Parallel ID3 Algorithm Based on Granular Computing and Hadoop

Liu Ping[1], Wu Zhenggang[1], Zhou Hao[1], Yang Junping[2] and Taorong Qiu[1,*]

[1]The School of Information Engineering, Nanchang University, Nanchang 330031, China

[2]Jiangxi University of Traditional Chinese Medicine, China

Abstract: Large data processing has become a hot topic of current research. How to efficiently dig out useful information from large amounts of data has become an important research direction in the field of data mining. In this paper, firstly, based on the idea of granular computing, some granular concepts about the decision tree are introduced. Secondly, referring to granular computing, the improvement and parallelization of ID3 algorithms are presented. Finally, the proposed algorithms are tested on two data sets, and it can be concluded that the algorithm's classification accuracy is improved. From the test on a Hadoop platform, the results demonstrate that parallel algorithms can efficiently process massive datasets.

Keywords: Data mining, granular computing, hadoop, ID3, large data processing.

1. INTRODUCTION

With the advent of the era of big data, data processing has drawn more and more attention of scholars. How to extract better very important information behind the data has become a topic of growing concern. In the field of data mining, decision tree technology is considered the key technology, and there exist common classification algorithms such as the ID3 algorithm, C4.5 algorithm, KNN, SVM and similar algorithms.

In order to adapt to the demand of large data processing, in recent years the Google Company has put forward a distributed file system (Google file system, GFS) and a Map-Reduce parallel programming model [1-5], which provide the infrastructure for massive data mining, but traditional data mining algorithms cannot be directly applied to the parallel platform. Therefore, parallelization of the traditional data mining algorithm has become a research focus of scholars at home and abroad.

Granular computing (GrC) [6-10] is a new concept of computational model, and may be regarded as a label for a family of theories, methodologies, and techniques that make use of granules. In the domain of data mining, GrC provides a conceptual framework for studying a number of issues.

This thesis is arranged as follows: In section 2, the traditional ID3 algorithm is briefly introduced, and the concept of GRC is also discussed, including, the defects of the traditional ID3 algorithm; an improved method is put forward. In section 3, the parallelization of the improved ID3 algorithm is discussed. In section 4, some algorithm accuracy testing and parallel efficiency testing are carried out and the results of the tests are discussed. Conclusions are given in the last section.

2. THE TRADITIONAL ID3 ALGORITHM

2.1. The Basic Principle

ID3 [11-15] is a decision tree algorithm based on information entropy. A decision tree is a flowchart like a tree structure, where: each internal node (non-leaf node) denotes a judge on an attribute, each branch represents an attribute value, and each leaf-node (terminal node) holds a class label. The core of the algorithm is to use the information entropy as the training sample set splitting measurement standard in the generation of the decision tree.

The algorithm proceeds as follows [11-15]: firstly, it calculates each attribute's information entropy which in turn selects the best attribute as the splitting attribute. Then, the data are partitioned into subsets in accordance with the values of the splitting attribute. For each subset, recursive implementation of the above process is carried out until the each tuple is correctly classified.

2.2. The Shortcomings of the ID3 Algorithm

Although ID3 is a typical decision tree classification algorithm, yet there are still shortcomings:

1) Because the attribute that has the maximum information entropy is selected as the splitting attribute, so the choice will be more inclined to the one that has more different attribute values, but not more attribute values is the best attribute.

2) The bigger the data set is, the more increased the computation of the algorithm.

The traditional calculation formula for information entropy is expressed as:

The given probability distribution $P=(p_1,p_2,...,p_n)$, then the information entropy of P is defined as $I(P)=-\sum_{i=1}^{n} p_i \log_2 p_i$.

