Multivariable Statistical Correlation Measure Applied to Association Rules Mining

J.Hu, H.F. Jian, J.H. Sun

School of Management
Chongqing University of Technology
Chongqing, 400054, China

Abstract—**Correlation is usually used in the context of real-valued sequences. However, in data mining, the values range may be of various types-real, nominal or ordinal. Regardless of their type, the methods on measuring correlation between multivariable sequences of data are reviewed. In particular, a new method on measuring the statistical correlation of multivariable sequences is proposed. As the method relies on the geometrical meaning of dot conduct to get the degree of multivariable correlation, it is called M-correlation. M-correlation is used to cut redundancy association rules in this paper. In order to enhance mining efficiency, a novel algorithm, namely FT-Miner, is presented to find all frequent sub-trees in a forest, using two new data structures called UFP-Tree and FP-Forest. The experimentation shows that the algorithm not only reduces a lot of unavailable rules, but also has better capability than classical algorithms.**

Keywords-data mining; association rules; M-correlation; FP-Forest

I. INTRODUCTION

The analysis of correlation between variables is a fundamental task at the heart of many data mining problems.

When data mining databases are handled where the fields can be real value, nominal value and ordinal value. It is always important for the data miner to recognize associations between fields. Correlation is a powerful technique and is used to measure the way that the values of one field tend to vary with respect to the values of another. Given two fields, *X* and *Y*, the database of n records defines a sequence of values, $X=(x_1,x_2,\ldots,x_n)$, for the first field and a sequence of values, $Y=(y_1,y_2,\ldots,y_n)$ for the second field. The correlation, corr (*X*, *Y*) produces a value, generally between −1 and 1, which measures how they change together. Although the term correlation is usually only used formally between two real-valued sequences and its use and interpretation is extended to sequences of all types of data.

However, in many cases, the data are not two sequences but are multivariable sequences. So M-correlation is put forward to get the correlation of multivariable sequences regardless of their datatype.

II. RELATED WORK

At present, the major research results are classified in term of correlation measure, including the following factors:

Measuring correlation should be an important part of any data mining project. By recognising significant correlations, the data miner will not only have an important result in itself but will also be able to exploit this information in further work. Classic examples of data mining techniques based on correlation include principal component analysis [1], and the use of the mahalanobis metric in clustering [2]. Both of these techniques are currently only used with real valued data. A metric called the DCV metric is proposed, which is similar to mahalanobis and used when all the data is nominal-valued [3].

Brin et al test the prominence by χ^2 and applied P $(X\cap Y)$/P(X) *P(Y) to judge the positive or negative correlation between A and B. But as Brin used improved apriori-like algorithm to seek for the frequent items defined by P-support, the rule based on low-high combination with support and correlation was not settled, once the min-supp was set low, the effectiveness of pruning could not been realized[4].

Xiong et al made a study on the mining of strong correlation item pairs by using Pearson's correlation coefficients as measuring criteria and expressed Taper algorithm. Their analytical and experimental results showed that Taper algorithm can filtrate a large number of backup item pairs in way of pruning technology. Whereas, when the items and trade numbers are great, the algorithm still need a mass of calculation, even if only testing remaining backup item pairs[5].

III. MULTIVARIABLE STATISTICAL CORRELATION MEASURE

A. Correlation Degree between Two Variables

Considering two variables *X*, *Y*, making *n* observation to them to get data $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$, then Pearson correlation coefficient

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad (1)$$

B. Statistical Correlation of P random Variables

Supposing $x_i^{(1)},x_i^{(2)},\ldots,x_i^{(n)}$ is the observed values of random variable X_i, letting $x_i=(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(n)})^T$, therein, $x_i^{(j)}(j=\overline{1,n})$ is data get after similar transformation. Therefore

$$\|x_i\|=1(i=\overline{1,p}) \quad (2)$$

In sub-space $\mu(t_1,t_2,\ldots,t_p)$ considering $2p$ surface with p-dimensional parallel, in which $x_1,x_2,\ldots,x_p$ is included.

Letting Ω become the volume of $2p$ surface with p-dimensional parallel combined with $x_1,x_2,\ldots,x_p;t_1,t_2,\ldots,t_p$ is standard orthogonal products of $\mu(t_1,t_2,\ldots,t_p)$ and x1 and t_1 are coincident. Supposing the projection of $x_i\ (i=\overline{2,p})$ exists in $\mu(t_1,t_2,\ldots,t_{i-1})\supset\{x_1,\ x_2,\ \ldots,\ x_{i-1}\}$.

Definition 1 Supposing the volume of 2p surface with p-dimensional parallel is Ω, then $\Omega^2=(x_1,e_1)^2(x_2,e_2)^2\ldots(x_p,e_p)^2$ is geometrical correlation coefficient between p random variables $X_1,X_2,\ldots,X_p$.

