Quick Knowledge Reduction Based on Divide and Conquer Method in Huge Data Sets^{*}

Guoyin Wang^{1,2} and Feng Hu^{1,2}

¹ Institute of Computer Science and Technology,
Chongqing University of Posts and Telecommunications,
Chongqing 400065, P.R.China

² School of Information Science and Technology,
Southwest Jiaotong University,
Chengdu 600031, P.R.China
{wanggy,hufeng}@cqupt.edu.cn

Abstract. The problem of processing huge data sets has been studied for many years. Many valid technologies and methods for dealing with this problem have been developed. Random sampling [1] was proposed by Carlett to solve this problem in 1991, but it cannot work when the number of samples is over 32,000. K. C. Philip divided a big data set into some subsets which fit in memory at first, and then developed a classifier for each subset in parallel [2]. However, its accuracy is less than those processing a data set as a whole. SLIQ [3] and SPRINT [4], developed by IBM Almaden Research Center in 1996, are two important algorithms with the ability of dealing with disk-resident data directly. Their performance is equivalent to that of classical decision tree algorithms. Many other improved algorithms, such as CLOUDS [5] and ScalParC [6], are developed later. RainForest [7] is a framework for fast decision tree construction for large datasets. Its speed and effect are better than SPRINT in some cases. L. A. Ren, Q. He and Z. Z. Shi used hyper surface separation and HSC classification method to classify huge data sets and achieved a good performance [8, 9].

Rough Set (RS) [10] is a valid mathematical theory to deal with imprecise, uncertain, and vague information. It has been applied in such fields as machine learning, data mining, intelligent data analyzing and control algorithm acquiring, etc, successfully since it was proposed by Professor Z. Pawlak in 1982. Attribute reduction is a key issue of rough set based knowledge acquisition. Many researchers proposed some algorithms for attribution reduction. These reduction algorithms can be classified into two categories: reduction without attribute order and reduction with attribute order. In 1992, A. Skowron proposed an algorithm for attribute reduction based on discernibility matrix. It's time complexity is $t = O(2^m \times n^2)$, and space complexity is $s = O(n^2 \times m)$ (m is the

^{*} This work is partially supported by National Natural Science Foundation of China under Grant No.60573068, Program for New Century Excellent Talents in University (NCET), Natural Science Foundation of Chongqing under Grant No.2005BA2003, Science & Technology Research Program of Chongqing Education Commission under Grant No.KJ060517.

number of attributes, n is the number of objects) [11]. In 1995, X. H. Hu improved Skowron's algorithm and proposed a new algorithm for attribute reduction with complexities of $t = O(n^2 \times m^3)$ and $s = O(n \times m)$ [12]. In 1996, H. S. Nguyen proposed an algorithm for attribute reduction by sorting decision table. Its complexities are $t = O(m^2 \times n \times \log n)$ and $s = O(n + m)$ [13]. In 2002, G. Y. Wang proposed an algorithm for attribute reduction based on information entropy. Its complexities are $t = O(m^2 \times n \times \log n)$ and $s = O(n \times m)$ [14]. In 2003, S. H. Liu proposed an algorithm for attribute reduction by sorting and partitioning universe. Its complexities are $O(m^2 \times n \times \log n)$ and $s = O(n \times m)$ [15]. In 2001, using Skowron's discernibility matrix, J. Wang proposed an algorithm for attribute reduction based on attribute order. Its complexities are $t = O(m \times n^2)$ and $s = O(m \times n^2)$ [16]. In 2004, M. Zhao and J. Wang proposed an algorithm for attribute reduction with tree structure based on attribute order. Its complexities are $t = O(m^2 \times n)$ and $s = O(m \times n)$ [17].

However, the efficiency of these reduction algorithms in dealing with huge data sets is not high enough. They are not good enough for application in industry. There are two reasons: one is the time complexity, and the other is the space complexity. Therefore, it is still needed to develop higher efficient algorithm for knowledge reduction.

Quick sort for a two dimension table is an important basic operation in data mining. In huge data processing based on rough set theory, it is a basic operation to divide a decision table into indiscernible classes. Many researchers deal with this problem using the quick sort method. Suppose that the data of a two dimension table is in uniform distribution, many researchers think that the average time complexity of quick sort for a two dimension table with m attributes and n objects is $O(n \times \log n \times m)$. Thus, the average time complexity for computing the positive region of a decision table will be no less than $O(n \times \log n \times m)$ since the time complexity of quick sort for a one dimension data with n elements is $O(n \times \log n)$. However, we find that the average time complexity of sorting a two dimension table is only $O(n \times (\log n + m))$ [18]. When $m > \log n$, $O(n \times (\log n + m))$ will be $O(n \times m)$ approximately.

Divide and conquer method divides a complicated problem into simpler sub-problems with same structures iteratively and at last the sizes of the sub-problems will become small enough to be processed directly. Since the time complexity of sorting a two dimension table is just $O(n \times (m + \log n))$, and quick sort is a classic divide and conquer method, we may improve reduction methods of rough set theory using divide and conquer method.

Based on this idea, we have a research plan for quick knowledge reduction based on divide and conquer method. We have two research frameworks: one is attribute order based, and the other one is without attribute order.

(1) Quick knowledge reduction based on attribute order.

In some huge databases, the number of attributes is small and it is easy for domain experts to provide an attribute order. In this case, attribute reduction algorithms based on attribute order will be preferable.