*Address correspondence to this author at the School of Information Engineering, Nanchang University, Nanchang 330031, China;
Tel: +86079183969684; E-mail: liuping@ncu.edu.cn

Suppose that $S=(U,A,D)$ is a set of data samples. Assume that class label attribute has m different values, expressed as: $U/D=\{D_1,D_2,...D_m\}$, $U/A=\{a_1,a_2,...a_v\}$, s_i is the number of samples in $D_i(i=1,2,...m)$, $s_i=|D_i|$, Equation (1) gives sample classification based on the expected information.

$$I(s_{1,}s_2,...,s_m)=-\sum_{i=1}^{m}p_i\log_2(p_i)\text{ , }p_i=s_i/s;s=|U| \quad (1)$$

$$E(A)=\sum_{i=1}^{v}\frac{s_{1j}+...+s_{mj}}{s}I(s_{1j,...,}s_{mj}) \quad (2)$$

$$I(s_{1j},s_{2j},...,s_{mj})=-\sum_{i=1}^{m}p_{ij}\log_2(p_{ij}) \quad (3)$$

Where $p_{ij}=s_{ij}/|s_{ij}|$ is a sample of S_j in the probability of belonging to class D_i.

The specific definition of the above three formulae can reference literature [11-15].

The most important aspect in the process of constructing a decision tree is the best splitting attribute selection, and the key to the best splitting attribute selection is to calculate the information entropy. From the above three formulae, it can be seen that the calculation of critical information entropy is the probability distribution of the condition attribute and the class attribute.

Assuming there are n objects in the sample set and m attributes of each object. In order to derive the probability distribution of attribute values, the time complexity of the system is at least $O(mn)$.If the algorithm is to deal with large data sets, one can imagine the time complexity. Therefore, in consideration of the above, a new method of calculating the information entropy is proposed.

2.3. An Improved Information Entropy Based on Granule Computing

Hypothesis that $S=<U,C\cup D,V,f>$, $A=C\cup D$ be an information system(IS),Where $U=\{u_1,u_2,...,u_{|U|}\}$ is a set of objects. It is a non-empty finite set of objects, called the universe. $A=C\cup D$ is a set of attributes. C is the condition attribute set, and D is the decision attribute set. An information granule [6] is defined as the tuple $G_r=(\phi,m(\phi))$, where ϕ is expressed as the intension of information granule G_r, and $m(\phi)$ is expressed as the extension of information granule G_r.

Definition 1; (Elementary granule):

Let $G_r=(\varphi,m(\varphi))$ be an information granule [6]. If $\varphi_{ij}=(a_i,v_i^j)$, $m(\varphi_{ij})=\{u\mid f(u,a_i)=v_i^j,u\in U,a_i\in A\}$, where $m(\varphi_{ij})$ refers to the object set whose attribute values of $a_i(a_i\in A)$ are v_i^j. v_i^j is the $j-th$ attribute value of attribute a_i, then G_r is called an elementary information granule.

$G_r=((a_i,v_i^j),m(a_i,v_i^j))$, if $a_i\in C$, then it is called the conditional granule; if $a_i\in D$, then called the decision granule.

Definition 2; (Size of information granule).

For $G_r=(\varphi,m(\varphi))$, the size of the information granule [6] can be defined as the cardinality of the extension of the information granule, and is expressed as $\|m(\varphi)\|$.

Definition 3;($\otimes$ operation)

Suppose that $G_{r1}=(\varphi,m(\varphi))$ and $G_{r2}=(\phi,m(\phi))$ are the elementary granules, the operation of intersecting between them is defined as: $G_{r1}\otimes G_{r2}=(\varphi\cup\phi,m(\varphi)\cap m(\phi))$ (4)

Definition 4; (binomial granule)

Suppose that $G_{r1}=(\varphi,m(\varphi))$ and $G_{r2}=(\phi,m(\phi))$ are the elementary granules, the binomial granule is defined as; $G_r=G_{r1}\otimes G_{r2}=(\varphi\cup\phi,m(\varphi)\cap m(\phi))$

Definition 5; (A combination of granules)

