Association Rules Mining Using Multi-objective Coevolutionary Algorithm

Jian Hu
Research Center of Technology, Policy and Management
School of Management, Harbin Institute of Technology, Harbin 150001
jianhu-hit @163.com

Xiang Yang-Li
Research Center of Technology, Policy and Management
School of Management, Harbin Institute of Technology, Harbin 150001
xiangyangli@hit.edu.cn

Abstract

Association rule mining can be considered as a multi-objective problem, rather than as a single objective one. To enhance the correlation degree and comprehensibility of association rule, two new measures, including statistical correlation and comprehensibility, as objection functions are proposed in this paper. Their calculating formulas and primary characteristics are given. Association rule mining is generally solved by lexicographic order method. On the basis of discussing the weakness of above method, a new coevolutionary algorithm is put forward in this paper to solve multi-objective optimization problem of association rule. Three coevolutionary operators are designed and the mining algorithm is realized in this paper. According to experimentation, the algorithm has been found suitable for association rule mining of large databases.

1. Introduction

It is an important issue to mine association rule from transaction databases through data mining. Agrawal[1] firstly put forward association rule problem of item sets in transaction database.

From the viewpoint of multi-objective optimization, association rule mining in classical support-confidence considered support and confidence as optimization objects and most methods adopted lexicographic order. However, there are many problems to solve.

- In most previous research, support was always one of the object functions. However, the real correlation of data was not founded.
- The comprehensible degree was too weak to use by decision makers in previous algorithms.

On this basis, this paper uses statistical correlation to substitute for support, which has real statistical meaning and puts forward comprehensibility as an object function to improve the comprehensible degree.

2. Relative work

The integration of data mining and evolutionary algorithms is currently an active research area. At present, a lot of research results have been published.

Beatriz de la et.[2] proposed the use of multi-objective optimization evolutionary algorithms to allow the user to interactively select a number of interest measures and deliver the nuggets according to those measures and implemented the Fast Elitist Non-Dominated Sorting Genetic Algorithm.

Tan.k.c[3] presented a dual-objective evolutionary algorithm for extracting multiple decision rule lists in data mining.

Jing Liu et[4] put forward organizational co-evolutionary algorithm for classification. Tan.K.C[5] proposed a co-evolutionary-based technique to discover classification rules in data mining.

In addition, Xavier Llorà[6] put forward Fine-Grained Parallel Evolutionary Algorithms. Alex A.Freitas[7] discussed the use of evolutionary algorithms, particularly genetic algorithms and genetic programming, in data mining and knowledge discovery.

3. Multi-objective optimization model

In the multi-objective optimization model of association rule mining, three objection functions statistical correlation, comprehensibility and confidence are adopted in this paper.

3.1. Definition of statistical correlation

0-7695-3073-7/07 $25.00 © 2007 IEEE
DOI 10.1109/CIS.Workshops.2007.22

To eliminate those rules without relativity, this paper puts forward the statistical correlation to substitute for support.

Definition 1 From the viewpoint of statistics, the statistical correlation is defined as followed.

$$scorrelation(X \cup Y) = \frac{|D| support(X \cup Y) - |D| \prod_{i \in X \cup Y} support(i)}{\sqrt{|D| \prod_{i \in X \cup Y} support(i)(1 - \prod_{i \in X \cup Y} support(i))}} \quad (1)$$

The statistical meaning of scorrelation is that if considering the cube with $|X \cup Y|$ dimension composed of items among $X \cup Y$, which are inter-independent, then $\prod_{i \in X \cup Y} support(i)$ is the ratio of the expected exchange number of this cube to the entire database. If the database has $|D|$ data and those data is proportional distribution, whether any data point is in cube or not can be regarded as a Bernaulli random variables with $\prod_{i \in X \cup Y} support(i)$ probability. According to central limit theorem, the distribution of point number in cube is approximate to Gaussian distribution on the basis of above-mentioned assumption. Thus, the expected exchange number in cube is $|D| \prod_{i \in X \cup Y} support(i)$, standard deviation is $\sqrt{|D| \prod_{i \in X \cup Y} support(i)(1 - \prod_{i \in X \cup Y} support(i))}$ .In brief, if scorrelation ($X \square Y$) is increasingly great, then the association is better.

3.2. Definition of comprehensibility

The research on association rule has proved that if the item number of the rule condition and conclusion parts is less, then this rule can be comprehended more easily.

