International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 6 1241– 1244

Preprocessing and Feature Selection on Group Structure Analysis using Entropy and Thresholding

Ms. Prajakta Deshmukh[1] , Prof. Jayant Adhikari[2]

Department of CSE, TGPCET Nagpur

Abstract: Many real data increase dynamically in size. We have been observing in many fields that data grow with time in size. This has led to the development of several new analytic techniques. This phenomenon occurs in several fields including economics, population studies, and medical research. As an effective and efficient mechanism to deal with such data, incremental technique has been proposed in the literature and attracted much attention, which stimulates the result in this paper. When a group of objects are added to a decision table, we first introduce incremental mechanisms for three representative information entropies and then develop a group incremental rough feature selection algorithm based on information entropy.When multiple objects are added to a decision table, the algorithm aims to find the new feature subset in a much shorter time. Experiments have been carried out on eight UCI data sets and the experimental results show that the algorithm is effective and efficient.

Keywords: *Feature Selection, Group Structure Analysis, UCI*

_____*****_____

I. Introduction

It has been observed in many fields that data grow with time in size. This has led to the development of several new analytic techniques. Among these techniques, as an effective and efficient mechanism, incremental approach is often used to discover knowledge from a gradually increasing data set, which can directly carry out the computation using the existing result from the original data set . In recent years, feature selection, as a common technique for data preprocessing in pattern recognition, machine learning, data mining, and so on, has attracted much attention. In this paper, we are concerned with incremental feature selection, which is an extremely important research topic in data mining and knowledge discovery.On feature selection, a specific theoretical framework is Pawlak's rough set model Feature selection based on rough set theory is also called attribute reduction .The feature subset obtained by using an attribute reduction algorithm is called a reduct . Attribute reduction is able to select features that preserve the discernibility ability of original ones, but do not attempt to maximize the class separability In the last two decades, based on rough set theory, many techniques of attribute reduction have been developed However, most of them can only be applicable to static data tables. When the number of objects increases dynamically in a database, these approaches often need to carry out an attribute reduction algorithm repeatedly and thus consume a huge amount of computational time and memory space .Hence, it is very inefficient to deal with dynamic data tables using these reduction algorithms. To deal with a dynamically increasing data set, there exists some research on finding reducts in an incremental manner based on rough set theory. Several incremental reduction algorithms have been proposed to deal with dynamic data sets .A common character of these algorithms is that they were only applicable when new data are generated one by one, whereas many real data from applications are generated in groups. When multiple objects are generated at a time in a database, these algorithms may be inefficient since they have to be executed repeatedly to deal with the added group of objects. In other words, when M (e.g., M =10,000) objects are generated at a time, one has to execute these algorithms M times. This is obviously very time consuming. If the size of an added object group is very small (e.g., M =10), the existing incremental algorithms may also be effective, of course. However, when massive new objects are generated at a time, this gives rise to much more waste of computational time and space when the existing reduction incremental algorithms are applied. With the development of data processing tools, the speed and volume of data generation increase dramatically.

II. Proposed system

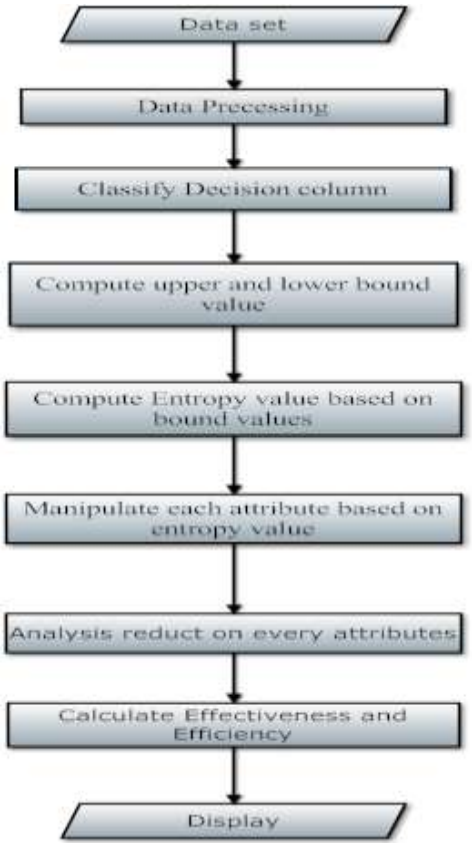
In this paper, we introduced group incremental approach algorithm to handle multiple data at a time. To select effective features from a dynamically increasing data set, an efficient group incremental feature selection algorithm is proposed in the framework of rough set theory. In the process of selecting useful features, this algorithm employs information entropy to determine feature significance, and significant features are selected as a final

feature subset .Experiments show that, compared with both the classical heuristic feature selection algorithms based on information entropy and existing incremental feature selection algorithms ,the proposed algorithm can find a feasible feature subset in a much shorter time. Knowledge updating for dynamically increasing data sets has attracted much attention. By integrating the changes of discernibility-matrix, Shan and Ziarko