If $G_{r1}=(\varphi_1,m(\varphi_1))$, $G_{r2}=(\varphi_2,m(\varphi_2))$,… $G_{rn}=(\varphi_n,m(\varphi_n))$ are the elementary granules, A combination of granules is defined as;

$$G_r=G_{r1}\otimes G_{r2}\otimes...\otimes G_{rn}=(\varphi_1\cup\varphi_2\cup...\cup\varphi_n,m(\varphi_1)\cap m(\varphi_2)...\cap m(\varphi_n)) \quad (5)$$

Definition 6; (The information entropy of attribute: $Gain(U,c_i)$)

Without loss of generality, assume that there is only one decision attribute. Suppose that the elementary granule $\varphi_{ij}=(c_i,v_i^j)$, $m(\varphi_{ij})=\{u\mid f(u,c_i)=v_i^j,u\in U,c_i\in A\}$

$G_r(\varphi_{ij})=((c_i,v_i^j),m(c_i,v_i^j))$ $G_r(d_k)=((d_k,v_d^k),m(d_k,v_d^k))$

$\omega_{ijk}=G_r(\varphi_{ij})\otimes G_r(d_k)$

$$Gain(U,c_i)\equiv Entropy(U)-\sum_{j\in\{1,2,...,|v_i|\}}\frac{|m(\phi_{ij})|}{|U|}Entropy(\omega_{ijk}) \quad (6)$$

$$Entropy(\omega_{ijk})\equiv\sum_{k=1}^{h}-p_k\log_2^{p_k}\text{ , }p_k\equiv\frac{|m(\omega_{ijk})|}{|m(\phi_{ij})|} \quad (7)$$

3. THE PARALLEL ID3 ALGORITHM BASED ON GRANULAR COMPUTING

Through the above analysis, the time complexity calculation of each attribute's information entropy is relatively high in the construction of the decision tree. So, computing attribute information entropy can be paralleled based on a Hadoop cloud computing platform.

The key to the algorithm's parallelization is the design of the map function and reduced function.

3.1. The Granulation of the Information System

Algorithm 1：Map (object key, value)

Input：Decision information system

$S=(U,A,V,f),A=C\cup D$, object $x\in U$, attribute $a\in A$。

```
Output: elementary granule space $G_{rs}=\{G_r\}$。$G_r=(\varphi,m(\varphi))$
(key: $\varphi=(a_i,v_i^j)(1\le i\le m)$ ; Value: $m(\varphi)$
//$m=|A|$, $a_i\in A$ . $v_i^j$ is the $j-th$ attribute value of attribute $a_i$
//$m(\varphi)=\{u\mid f(u,a_i)=v_i^j,u\in U,a_i\in A\}$
)
for $u\in DS_i$ do // $DS_i$ :the $i-th$ data slice
for ( $a_i\in A(1\le i\le|A|)$ ) do
if ( $f(u,a_i)=v_i^j$ )
 { $\varphi=(a_i,v_i^j)$;
   $m(\varphi)=\{u\}$;
Emit $<\varphi,m(\varphi)>$
}
Algorithm 2 : Reduce (Text $\varphi$ , Iterable <Text> values)
//The final result of processing the map function will be sent to the merger of the reduce function, //when combined, has the same key as is sent to the same reducer.
Input : granule $\varphi=(a_i,v_i^j)(1\le i\le m)$ ,   $m(\varphi)=\{u\}$;
Output : < granule $\varphi=(a_i,v_i)(1\le i\le m)$ , the union of $m(\varphi)$>
For $i=1$ to values .size()
 $m(\varphi)=m(\varphi)+values[i]$;
Emit $<\varphi,m(\varphi)>$
```

Through the parallel processing process, the information system has been converted into a particle space. Next, based on the particle space, each attribute's information gain can be calculated, and thus the best splitting property can be found.