Definition 2 If $|B|$ represents the attributes number of the consequent part Y, $|A \cup B|$ expresses the attributes number of whole rule $X \Rightarrow Y$, the comprehensibility of rule $X \Rightarrow Y$ is defined as followed:

$$compreh(X \Rightarrow Y) = \frac{log(1 + |B|)}{log(1 + |B \cup A|)} \quad (2)$$

Apparently, with $|B|$ and $|B \cup A|$ decreasing, $compreh(X \Rightarrow Y)$ is increasing.

In addition, confidence can be considered as accurate rate of X forecasting Y. So it has practicality as one of objection function.

4. Coevolutionary algorithm describing

Due to above three objections having the same importance, this paper adopts Pareto method based multi-objective coevolutionary algorithm to get association rules, which can avoid prematurity convergence and enhance convergence speed.

4.1. Denotation of individual

This paper adopts Michigan approach to denote each individual, where each individual represents a single rule. This paper uses dual bit binary number to express each attribute. If dual bit is 10, then the attribute appears in the antecedent part; if it is 01 then the attribute appears in the consequent part. And the combination, 00 will indicate the absence of the attribute in rule. For example, the attributes of a data set are {A, B, C, D, E, F}, then the rule ACD→F will look like 100010100001. In this way, we can handle variable length rules with higher storage efficiency.

4.2. Fitness value definition and selection

Zitzler's fitness evaluation value method[8] is adopted to calculate fitness of individual. This paper proposes method to use an external population to perform elitism. Supposing the number of current population P_t is N, and fitness value calculation method is as follows.

Step1: defining the Pareto optimal solution set as $p(P_t)$, and calculating the fitness value for random $i \in p(P_t)$.

$$S(i) = \frac{|\{ j \mid j \in p_t \wedge i \succ j|}{N + 1} \quad (3)$$

Therein, j is a random individual of P_t; $i \succ j$ means i dominating j.

Step2: calculating fitness value of P_t| $p(P_t)$.

$$F(j) = 1 + \sum_{i \in p(P_t) \wedge i \succ j} S(i) \quad (4)$$

Individual selection method is as follows.

Step1: selecting i and j from P_t at random.

Step2: if $F(i) < F(j)$, letting $P_{t+1} \leftarrow P_{t+1} + (i)$, or else $P_{t+1} \leftarrow P_{t+1} + (j)$.

External population number is $\overline{N}$, Supposing the Pareto optimal solution is S_t for current population P_t, and external population renovation is as follows.

Step1: copying S_t to $\overline{P_t}$, That is $\overline{P_t'} \leftarrow \overline{P_t'} + (S_t)$.

Step2: finding the Pareto optimal solution of $\overline{P_t'}$, then getting external population of t+1 generation. That is $\overline{P_t'} \leftarrow p(\overline{P_t})$.

4.3. Coevolutionary operator design

Supposing current population is Pt, in which St is considered as the optimal solution set of Pareto.

4.3.1. Pareto neighborhood crossover operator. Supposing $(x_1, x_2, ..., x_n)$ belongs to P_t, an individual $(r_1, r_2, ..., r_n)$ is chosen randomly in S_t, which is substituted by an individual $(z_1, z_2, ..., z_n)$ generated according to the following formula.

$$z_i = r_i + U(0,1) \times (r_i - x_i), i = 1, 2, \cdots, n \quad (5)$$

Therein, $U(0, 1)$ represents the random number of 0 or 1. Sign "+" expresses logic and operation; Sign "-" expresses logic or operation.

4.3.2. Combination operator. Supposing two parents population $P_t^a = (X_1^a, X_2^a, \cdots, X_M^a)$ and $P_t^b = (X_1^b, X_2^b, \cdots, X_N^b)$, which Pareto optimal solution sets are respectively S_t^a and S_t^b, for any individual $(x_1^a, x_2^a, \cdots, x_n^a)$ of P_t^a, we randomly choose $(r_1^a, r_2^a, \cdots, r_n^a)$ in S_t^b, and generate a new individual $(z_1^a, z_2^a, \cdots, z_n^a)$ according to the formula (6), which will produce offspring population P_{t+1}^a.

$$z_i^a = r_i^b + U(0,1) \times (r_i^b - x_i^a), i = 1, 2, \cdots, n \quad (6)$$

Similarly, we also generate offspring population P_{t+1}^b.

