2016 International Conference on Computational Science and Computational Intelligence

MH-ARM: a Multi-mode and High-value Association Rule Mining Technique for Healthcare Data Analysis

Libao Yang[*]
Beijing Institute of Technology
Beijing, China
2120141550@bit.edu.cn

Zhe Li
Tianjin University
Tianjin, China
tywzlizhe29121@126.com

Guan Luo[†]
NLPR, Institute of Automation, CAS
Beijing, China
gluo@nlpr.ia.ac.cn

Abstract: The association rules mining process enables the end users to analyze, understand, and use the extracted knowledge in an intelligent system or to support the decision-making processes. To find valuable association rules from a large number of redundant rules, this paper proposes a deeper mining process, multi-mode and high value association rules mining (MH-ARM). This method takes into account the category information, the size of the item set, natural semantics, various metrics, and effective visualization of results. The process can effectively reduce the number of rules and improve the value and accuracy of the rules screened out for auxiliary diagnosis. In the end, the experimental data of rhinitis were analyzed and the effectiveness of the process was verified.

Keywords: association rule mining; multiple modes; auxiliary diagnosis; MH-ARM

I. INTRODUCTION

Association rules mining is widely used in data mining due to its simplicity and comprehensibility, which has been applied to various fields, including marketing [1] and clinical diagnosis [2-4]. In recent years, with the development of the big data research, people rely more heavily on association rules for the understanding of large data sets, and this has led this field to being rather hot [5]. Previous research has focused on the effect of association rules on medical informatics for a long time, but most of them failed to do the data mining process based on different modes and perspectives [6-8].

By mining the relevant information from medical data sets, we can extract valuable knowledge about disease diagnosis and prevention, and provide scientific decision-making for end users. This paper proposes a multiple model of association rule mining and analyzes its application in the data of rhinitis. It will play a very important role in disease prevention, diagnosis, and medical research. High value rules will be accumulated to form a rule base, and ultimately to prepare for the automatic diagnosis and treatment system.

II. Multi-mode and High-value Association Rules Algorithm

It is a well-known fact that association rules mining on real data often results in a huge set of rules with redundant ones. Interesting and useful rules must be picked from the set of generated rules. As so far, the re-mining algorithms of association rules are usually based on the data set itself. In addition, from users' perspectives, they are interested only in a specific topic. It is important to locate desired patterns, evaluate, and select them before presenting them to the users. Therefore, this paper proposes an association rules re-mining algorithm Multi-mode and High-value Association Rules Mining (MH-ARM) based on both data characteristics and user's intention and knowledge.

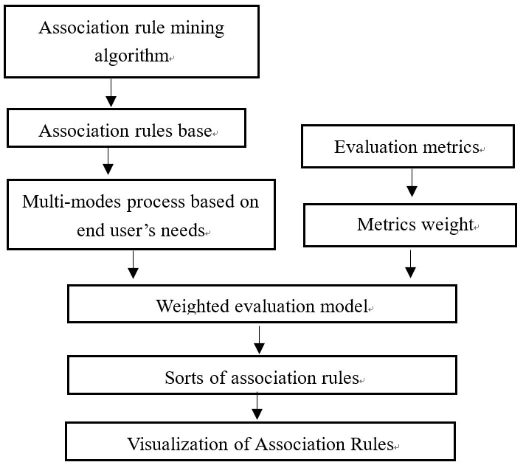
Association rule mining algorithm
Association rules base
Evaluation metrics
Multi-modes process based on end user's needs
Metrics weight
Weighted evaluation model
Sorts of association rules
Visualization of Association Rules

Figure 1. The framework of MH-ARM

[*] Supported by National Natural Science Foundation of China (Grant No. 71432002)
[†] Corresponding author Email: gluo@nlpr.ia.ac.cn

978-1-5090-5510-4/16 $31.00 © 2016 IEEE
DOI 10.1109/CSCI.2016.29

The framework of MH-ARM is shown in Figure 1. Part A is an introduction of aspects of the Re-mining approach in general. Part B mainly focuses on multiple modes of the Re-mining approach, and several solutions to key problems in practice are analyzed in part C. In part D, we construct the value metrics system of association rules.

