A Fuzzy Multi-Objective Evolutionary Subgroup Discovery Algorithm Based on MapReduce

Jian Nong*
Guangxi Colleges and Universities Key Laboratory of Professional Software Technology, Wuzhou University, WuZhou543002, Guangxi, China

Pengyu Xie
Guangxi Power Grid Co., Ltd, NanNing530023, Jiangsu, China
*Corresponding author

Abstract
Subgroup discovery algorithm is a new data mining technology, which is used to extract valid information from data, and the knowledge acquired by it is usually expressed in the form of rules. MapReduce can process and generate an algorithm model for extremely large data sets, which can realize the parallel processing on a large number of general configured computers. Fuzzy systems are usually used to optimize or search problems, and have demonstrated the effectiveness in solving problems in different application areas. Combined with the advantages of fuzzy system and genetic algorithm, fuzzy logic and genetic algorithm are combined to produce a genetic fuzzy system, which can be used for subgroup discovery. This paper studies the fuzzy multi-objective evolutionary subgroup discovery algorithm based on MapReduce, and through the analysis of its characteristics, this paper finds that the evolutionary distributed parallel design of fuzzy multi-objective evolutionary algorithm based on MapReduce is an effective way to reduce the running time of the algorithm.

Key words: MapReduce, Multiple Objective, Subgroup Algorithm, Fuzzy Rule, Evolutionary Process

1. INTRODUCTION
The massive data processing on the C server with the rapid expansion of the Internet cyber source, the Internet contains various types of data PU mass, IO throughput is a severe test, whether the processing speed, storage space, fault tolerance, or in the access speed and so on, the traditional architecture and relying on a single computer based on the serial way more suited to the current massive data processing requirements.

Scholars at home and abroad put forward a lot of massive data processing methods to improve the massive data processing problems (Zheng, 2008). Currently, some data processing method is easy to understand in concept, but due to the huge amount of data, to complete the corresponding processing within an acceptable period of time, only the parallel processing of these calculations, extracted by the amount of parallel work can be exist in the process, to achieve the parallel component of the parallel execution the process of using the distributed model. With the development of technology, the single performance has been changing rapidly, especially the memory and processor hardware technology, but the development of the hardware technology in theory is limited, if the development of hardware and improve the performance of the system in the vertical direction, so the development of the parallel technology is extended from the horizontal processing the way. MapReduce is a programming model launched by Google, and also an algorithm model that can process and generate super large data sets, which can be parallelized on a large number of common configured computers.

In recent years, the multi-objective evolutionary algorithm has entered the stage of comprehensive development, in order to make a breakthrough in the performance of the algorithm, to make up for lack of basic genetic algorithm under the framework of similar chaotic search, differential evolution, CO evolutionary strategy was introduced into the MOEA, the hybrid multi-objective evolutionary algorithm has become the mainstream of the development direction objective evolutionary algorithm. At the same time, with the advent of the information age, the emergence of complex optimization problems more and more large, complex and interactive branch miscellaneous wrong in real life, it also needs to consider the optimization of several or even dozens of objects, this kind of problem is also known as a multi-objective optimization problem of high dimension (MOOP), classical multi-objective evolutionary the algorithm is not ideal in the efficiency of the solution of MOOP, a target growth is often accompanied by the time complexity of the algorithm exponentially, so many excellent multi-objective evolutionary algorithm in high dimension under the goal of not always

Effective promotion and use. Therefore, the solution of MOOP has become one of the hot topics in multi-objective evolutionary algorithms. Subgroup found selection of evaluation criterion to evaluate the obtained rules of quality of various quality algorithm, so the subgroup discovery is a multi-objective problem, the quality
evaluation standard of different rules in the evolutionary process can be regarded as the goal of evolution, different genetic algorithm in multi-objective evolutionary algorithm to solve the multi-objective optimization problems found in subgroup.

2. THE WORKING MECHANISM OF MAPREDUCE

2.1. The Working Principle of MapReduce

MapReduce is used by Hadoop to process ultra large amounts of data. It is divided into map processing phase and reduce processing phase, and each stage of input and output is the developer can define the type of key value pair format. Developers can write map function and reduce function of their own to achieve their goals, without the need for distributed too much attention, because the MapReduce framework itself is a distributed framework, distributed development work for developers to provide a great convenience. Data is a key from the HDFS file system on the value of the input format to the map function, developers get the data, and can be customized in the inside of the map operation is applied to the input data output data need, after the end of map, its output data will be saved to the HDFS, then to reduce by stages read into the program, developers can also require custom actions in reduce.

![Figure 1. The working principle of MapReduce operation](image)

As shown in Figure 1, the working principle of the MapReduce job is that after the program runs, the first step is that the job client requests a new job ID from the job tracker, and then gets the input data from the HDFS and submits the job to the job tracker. When the job tracker receives the submitted job, it initializes the job and creates the corresponding map and reduce tasks to perform the job.

