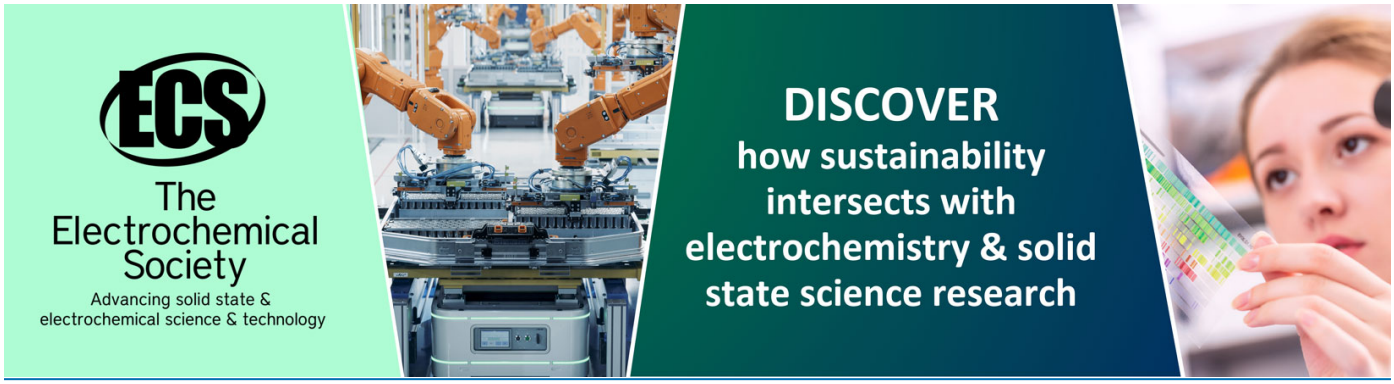
Algorithm Study of New Association Rules and Classification Rules in Data Mining



Algorithm Study of New Association Rules and Classification Rules in Data Mining

Xiaobo Yang[1,*]

[1]Zhejiang Shuren University, Hangzhou, China

*Corresponding author e-mail: yxb71520@163.com

Abstract. In order to further improve the efficiency of data mining, it proposes a kind of data mining algorithm based on new association rules and classification rules. Specific research process is as follows. Firstly, MDML-PP algorithm is analyzed and applied to the multi-dimensional multi-level association rules in data mining, and then selects the test data sets for performance evaluation, meanwhile, it also studies the multi-support rate cut and classification rule mining algorithm, which is applied to multi-support rate classification rules in data mining, finally, using the compare experiment to verify the effectiveness of multi-support rate classification rules. The results show that the proposed algorithm in this paper can be better used in data mining with multi-dimensional and multi-layer association rules, which can improve the efficiency and quality of data mining.

1. Introduction

Association rules is one of the most common types of knowledge, which has a wide application. However, there exist three bugs in the current algorithm of association rules. Firstly, the most algorithms are based on Apriori, which need high cost to dig multi-layer and multi-dimension association rule. Second, the concept layers are digged by using top-down mode, not support the cross-layer digging. Third, the obtained association rules have larger redundancy. In addition, the classification rules can be regarded as a kind of special association rule, due to the uneven distribution of data in real world application, the general algorithm often uses two-stage mining method, which can only use the single support rate, the mining efficiency and system scalability are not high, it is necessary to use multiple support rate. In view of this, it proposes a new data mining algorithm with association rules and classification rules in this paper, in order to solve the related problems.

2. Multi-dimension and multi-layer association rule mining

Association rules can be divided into one dimension and multi dimension according to the number of attributes, and be divided into single layer and multi layer according to the attribute concept hierarchy, and also be divided into Boolean type, quantity type according to the attribute data types. Relative scholars have proposed Boolean association rules to dig monolayer and multilayer [3], as well as multi-dimension and multi data types association rule algorithms, such as Cumulate algorithm [4] and ML-TmLn algorithm [5]. After analysis, these algorithm are mostly based on Apriori, its defect is that the times of database scanning is proportional to the maximum mode length, and mining multi-layer and multi-dimension association rules need 20 mode length, the cost of this algorithm is very high. In addition, these algorithms do not support cross-layer mining, only to create a limited

concept layer by merging adjacent numeric attribute intervals, which can not meet the needs of practical application.

In view of these problems, it is automatically generated by information entropy and combined with interactive human-machine to determine the numerical type and categorical type attributes conceptual level, the new algorithm is designed for mining multi-dimension and multi-layer association rules, that is MDML-PP algorithm, which is verified by experiment.