A. Aspects of the Re-mining approach

The collected medical phenotype data will be mined from the following aspects, generating high value association rules for assistant diagnosis.

- Attribute classification

Classification information is fused with the concrete attributes. This processing can help mine the association relations from a higher level.

- Multi-mode rules mining

By designing multiple mining models, the association rules from different aspects with the interaction of user needs can make the mining results more valuable. This procedure will be introduced in detail in part B.

- The size of item sets of rules

By controlling the size of item sets of rules, lots of obvious rules can be filtered out to some extent. Usually, such rules with a large number of items are not easy to be perceived, and these rules are more innovative and of high value.

- Natural Language Processing

Stop words filtering mechanism of natural language processing can filter out rules with the relevant positive words of the diseases diagnosis, especially for the consequent of rules, such as ''normal hearing''. In this paper, the stop words thesaurus are constructed for our data manually.

- Various metrics of the association rules

On the basis of the measurement of the support-confidence framework, more metrics will be considered, such as Kulczynski (KULC) [9], Imbalance ratio (IR) [10].

- Effective visualization of rules

The results of the current association rules are presented mainly with data tables. The effective visualization of association rules will directly affect the perceived values and use these values for decision making.

B. Multi-mode Association Rules Mining

In order to find the potential cause and effect relationship, statistical convergence, and unique attributes distribution, this paper proposes the following Multi-mode Association Rules Mining based on aspects of Session A.

The traditional association rules mining methods are commonly based on support and confidence. We set the minimum support *minsup* at 0.15, and the minimum confidence *minconf* at 0.7. These two thresholds can be adapted to the user needs.

A, B can be the shared attributes for every instance in the same class, such as diagnosis, or a specific attribute that can be different for each instance, such as gender.

- Mode 1 is designed to find the relationship like if A then B, $R(A \Rightarrow B)$ meets the conditions of *minsup* and *minconf*. At this point, A and B are all selected variables. Therefore, all based-on user needs association rules are found.
- Mode 2, similar to Mode 1, is denoted by Z rather than A for differentiation. The mixed antecedent of attribute classes set can be more practical. When B is a focus on variables (i.e., diagnosis), it can be effective to dig out the corresponding rules with mixed antecedent.
- Mode 3 exploits the data set with a fixed attribute value, to find the relationship if A then B. For example, the fixed attribute: sex, the value: male, $R(A \Rightarrow B, \mathrm{sex} = male)$ are rules that meet the *minsup* and *minconf*. In the same way, $R(Z \Rightarrow B, \mathrm{sex} = male)$ are found by Mode 4 based on Mode 2 & 3.
- Mode 5 is designed to find the convergence value of subsequent attribute B with A for different values (i.e., population). While Mode 6 is to find the unique value of subsequent attribute B with A for different values.

C. Solutions to key problems in practice

- Logical relations among attributes

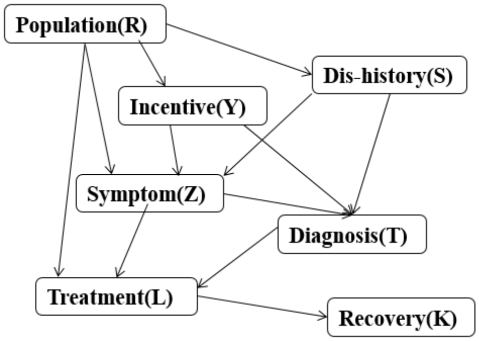

Figure 2 Logical relations among attributes

It is important that a scientific phenotype attribute classification and logical relation network should be acquired before the mining process. Since it is difficult find the logical relations automatically, we have to rely on users with medical knowledge to provide the logical relationships shown in Figure 2.

- The construction of Stop words Thesaurus

The following symptoms stop words thesaurus is constructed based on the phenotypic data of rhinitis patients:

{ ‘No stuffy nose’, ‘Little nose’, ‘No postnasal drip’, ‘Not dry-itchy eyes’ , ‘Not sneeze’, ‘No headache’, ‘Normal sense of smell’, ‘Normal vision’, ‘Hearing-normal’,…}.