2.2. Hadoop Open Source Distributed Framework

Hadoop is an open source distributed framework that provides support for data intensive distributed applications, including data storage and data processing capabilities. Hadoop is not a single item, it is a general term for a group of related projects, which contains multiple and distributed computing and massive data processing related projects, such as HDFS, MapReduce etc.. Users don't need to understand the distributed details of Hadoop, but they can still easily use it to develop distributed applications. The emergence of Hadoop has brought great benefits to developers of distributed applications. Hadoop supports expansion in a fault-tolerant way, this feature enables the Hadoop to handle a large number of cheap personal computers with large amount of data, greatly improving the ability of processing data in the mass data processing at the same time, greatly reduces the cost of hardware overhead, and made great contributions to the development and promotion of large data processing. The birth of Hadoop has promoted the wave of big data sweeping the world, and opened up a new situation for big data processing and distributed computing.

2.3. Hadoop Distributed File System

When the amount of data is too large, so that a single physical computer alone cannot store, makes the data processing work can not be on a separate physical machine smoothly, at this time will need to handle data segmentation, respectively to a plurality of separate physical machines for operation. But when we partition data that needs to be processed to different machines, we need an additional operation to manage the data that is distributed across different machines, and distributed file systems emerge as the times require. In order to better
manage the data on each independent computer, Hadoop has its own abstract distributed file system, in which HDFS is a more widely used implementation, this experiment is built on the HDFS.

HDFS believes that the most efficient access mode should be to write data once, and then read the data multiple times. This is because the one-time write data data set after a long period of time will be the most data of the data set and all the data were all related to the operation, the characteristics of the data write result if repeatedly writing will spend a lot of time in the data, and according to the characteristics, write once reading can greatly reduce the time loss.

Another feature of HDFS is its reliability. Because one of the features of Hadoop is to reduce operating costs, the clusters it runs are made up of cheap but unreliable hardware. However, it also brings some problems while reducing costs, one of the most serious point is composed of cheap hardware large cluster, node failure rate is high, and this phenomenon will seriously affect the smooth progress of the project, which will lead to more losses. In order to solve this problem, HDFS has been designed to add such a feature, when the node fails, HDFS does not stop working, but still can continue to run normally.

3. Multi objective evolutionary subgroup discovery algorithm

3.1. Subgroup Discovery

The subgroup finds that the group with the target attribute is described in the form of rules, and the subgroup discovery is very suitable for finding the relationship between the attribute of the target and the attributes of other variables. The rules extracted from subgroup discovery have the following form

\[ Cond \rightarrow \text{Class} \]  

Among them, \( \text{Class} \) is the target attribute value, said rules describing the group category, called the rule, the \( Cond \) value of a variable, described with unusual distribution groups for the target attribute.

The related processing steps of subgroup discovery are as follows

1. Input data preprocessing.
2. Discretization of continuous numerical variables.
3. Use of expert knowledge.
4. Regular hoof selection.

3.2. Fuzzy System

Fuzzy rules show the knowledge that the user is interested in in the form of close to the human language, and it is easier to acquire knowledge from the rules (Li and He, 2010). Genetic algorithm is a search algorithm based on natural inheritance, which can provide better search ability in complex space. The system combines fuzzy logic and genetic algorithm called genetic fuzzy system, genetic fuzzy system is applied to the sub group, rules and consequent by fuzzy logic statement, to obtain a series of novel form, strong interpretation of fuzzy rules, describe the different groups of data set (Ren, 2012). In the subgroup discovery algorithm, fuzzy rules can be used to obtain knowledge in a way similar to human language. Most of the logic variables in human language are not purely two values or multiple logical values, but have fuzzy true and false connections. Fuzzy rules tend to replicate language knowledge, so fuzzy rules can be used in subgroup discovery, data analysis and pattern recognition to obtain knowledge with high interpretability and flexibility.

Fuzzy systems are usually used to optimize or search problems, and have demonstrated the effectiveness of the problem in solving different applications. Combined with the advantages of fuzzy system and genetic algorithm, genetic fuzzy system is generated by combining fuzzy logic with genetic algorithm, which can be used for subgroup discovery (Liu, 2010).

![Diagram of Genetic Fuzzy System](image-url)

**Figure 2.** The main structure of genetic fuzzy system
3.3. Objective Function of Evolutionary Process

Based on fuzzy rules and non dominated multi objective evolutionary subgroup discovery algorithm, combined with the reasoning ability of fuzzy system and learning ability of multi-objective evolutionary algorithm, a series of fuzzy rules with high accuracy and interpretability are obtained based on the algorithm, which is used to describe the groups of users interested (Zhu and Chen, 2007). The initialization process reduces the number of variables contained in the algorithm, and introduces the overlay based re initialization to improve the diversity of the rule population. In order to obtain the accurate rules covering most of the standard samples, the support, confidence and accuracy of the rules are chosen as the objective functions of the evolutionary process (Zhan, Wang and Bao, 2012). The purpose of the rule population evolution process is to discover the universal rules with high confidence and describe most of the target samples in the target population.