Multi-dimension and multi-layer association rules are the process to find mode in the relational database. The data types of relational table attribute are more abundant, which generally can be attributed to numerical type and category type. The value of category type can be mapped into a set of consecutive integers, and the value of numerical type attributes can be mapped to a set of consecutive integers to maintain attribute order. In order to solve the problems that attribute interval or subset partition is too big or too small to cause low support rate, some scholars have proposed the methods and principles to divide interval [6]. However, most methods adopt mechanical method to divide interval, only allowing the adjacent interval merging, category type attribute has no concept hierarchy, which can not fully reflect the actual application. It proposes the concept formation by generalizing attribute values, one concept is the union of same attribute and mutually non-overlapping value intervals, therefore, each attribute can have a concept hierarchy, and the connotation of different concept can overlap.

MDML-PP algorithm is an extension to the virtual projection algorithm PP, the property is pretreated by information entropy method, and established various properties of concept hierarchy, which transform the problem of multi-dimension and multi-layer into single layer Boolean problems. Here are the main steps to achieve MDML-PP algorithm.

- Pretreatment based on information entropy

Firstly, the value of each attribute is mapped into integer, then, using information entropy to segment the attribute value interval of numeric type recursively, and the attribute value set of categorical type, each partition can divide interval or set into two parts, the basic principle is that each segmentation can increase the entropy with maximum amount, till the entropy increase amount below the threshold.

Assuming the i-th value a_i of attribute $a \in A$ with sample number $count(a_i)$ in the database R. Set S is composed of a certain values, which divides the S into two sets S_1 and S_2, the expression is as follows.

$$\sum_{a_i \in S} count(a_i) = n, \quad \sum_{a_i \in S_1} count(a_i) = n_1, \quad \sum_{a_i \in S_2} count(a_i) = n_2, \quad n_1 + n_2 = n \tag{1}$$

Segmentation aims are:

$$\max_{\Delta E \geq \delta} \Delta E = E(S) - \frac{n_1}{n} E(S_1) - \frac{n_2}{n} E(S_2) \tag{2}$$

It can be proved that $\Delta E = \frac{n_1}{n} \log_2(\frac{n_1}{n}) - \frac{n_2}{n} \log_2(\frac{n_2}{n})$. Therefore, each partition should make the two sets or interval corresponding to the sample numbers closer, until the increase entropy of the partition is below the threshold value δ.

- Determine the conceptual level

It can automatically create binary tree concept hierarchy according to the step (1). Through human-computer interaction, let user restructure the concept level according to the actual application meaning, and set up the concept level for each attribute.

- Statistical support number and sort

for each record r in record library R do
for each item i in record expand E(r) do
supp(i,R) add 1

Items grouped by attributes, inter group sort according to the number of items in ascending order, inside group sort according to the support rate in descending order.

- Create TTF tree

for each record r in database R do begin
E(r) frequent items are arranged in descending order and get sub sequence s
Let pointer v point to TTF root node
for each item i in sequence s do begin
if exists the child u of v and u.item==i
then u.weight add1;
else set up the child $u \leftarrow \langle i,1 \rangle$ of v
$v \leftarrow u$;
end
end

- Generate frequent pattern tree

Create FIST root node, and labeling < >
$\forall (i, w, link) \in TTF.IL \quad w \geq \min \sup \Rightarrow$ set up the child of FIST root, labeling (i,w)
The first sub node of FIST root node enters stack
While stack not empty do begin
Pop off a FIST node x from stack, let $i = x.item$
$\forall v \in TTF.forest(v.item \prec i \wedge v.weight > 0 \Rightarrow v.weight)$ clear 0
$\forall u, v \in TTF.forest(u.item = i \wedge u.weight > 0 \wedge field(u.item) \neq field(v.item) \wedge v \in anc(u, TTF) \Rightarrow v.weight + u.weight)$
$\forall (i_c, w_c) \in TTF.IL(w_c \geq \min \sup \Rightarrow set\ up\ x\ child, labeling(i_c, w_c))$
The next brother x is pushed on the stack, then the first child into stack
End