3.2. The Attribute Information Entropy Calculation Based on Granule

```
Algorithm 3: The formation of conditions granule space $G_{rc}$ and decision granule space $G_{rd}$
Input : Elementary granule space $G_{rs}=\{G_r\}$。$G_r=(\varphi,m(\varphi))$
Output: Condition granule space $G_{rc}=\{G_r\}$。
Decision granule space $G_{rd}=\{G_r\}$ ; $G_r=(\varphi,m(\varphi))$
Step1: $G_{rc}=null$; $G_{rd}=null$
Step2: for $G_r=(\varphi,m(\varphi))\in G_{rs}$ do //$\varphi=(a,v)$
 {if ( $a\in C$ )
    $G_{rc}=G_{rc}+\{G_r\}$ ;
  if ( $a\in D$ )
    $G_{rd}=G_{rd}+\{G_r\}$ ;
 }
```

```
Algorithm 4: According to the formula, calculate conditions of information gain
input: Condition granule space $G_{rc}=\{G_r\}$ 。
Decision granule space $G_{rd}=\{G_r\}$ ; $G_r=(\varphi,m(\varphi))$
Output: The information entropy of each condition attribute $Gain(U,c_i)$
step1 : According to definition 6,
  for  $G_r(\varphi_{ij})\in G_{rc}$
   for $G_r(d_k)\in G_{rd}$
  compute $\omega_{ijk}=G_r(\varphi_{ij})\otimes G_r(d_k)$ ;
step2: compute the condition information entropy
 $\varphi_{ij}=(c_i,v_i^j)$ $m(\varphi_{ij})=\{u\mid f(u,c_i)=v_i^j,u\in U,c_i\in A\}$
$G_r(\varphi_{ij})=((c_i,v_i^j),m(c_i,v_i^j))$
$G_r(d_k)=((d_k,v_d^k),m(d_k,v_d^k))$
```

$$Gain(U,c_i)\equiv Entropy(U)-\sum_{j\in\{1,2,\ldots,|v_i|\}}\frac{|m(\phi_{ij})|}{|U|}Entropy(\omega_{ijk})$$

$$Entropy(\omega_{ijk})=\sum_{k=1}^{h}-p_k\log_2^{p_k}\quad p_k=\frac{|m(\omega_{ijk})|}{|m(\phi_{ij})|}$$

4. THE EXPERIMENTAL ANALYSIS

(1) The experimental environment：

The software environment: Hadoop-1.0.4 , Ubuntu Linux 10.04.4, Jdk1.6.0_41。

The hardware environment：4 computer, 1 master, 3 slaves。

(2) Data source

The experimental data from the UCI (http://archive.ics.uci.edu/ml/datasets.html) data set.

(3) The experimental analysis：

a) Test and analysis of algorithm accuracy

In testing the classification accuracy, randomly selected 90% of the data from the original data as training set to build a decision tree, select 10% of the data as a test set, and repeat the test 10 times for each data set, taking the average value as the test accuracy rate.

From the analysis of Table **1** and Table **2** of the experimental data, the conclusion can be drawn that the proposed algorithm's classification accuracy is higher than that of the traditional algorithm, which also shows the effectiveness of the algorithm.

b) The running time and speedup ratio：

Some standard test data are from the UCI machine learning database set; in this experiment, each data set is amplified to 100M, 300M, 500M, and 1000M using the duplication means and running on clusters whose slaves number is 1, 2, 3 machines respectively. The running results are shown in the table below.

Table 1. UCI Data set.

Data set name	Tuples number	Attribute number	Number class	The type of attribute	Missing values
Mushroom	8124	22	2	Discrete	include
Nursery	12960	8	5	Discrete	Not include
Promoters	106	58	2	Discrete	Not include

Table 2. The results of accuracy.

