Study and Application on the Method of Association Rules Mining Based on Genetic Algorithm

Xinhang Xu
Hebei Electric Power Research Institute,Shijiazhuang 050021,China

Qiuhong Sun
Hebei University of Science and Technology, Shijiazhuang 050000,China
YanShan University,Qinhuangdao 066004,China

Hongtao Zhang , Lei Wang , Yonghong Liu
Hebei Electric Power Research Institute,Shijiazhuang 050021,China

***Abstract*-This article is mainly discussing the application of genetic algorithms to data mining of association rules, and proposing method of extract association rules using genetic algorithm, at the same time discussing genetic algorithm coding methods and construction of fitness function and improvement of genetic operator and so on. Given extraction algorithm of association rules based on genetic algorithm, and would be applied in hospital medical database in data mining.**

Keywords-Association Rules, Genetic Algorithm, Fitness Function

I. CODING AND IMPROVEMENT OF THE FITNESS FUNCTION

Based on the characteristics of mining for association rules , this article make the following improvements on genetic algorithm.

1)Coding

We use a coding method of an array of real numbers because of the genetic operators, association rule mining and project needs. This encoding method has high precision, easy space advantage of search, and most importantly related to this issue, implement relatively simple. In the process of encoding method using an array of real numbers. The number of elements of an array of real numbers corresponds to the number of fields in a database with a transaction, and real numbers elements of an array value represent the property value for the filed.

We use an array of N tuples to represent such as transactional database that is shown on individual coding. A[1] indicates that the field 1, and A[2] indicates that the field 2, and A[n] indicates that the field N ; And we property value expressed as a numeric value, for example ,we use the value 1 indicates that the attribute value 1, and use value 2 indicates that the attribute value 2, and use value E indicates that the attribute value E, so we can use array A[N] element value to represent the property value for the filed that corresponds to . At the same time, we need to use the value 0 indicates that the property is not associated with other properties.

2) Calculation of individual fitness

In the evolution of genetic algorithms in search generally does not require other external information, but only requires fitness function with values to access the merits of the individual or solutions, and the genetic algorithms as the basis for later genetic manipulation. On its only requirement is that the input can be calculated can be non-negative result of the comparison. In special issues, the design of the fitness function should be connected with the question.

For the mining association rules based on genetic algorithm, support is a measure of the importance of association rules. We consider support of association rules to define its fitness function. And such we can filter the rules. First with support to filter rule first, and then meet the minimum level of support set out in the rules of its associated and relevance. So the adaptation of the rule values such as (1)

$$fitness(R_i) = s'/s \begin{cases} p & s' > s \\ q & s' < s \end{cases} \quad (1)$$

In the formula, S' to go through genetic manipulation in support of the formation of a new rule, and S given threshold of support for the user. When the Ri to meet the requirements of the rule, its Adaptive function value should be greater than 1; Otherwise the adaptability function value will be less than 1, so this rule will be eliminated in the next - generation genetic.

II. THE APPLICATION OF THE IMPROVED ADAPTIVE PC AND PM

This article provides an analysis of a new population of "precocious " evaluation indicators, and finally combining genetic algorithm with adaptive adjustment of control parameters of the mind, have come up with a modified adaptive genetic algorithm. Experimental results show that the algorithm can not only speed up the speed of genetic evolution, but also enhanced global convergence algorithm performance, to be satisfied with the globally optimal value.

Early Evaluation Index

Set t's population made up of individuals $X_t^1, X_t^2, ..., X_t^M$, fitness

$$F_{av} = \frac{1}{M}\sum_{i=1}^{M} F_t^i \quad (2)$$

respectively $F_t^1, F_t^2, ..., F_t^M$, population average fitness of the individual

Optimum fitness for the individual $F_{\max}, \overline{F_{t\max}}$, fitness is greater than the average fitness of the F_{av} individual representative, defines the difference between $F_{\max}$ and $\overline{F_{t\max}}$:

$$\Delta' = F_{\max} - \overline{F_{t\max}} \quad (3)$$

So the index Δ' it can be used for characterization of population "premature" level.

1)Improved Pc and Pm Adaptive Algorithm

Crossover probability and mutation probability of genetic algorithms convergence performance has a material effect. We describe Pc and Pm Adaptive Algorithm as follows: First , when the Group's largest fitness Fmax and fitness Fav approached the Group tends to average convergence then should be increased Pc and Pm, Otherwise, strong group diversity, should be reduced Pc and Pm, that means Pc and Pm is inversely proportional to the value of (Fmax-Fav). Secondly, to prevent the good gene structure is compromised, must make fitness solution for smaller Pc and Pm, so that fitness is larger Pc and Pm, and small solutions, Pc and Pm, proportional to the value (Fmax-Fav). In the following, for a time X1 involved in crossover operation of larger of the two individual fitness, Y fitness to participate in the variation of the individual.

$$P_c = \begin{cases} P_c * \dfrac{F_{\max} - X_1}{F_{\max} - F_{av}} & X_1 \geq F_{av} \\ P_c & X_1 < F_{av} \end{cases} \quad (4)$$

$$P_m = \begin{cases} P_m * \dfrac{F_{\max} - Y}{F_{\max} - F_{av}} & Y \geq F_{av} \\ P_m & Y < F_{av} \end{cases} \quad (5)$$

Type (4) and (5) showed that the smaller the (Fma--Fav) reach , Groups reach likelihood of local optimization, and greater the likelihood of early , let Pc and Pm the larger, to enhance the Group's ability to produce new individual .On the other hand, the larger the (Fmax-Fav) reach, the population divergence ,then should lower Pc and Pm to improving the convergence of individual capacities, to maintain its convergence, we should let Pc and Pm and (Fmax-Fav) is inversely proportional to.