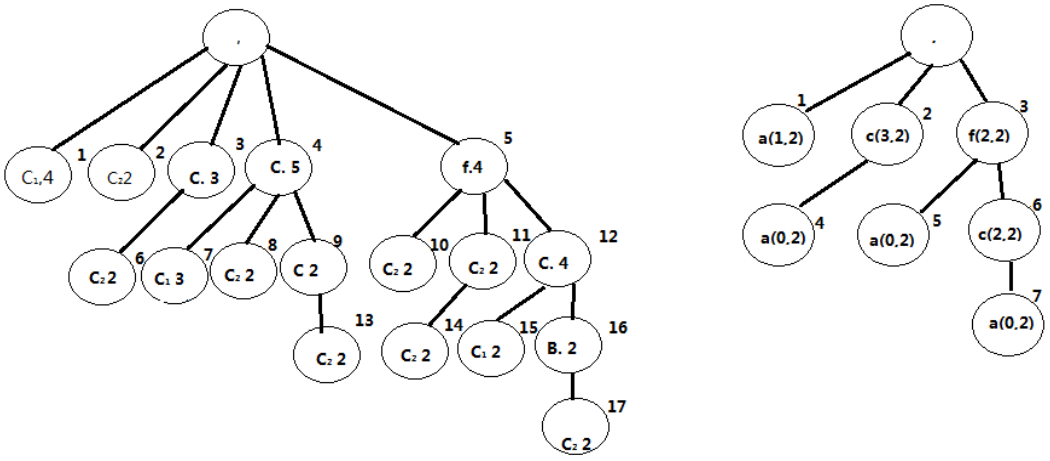
- Create association rules

3. Multiple supports classification rule mining

Classification rules can be regarded as a special association rule, classification rule mining is an important data mining problem. Frequency set mining is the basis for classification data mining, which can only use a single support rate, but the data distribution is not uniform in the practical application problems, so it is necessary to adopt the multi support rate of classification rules mining.

In practical applications, the probability difference of different classification affair is very large, if using the larger and single support rate, a lot of classification information will be lost. Whereas the smaller single support rate will produce a lot of meaningless rules. Therefore, setting different support rate threshold is more reasonable for different classification.

Assuming classified item sets to $C = \{c_1, c_2, \cdots, c_l\}$, the support rate threshold of c_k is minsupk, take minsup1=3, minsup2=2. it can get two kinds of multi-support rate cutting mode, as shown in Figure 1.



(a) Basic type FIST (b) Extended type FIST

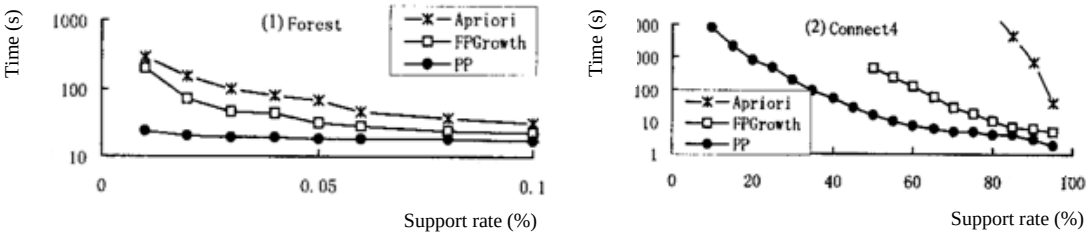
Figure 1 Basic type FIST and extended type FIST

It can be seen from Fig.1, if using the basic type FIST to express the classification database and two stage mining, firstly it can take frequent item sets mining according to the all support rate threshold lower bound minsup, then extract the classification rules according to the classification support rate minsupk and confident minconf. When using extended FIST, it can be cut according to the multi-support rate in the same stage directly, which can improve the efficiency of mining.

On the basis of the frequency set mining algorithms, it directly proposes a multi-support ratio classification rules for single-phase algorithm in this paper, referred to CRM-PP. The algorithm express mode support set from extended TTF, using multi-support ratio cut extended FIST, the frequency set mining and rules extraction are integrated in a single stage, the core is still the virtual projection operation.

In order to verify the effectiveness of multi-support rate of classification rules, the comparative experiment is applied in this paper, the experiment tests various types of data sets, where Forest is sparse data sets, Connect4 is dense data sets, both from UCI machine learning data warehouse.

Firstly, it takes the performance comparison experiment of mining frequency sets, experimental result is shown in Figure 2, the vertical axis represents the running time (logarithmic scale), the horizontal axis represents the minimum support rate (percentage).



