ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, May 2017

Feature Selection with Group Structure Analysis on Rough Dataset

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

Department of CSE, TGPCET, Nagpur, India

ABSTRACT: Numerous genuine information increment powerfully in measure. We have been seeing in many fields that information develop with time in estimate. This has prompted the improvement of a few new explanatory strategies. This marvel happens in a few fields including financial aspects, populace studies, and therapeutic research. As a compelling and effective instrument to manage such information, incremental method has been proposed in the writing and pulled in much consideration, which animates the outcome in this paper. At the point when a group of objects are added to a choice table, we initially present incremental systems for three delegate data entropies and after that build up a group incremental unpleasant feature choice algorithm in light of data entropy. At the point when different objects are added to a choice table, the algorithm plans to locate the new feature subset in a substantially shorter time. Analyses have been completed on eight UCI informational indexes and the exploratory outcomes demonstrate that the algorithm is compelling and effective.

KEYWORDS: Feature Selection, Group Structure Analysis, UCI

I. INTRODUCTION

It has been seen in many fields that information develop with time in estimate. This has prompted the improvement of a few new logical systems. Among these systems, as a successful and productive instrument, incremental approach is often used to find information from a bit by bit expanding informational collection, which can straightforwardly complete the calculation utilizing the current outcome from the first informational collection . As of late, feature choice, as a typical system for information preprocessing in design acknowledgment, machine learning, information mining, et cetera, has pulled in much consideration. In this paper, we are worried about incremental feature determination, which is a critical research theme in information mining and learning discovery.On feature choice, a particular hypothetical structure is Pawlak's harsh set model Feature choice in light of unpleasant set hypothesis is likewise called quality lessening .The feature subset gotten by utilizing a property decrease algorithm is known as a reduct . Credit diminishment can choose features that safeguard the perceptibility capacity of unique ones, yet don't endeavor to boost the class detachability Over the most recent two decades, in light of harsh set hypothesis, numerous systems of quality lessening have been produced However, the vast majority of them must be relevant to static information tables. At the point when the quantity of objects increments powerfully in a database, these methodologies often need to do a characteristic diminishment algorithm over and again and in this way expend a tremendous measure of computational time and memory space .Hence, it is extremely wasteful to manage dynamic information tables utilizing these lessening algorithms. To manage a progressively expanding informational collection, there exists some examination on discovering reducts in an incremental way in view of unpleasant set hypothesis. A few incremental lessening algorithms have been proposed to manage dynamic informational collections .A typical character of these algorithms is that they were just material when new information are produced one by one, though numerous genuine information from applications are created in groups. At the point when different objects are produced at once in a database, these algorithms might be wasteful since they must be executed more than once to manage the additional group of objects. At the end of the day, when M (e.g., M =10,000) objects are created at once, one needs to execute these algorithms M times. This is clearly extremely tedious. On the off chance that the span of an additional question group is little (e.g., M =10), the current incremental algorithms may likewise be compelling, obviously. Be that as it

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, May 2017

may, when enormous new objects are produced at once, this offers ascend to substantially more misuse of computational time and space when the current diminishment incremental algorithms are connected. With the advancement of information handling apparatuses, the speed and volume of information era increment drastically.

II. RELATED WORK

Information Entropy, Rough Entropy And Knowledge Granulation In Incomplete Information Systems
In this paper, the concepts of information entropy, rough entropy and knowledge granulation in incomplete information systems are introduced, their important properties are given, the relationships among those concepts are established. The relationship between the information entropy E(A) and the knowledge granulation GK(A) of knowledge A can be expressed as E(A) þ GK(A)= 1, the relationship between the granularity measure G(A) andthe rough entropy Er(A) of knowledge .Furthermore, two inequalities aboutthe measures GK, G, E and Er are obtained. Information entropy, rough entropy and knowledge granulation characterize the significance of knowledge in different ways, and make more profound explanation. These results have a wide variety of applications, such as measuring the significance of attributes, constructing decision trees and measuring uncertainties of rules, etc. They will play a significant role in further researches in incomplete information systems.