4.3.3. Annexing operator. Supposing $P_t^a = (X_1, X_2, \cdots, X_M)$, $P_t^b = (Z_1, Z_2, \cdots, Z_N)$ are two parents population, their Pareto optimal solution sets are respectively S_t^a and S_t^b, if meeting the following conditions: $\forall x \in S_t^a, \exists z \in S_t^a, z \succ x$ is shown as ($S_t^b \rhd S_t^a$), then when the population P_t^b annexes P_t^a, it can produce a new population $P_{t+1}^b = (L_1, L_2, \cdots, L_{N+M})$. Therein, L_i=Z_i, i=1,2,…,N and $L_i, i = N, N+1, \cdots, N+M$ is calculated by the formula (7).

$$l_{ij} = r_j + U(0,1) \times (r_j - x_{i\text{-}Nj}), j = 1, 2, \cdots, n \quad (7)$$

4.4. Algorithm realizing

Supposing two populations P^a and P^b, which numbers are N; S^a and S^b are the Pareto optimal solution sets of P^a and P^b; $\overline{P}$ is external population, which size is $\overline{N}$; P_m is mutation probability.

Step1: scanning database, loading records of data in the memory, ordering data attributes according to the appearance number, and using attributes to denote individuals and initialize P_0^a and P_0^b.

Step2: scanning loaded data, calculating above three objective function value of P_0^a and P_0^b, getting S_0^a and S_0^b, and giving $\overline{P_0} \leftarrow \varnothing$, $t \leftarrow 0$.

Step3: getting $\overline{P_{t+1}}$ by using S_t^a and S_t^b to update external population $\overline{P_t}$.

Step4: $\overline{P_t}$ is divided into $\overline{P_t^a}$ and $\overline{P_t^b}$, for which we respectively implement optimization preserving policy and get the medium population $\overline{P_t^{ta}}$ and $\overline{P_t^{tb}}$, namely, $\overline{P_t^{ta}} = P_t^a + \overline{P_t^a}$, $\overline{P_t^{tb}} = P_t^b + \overline{P_t^b}$.

Step5: scanning loaded data, calculating the fitness value of population $\overline{P_t^a}$ and $\overline{P_t^b}$, and selecting new population $\overline{P_{t+1/2}^a}$ and $\overline{P_{t+1/2}^b}$.

Step6: if meeting stopping condition, and outputting $\overline{P_{t+1}}$, then outputting association rules by population of $\overline{P_{t+1}}$. Otherwise, go to the next step.

Step7: respectively implementing the neighborhood mutation operation for the population $\overline{P_{t+1/2}^a}$ and $\overline{P_{t+1/2}^b}$, and getting $\overline{P_{t+1/2}^{ta}}$ and $\overline{P_{t+1/2}^{tb}}$.

Step8: respectively going on mutation operation for every individual of $\overline{P_{t+1/2}^{ta}}$ and $\overline{P_{t+1/2}^{tb}}$ according to possibility P_m: randomly choosing an individual's some dimension $j \in (1, 2, \cdots n)$ to generate a random number to substitute for initial value, and getting $P_{t+1/2}^{na}$ and $P_{t+1/2}^{nb}$. Therein, the feasible value of one-dimension belongs to $U(0, 1)$.

Step9: if $S_{t+1/2}^{na} \rhd S_{t+1/2}^{nb}$, then using $P_{t+1/2}^{na}$ to annex $P_{t+1/2}^{nb}$, and getting $P_{t+1/2}^{mb}$; if $S_{t+1/2}^{nb} \rhd S_{t+1/2}^{na}$, then making $P_{t+1/2}^{nb}$ annex $P_{t+1/2}^{na}$ and getting $P_{t+1/2}^{mb}$. Or else, we implement combination operator for $P_{t+1/2}^{nb}$ and $P_{t+1/2}^{na}$ to get $P_{t+1/2}^{mb}$ and $P_{t+1/2}^{ma}$. Afterward, returning Step 3.

5. Algorithm capability analysis

In order to test coevolutionary algorithm capability, experiment adopts the real data. Those data come from UCI[9]. The platform of experiment is an IBM pc with 512 MB RAM and 3.0 GHZ CPU.

5.1. Association rule mining result

The proposed algorithm is implemented on different data sets. Here we test on Mushroom, Connect data sets. Parameter setting is as followed: N=200, $\overline{N}$ =200, P_m=0.25.