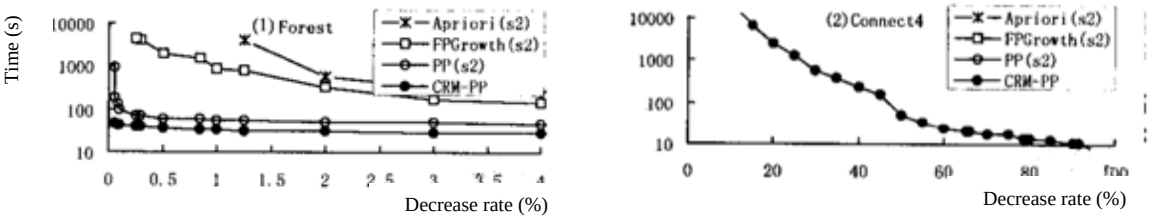
2(a) Forest data set 2(b) Connect4 data set

Figure 2 Performance comparison chart of mining frequency sets

Experimental results show that the algorithm PP mining frequent sets is 1 to 3 times higher efficiency than Apriori and FPGrowth. For example, mining Forest, minimum support rate less than 0.1%, PP efficient is 2 to 12 times higher than Apriori, and 1.5 to 8 times higher than FPGrowth, such as taking the minimum support rate of 0.05%, PP, FPGrowth and Apriori run 18, 31 and 65 seconds separately. When mining Connect4, PP efficient is over 3 times higher than Apriori, PP efficiency is 1 to 2 orders of magnitude higher than FPGrowth.

In the following, performance comparison test is taken to excavate multi-support rate classification rules. During the experiment, each classification support rate is determined from the share multiplied by a given lapse rate. For example, the data set has two categories, accounted for 30% and 70%

respectively, the decreasing rate is 10%, the minimum support rate of two categories are 3% and 7% respectively. The applied comparison algorithm are CRM-PP, PP (s2), FPGrowth (s2) and Apriori (s2), where CRM-PP is the classification rules mining algorithm from multi-support rate cut , in the process of algorithm implementation, firstly, according to the minimum support rate lower bound of mining frequent sets, then according to the minimum confidence and minimum support rate of each category to generate classification rules. The results is shown in Figure 3.



3(a) Forest data set 3(b) Connect4 data set

Figure 3 Performance comparison chart for mining classification rules

Figure 3 shows that the efficiency of the algorithm is in descending order. CRM-PP, PP (s2), FPGrowth (S2) and Apriori (S2), the efficiency of CRM-PP is 1~3 magnitude higher than other algorithms. When mining Forest, minimum confidence is 70%, when the decline rate is 2%, four algorithms are running 31s, 52s, 322s and 589s respectively, when the decline rate is 0.1%, CRM-PP and PP (S2) are running 43s and 92s respectively, while the other algorithms have been unable to run, when miningConnect4, minimum confidence is 80%, except CRM-PP, other algorithms can not run.

4. Conclusions

In this paper, in the basis of frequent pattern mining algorithms, it analyzes the mining problem of multi-dimension and multi-layers association rules with classification rules, and gets the following conclusions.

- It proposes an effective method to decompose the composite node in this paper, which realizes the effective mining with non redundant association rules.
- It proposes a method to determine the hierarchy according to information entropy partition attribute or combination with sets, automatic generation and human-computer interactive, which realizes the new algorithm MDML-PP to mining multi-dimension and multi-layer association rules effectively, the algorithm has not only higher efficiency and scalability than classical methods, but also the ability to mining cross-layer association rules, which support a more flexible definition of the concept hierarchy.
- It proposes the classification rules mining algorithm CRM-PP with multi support rate tailoring, which can let the multi support rate tailoring be integrated into frequent set discovery phase, frequent set mining and rules extraction can be completed in single stage. Experimental results show that CRM-PP time efficiency is 1~3 orders of magnitude higher than the Apriori and FPGrowth two rank algorithm.

5.Acknowledgments

This work was financially supported by Zhejiang Nature Scientific fund.

References

[1] R.Agrawal,H.Mannila, R.Sfikant,H.Toivonen, and A.I.Verkamo. Fast discovery of association rules.In U.M.Fayyad,G.Piatetsky—Shapiro,P.Smyth,and R.Uthurusamy,editors,Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press,1996, pp. 307-328

[2] J.Hart,J.Pei,and Y.Yin,Mining frequent patterns without candidate generation.In SIGMOD'2000,Dallas,TX,May 2000.
[3] J.Hart and Y.Fu,Discovery of multipte—level association rules from large databases,IEEE Transactions on Knowledge and Data Engineering,1999,11(5), pp. 420-431
[4] R.Srikant and R.Agrawal.Mining generalized association rules.In VLDB'95.Zurich,Switzerland,Sept.1995.pp.407—419
[5] J.Han and Y.Fu.Discovery of multiple—level association rules from large databases.In VLDB'95,Zuich,Switzerland,Sept.1995, pp.420—431
[6] R.J.Miller and Y.Yang.Association Rules over Interval Data.In SIGMOD'97, Arizona,USA,1997, pp. 452—461
[7] Information on http://fuzzy.CS.uni-magdeburg,de/～borgelt/src/apriori.exe