Data set name	The accuracy rate	
	The algorithm in this paper	The traditional algorithm
Mushroom	99.80%	100%
Promoters	79%	76.4151%
Nursery	94.1%	78.9%

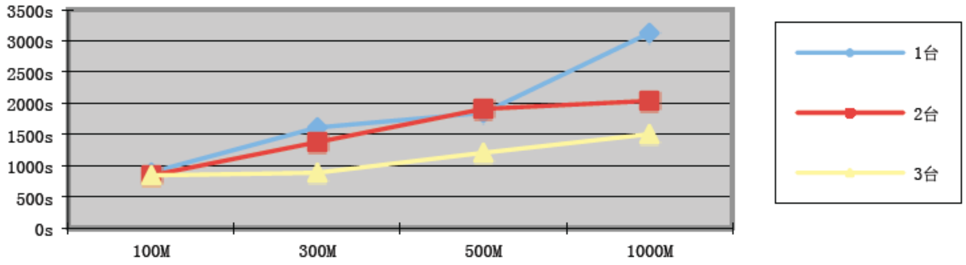

Fig. (1). Running time of nursery data set.

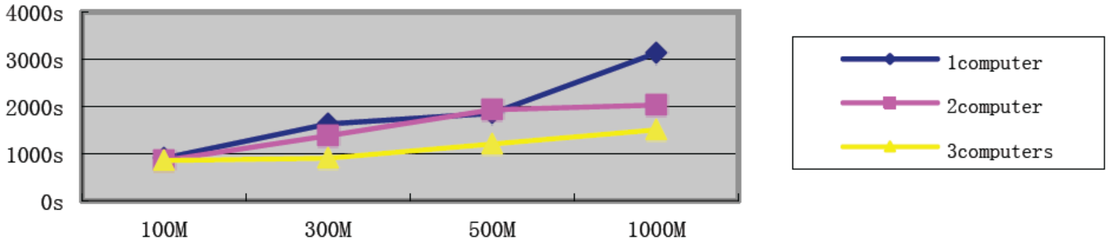

Fig. (2). Running time of mushroom data set.

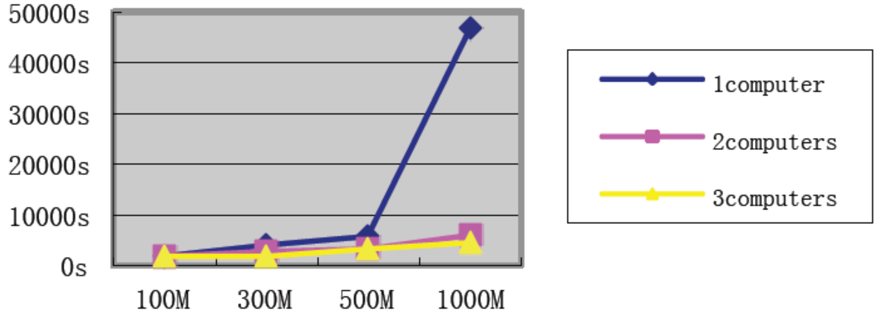

Fig. (3). Speedup ratio of nursery data set.

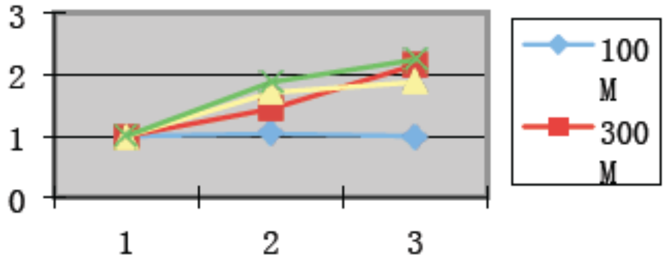

Fig. (4). Speedup ratio of mushroom data set**.**

The speedup ratio $S = T_s / T_m$ is an important measure of the performance of the parallel algorithm. T_s refers to the running time spent on a single node, while T_m refers to the running time spent on m nodes. It refers to the performance scale of a parallel algorithm; in which case the data set size being fixed while the number of nodes is increasing; an ideal speedup of the parallel algorithm is linear.

It can be seen from Figs. (**1**, **2**, **3** and **4**), when dealing with the same data set, along with the increase of the cluster nodes the processing time is reducing. When processing the data set whose size is 100M, the processing time of the cluster with only 1 node is nearly similar to that with 2 nodes or 3 nodes. However, the processing of 1000M data set is very different, and the speedup is almost a straight line. So, it can be seen that the algorithm should be used to solve the problem of massive data mining.

