set X is weighted frequent if the weighted support of X is greater than a predefined minimum weighted support threshold (*wminsup*):

$$\text{Wsupport(X)} \geq wminsup \tag{4}$$

The weighted support of a rule X→Y can be defined as:

$$\text{Wsupport}(\text{X} \rightarrow \text{Y}) = \text{support}(\text{X} \rightarrow \text{Y}) \times \left(\frac{1}{m} \times \sum_{i_j \in (X \cup Y), j=1}^{m} w_j\right) \tag{5}$$

in which m is the total number of items in the set of (X ∪ Y).

The weighted association rule mining algorithm will retrieve all rules X→Y, where X ∪ Y is weighted frequent and whose confidence is greater than or equal to a minimum confidence threshold [18].

In order to improve the algorithm efficiency, we adopt the frequent pattern (FP) tree structure to optimize the weighted association rule mining algorithm [19]. At first, by scanning transaction database and define the minimum weighted support threshold (*wminsup*), the weighted FP-tree is constructed with the weighted potential frequent 1-itemsets. Then the list of potential rules is mined by the weighted association mining approach.

3.4 Complication Prediction

The mined complication association rules among diseases provide valuable information for patient diagnosis. Based on patients' medical history, we can predict patients' possible complication by applying the mined complication rules.

When a patient has a new visit to the hospital and the doctor identifies his/her disease, the patient's main diseases in several latest visits are considered as a disease set, which includes patients' medical history information. The antecedents of mined complication association rules will be browsed to identify whether it contains all diseases in the set. If some rules are matched, the consequents will be displayed as the possible complications. If no rules have antecedents that contain all diseases in the set, the oldest diagnosis will be excluded. Suppose there have *n* diseases in the original set, the set will be *n-1* diseases after the exclusion. Then the antecedents of mined

complication association rules will be browsed again to find the matched rules. Iterate the above steps until some complication rules are matched or the set is empty. If the set is empty, no prediction will be given. Because only few rules include more than 6 diseases, we consider patients' 6 latest visits for prediction in the first step. For the matched rules, the possible complications are listed in the order of confidence of the complication association rules.

4 Evaluation

We have conducted an empirical evaluation of the proposed approach by using electronic medical records from a hospital in China and using the methods proposed by Wright et al. [14] and Hoque et al. [20] as the benchmarks.

4.1 Data Preprocessing

The medical dataset we used is from a hospital in China. The dataset includes the information of inpatients and outpatients. Each patient gets a descriptive and longitudinal record to describe what happened during each visit. The record covers the information of diagnoses, lab test results, medications and procedures. The total number of records is about one million. Because we focus on the disease comorbidity relationship mining, we exclude the patients' data with only one visit to the hospital.

Before mining complication association rules from the dataset, we firstly clean the data. First, the outpatient information is excluded. In the dataset, some medical information of outpatients is missing or incomplete. Moreover, the treatment outcome of outpatients is not recorded and the correctness of the diagnosis can't be evaluated. Second, some doctors may fail to diagnose patients' diseases and patients don't get better after the inpatient treatment. Thus we remove the inpatients information with unclear or uncured treatment results. Third, for those records that missed some important information, we mark them as invalid and exclude from the experiment.

After the data preprocessing, the final qualified diagnosis data includes 253,271 records, and it is related with 24,754 patients and 6,698 diseases.

4.2 Metric and Benchmarks

We used precision (P) as the metric to assess the effectiveness of the proposed approach. Specifically, precision is the fraction of complication predictions that are correct. Higher value of P indicates better performance.

We use the diagnosis with the complication information as test dataset. The dataset includes 1,410 diagnosis records, which include the main disease and complication information. Based on the mined complication association rules, a list of possible complications for each patient can be generated. If the list includes the actual complications, we count that as a correct predication. Thus, the metric P is defined as:

$$P = \frac{Number\ of\ corrected\ prediction}{Number\ of\ test\ data} \times 100\% \quad (6)$$

To evaluate the effectiveness of the proposed approach, the association rule mining (ARM) algorithm introduced by Wright et al. [14] and a rare association rule (RAR) mining approach proposed by Hoque et al. [20] are chosen as benchmarks. Wright et al. applied the tradition association rule mining algorithms in the medical data and confirmed the validity of the association rule mining algorithms. Hoque et al. focused on the improved low-frequency association rule mining and the effectiveness of the generated rules has been validated over several real life datasets.

