AN OPTIMIZATION GENETIC ALGORITHM FOR IMAGE DATABASES IN AGRICULTURE

Changwu Zhu [1], Guanxiang Yan [2], Zhi Liu [3], Li Gao [1,*]

[1] *Department of Computer Science, Hua Zhong Normal University, Wuhan 430079, China*
[2] *School of Information Management, Wuhan University, Wuhan 430072,China*
[3] *Wuhan Junxie Shiguan School, Wuhan 430079, China*
[*] *Corresponding author, Tel: 027-62265691, E-mail: lgao@mail.ccnu.edu.cn*

Abstract: Data Mining is rapidly evolving areas of research that are at the intersection of several disciplines, including statistics, databases, pattern recognition, and high-performance and parallel computing. In this paper, we propose a novel mining algorithm, called ARMAGA (Association rules mining Algorithm based on a novel Genetic Algorithm), to mine the association rules from an image database, where every image is represented by the ARMAGA representation. We first take advantage of the genetic algorithm designed specifically for discovering association rules. Second we propose the Algorithm Compared to the algorithm in, and the ARMAGA algorithm avoids generating impossible candidates, and therefore is more efficient in terms of the execution time.

1. INTRODUCTION

The image databases contain an enormous amount of information, and it is becoming more and more complex while its size continues to grow at a remarkable rate. So it can be exceedingly difficult for users to locate resources that are both relevant to their information needs and high in quality. Vast numbers of images have accumulated on the Internet and in entertainment, agriculture, education, and other multimedia applications. Therefore, how to mine interesting patterns from image databases has attracted more and more attention in recent years (G.chen, 2002).

Many additional algorithms have been proposed for association rule mining. Also, the concept of association rule has been extended in many different ways, such as

generalized association rules, association rules with item constraints, sequence rules etc. Apart from the earlier analysis of market basket data, these algorithms have been widely used in many other practical applications such as customer profiling, analysis of products and so on(gaoli, 2003).

Genetic Algorithm (GA) is one self-adaptive optimization searching algorithm. GA obtains the best solution, or the most satisfactory solution through generations of chromosomes' constant evolution includes the reproduce, crossover and mutation etc. operation, until a certain termination condition is coincident (K. Koperski, 1995).

Association rules mining Algorithm Based on a novel Genetic Algorithm (ARMAGA) is an optimal algorithm combing GA with ARMA.

In this paper we first take advantage of the genetic algorithm designed specifically for discovering association rules. Second we propose a novel spatial mining algorithm, called ARMAGA, Compared to the algorithm in (G.chen, 2002), and the ARMAGA algorithm avoids generating impossible candidates, and therefore is more efficient in terms of the execution time.

2. ASSOCIATION RULES

Definition 1 confidence

Set up $I = \{i_1, i_2, i_m\}$ for items of collection, for item in $i_j (1 \le j \le m)$, $(1 \le j \le m)$ for lasting item, $D = \{T_1, T_N)$ it is a trade collection, $T_i \subseteq I (1 \le i \le N)$ here T is the trade.

Rule $X \to Y$ is probability that $X \bigcup Y$ concentrates on including in the trade.

The association rule here is an implication of the form $X \to Y$ where X is the conjunction of conditions, and Y is the type of classification. The rule $X \to Y$ has to satisfy specified minimum support and minimum confidence measure (Shijue Zheng ,2006).

The support of Rule $X \to Y$ is the measure of frequency both X and Y in D

$$S(xy) = |xy| / |D| \qquad (1)$$

The confidence measure of Rule $X \to Y$ is for the premise that includes X in the bargain descend, in the meantime includes Y

$$C(x \to y) = S(xy) / S(x) \qquad (2)$$

Definition 2 Weighting support

Designated ones project to collect I = {i_1, $i2$, i_m}, each project i_j is composed with the value w_j of right $(0 \le wj \le 1,\ 1 \le j \le m)$. If the rule is $X \to Y$, the weighting support is

$$S_w(xy) = \frac{1}{k} \sum_{i \in xy} w_j S(xy) \qquad (3)$$

And, the K is the size of the Set XY of the project. When the right value Wj is the same as i_j, we calculating the weighting including rule to have the same support.