As we can see from Table 1, rule number is gradually reduced with the generations increasing. It is known that after 300 generations it ceases to generate more rules. It denotes that we have got the optimal solution.

Table 1. Association rules mining result

Data set	Data attribute number	Data record number	Number generation	Number of association rule
Mushroom	119	8124	100	146
			200	78
			300	24
			400	23
Connect	130	67557	100	200
			200	126
			300	60
			400	60

5.2. Running time

A classical algorithm F-growth is selected to compare running time with coevolutionary algorithm. This experiment is selected from mushroom data set, at random. Parameter setting of F-growth is as followed: minsupp=0.5, minconf=0.6. Parameter setting of coevolutionary algorithm is as followed: N=200, $\overline{N}$ =200, P_m=0.25. According to Figure 1, coevolutionary algorithm's running time is less than F-growth.

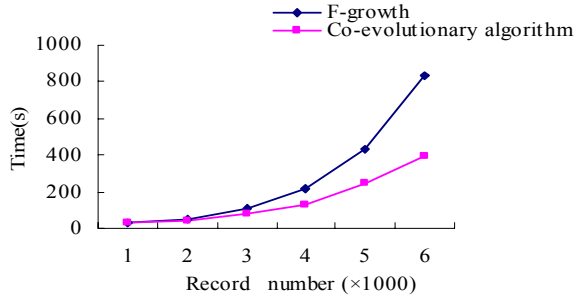

Figure 1. Running time

6. Conclusion

The paper adopts multi-objective coevolutionary algorithm to solve association rule problem, which don't require users to input the minimum threshold of measures, so it is an intelligent data mining way. Furthermore, statistical correlation is presented to help improve the correlation of association rule, and comprehensibility is proposed to get the applicability rules. Based on experimentation, weak rules and negative rules are greatly reduces, and comprehensible degree is enhanced.

7. Acknowledgement

The authors thank the CIS' 2007 anonymous referees for their substantive suggestions which have improved the paper. This work is partially supported by the National Natural Science Foundation of China (Grant No. 70571019) and the Research Center of Technology, Policy and Management, Harbin Institute of Technology.

8. Reference

[1] R.Agrawal, T.Imielinski and A.Swami., Ming Association Rules Between Sets of Items in Large Databases, Proceeding of the ACM SIGMOD Conference on Management of Data, Washington DC, 1993, pp.207–216.

[2] Beatriz de la Iglesia, S. P. Mark, J. B. Anthony and J.R.S. Vie, Data Mining Rules Using Multi-Objective Evolutionary Algorithms, The 2003 Congress on Evolutionary Computation, 2003, pp.1552–1559.

[3] K.C. Tan, Q. Yu and J. H. Ang, A Dual-Objective Evolutionary Algorithm for Rules Extraction in Data Mining, Computational Optimization and Applications, Vol.34, Springer-Verlag, Heidelberg New York 2, 2006, 273–294.

[4] J. Liu, W.C. Zhong, F. Liu and L.C. Jiao, A New Data Mining Method Using Organizational Coevolutionary Mechanism, Lecture Notes in Computer Science, Vol.3056, Springer-Verlag, Heidelberg New York, 2006, pp.196–200.

[5] K.C. Tan, Q. Yu and J.H. Ang, A Coevolutionary Algorithm for Rules Discovery in Data Mining, International Journal of Systems Science, Vol. 37.12/10, 2006, pp.835–864.

[6] Xavier Llorà and Josep M.Garrell, Knowledge-Independent Data Mining With Fine-Grained Parallel Evolutionary Algorithms. PhD thesis, Enginyeria i Arquitectura La Salle, Ramon Llull University, Barcelona, Catalonia, European Union, 2002.

[7] A.F. Alex, A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery. A.Ghosh, S. Tsutsui (Eds.), Advances in Evolutionary Computing, Springer-Verlag, Berlin Heidelberg New York, 2003.

[8] E. Zitzler and L.Thiele, Multiobjective Evolutionary Algorithms: A comparative Case Study And the Strength Pareto Approach, IEEE Trans, Evolutionary Computation, Vol.3.4, 1999, pp.257–271.

[9] C. J. Merz and P. Merphy, UCI Repository of Machine Learning Databases, 1996. (Http://www.ics.uci.edu)