CONCLUSION

In this article, an improved ID3 algorithm was presented based on granular computing. The algorithm's classification accuracy is better than traditional methods. At the same time, the ID3 algorithm is parallelized, and the parallel algorithm is proved to be efficient through experimental verification. The parallel algorithm provides solution for parallelization of other data mining algorithms.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This research is supported by the National Natural Science Foundation of China under grants No. 81260578, No. 81460769, No.61163023 and No.61070139.

REFERENCES

[1] Q. Lu, and X. Cheng, "Parallelization of decision tree algorithm based on MapReduce," *Journal of Computer Applications*, vol. 32, no. 9, pp. 2463-2465, 2469, 2012. (in Chinese)

[2] A. Nasridinov, Y. Lee, and Y.H. Park, "Decision tree construction on GPU: ubiquitous parallel computing approach,"*Computing*, vol. 96, no. 5, 2014, pp. 403-413.

[3] J. Zhang, T. Li, D. Ruan, Z. Gao, and C. Zhao, "A parallel method for computing rough set approximations," *Information Sciences,* vol. 194, 2012, pp. 209-223.

[4] X. Xiang, Y. Gao, L. Shang, and Y. Yang, "Parallel text categorization of massive text based on hadoop,"*Computer Science*, vol. 38, no. 10, 2011, pp. 184-188.

[5] J. Qian, D. Q. Miao, and Z. H. Zhang, "Knowledge reduction algorithms in cloud computing," *Chinese Journal Of Computers*, vol. 34, no. 12, 2011, pp. 2332-2343. (in Chinese)

[6] T. R. Qiu, Q. Liu, and H. K. Huang, "A granular computing approach to knowledge discovery in relation databases," *Acta Automatica Sinica*, vol. 35, no. 8, 2009, pp. 1071-1079.

[7] A. Bargiela, and W. Pedrycz, "Granular computing: an introduction,*" Boston: Kluwer Academic Publishe*rs, 2003, pp. 1-17.

[8] D. Q. Miao, G.Y. Wang, Q. Liu, T.Y. Lin, and Y.Y. Yao, "Granular computing: past, present, and the future perspectives," Beijing: *Academic Press*, 2007, pp. 1-178. (in Chinese)

[9] X. Li, T. Qiu, Q. Liu, and X. Bai, "Ontology building from incomplete information system based on granular computing," In: *Proceedings of 2010 IEEE International Conference on granular Computing*, san Jose, California pp. 14-16, pp. 292-296, 2010

[10] T. Qiu, L. Wang, S. Xiong, and X. Bai, "A granular computing approach for knowledge hiding," *Journal of Shandong University*, vol. 45, no. 7, pp. 60-64, 75, 2010. (in Chinese)

[11] L. Zhang, Y. Chen, T. Li, *et al.* "Decision tree classification algorithm research," *Computer Engineering,* vol. 37, no. 13, 2011, pp. 66-67. (in Chinese).

[12] W. Xiaohu, W. Lele, L. Nianfeng, ”An application of decision tree based on ID3,” In: *International Conference on Solid State Devices and Materials Science (SSDMS*), Macao, Peoples R China 01-02 APR 2012: 1017-1021 vol: 25

[13] Y. Jinping, H. Ximei, L. Kangshun, “Improved ID3 algorithm based on new attributes selection criterion,” *Application Research of Computer*, vol. 29, no. 8, 2012, pp. 2895-2908. (in Chinese)

[14] A.B. Kumar, R. Ramesh, C. Phani, E. Madhusudhan, M. Padmavathamma, “Threshold extended ID3 algorithm,” In: *Proceedings of SPIE - The International Society for Optical Engineering,* v 8334, 2012, Fourth International Conference on Digital Image Processing, ICDIP 2012.

[15] Fresku, E. and Anamali, “A. decision-tree learning algorithm,” *Journal of Environmental Protection and Ecology*, vol. 15, no. 2, 2014, pp. 686-696.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014