Property 1 $0\le\Omega^2\le1$

Proof $\Omega^2=[(x_1,t_1)^2(x_2,t_2)^2\ldots(x_p,t_p)^2]=(\cos\theta_1\cos\theta_2\ldots\cos\theta_p)^2$

Therefore, $0\leqslant\Omega^2\leqslant1$

According to property 1, when Ω is equal to 0, $x_1,x_2,\ldots,x_p$ exists in p-1 dimensional sub-space, which shows that the correlation between $x_1,x_2,\ldots,x_p$ is very closed; when Ω^2 is equal to 1, the orthogonal vector system is combined with $x_1,x_2,\ldots,x_p$, which shows that the correlation between $X_1,X_2,\ldots,X_p$ air related. when Ω^2 increases from 0 to 1, the correlation degree between $X_1,X_2,\ldots,X_p$ becomes increasingly weak. Thus, it is reasonable to regard Ω^2 as the character parameter of correlation degree between random variables $X_1,X_2,\ldots,X_p$.

Property 2 $\Omega^2=\left|A^TA\right|=\left|B^TB\right|$**, There in,**

$$A^T=\begin{bmatrix}x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(n)}\\ x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(n)}\\ M & M & M & M\\ x_p^{(1)} & x_p^{(2)} & \ldots & x_p^{(n)}\end{bmatrix}\qquad B^T=\begin{bmatrix}(x_1,e_1) & 0 & \ldots & 0\\ (x_2,e_1) & (x_2,e_2) & \ldots & 0\\ M & M & M & M\\ (x_p,e_1) & (x_p,e_2) & \ldots & (x_p,e_p)\end{bmatrix}\begin{bmatrix}e_1\\ e_2\\ e_3\\ e_4\end{bmatrix}\tag{3}$$

Property 3 Supposing $\lambda_1,\lambda_2,\ldots,\lambda_p$ are characteristic roots of A^TA, then $\Omega^2=\prod_{j=1}^{p}\lambda_j^2$. Apparently, this property is tenable.

So *M-correlation*$=\Omega^2$ in this paper.

IV. Algorithm Description

In order to enhance efficiency of association rules mining, a new algorithm, called FT-Miner, is present in this paper. The FT-Miner algorithm uses two new data structures called UFP-Tree and FP-Forest [6]. A transaction database is stored by many UFP-Trees, those UFP-Trees form FP-Forest. The FT-Miner algorithm traverses FP-Forest in top-down depth-first order.

A. *Designing of UFP-Tree and FP-Forest*

UFP-Tree is a mutation of FP-Tree. Frequency 1-itemsets are arranged in frequency ascending order when we construct an UFP-Tree. The transaction database is built in FP-Forest, which is shown in Table 1 and Figure 1.

TABLE I. TRANSACTION DATABASE.

T_{id}	Item-set	T_{id}	Item-set	T_{id}	Item-set
100	{I1, I2, I5}	400	{I1, I2, I4}	700	{ I1, I3}
200	{I2, I4}	500	{ I1, I3}	800	{I1, I2, I3, I5}
300	{ I2, I3}	600	{ I2, I3}	900	{I1, I2, I3}

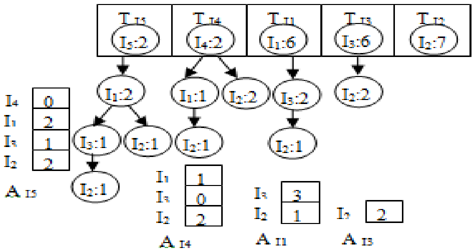

FIGURE I. FP-FOREST.

B. *Realizing of FT-Miner Algorithm*

Definition 2 If the count of a node is less than the minimum threshold value of support and the count of father node is greater than or equal to the minimum threshold value of support when traversing an UFP-tree in top-down, every ancestor node only have one child node, then the UFP-tree is defined as single frequent tree.Figure2, 3shows the process of FT-Mineralgorithm.

```
Input: MinM-correlation, Min-confidence
Output: Association rules
  FT-Miner (FP-Forest)
   {For each UFP-TreeT_Ii in FP-Forest
    {Scan UFP-Tree T_Ii;
      If T_Ii is single frequent tree
      {Get combination of root node and T_Ii'nodes ;
        Calculate the M-correlation of every combination and confidence;
          If M-correlation>MinM-correlation
             Confidence>Min-confidence
              Output association rules;
       }
      Else{F-Ngrowth(T_Ii);
          Calculate the M-correlation of frequent item-set and confidence;
          If M-correlation>MinM-correlation
                Confidence>Min-confidence
            Output association rules;
          }
       }
      }
   F-Ngrowth(T_Ii)
   {Scan array A of UFP-Tree T_Ii;
     Get the frequent items, form frequent 1-itemset;
     Build new FP-Forest, FP-Nforest;
     FT-Miner(FP-Nforest);}
```

FIGURE II. FT-MINER ALGORITHM.