3. GENETIC ALGORITHM (GA)

Genetic Algorithm (GA) is a self-adaptive optimization searching algorithm. GA obtains the best solution, or the most satisfactory solution through generations of chromosomes constant evolution includes reproduction, crossover and mutation etc.

Here is the general description of this problem:

$$F(x) = a \times S(x) + b \times C(x) \qquad (4)$$

a, b is constants, a ≥*0, b*≥*0, S(x)* is the support, and *C(x)* is the confidence.

4. ASSOCIATION RULES MINING BASED ON A NOVEL GENETIC ALGORITHM

4.1 Encoding

This paper employs natural numbers to encode the variable A_{ij}. That is, the number of the lines of every range in the matrix A in which the element 1 exists is regarded as a gene. The genes are independent of each other. They are marked by A_1, A_2... A_j,..., A_n, in which $A_j \in [1, m], j \in [1, n]$ and A_n may be a repeatedly equal natural number.

When the distributive method at random is employed to produce the initial population comprised of certain individuals, the population must be in a certain scale in order to achieve the optimal solution on the whole. The best way is the generated *M* individuals randomly that the length is N , then the chromosome bunch encoded by the natural number is calculated as the initial population.

4.2 The Fitness

Formula (3) is properly transformed into:

$$F(xy) = W_s \times \frac{S(xy)}{S_{\min}} + W_c \times \frac{C(xy)}{C_{\min}} \qquad (5)$$

Here, $W_{C+}W_s=1$, $W_c \geq 0$, $Ws \geq 0$, $S_{min,}$ is minimum support, and C_{min} is minimum confidence.

4.3 Reproduction Operator

Reproduction is the transmission of personal information from the father generation to the son generation. Each individual in each generation determines the probability that it can reproduce the next generation according to how big or small the fitness value is. Through reproducing, the number of excellent individuals in the

population increases constantly, and the whole process of evolution head for the optimal direction. We are adopting roulette selection strategy; each individual reproduction probability is proportion to fitness value.

1) Compute the reproduction probability of all the individuals

$$P(i) = \frac{f(i)}{\sum_{i=1}^{M} f(i)} \qquad (6)$$

2) Generate a number r randomly, r=random [0, 1] ;

3) If $P(0)+ P(1)+\ldots+ P(i-1)<r< P(0)+ P(1)+\ldots+ P(i)$,the individual i is selected into the next generation.

4.4 Crossover Operator

Crossover is the substitution between two individuals of the father generation that is to generating new individual .The crossover probability Pc directly influences the convergence of the algorithm. The larger *Pc* is the most likely is the genetic mode of the optimal individual to be destroyed .However, the over-small of *Pc* can slow down the research process (Wu zhaohui,2005) .Here is the definition of the crossover operator:

Computing crossover probability Pc

$$Pc = \begin{cases} 0.9 - \dfrac{0.3(f(x) - \overline{f(x)})}{f_{\max}(x) - \overline{f(x)}} & f(x) \geq \overline{f(x)} \\ 0.9 & f(x) \prec \overline{f(x)} \end{cases} \qquad (7)$$

In which, $f_{max}(X)$ is the maximum fitness value of the population, $\overline{f(X)}$ is the average fitness value of the population.

4.5 Mutation Operator

The role of the mutation operator lies in that it enables the whole population to maintain a certain variety through the abrupt change of the mutation operator when a local convergence occurs in the population. The selection of the mutation probability P_m is the vital point because it influences the action and performance of the ARMNGA. If P_m is over-small, the ARMNGA will become a pure random research .Here is the definition of the mutations operator, computing the mutation probability P_m

$$Pm = \begin{cases} 0.1 - \dfrac{0.009\,(f(x) - \overline{f(x)})}{f_{\max}(x) - \overline{f(x)}} & f(x) \geq \overline{f(x)} \\ 0.1 & f(x) \prec \overline{f(x)} \end{cases} \quad (8)$$

In which, $f_{max}(X)$ is the maximum fitness value of the population, $\overline{f(X)}$ is the average fitness value of the population.