Introduced an incremental approach to obtain all maximally generalized rules of a changed decision table. Bang and Zeungnam introduced an incremental learning algorithm to find a minimal set of rules of a decision table. TongandAn constructed the concept of _-decision matrix, and presented an algorithm for incremental learning of rules. Zheng and Wang developed an effective incremental algorithm which was called RRIA. This algorithm can learn from a domain data set incrementally. Guo et al. Proposed an incremental rules extraction algorithm based on the search tree, which is one kind of the first heuristic search algorithms. Furthermore, under variable precision rough-set model (VPRS), Chen et al. Introduced a new incremental method for updating approximations of VPRS while objects in the information system dynamically alter. Feature selection is a common technique for data preprocessing. For incremental feature selection, researchers have also proposed several approaches. Liu proposed an incremental reduction algorithm for the minimal reduct. This algorithm can only be applied to information systems without decision attribute. For decision tables, a reduction algorithm was presented to update reductin , but it was very time consuming. To overcome the deficiencies of these two algorithms, Hu et al. presented an incremental reduction algorithm based on the positive region, and pointed out that this one was more efficient than those two algorithms. Moreover, an incremental reduction algorithm based on the discernibility matrix was proposed by Yang .

Rough set theory has been conceived as a powerful soft computing tool to analyze various types of data , and is also a specific framework of selecting useful features. Based on rough set theory, to select useful features, a kind of common approaches is using information entropy to measure the feature significance and selecting significant features as a final feature subset . Liang et al. proposed complementary entropy and combination entropy, respectively. These two entropies have been used to determine feature significance in a feature selection algorithm .In information entropy is employed to determine feature significance in an accelerated feature selection algorithm. In , Liang et al. proposed an effective feature selection algorithm from a multi-granulation view. This algorithm was also designed based on information entropy. In this paper, to select useful features from a dynamically increasing data set, we focus on incremental feature selection in the framework of rough set theory. In view of that many real data from applications are generated in groups, a group incremental feature selection algorithm isproposed in the framework of rough set theory. And this algorithm employs information entropy to measure the feature significance.

Flow of System



Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set. Kotsiantis et al. (2006) present a well-known algorithm for each step of data pre-processing

Data Cleaning: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.In this paper,Weka tool is used for preprocessing .

Classifying decision table

Heuristic attribute reduction algorithms based on information entropy for decision tables. The main idea of these algorithms is to keep the conditional entropy of target decision unchanged. the classic attribute reduction algorithm based on information entropy.

In the process of selecting useful features, this algorithm employs information entropy to determine feature significance, and significant features are selected as a final feature subset .Experiments show that, compared with both the classical heuristic feature selection algorithms based on information entropy and existing incremental feature selection algorithms ,the proposed algorithm can find a feasible feature subset in a much shorter time. Knowledge updating for dynamically increasing data sets has attracted much attention. By integrating the changes of discernibility-matrix, Shan and Ziarko introduced an incremental approach to obtain all maximally generalized rules of a changed decision table.

. Feature selection is a common technique for data preprocessing. For incremental feature selection, researchers have also proposed several approaches. Liu proposed an incremental reduction algorithm for the minimal reduct. This algorithm can only be applied to information systems without decision attribute. For decision tables, a reduction algorithm was presented to update reductin , but it was very time consuming.

Compute entropy

Compute upper and lower bound value of decision table.Based on upper and lower bound value of decision table , find entropy value to find reduct on given dataset**.**This paper mainly develops an efficient group incremental reduction algorithm based on the three entropies. In view of that a key step of the development is the computation of entropy, we first introduce in this paper three incremental mechanisms of the three entropies, which determine an entropy by adding objects to a decision table in groups. When a group of objects are added, instead of recomputation on a given data set, the incremental mechanisms derive new entropies by integrating the changes of conditional classes and decision classes into existing entropies .With these mechanisms, a group incremental reduction algorithm is proposed for dynamic decision tables .After a group of objects is added to a decision table, the proposed algorithm generates a reduct for this expanded decision table by fully exploiting the reduct of the original decision table.

