Design of Association Rules Data Mining System Based on Improved Ant Colony Algorithm

Xiaoying Sun[1, a *]

[1]School of Computer and Information Engineering, Nanyang Institute of Technology, Henan Nanyang, 473004, China

[a]shenruiminsjd@163.com

Keywords: Ant Colony algorithm; Data mining; Association rule; Group of wisdom; Path

Abstract. Data Mining is from large, incomplete, noisy, fuzzy and random Data, extract implicit in it, people don't know in advance, but it is potentially useful information and knowledge of the process. The Ant Colony algorithm is actually positive feedback principle, and it is an algorithm combining the heuristic algorithm. The Ant Colony algorithm is easy to fall into local optimum and slow convergence, many new models are put forward, such as ACA based on cloud model. The paper presents design of association rules data mining system Based on improved ant colony algorithm.

Introduction

At present, the group of intelligence theory research field includes two main algorithms: Ant Colony algorithm (ACO) and Particle Swarm Optimization (PSO).And represented by Ant colony algorithm of swarm intelligence has become A hot spot in today's distributed artificial intelligence research, it is by the Italian scholar m. Dorigo, v. Maniez zo, a. Colorini (three, four, five) and others was inspired from the mechanism of biological evolution, by simulating the nature after the behavior of ants search path, also known AS the Ant System, Ant System. M. Dorigo et al., to make full use of the ant colony search food process with the famous Traveling Salesman Problem (Salesman Problem) between the similarities, absorb the behavior characteristics of the ants, the design of the virtual "ant" grope for different routes, and leave will gradually disappear with time virtual "information content". Virtual "information content" will evaporate when ants randomly choose the path to go, tend to choose the path.

Data mining: (definition) from a large number of data mining, incomplete, noisy, fuzzy and random data, extract implicit in it, people don't know in advance, but is potentially useful information and knowledge of the process is called data mining.(function) concept description, correlation analysis, classification and prediction, cluster analysis, trend analysis, outlier analysis and deviation analysis, etc.(typical data mining system of database, data warehouse, or other information database; Database or data warehouse server; The knowledge base. Data mining engine; Graphical user interface.

Ants, bees and birds such as social animals can be done through collaboration, such as the discovery of new food source, building complex nests, thousands of kilometers across complex tasks such as migration to the designated area. Through the study of these animals have the concept of swarm intelligence [1]. Refers to the group of intelligence is to the body of the "no intelligence through cooperation showed the characteristics of intelligent behavior", is a kind of computing technologies based on biological group behaviors.

For the data set of data, there are two situations like this: there is a tremendous amount of missing data attribute values, we usually take the measures is directly deleted, but in some systems during the ETL process, can not deal directly with a large number of missing value. For the more important attributes, but also there are a few missing value, need to supplement the data after complete a series of data mining.

People workers imitate behavior, sensation and perception of the real ant setting reasonable absolute sense threshold to overcome the ants in the initial choice easy to lose the diversity of solution, to improve the selection strategy with adaptive change of information on the path, through the simulation of different size and asymmetric TSP algorithm has good convergence and stability, new I/ACA heuristic search method, intelligence, by eliminating pheromone, automatically adjust the proportion of the optimal path choice, change the basis and the introduction of disturbance to select target cities, the simulation results show that the decrease in the amount of calculation, at the same time can get better search results, but also pointed out that through the experiment to determine the relevant physical factor against algorithm promoted but this paper only TSP, other problems will be application is still not clear. The paper presents design of association rules data mining system Based on improved ant colony algorithm.

Typical Example of Association Rule Mining and Frequent Items

Although the ant colony algorithm research time is not long, but the preliminary research has shown that it has a great advantage in solving complex optimization problems, especially in Brussels, Belgium in 1998 specially convened the first international symposium on ant optimization, held once every two years, so it is now the international symposium on optimization of ants. This marked the study of ant colony algorithm has been widely international support, make this emerging intelligent bionic evolution algorithm shows vitality.