![Flowchart of multi objective evolutionary subgroup discovery algorithm](image)

**Figure 3.** Flowchart of multi objective evolutionary subgroup discovery algorithm

Support of rules expressed as class type is due to the underlying rules covering the proportion of samples, the formula is as follows

\[
Sup \_2 (R) = \frac{n(ClassCond)}{n(Class)}
\]  \hspace{1cm} (2)

The rules of confidence is a measurement standard, that is covered by rules in the sample meet the complete rules of proportion, using fuzzy rules "", will the reliability are calculated for the complete coverage of the sample membership rules and sum of all meet before rule of the ratio of total sample membership.

\[
Conf \_1 (R) = \sum_{E_k \in E \_k \cap Class} \frac{APC(E_k, R)}{\sum_{E_k \in E} APC(E_k, R)}
\]  \hspace{1cm} (3)
The accuracy of the rule is proportional to the sample size in the sample covered by the rule

\[ Accu(R) = \frac{n(ClassCond)}{n(ClassCond) + k} \]  

(4)

3.4. Analysis Measure

The measure is used to measure the uniformity of the distribution in the Pareto front,

\[ E = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n-1} (d_i' - \overline{d})^2} \]  

(5)

\[ \overline{d} = \frac{1}{n-1} \sum_{i=1}^{n-1} d_i' \]  

(6)

For a Pareto frontier set, the \( i \) is the ordinal number of each solution, which is the distance between the \( i \) individual and the \( i+1 \) individual (Euclidean distance), which is based on the interval value of the first dimensional direction. The formula (5) shows that if the distribution of the Pareto front is uniform, the \( E \) measure is needed to be reduced.

\( D \) dispersion can be used to measure the dispersion of Pareto front,

\[ D = \sqrt{\sum_{m=1}^{S} \left( \max(f_m(x_i,u)), \min(f_m(x_i,u)) \right)^2} \]  

(7)

4. IMPROVEMENT OF MULTI OBJECTIVE EVOLUTIONARY SUBGROUP DISCOVERY ALGORITHM

4.1. Weighted Covering Algorithm

The biggest difficulty of using classification algorithm to find subgroups is how to use the covering algorithm to obtain the rules to meet the requirements. For the description of user interest groups can be only a small part of the classification rules, because in the process of running the algorithm, some rules were originally obtained with high coverage and importance of the classification algorithm in order to maximize classification accuracy, will continue to extract rules from a small amount of uncovered target in the sample at this time, the new rules cover only a small amount of the target sample, because subgroup discovery aims to obtain simple rules to describe the target sample, most of the target group so that this part of the rule is not a sub group found hope to get.

In order to improve the coverage algorithm, MapReduce algorithm using the weighted coverage algorithm based on rule induction in the process still can ensure that important degree rule is high, and can cover most of the target groups in the sample. The classification algorithm in the algorithm will grant cover rules cover from the current grant wood samples removed, weighted covering algorithm will be no rule coverage from the current sample wood samples removed, with a variable count record of the current sample trees covered the number of samples in each process but the rules, through the calculation of sample count weight.

4.2. Quality Evaluation Method

A fuzzy multi-objective subgroup discovery algorithm based on MapReduce is used to evaluate the quality of the rules obtained by descriptive and predictive methods. Because the main task is to extract rules of subgroup discovery is used to describe the target group, so, in the subgroup discovery algorithm, descriptive evaluation methods occupy a large proportion of the algorithm using the predictive evaluation method is to obtain classification rules that has high accuracy, can be used to compare with other classification algorithms.

The descriptive evaluation method is used to evaluate the quality of the single rule obtained, and the average quality of all rules can be evaluated to evaluate the quality of a series of rules obtained (Wu and Liu, 2013). Although the multi-objective fuzzy subgroup discovery algorithm to obtain the final purpose is not to have the maximum classification accuracy of the rules, but in the subgroup discovery using the predictive evaluation method, one can assess the classification accuracy of the rules, on the other hand can be compared between the obtained results and other classification results.

In order to test the performance of the algorithm, this paper selects the following interval multi objective function as the test function to test (Wang, Wang and Cao, 2014). As can be seen from the chart, the convergence effect is very good, and can get the expected results.
Figure 4. The simulation result
Figure 5. The influence of uncertainty on the Pareto front
5. CONCLUSIONS

Fuzzy systems are usually used to optimize or search problems, and have demonstrated the effectiveness in solving problems in different application areas. Combined with the advantages of fuzzy system and genetic algorithm, fuzzy logic and genetic algorithm are combined to produce a genetic fuzzy system, which can be used for subgroup discovery. This paper studies the fuzzy multi-objective evolutionary subgroup discovery algorithm based on MapReduce, and through the analysis of its characteristics, this paper finds that the evolutionary distributed parallel design of fuzzy multi-objective evolutionary algorithm based on MapReduce is an effective way to reduce the running time of the algorithm.

ACKNOWLEDGEMENT

This research work is supported by National Natural Science Foundation of China(Grant No.61502343),Natural Science Foundation for Young Scholars of Guangxi (Grant No.2015GXNSFBA139262), basic ability improvement project of young college teachers of Guangxi (Grant No. KY2016LX366).

REFERENCES