D. Value Metrics of Association Rules

The traditional association rules mining methods are commonly based on support and confidence. However, support as a measure index has a shortcoming that is subject to the sparsity of item. Researches show that only the support and confidence metrics are not adequate to filter out the non-interesting association rules [9]. Because of this, we use the correlation and novelty measures to extend the support-confidence framework of the association rules. That is

$$X \Rightarrow Y[\mathrm{sup}, \mathit{conf}, \mathit{corr}, \mathrm{novelty}] \qquad (1)$$

After comparing various correlation metrics, the correlation metrics of coordination of *KULC* and *IR* with zero invariance are appropriate.

KULC: Given two item sets of X and Y, KULC of X and Y can be calculated by (2)

$$KULC(X, Y) = \frac{1}{2}(P(X \mid Y) + P(Y \mid X)) \qquad (2)$$

The *KULC* is influenced by the conditional probability P (X|Y) and P (Y|X), but not by the total number of records. The value of KULC ranges between 0 and 1. The greater the value is, the more relevant X and Y are.

IR: The imbalance ratio between X and Y can be calculated by formula (3).

$$IR(X, Y) = \frac{|\sup(X) - \sup(Y)|}{\sup(X) + \sup(Y) - \sup(X \cup Y)} \qquad (3)$$

If X and Y share the same directions, then *IR* (X, Y) is 0; otherwise, the greater difference between the two directions is, the greater *IR* will be. This ratio is independent of zero transaction as well as the total number of records.

Novelty: These rules are not known to the user, and cannot be inferred from other known rules. A rule $X \Rightarrow Y$ is regarded as novel when P (XY) cannot be inferred from P (X) and P (Y) [11]. The novelty can be redefined as (4).

$$Nov(X \Rightarrow Y) = \frac{P(XY) - P(X)P(Y)}{P(X)P(Y)} \qquad (4)$$

In this paper, we use *support, confidence*, *KULC, IR* and *Novelty* as the value metrics to construct the evaluation system of the association rules model.

The determination of weights is a complicated problem. Different scenes have different weights. And there are many methods to determine the weights such as Delphi Method, analytic hierarchy process (AHP), and the neural network method, etc. And AHP is used here. This paper adopts the "support - confidence - KULC-Nov" making comparison between any two indexes, and construct judgment matrix between two indexes.

Obtained by AHP method, four weights of *C, S, K, N* indexes are 0.482, 0.11, 0.19 and 0.218. Comprehensive assessment of each index also has a variety of methods, namely, the simple weighted sum method, weighted quadrature, and weighted geometric average method. As a result, the comprehensive evaluation coefficient obtained is shown in formula (5).

$$R = 0.482 * C + 0.11 * S + 0.19 * K + 0.218 * N \qquad (5)$$

The total ranking of association rules can be calculated according to the model, getting a unique comprehensive evaluation result, which may facilitate the decision makers to choose more valuable rules for application.

III. Application in Assistant Diagnosis

We realized some experiments using the MH-ARM algorithm to demonstrate it can find more valuable association rules and easier to analyze and visualize the rules. Part A introduces the data of rhinitis patients. Parts of results are analyzed in part B. And part C displays the association rules in effective ways.

A. Data preparation

The data set provided by an internet medical company is a survey data set of rhinitis patients, in which there are 20,000 records with 55 dimensions. Attributes can be divided into the classes and the logical relationship between the classes shown in Figure 2.

B. Results of Multi-mode Association Rules

Our analysis is performed based on the logical relation graph. The parameters of models can be adjusted according to user needs. Each mode has a specific application scene for clinical assistant diagnosis.