Swarm intelligence has the following features and benefits:

(1) Mutual cooperation among individuals is the distribution of (Distributed), it is able to adapt to the current working condition under the network environment.

(2) There is no central control and data, this system is more Robust (Robust), and not because of one or a few individual fault and affects the whole solution of the problem.

(3) Can not through direct communication between individuals, but by an indirect communication cooperation, such a system has better Scalability (Scalability).

For basic ACA is easy to fall into local optimum and slow convergence, many new models are put forward, such as ACA based on cloud model, the limitation on the pheromone, regression model and so on, and even a lot of researchers trying to from a new perspective to re-examine and tentative ACA, a novel has from the "polymorphism of ant colony society, try to closer to the real world of ant behavior to study ACA, found that more adapt to a larger problem, and will look at ant colony's overall research, speed switch to focus on individual ants from the Angle of the impact on the algorithm.

When the entire n city to join tabuk, ant k is complete a travel, the path through the ant k is a feasible solution of the TSP problem [2]. The eta in type 1 ij is a heuristic factor, said the ant from city I moved to the city's expectations of the j. In the AS algorithm, eta ij usually take the reciprocal of distance between city I and j. Alpha and beta respectively pheromone and heuristic factor of relative importance. After completing a travel around when all the ants, each path pheromone update according to type 2 [3].

$$\begin{aligned} z_1^k &= \{z(1), z(2), \cdots, z(k)\} \\ &= \{Z(1), Z(2), \cdots, Z(m-1); z((m-1)M+1), \cdots z((m-1)M+s)\} \\ &= \{Z_1^m; z(m,1), z(m,2), \cdots, z(m,s)\} \end{aligned} \tag{1}$$

Among them, the Q as normal number, Lk said the first k ant in the travel through the length of the path. M. Dorigo proposed three kinds of AS algorithm model, type (2.4) week called ant system, the other two models respectively called ant system and ant system.

To fundamentally solve the lack of ACA, its convergence analysis has been launched, such as using dynamic phases, and the concrete influence the parameters of the ACA also more and more attention, if you have the discussion, but there is little theoretical basis for how to set parameters, how to establish common standards to the optimal parameter setting is still the difficulty in the study were also from the beginning of the application scope of discrete domain extensions to the continuous domain, continuous domain study of convergence, and the design of the new model also.

The AS algorithm is actually positive feedback principle, and it is an algorithm combining the heuristic algorithm. When choosing a path, and it is not only use the ant pheromone trails, and it is with the help of the reciprocal of the distance between the cities as a heuristic factor. The experimental results show that the ant - cycle model than ant - quantity and ant - density model has a better performance [4].

This is because the ant - cycle model using global information updating the pheromone trails, and ant - quantity and ant - density model using local information. AS the time complexity of the algorithm for O n2 (NC * * m), the space complexity of the algorithm for S (n) = O (n2) + O (n * m), including NC said the number of iterations, n for city number, m is the number of ants, the calculation formula is as follows [5].

$$\tau_{ij}(t+1)=(1-\rho)*\tau_{ij}(t)+\Delta\tau_{ij}+\tau_{ij}^{*} \tag{2}$$

At the same time it also has some defects:

(1) Limited to local optimal solution, the nature of the solution from the algorithm, ant colony algorithm is also looking for a better local optimal solution, not importune is global optimal solution.

(2) the stagnation of the middle of the working process of the problem, and the algorithm convergence speed, at the beginning of the working process of the algorithm of iteration to a certain number of times after the ants may also be in a certain or some local optimal solution of the neighborhood near stagnation.

ACA of the convergence speed and the global optimal solution is a pair of contradiction, fast convergence speed, can lead to premature, trapped in local optimal solution, and when the pheromone update time algorithm to calculate the amount is too large, will lead to the slow convergence speed and application is not reality, in order to overcome these problems. The corresponding improved ACA is put forward.