A New Method For Measuring Uncertainty And Fuzziness In Rough Set Theory
A new definition of information entropy based on the complement behavior of information gain has been proposed along with its justification in rough set theory. Based on this concept, conditional entropy and mutual information have been introduced. In particular, the new information entropy can measure both uncertainty and fuzziness in rough set theory. Now we are studying information measures in generalized rough set model for data mining applications, which will be reported in another paper.

Hybrid Attribute Reduction Based On A Novel Fuzzy-Rough Model And Information Granulation
Selecting optimal numerical and categorical features is an important challenge for pattern recognition and machine learning. There are two strategies for selecting mixed attributes. One is to discretize numerical attributes into several intervals and take discretized attributes as categorical one. The other is tocode a categorical attribute with some integral numbers and look it as a numerical variable. These techniques lose the original information in the data.In this paper, we show a simple but efficient feature subset selection technique based on a proposed fuzzy-rough model. This approach does not require discretizing the numerical data, whereas classical rough set just work on categorical data. We introduce a symmetric function to compute fuzzy similarity relations between the objects with a numerical attribute and transform the similarity relation into a fuzzy equivalence one. We compute the positive region of the decision by fuzzy inclusion and variable precision fuzzy inclusion. Four attribute significance measures are defined. Based on the measure, we constructa forward hybrid attribute reduction algorithm, named FAR-VPFRS.
With 10 UCI data sets, a series of experiments are conducted for evaluating the proposed method. The results show that most of the features in raw data can be eliminated without decreasing classification performances. What is more, most of classification are improved. We also find that the optimal thresholds for computing variable precision positive regions depend on the learning algorithms. The experiments show a default threshold should be in the interval [0.75, 0.85].
An efficient rough feature selection algorithm with a multi-granulationView
At present, feature selection for large-scale data sets is still a challenging issue in the field of artificial intelligence. In this paper, with some concepts in statistics, an efficient rough feature selection algorithm has been proposed to deal with large-scale decision tables. The algorithm found a valid feature subset though dividing a large-scale table into small ones and fusing the feature selection results of small tables together. The experimental analysis shows that the proposed algorithm is feasible and efficient. Note that the proposed algorithm not only saves computational time, but also can handle some large scale data sets that are very difficult to deal with on a PC because of the high computational time. It is our wish that the idea of dividing and fusing on data sets provides a new view and thinking on dealing with large-scale data sets in applications.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, May 2017

III. 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. Tongand An 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 reduct in , 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.

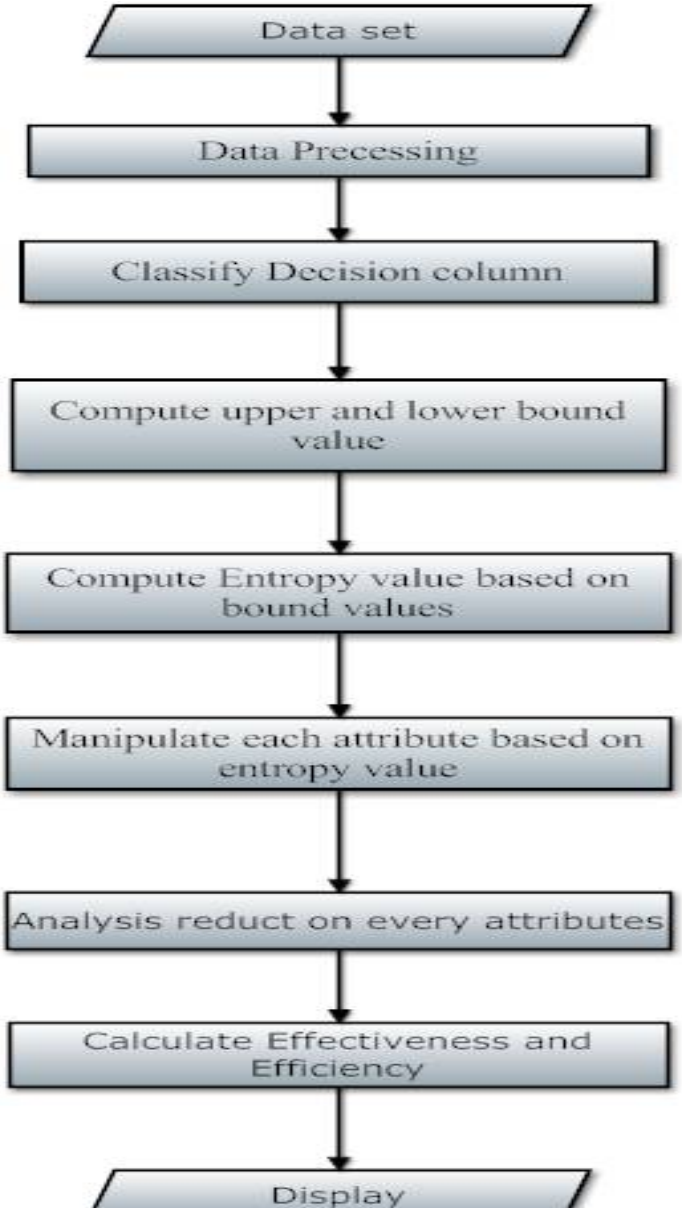
ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, May 2017