1) Mode 1: In the diagnosis of patients with rhinitis, there is a research or diagnostic need to catch the relationship between symptoms. In this case, A and B of mode 1 are symptoms. We got 108 rules under the current experimental framework in total. Figure 3 shows the change of these rules on metrics. Part of the rules extracted are shown in table 1:

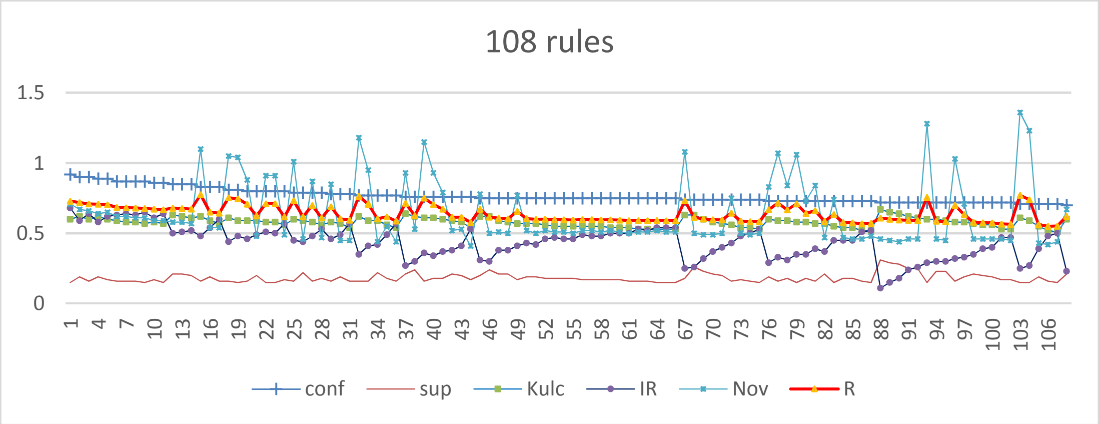

Figure 3 Metrics of 108 Rules

The antecedent and consequent were set as symptoms, so as to find the relationship among attributes of symptoms. The item set of antecedent was set to more than two items but not limited to the consequent. Figure 3 shows the change of metrics of the 108 rules, in which data were sorted by confidence and *KULC* in the descending order. It can be seen that there is a certain negative correlation between *KULC* and *IR*. At some points, *KULC* reaches the local maximum while the IR reaches its local minima. This result suggests that the rules are good on the correlation and the balance, and they are more valuable among the same confidence batch of rules, such as point 37,45,95, etc. And *R* reflects the changes of other indicators from the overall while Nov makes a great contribution.

Table 1 Symptom- Symptom of Mode 1

antecedent	consequent	C	S	K	IR	N	R
'Hearing-loss', 'Sneezing']	['Headache']	0.83	0.16	0.62	0.48	1.10	0.78
['Olfactory-hallucination', 'Sneezing']	['Dry-itchy-eyes', 'Headache']	0.71	0.15	0.61	0.25	1.36	0.77
['Olfactory-hallucination', 'Dry-itchy-eyes', 'Headache']	['Hearing-loss']	0.77	0.16	0.62	0.35	1.18	0.76
['Sneezing', 'Headache']	['Olfactory-hallucination', 'Dry-itchy-eyes']	0.72	0.15	0.60	0.29	1.28	0.76
['Olfactory-hallucination', 'Dry-itchy-eyes', 'Hearing-loss']	['Headache']	0.81	0.16	0.61	0.44	1.05	0.75
['Sneezing', 'Headache']	['Hearing-loss']	0.76	0.16	0.61	0.36	1.15	0.75
['Olfactory-hallucination', 'Sneezing', 'Headache']	['Dry-itchy-eyes']	0.92	0.15	0.60	0.68	0.71	0.73
['Sneezing', 'Headache']	['Dry-itchy-eyes']	0.90	0.19	0.62	0.59	0.67	0.72

In order to illustrate the superiority of the comprehensive evaluation system compared with the single evaluation indicators. In Table 1, the first 8 rules sorted by *R* are displayed clearly with six metrics. Weighted comprehensive evaluation method and the traditional evaluation methods have very big difference in the evaluation results. In addition, the 7th and 8th are considered the most valuable by confidence and the 2nd and 4th are of no value while it is the opposite by Nov, There are some differences in the evaluation results of other rules. It is metaphysical to evaluate rules with the single evaluation method merely from a perspective, therefore, the comprehensive evaluation method R makes a good evaluation method.