III. IMPROVEMENT OF GENETIC MANIPULATION

A. Selecting (copy) operation

Fitness proportional selection mechanisms used in this article and the best individual retention mechanisms with the collection of methods. First fitness value to individual selection probability proportional to copy individual, set group size M, individual i fitness value of fi , then i selected the probability is

$$P_{s_i} = f_i / \sum_{j=1}^{M} f_i \quad (6)$$

Select full, we can let the next - generation compared to the best of the best individual and parent individual, full of individual institutions to adapt to most of them copied to a new generation.

B. Interlace Operation

This article uses one cross: Individuals will be selected in the previous step out for pairs, and with the probability of a random set in each individual Ps cross point. Before or after the point part of the structure of two individual swaps to generate a new individual. For example,

Pairing individual A 1001|111-→1001000 new individual A

Pairing individual B 0011|000-→0011111 new individual B

Crossing points are located in between the fourth and fifth loci. Intersections are random set. When L is length of the chromosome may have L-1 Cross and the location. One point cross may implement L-1 different cross results.

C. Mutation in Operation

Mutation

This article is based on the improved method of uniform variations. We use the mutation operator is designed for:

To a certain mutation probability random variations in the populations of individual. After you select it, turn it at every level of the individual loci variation, and gene's is in its turn within the range of allowed values. So this mutation in turn ensures that after each property values will exist.

IV. APPLICATIONS

Based on a hospital physical examination database, this article set up an association rules mining system, and through the establishment of the system of medical personnel in information management and analysis, in which application of association rule mining algorithm based on genetic algorithm for data mining. Look forward to the establishment and implementation of systems to help hospitals manage medical information. And also in the medical information from hospital, the Mining Association rule has practical significance; you can use these rules to guide the daily lives of medical staff, and makes recommendations for public health.

Rule 1200100 indicated that 86% of the medical person aged between 30-40 is normal weight. And 56% support means this situation has a high proportion in teachers. Rule 4000001 indicated that 90% of the medical person that older than 60 have bad ultrasound examination results of liver and gallbladder. 2% support that is because of the proportion of the age over 60 has only a small percent. Rule 0100202 indicated that intermediate title in the medical personnel in both weight and ultrasound examination check results are good. Rule 0020002 indicated that 97% of the teachers are have normal weight, but cause of the low ratio of senior title in the physical examination , the support is low, only 2%. The rule 0040200 indicated that most of the workers do not

attach too much importance to weight control. The application indicates that the above algorithm is effective.

REFERENCES

[1] Jiuyong Li,Hong Shen and Rodney Topor,Mining the smallest association rule set for predictions

[2] R.Agrawal,T.lmielinski,and A.Swami.Mining associations between sets of items inmassive databases.In Proc.Of the ACM SIGMOD Int'l Conference on Management of Data.1993

[3] R.Agrawal and R.Srikant.Fast algorithms for mining association rules in large databases. In Proceedings of the Twentieth International Conference on Very Large Databases. Pages 487-499,Santiaga,Chile,1994.

[4] R.Vayardo and R.Agrawal.Mining the most interesting rules.In S.Chaudhuri and D. Madigan,editors,Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,pages 4145-154,N.Y.,Aug.15-18 1999,ACM Press.

[5] Xingping Wang,"The influence of parameters and system to turbo-unit operation efficiency",Power Engineering,Vo 17,No.3,pp.48-52,June.1997.

[6] Jing He,"Multi-space association rule algorithm",Yunan University,May,2003.

[7] Yong Lu,Xiangdong Xu,Ming Chen,"Power Plant Process Control based on data mining",Power System Engineering,Vo 19,No.2,pp.48-50,Mar.2003.

[8] Wei Li,Changdong Liu,Deren Sheng,"Present status and forecasting of operation optimization system in power plant",Power System Engineering,Vo 20,N0.1,pp.59-61, January,2004.

[9] Wen Bao,Daren Yu,Wei Wang,"Power Plant feeler unit fault detection based on association rule",Proceedings of the CSEE,Vo 23,No.12,pp.170-174,Dec.2003.

[10] Tunpei Hong,"A fuzzy AprioriTid algorithm with reduced computation time",Power Engineering,Vo 17,No.3,pp.48-52,June.1997.

[11] Chunlei Luo,"Optimum burning system of boiler based on BP network",Electric Power, Vo 34,No.10,pp.31-34,October,2001

TABLE I. GENERATION RULES

rules code	parameter
1200100	56% support,86% confidence
4000001	2% support,90% confidence
0100202	3% support,93% confidence
0020002	2% support,97% confidence
0040200	2% support,91% confidence
……	……