Combining the divide and conquer method, a quick attribute reduction algorithm based on attribute order is developed. Its time complexity is $O(m \times n \times (m + \log n))$, and its space complexity is $O(n + m)$ [19].

(2) Quick knowledge reduction without attribute order.

In some huge databases, the number of attributes is big, even over 1000. It is very difficult for domain experts to provide an attribute order. In this case, attribute reduction algorithm without attribute order will be needed. Although many algorithms are proposed for such applications, their complexities are too high to be put into industry application. Combining the divide and conquer method, we will develop new knowledge reduction algorithms without attribute order. The aim of complexities of such algorithms will be: $t = O(m \times n \times (m + \log n))$ and $s = O(n + m)$. In this research framework, we have had some good results on attribute core calculation, its complexities are $t = O(m \times n)$, $s = O(n + m)$ [20].

Keywords: Huge data sets processing, rough set, quick sort, divide and conquer, attribute reduction, attribute order.

References

1. Catlett, J.: Megainduction: Machine Learning on Very Large Databases. PhD thesis, Basser Department of Computer Science, University of Sydney, Sydney, Australia (1991)
2. Chan, P.: An Extensible Meta-learning Approach for Scalable and Accurate Inductive Learning. PhD thesis, Columbia University, New York, USA (1996)
3. Mehta, M., Agrawal, R., Rissanen, J.: SLIQ: A fast scalable classifier for data mining. In: Apers, P.M.G., Bouzeghoub, M., Gardarin, G. (eds.) EDBT 1996. LNCS, vol. 1057, pp. 18–32. Springer, Heidelberg (1996)
4. Shafer, J., Agrawal, R., Mehta, M.: SPRINT: A scalable parallel classifier for data mining. In: VLDB. Proceedings of 22nd International Conference on Very Large Databases, pp. 544–555. Morgan Kaufmann, San Francisco (1996)
5. Alsabti, K., Ranka, S., Singh, V.: CLOUDS: A Decision Tree Classifier for Large Datasets. In: KDD 1998. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, New York, USA, pp. 2–8 (1998)
6. Joshi, M., Karypis, G., Kumar, V.: ScalParC: A new scalable and efficient parallel classification algorithm for mining large datasets. In: Proceedings of the 12th International Parallel Processing Symposium (IPPS/SPDP 1998), Orlando, Florida, USA, pp. 573–580 (1998)
7. Gehrke, J., Ramakrishnan, R., Ganti, V.: RainForest: A Framework for Fast Decision Tree Construction of Large Datasets. In: Proceedings of the 24th International Conference on Very Large Databases (VLDB), New York, USA, pp. 416–427 (1998)
8. Ren, L.A., He, Q., Shi, Z.Z.: A Novel Classification Method in Large Data (in Chinese). Computer Engineering and Applications 38(14), 58–60 (2002)
9. Ren, L.A., He, Q., Shi, Z.Z.: HSC Classification Method and Its Applications in Massive Data Classifying (in Chinese). Chinese Journal of Electronics, China 30(12), 1870–1872 (2002)
10. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)

11. Skowron, A., Rauszer, C.: The discernibility functions matrices and functions in information systems. In: Slowinski, R. (ed.) *Intelligent Decision Support - Handbook of Applications and Advances of the Rough Sets Theory*, pp. 331–362. Kluwer Academic Publisher, Dordrecht (1992)
12. Hu, X.H., Cercone, N.: Learning in relational database: A rough set approach. *International Journal of Computational Intelligence* 11(2), 323–338 (1995)
13. Nguyen, H.S, Nguyen, S.H.: Some efficient algorithms for rough set methods. In: *Proceedings of the Sixth International Conference, Information Processing and Management of Uncertainty in Knowledge-Based Systems(IPMU 1996)*, Granada, Spain, July 1-5, vol. 2, pp. 1451–1456 (1996)
14. Wang, G.Y., Yu, H., Yang, D.C.: Decision table reduction based on conditional information entropy (in Chinese). *Chinese Journal of computers* 25(7), 759–766 (2002)
15. Liu, S.H., Cheng, Q.J., Shi, Z.Z.: A new method for fast computing positive region(in Chinese). *Journal of Computer Research and Development* 40(5), 637–642 (2003)
16. Wang, J., Wang, J.: Reduction algorithms based on discernibility matrix: the ordered attributed method. *Journal of Computer Science and Technology* 11(6), 489–504 (2001)
17. Zhao, M., Wang J.: *The Data Description Based on Reduct* PhD thesis, Institute of Automation, Chinese Academy of Sciences, Beijing, China (in Chinese) (2004)
18. Hu, F., Wang, G.Y.: Analysis of the Complexity of Quick Sort for Two Dimension Table (in Chinese). *Chinese Journal of computers* 30(6), 963–968 (2007)
19. Hu, F., Wang, G.Y.: Quick Reduction Algorithm Based on Attribute Order(in Chinese). *Chinese Journal of computers* 30(8) (to appear, 2007)
20. Hu, F., Wang, G.Y, Xia, Y.: Attribute Core Computation Based on Divide and Conquer Method. In: Kryszkiewicz, M., et al. (eds.) *RSEISP 2007. Lecture Note in Artificial Intelligence*, Warsaw, Poland, vol. 4585, pp. 310–319 (2007)