Reduct on decision table

The feature subset obtained by using an attribute reduction algorithm is called a reduct.when multiple objects are added to a given decision table, the reduct can be obtained by the proposed algorithm in a much shorter time. By doing so, when multiple objects are added to a given decision table, the new reduct can be obtained by the proposed algorithm in a much shorter time. Further more ,in view of that incremental reduction algorithms based on entropies have not yet been discussed so far, this paper also introduces an incremental reduction algorithm for adding a single object to a decision table. Experiments have been carried out on eight data sets downloaded from UCI. The experimental results show that the proposed algorithm is effective and efficient .For the convenience of following discussion, here is a description of the main idea in this paper. To select effective features from a dynamically increasing data set, an efficient group incremental feature selection algorithm is proposed in the framework of rough set theory. In the process of selecting useful features, this algorithm employs information entropy to determine feature significance, and significant features are selected as a final feature subset .Experiments show that, compared with both the classical heuristic feature selection algorithms based on information entropy and existing incremental feature selection algorithms, the proposed algorithm can find a feasible feature subset in a much shorter time.

III. Conclusion

In this paper, in view of that many real data in databases are generated in groups, an effective and efficient group incremental feature selection algorithm has been proposed in the framework of rough set theory. Compared with existing incremental feature selection algorithms, this algorithm has the following advantages:

1. Compared with classic heuristic feature selection algorithms based on the three entropies, the proposed algorithm can find a feasible feature subset of a dynamically-increasing data set in a much shorter time.
2. When multiple objects are added to a data set, the proposed algorithm is more efficient than existing incremental feature selection algorithms.
3. With the number of added data increasing, the efficiency of the proposed algorithm is more and more obvious.
4. This study provides new views and thoughts on dealing with large-scale dynamic data sets in applications.

 ISSN: 2321-8169
1241– 1244

References

[1] A. An, N. Shan, C. Chan, N. Cercone, and W. Ziarko, "Discovering Rules for Water Demand Prediction: An Enhanced Rough-Set Approach," Eng. Application and Artificial Intelligence, vol. 9, no. 6, pp. 645-653, 1996.

[2] W.C. Bang and B. Zeungnam, "New Incremental Learning Algorithm in the Framework of Rough Set Theory," Int'l J. Fuzzy Systems, vol. 1, no. 1, pp. 25-36, 1999.

[3] C.C. Chan, "A Rough Set Approach to Attribute Generalization in Data Mining," Information Sciences, vol. 107, pp. 177-194, 1998.

[4] H.M. Chen, T.R. Li, D. Ruan, J.H. Lin, and C.X. Hu, "A Rough-Set Based Incremental Approach for Updating Approximations under Dynamic Maintenance Environments," IEEE Trans. Knowledge and Data Eng., vol. 25, no. 2, pp. 274-284, Feb. 2013.

[5] I. Du¨ ntsch and G. Gediga, "Uncertainty Measures of Rough Set Prediction," Artificial Intelligence, vol. 106, no. 1, pp. 109-137, 1998.

[6] D. Dubois and H. Prade, "Rough Fuzzy Sets and Fuzzy Rough Sets," Int'l J. General Systems, vol. 17, pp. 191-209, 1990.

[7] I. Guyon and A. Elisseeff, "An Introduction to Variable Feature Selection," Machine Learning Research, vol. 3, pp. 1157-1182, 2003.

[8] S. Greco, B. Matarazzo, and R. Slowinski, "Rough Sets Theory forMulticriteria Decision Analysis," European J. Operational Research,vol. 129, pp. 1-47, 2001.

[9] X. Wu, X. Zhu, G.Q. Wu, and W. Ding, "Data mining with big data," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97–107, 2014.

[10] Guyon and A. Elisseeff. "An introduction to variable and feature selection," Journal of Machine Learning Research, 3:1157–1182, 2003

[11] Daphne Koller, MehranSahami, "Toward Optimal Feature Selection," Computer Science Department, Stanford University, Stanford, CA 94305-9010.1996

[12] Haiguang Li, Xindong Wu, Zhao Li, Wei ding"Group feature selection with streaming features," IEEE 13th international conference on data mining. 2013

[13] Jennifer G. Dy, Carla E. Brodley "Feature Selection for Unsupervised Learning," Journal of Machine Learning Research, 845–889.2004

[14] H. Liu and H. Motoda, "Computational methods of feature selection," CRC Press, 2007.

[15] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," The Journal of Machine Learning Research, vol. 5, pp. 1205–1224, 2004.