Design of Association Rules Data Mining System Based on Improved Ant Colony Algorithm

For now, association rule mining technology has been widely used in western financial industry enterprises; it can successfully predict bank customers' requirements. Once you get the information, the bank can improve their marketing [6]. Banks in their own an ATM is bound to the customer might be interested in our product information, for the use of bank ATM users understand. At the same time, some well-known e-commerce sites also benefit from strong in association rules mining. These rules for electronic shopping website using association rule mining, and then set the user to buy bundle together. There are also some shopping websites use them to set the corresponding cross selling, is to buy a commodity customers will see another commodity advertisement.

A basic principle is that when a transaction does not include the length of k, is not necessarily contains the length of k + 1 set of categories. So we can move these transactions, so the next time the scan can be used in less number of sets of transactions [7]. This is the basic idea AprioriTid..

From the research trend of frequent itemsets mining algorithm in recent years, in order to improve the efficiency of the algorithm, proposed a series of hybrid search strategy and efficient pruning strategy. On the basis of nature based on Apriori algorithm Based on the pruned concept lattice model said and mining frequent itemsets is proposed based on the photo model algorithm to solve the frequent itemsets, improved the performance of time and space of frequent itemsets mining algorithm, as is shown by equation (3).

$$\begin{cases} C_{j+1} = HC_j H' \\ D_{j+1}^{h} = GC_j H' \\ D_{j+1}^{v} = HC_j G' \\ D_{j+1}^{D} = GC_j G' \end{cases} \qquad (j = 0,1,2,\cdots,J-1) \tag{3}$$

Among them, the Q as normal number, Lk said the first k ant in the travel through the length of the path. M. Dorigo proposed three AS algorithm of the model, the type (2.4) is called ant - cycle, the other two models respectively called ant - quantity and ant - density, the difference is mainly in the (2, 4) type, that is, the ant - quantity model.

If (r, s) belongs to the global optimal path delta tau (r, s) = 1 / Lgb, otherwise 0. The alpha for pheromone is volatilization parameters. Lgb for so far is to find the global optimal path [8]. Lgb here also can substitute the Lib, the former is the global optimal, the latter is the most optimal iteration. In the process of creating solutions to problems, local pheromone updating rule. Local pheromone updating is 3.6 type rules. Local update rules makes corresponding pheromone gradually reduce, can effectively avoid the ant converge to the same path.

System Experiments and Analysis

Data cleaning: including filling vacant values, identify outlier, remove noise and irrelevant data. B data integration: combine data from multiple data sources in a consistent data storage [9]. Need to pay attention to different sources of data matching problem, numerical conflict and redundancy, etc. C data transformation: to convert raw data into suitable for data mining form. Including the summary of the data, gathered, generalization, standardization, may also need to be property of refactoring. D data reduction: narrow the scope of the data, make it more suitable for the need of data mining algorithm, and can get is the same as the original data analysis results.

The parameters of the ACA is often through trial and error, by experience together, but the computation efficiency and convergence of the algorithm will produce adverse effect is the main basis of ACA and a heuristic algorithm based on the principle of positive feedback and information, type 1 fully illustrates this [10]. If on a path pheromone and the shorter the path, the more, the greater the probability of the path is selected, the transition probability, became the "transfer coefficient", since the ants always choose the path with maximum transfer coefficient, the value of certainty, and at this point, the random disturbance of ant colony algorithm is the same as the basic ant colony algorithm, is inevitably appear stagnation phenomenon, thus puts forward the variable disturbance factors, as is shown by equation (4).

$$\begin{cases} \Delta p = \vec{g}_{\Omega+\partial\Omega} & on \quad \Omega \\ p\,|_{G+\partial G} = q\,|_{\Omega+\partial\Omega} & on \quad \partial\Omega \end{cases} \tag{4}$$

Though it is a waste of time and in the record to rewrite the I/O overhead, and it is but with the increase of cycling times, this algorithm for later in the 'new generation database scan times soon comparing will reduce gradually. From the evaluation standard of data mining is tools (1). The number of species; (2) ability to solve complex problems, (3) The operating performance, 4. Data acquisition ability, 5 mining results output, 6. The noise robustness of data processing and mining tools), this algorithm can achieve a higher standard; From the Angle of the efficiency of the algorithm, the algorithm of mining efficiency is higher. Therefore the author believes that this method is worth using in practical application.