2) Mode 2: Mode 2 focuses more on the situation that the antecedent is not limited in symptoms, but contains other attributes such as population, disease history, and so on. Some results are shown in table 2:

Table 2 Diagnosis for the consequent of Mode 2：

antecedent	consequent	R
['sneezing',' Without-odor ']	['Allergic-rhinitis ']	0.85
['Rhinitis-3m~1y(S)', 'Hearing-normal(Z)', 'inter-nasal-obstruction (Z)']	['Chronic-simple-rhinitis ']	0.67

In Table 2, the consequent of Mode 2 is diagnosis, while the antecedents are former items of the logical relations. The antecedent of the second rule contains the information about disease history, which is more close to the real situation in medical diagnosis. The diagnosis dose usually needs to consider population, disease history, symptoms, and so on. What is more, Mode 2 will generate more valuable diagnostic rules when this model is applied to more dimensions of a single disease, more diseases, and a larger amount of data.

3) Mode 3: This mode studies Mode 1 with the data set with a fixed attribute value. Meanwhile, Mode 4 is based on Mode 2, and only consequent category is restricted. The results are omitted here.

4) Mode 5: When there is a demand to study what incentives will produce the same symptoms for diagnosis, Mode 5 will be used, the different incentives for the antecedent and the sane symptoms for the consequent. For example, we found different incentives for 'dry-itchy-eyes' are 'Physical fatigue' and ' Eat spicy, cold food '.

5) Mode 6: In order to find the unique value of symptom with disease history for different values, Mode 6 is used here, and some results are shown in Table 3:

Table 3 Disease History-Symptom of Mode 6

antecedent	consequent	R
['Rhinitis-3 m~1 y']	['inter-nasal-obstruction']	0.50
['Rhinitis-1~2 y ']	['nasal-drip ']	0.43
['Rhinitis-1~2 y ']	['dry-itchy-eyes ']	0.47

In Mode 6, *minconf* is set at 0.6, the disease history for the antecedent, the symptom for the consequent, and the three symptoms in Table 5 appear only behind values of the disease history above. Rhinitis lasts from 3 months to 1 year can appear the situation of intermittent nasal obstruction, while 'nasal drip' and 'eyes dry and itchy' for 1 to 2 years.

After six modes of association rule mining above, the primary algorithm of MH-ARM has been carried out based on the results of the original Apriori algorithm, making a lot of rules present their value and usage scenarios more easily. It can extract rules based on end user needs to help assist the decision-making better.

C. Visualization of Association Rules

Study on visualization of association rules is directly related to the effective display of association rules. The visualization process not only influences the usage of association rules for the decision-makers, but also directly affects the value of association rules. We chose two effective visual forms, namely, directed graph and parallel coordinates graph.

1) Directed graph of association rules

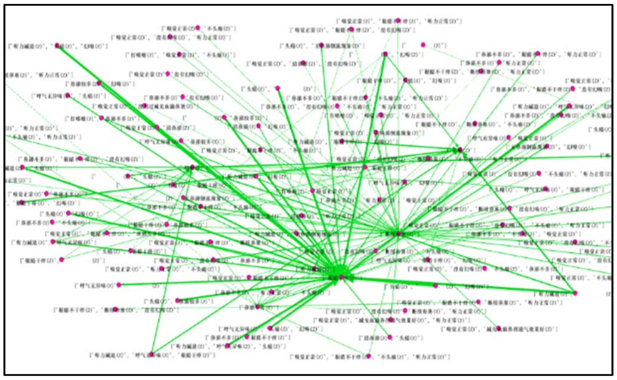

Figure 4 Directed graph of association rules

Figure 4 shows 125 rules of mode 1, based on confidence to display the relationship and the value of the rules. In this figure, nodes represent different attributes values, and the lines between these nodes stand for the association rules, in which the strength of these rules, which is related to the value of *R*, can be indicated by the thickness of lines. *R* is an ideal metric to show the diagnosis needs for its superiority on selecting points dynamically to show the related rules according to the user needs as well as highlighting strong connections.