4.3 Data Analysis and Results

Derive the Concurrent Weight. In the process of deriving disease concurrent weight, there has one important parameter which influences the effect of the BP network. It is the number of neurons in the hidden layer of BP network. Therefore, we choose the training time of the BP network as the evaluation metric and compare the performances with different values of the number of neurons in the hidden layer. By comparing the training time, the number of neurons in the hidden layer is set to 7.

After determining the parameters of BP network, we predict the concurrent weight of the whole 6,698 diseases. The known weights of 596 diseases are inputted to train the model. The weight of other diseases is predicted through the trained model. Finally, the weight of 3,340 diseases is calculated. For those diseases whose weights failed to be predicted, we set their weight as 0.

Weighted Association Rules Mining. Based on the derived concurrent weight, we develop the weighted association rule mining approach by Java language and mine lots of interesting complication relationship. For the mined complication association rules, the number of items in rules is varying. Fig. 1 demonstrates the distribution of the number of items in mined rules. Obviously, most rules include 4 items, which accounts for 21% of the total association rules. The rules that include 3, 4 or 5 items are more than half of the total rules. Surprisingly, there has a little of rules that only include two items.

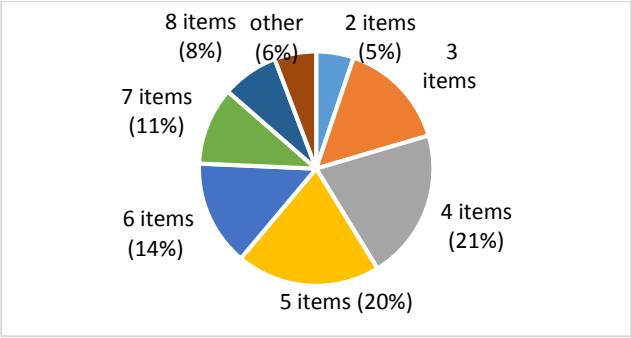

Fig. 1. Number of items in complication rules

We also apply the baseline methods to mine the complication association rules in the experiment dataset. The RAR method is trying the find the rare association rules and the number of generated rules is the biggest, i.e., 96,637. The ARM method derives 42,142 rules. Because some uncommon diseases are weighted, 83,029 rules are mined by the proposed method.

Performance Comparison. We compare three algorithms from two perspectives: processing time and accuracy. For processing time, three algorithms have significant differences on the time consumption for complication association mining. All tests were performed on a PC with 3.4 GHz Intel i7-4770 CPU and 12G RAM. The running time of RAR, RAM, and the proposed method are 262, 196, and 27 minutes, respectively. The predication accuracy of RAR, ARM and the proposed method are 45.3%, 38.5% and 80% respectively. The results show that the proposed method performs better than two baseline methods in processing time and accuracy.

As we mentioned before, the prediction step is to go through all mined complication rules and identify the consequent items of the matched rules as the predicted complications. And the prediction list includes several diseases. However, in the practical scenario, it is important to limit the list length, which can give more insightful suggestions for doctors. Thus we also compare the accuracy of three algorithms with different length of prediction list (Table 1). The results show that the proposed method is better than the baseline methods in three scenarios.

Table 1. The accuracy comparison of three algorithms with different list length

Methods	Accuracy		
	length =10	length =20	Unlimited length
The proposed method	32.4%	37.5%	80%
RAM	13.1%	17.5%	38.5%
RAR	14.9%	21.1%	45.3%

5 Conclusions

This paper focuses on mining disease complication association rules based on medical information. The concept of concurrent weight of diseases is proposed and defined. And the BP neural network model is introduced to predict the weight for all related diseases. Then, we adopt the weighted association rule mining algorithm and FP-tree structure to retrieve the complication relationship among diseases. Based on the mined rules, the potential list of patients' complication can be generated.

This research provides several research contributions. First, we propose a new index to evaluate the importance of different diseases on complication prediction. The defined index, i.e., concurrent weight, can describe the possibility that a disease become a complication of other diseases. Second, we introduce the BP network to predict the disease weight and design the appropriate input and output data. We define the disease information as a three dimension vector and the output of BP network is described by two neurons. By using the BP model, we can deduce a relatively

complete disease knowledgebase. Third, we adopt the weighted association rule approach to mine the diseases association rules. To the best of our knowledge, it is the first time to apply the weighted association rule mining approach in the medical field. And some interesting association rules are retrieved.

There are several limitations of this study, which provide opportunities for future research. First of all, we only focus on the mining of relationship among diseases, which can't describe the complication relationship accurately. Second, due to the scope and complexity of this study, we do not invite medical professionals to evaluate mined complication association rules. Although the derived rules are surely helpful for doctors' decision-making in the real practice, some rules maybe not meaningful or even wrong from the view of clinical research. Third, this research only uses the predication accuracy as the metric.