Support is greater than the minimum support of itemsets called frequent itemsets. When the association rules and satisfies (min_sup) is greater than the minimum support threshold and minimum confidence threshold (min_conf), known as the strong association rules, the association rules and vice called weak association rules. The above threshold needs to be set automatically according to the data mining. For convenience of calculation, value the support and confidence of the general use value between 0 ~ 100% rather than a value between 0 ~ 1.0.

For transactional database, some implied concepts have layers. For example, we say "down" goods, for an application analysis and decision may care for its higher level concepts: "winter clothes", "clothing", etc. Different users and it is which may be of certain level association rules more meaningful. At the same time, due to the distribution of data and efficiency considerations, data may be stored data on the multi-level granularity, therefore, multi-level association rule mining could be obtained is more common knowledge.

Summary

Ant colony algorithm was first used to solve the traveling salesman problem (TSP). Since the famous traveling salesman problem and artifacts to effective scheduling problem, already seeped into other areas. One of the most successful is its application in the combinatorial optimization problem, these applications can be divided into two kinds: one kind is static combinatorial optimization problem, its typical representative is TSP, QAP, workshop scheduling problem, vehicle routing problem, etc. Another kind is applied to dynamic combinatorial optimization problems, such as network routing problem. The paper presents design of association rules data mining system Based on improved ant colony algorithm. Internal database consists of a set of related data and a set of management software program and access data, and is an operational database, the source data is composed of data warehouse. It organizes the data in table, using the ER data model. (database) same they all provides the source data for data mining, is the combination of the data. (system structure) is the underlying data warehouse server, is always a relational database system.

References

[1] M. Dorigo, V. Maniezzo, and A. Colorni. The Ant System: Optimization by colony of cooperating agents [J]. IEEE Transactions on Systems, Man, and Cybernetics–Part B, 1996, 26(1): 29-41.

[2] M Dorigo and LM Gambardella. Ant colony system: A cooperative learning approach to the traveling salesman problem [J]. IEEE Transactions on Evolutionary Computation, 1997, 1(1): 53-66.

[3] B Bullnheimer, R F Hartl, and C. C Strauss. A new rank-based version of the Ant System: A computational study [J]. Central European Journal for Operations Research and Economics, 1999, 7(1):25-38.

[4] IRIDIA. Ant Colony 0ptimization——Artificial Ants as A Computational Intelligenee Technique [J].IRIDIA - Techinal Report Series,2006(6):1-9.

[5] Guoqing Chen, Hongyan Liu, Lan Yu, Qiang Wei, Xing Zhang. A new approach to classification based on association rule mining .Decision Support Systems, Volume 42, Issue 2, November 2006, Pages 674-689.

[6] Brin S, Motwani R, Ullman J D et al. Dynamic itemset counting and implication rules for market basket analysis. In: Proceedings of 1997 ACM-SIGMOD International Conference on Management of Data. Tucson, AZ, 1997:255-264.

[7] Ya-Han Hu, Yen-Liang Chen .Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism Decision Support Systems, Volume 42, Issue 1, October 2006, Pages 1-24.

[8] Somboon Anekritmongkol, Kulthon Kasamsan, "The Comparative of Boolean Algebra Compress and Apriori Rule Techniques for New Theoretic Association Rule Mining Model", IJACT, Vol. 3, No. 1, pp. 58 ~ 67, 2011.

[9] Park J S, Chen M S, Yu P S. An Effective Hash-Based Algorithm for Mining Association Rules. In: Proceedings of ACM SIGMOD International Conference Management of Data, San Jose,CA,1995:175-186.

[10] Shelly Salim, Sangman Moh, "Energy-Efficient Clustering Based on Game Theory for Apriori Rule Techniques ", IJEI, Vol. 5, No. 3, pp. 31 ~ 37, 2014.