2) Parallel coordinates of association rules

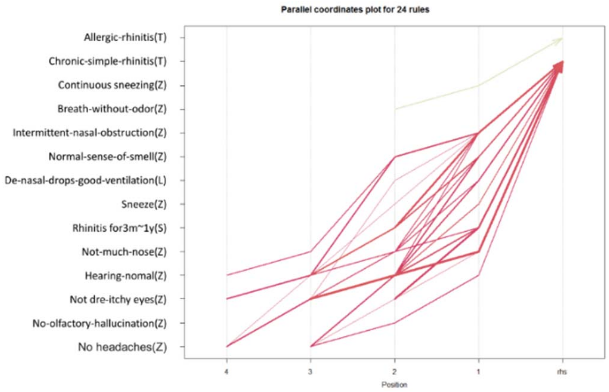

Figure 5 Parallel coordinates of association rules

Figure 5 shows the post diagnostic attributes, and we set *minconf* at 0.7 and *minsup* at 0.15. From the association rules of more than two items, we can see the parallel coordinate plot of diagnostic variables, only allergic rhinitis and chronic simple rhinitis meeting the condition, which effectively displays the number, the visualization and composition of rules, making it an effective presentation in front of end users.

IV. Conclusion

This paper describes an effective methodology for association rules re-mining. MH-ARM including six modes for post-processing based on end user needs is proposed to deal with a large number of results of the original rules mining. The whole process can be divided into mainly four steps. First, the classification information is fused with the original data; second, the size of the item set as well as the natural semantics of the rules are taken into account to improve the effectiveness of association rules mining. Then, we filtered out high value rules based on a comprehensive evaluation system built with several metrics including *support*, *confidence*, *KULC* and *Nov* as well as the effective visualization. Finally, the effectiveness of the process was verified by using the experimental data of rhinitis.

The results of association rule mining are improved in the algorithm, which can effectively filter out the redundant rules and make up for the deficiency of the traditional association rule mining. It is important that the new valuable rules can be stored in the knowledge base for the accumulation of knowledge and other subsequent usage to implement the experience accumulation and reuse theoretically. In the field of practical application, it can effectively reduce the number of rules, to find out more accurate rules for decision-making, which is of particularly high value in the medical field.

REFERENCES

[1] Weng C H. Identifying association rules of specific later-marketed products[J]. Applied Soft Computing, 2016, 38: 518-529.

[2] Yang H, Chen Y P P. Data mining in lung cancer pathologic staging diagnosis: Correlation between clinical and pathology information[J]. Expert Systems with Applications, 2015, 42(15): 6168-6176.

[3] Huang L, Yuan J, Yang Z, et al. Patterns Exploration on Patterns of Empirical Herbal Formula of Chinese Medicine by Association Rules[J]. The Scientific World Journal, 2015, 2015.

[4] Crane P K, Gibbons L E, Dams-O'Connor K, et al. Association of traumatic brain injury with late-life neurodegenerative conditions and neuropathologic findings[J]. JAMA neurology, 2016, 73(9): 1062-1069.

[5] Singh B, Miri R. A Review Paper on Parallel Association Rules Mining Algorithm in Data Mining and MapReduce Framework[J]. Data Mining and Knowledge Engineering, 2016, 8(5): 147-149.

[6] Nahar J, Imam T, Tickle K S, et al. Association rule mining to detect factors which contribute to heart disease in males and females[J]. Expert Systems with Applications, 2013, 40(4): 1086-1093.

[7] Thanigaivel R, Kumar K R. Boosted Apriori: an Effective Data Mining Association Rules for Heart Disease Prediction System[J]. Middle-East Journal of Scientific Research, 2016, 24(1): 192-200.

[8] Rashid M A, Hoque M T, Sattar A. Association rules mining based clinical observations[J]. arXiv preprint arXiv:1401.2571, 2014.

[9] T. Wu, Y. Chen, and J. Han, "Re-examination of interestingness measures in pattern mining: A unified framework," Data Mining Knowl. Discovery, vol. 21, pp. 371–397, 2010.

[10] Jiawei Han, Micheline Kamber, Jian Pei. Data Mining Concepts and Techniques Third Edition[M]. Morgan Kaufmann, 2011.

[11] Lavrac N, Flach P, Zupan B. Rule evaluation measures: A unifying biew[C]. Proceedings of the Ninth International Work shop on Inductive Logic Programming, Bled, Slovenia, June 24-27, 1999