Acknowledgments. This paper was funded by National Natural Science Foundation of China (Grant No. 71272057, 71572013) and Beijing Natural Science Foundation (Grant No. 9152015).

References

1. Esfandiari, N., Babavalian, M.R., Moghadam, A.M.E., Tabar, V.K.: Knowledge discovery in medicine: Current issue and future trend. Expert Systems with Applications 41(9), 4434–4463 (2014)
2. Shea, S., Hripcsak, G.: Accelerating the use of electronic health records in physician practices. New England Journal of Medicine 362(3), 192–195 (2010)
3. Coloma, P.M. and S. de Bie, *Data mining methods to detect sentinel associations and their application to drug safety surveillance.* Current Epidemiology Reports 1(3), 110–119 (2014)
4. Murff, H.J., FitzHenry, F., Matheny, M.E., Gentry, N., Kotter, K.L., Crimin, K., Dittus, R.S., Rosen, A.K., Elkin, P.L., Brown, S.H.: Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 306(8), 848–855 (2011)
5. Eshlaghy, A.T., Poorebrahimi, A., Ebrahimi, M., Razavi, A.R., Ahmad, L.G.: Using three machine learning techniques for predicting breast cancer recurrence. Journal of Health Medical Information 4(2), 124 (2013)
6. Jensen, P.B., Jensen, L.J., Brunak, S.: Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics 13(6), 395–405 (2012)
7. Kwon, J.H., Lee, , Park, J.Y., Yu, Y.S., Kim, J.P., Shin, J.H., Kim, D.S., Joh, J.W., Kim, D.S., Choi, K.Y.: Actionable Gene Expression-Based Patient Stratification for Molecular Targeted Therapy in Hepatocellular Carcinoma. PLoS One 8(6), e64260 (2013)
8. Roque, F.S., Jensen, P.B., Schmock, H., Dalgaard, M., Andreatta, M., Hansen, T., Søeby, K., Bredkjær, S., Juul, A., Werge, T.: Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS computational biology 7(8), e1002141 (2011)
9. Hanauer, D.A., Rhodes, D.R., Chinnaiyan, A.M.: Exploring clinical associations using '-omics' based enrichment analyses. PLoS One 4(4), e5203 (2009)
10. Holmes, A.B., Hawson, A., Liu, F., Friedman, C., Khiabanian, H., Rabadan, R.: Discovering disease associations by integrating electronic clinical data and medical literature. PloS One 6(6), e21132 (2011)

11. Tai, Y.M., Chiu, H.W.: Comorbidity study of ADHD: applying association rule mining (ARM) to National Health Insurance Database of Taiwan. International journal of medical informatics 78(12), e75–e83 (2009)
12. Kim, H.S., Shin, A.M., Kim, M.K., Kim, Y.N.: Comorbidity study on type 2 diabetes mellitus using data mining. The Korean journal of internal medicine 27(2), 197–202 (2012)
13. Shin, A.M., Lee, I.H., Lee, G.H., Park, H.J., Park, H.S., Yoon, K.I., Lee, J.J., Kim, Y.N.: Diagnostic analysis of patients with essential hypertension using association rule mining. Healthcare Informatics Research 16(2), 7–81 (2010)
14. Wright, A., Chen, E.S., Maloney, F.L.: An automated technique for identifying associations between medications, laboratory results and problems. Journal of Biomedical Informatics 43(6), 891–901 (2010)
15. Kuo, C.F.J., Hsu, C.T.M.Z., Liu, X., Wu, H.C.: Automatic inspection system of LED chip using two-stages back-propagation neural network. Journal of Intelligent Manufacturing 25(6), 1235–1243 (2014)
16. Wang, L., Zeng, Y., Chen, T.: Back propagation neural network with adaptive differential evolution algorithm for time series forecasting. Expert Systems with Applications 42(2), 855–863 (2015)
17. Cai, C.H., Fu, A.W., Cheng, C., Kwong, W.: Mining association rules with weighted items. in Proceedings of Database Engineering and Applications Symposium. IEEE (1998)
18. Pears, R., Yun, S.K., Dobbie, G., Yeap, W.: Weighted association rule mining via a graph based connectivity model. Information Sciences 218(1), 61–84 (2013)
19. Li, T., Li, X.: Novel alarm correlation analysis system based on association rules mining in telecommunication networks. Information Sciences 180(16), 2960–2978 (2010)
20. Hoque, N., Nath, B., Bhattacharyya, D.K.: A New Approach on Rare Association Rule Mining. International Journal of Computer Applications 53(3),1–6 (2012)