4.6 Termination Condition

When the matching error ε≈0 or the condition is not coincident, the process will naturally stop.

5. EXPERIMENTS ON SYNTHETIC DATA

To check the research capability of the operator and its operational efficiency, such a simulation result is given compared with the GA in [2] ,The platform of the simulation experiment is a Dell power Edge2600 server (double Intel Xeon 1.8GHz CPU,1G memory , Redhat Linux 9.0).

We first compare the performance of our proposed method with the algorithm in Fig. 1 shows the runtime vs. the size of an image for both algorithms, where the size of the image varies for the synthetic dataset. As the size of the image increases, the runtimes of both algorithms decrease; nevertheless, the runtime of the ARMAGA algorithm does not change very much.

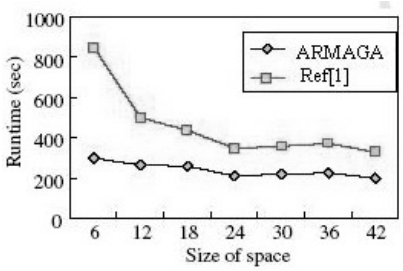

Fig.1 Runtime vs. image size

Fig.2 shows the runtime vs. number of objects for both algorithm, where the number of objects varies from 25 to 100 for the synthetic dataset .Since the average size of process and number of transaction are both fixed, the average support for the item sets decreases as the number of objects increases. Thus, the runtimes of both algorithms decrease slightly when the

number of objects increases. Nevertheless, our proposed algorithm is faster than the algorithm in.

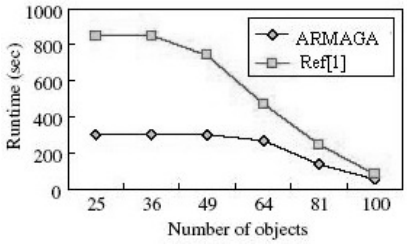

Fig.2 Runtime vs. number of objects.

From Fig.1 and Fig.2, we can educe that ARMAGA has a higher convergence speed and more reasonable selective scheme which guarantees the non-reduction performance of the optimal solution. Therefore, it is better than GA and ARMA through the theoretic analysis and the experimental results.

6. CONCLUSIONS

The image data mining in agriculture is a newly researching hot point in database area. But general data mining get knowledge from large quantities of data. We propose an Association rules mining based on a novel Genetic Algorithm, designed specifically for discovering association rules. We compare the results of the ARMAGA with the results of (G.chen,2002), and, it is better than GA and ARMA through the theoretic analysis and the experimental results.

REFERENCES

G. Chen, Q. Wei. Fuzzy association rules and the extended mining algorithms, Information Sciences 147 (2002) 201–228.

Gao Li, Li Dan, Dai Shangping. A mining Algorithm of constraint based association rules, journal of Henan University Vol.33 (2003) pp.55-58

K. Koperski, J. Han. Discovery of spatial association rules in geographic information databases, in: Proc. of International Symposium on Advance in Spatial Databases, SSD, LNCS, vol. 951, Springer Verlag, 1995, pp. 47–66.

P.Y. Hu, Y.L. Chen, C.C. Ling. Algorithms for mining association rules in bag databases, Information Sciences 166 (2004) 31–47.

Shijue Zheng, Wanneng Shu, Li Gao.Task Scheduling Using Parallel Genetic Simulated Annealing Algorithm ,2006 IEEE International Conference on Service Operations and Logistics, and Informatics Proceedings June 21-23, 2006, Shanghai, pp46-50

Wu zhaohui. Association rule mining based on simulated annealing genetic algorithm, Computer Applications Vol.25 (2005) pp.1009-1011