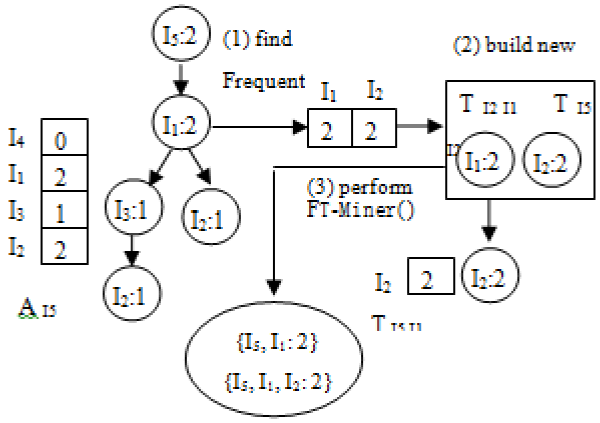

FIGURE III. THE PROCESS PRODUCING FREQUENT ITEM-SET.

V. ALGORITHM CAPABILITY ANALYSIS

In order to test FT-Miner algorithm capability, experiment adopts the real data. Those data come from UCI. Table 2 shows the characteristic of data.

TABLE II. UCI DATA SET.

Data set	Item number	Record number
Mushroom	119	8124
Connect	129	67557

A. *Running Time*

Running time is an important parameter of algorithm capability, so two classical algorithms F-Growth and Apriority are selected to compare running time with FT-Miner. F-Growth and Aprioria algorithms adopt support confidence framework, and FT-Miner uses M-correlation and confidence framework. This experiment is carried out on Connect data set, and parameter setting is as followed: Min-support is equal to 0.7; M-correlation is equal to 0.7; Min-confidence is equal to 0.8, which result is shown in Figure 4. According to Figure 4, FT-Miner's running time is least.

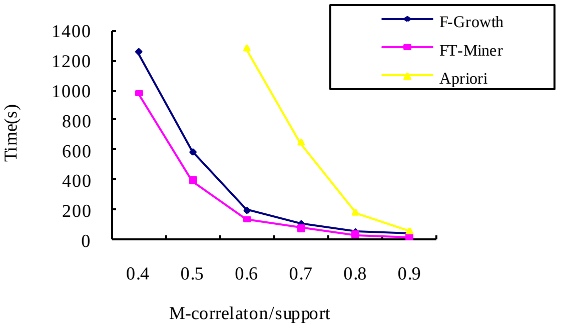

FIGURE IV. COMPARING OF RUNNING TIME.

B. *Comparison with Support-confidence*

This experiment on Mushroom data set mainly compares M-correlation-confidence framework with Support-Confidence framework in the number of producing association rules. Parameter setting is as followed: Min-support is equal to 0.7; Min M-correlation is equal to 0.7; Min-confidence is equal to 0.8. From the experiment result which is shown in Figure 5, M-correlation-Confidence framework is better than Support-Confidence framework in pruning negative association rules and weakness association rules.

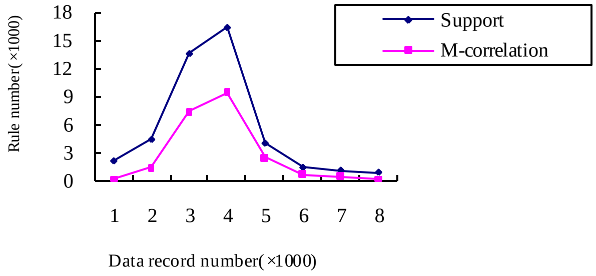

FIGURE V. COMPARISON WITH SUPPORT-CONFIDENCE.

VI. CONCLUSION

The statistical correlation of multivariable, called M-correlation is presented to help improve the correlation of association rules, which can measure multivariable correlation degree. At the same time, M-correlation greatly reduces rules number by pruning weak rules. FT-Miner is designed as a novel algorithm using two new data structures called UFP-Tree and FP-Forest which can highly improve efficiency of algorithm. By the testing on the UCI database, the method in this paper is too appropriate to mining association rules.

ACKNOWLEDGMENTS

This work is partially supported by the National Natural Science Foundation of China (Grant No.71301181, Grant No.71401021).

REFERENCES

[1] Jolliffe, I.T., Principal Component Analysis. Springer, Berlin,1986

[2] Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Comput. Surveys31 (3), pp.264–323, 1999.

[3] Schrijver, A., Combinatorial Optimization: Polyhedral and Efficiency. Springer, Berlin,2003

[4] S. Brin, R. Motwani, & C. Silverstein. Beyond market basket: Generalizing association rules to correlations. In Proc. SIGMOD'97, pp. 265–276, 1997

[5] H. Xiong, S. Shekhar, P-N. Tan, V. Kumar. Exploiting a Support-based Upper Bound of Pearson's Correlation Coefficient for Efficiently Identifying Strongly Correlated Pairs. In: Proc. of ACM SIGKDD'04, pp. 334–343,2004

[6] Li Long Jie., F-Miner: A New Frequent Itemsets Mining Algorithm. Lanzhou University Master Thesis,2006