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.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, May 2017

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.

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 computation 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.

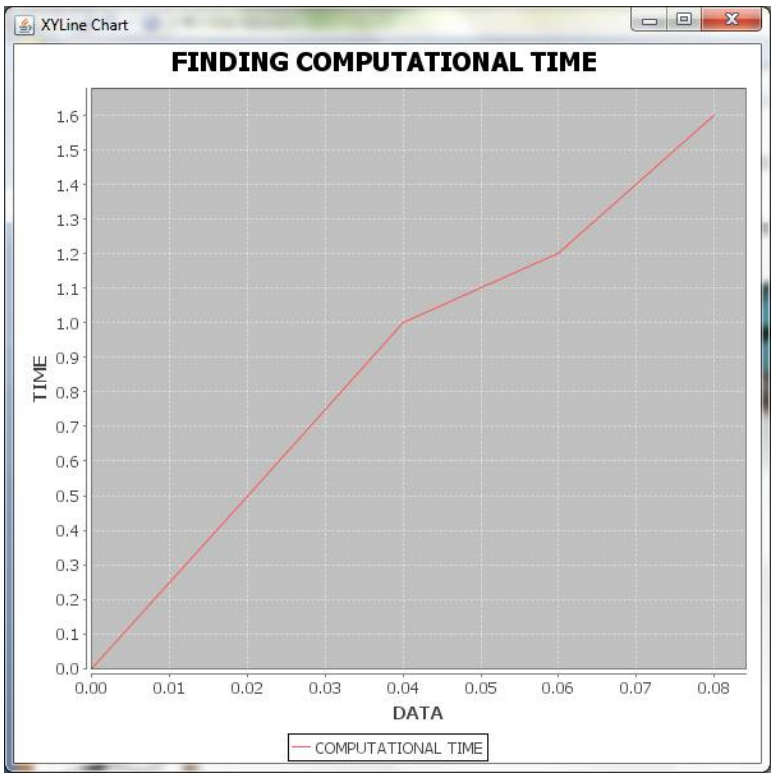
ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, May 2017

IV. RESULT



The above figure graph shows the computational time required for applying reduct on given dataset. Time variation is compared with data probabilities.

V. 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.

Future Enhancement

1.The incremental mechanism of data expanding in groups is in reality the fusion of two data tables. Thus, by generalizing the incremental mechanism, future work would include the information fusion of multidata tables or multigranularity.

2.Further analysis of dynamic data tables shows that the variation of data tables can also include the changes of data values. For data tables with data values changing dynamically, feature selection approaches based on rough set model will be introduced to discover knowledge from dynamic data tables.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, May 2017

3.With the variation of data sets, to predict the decision, the rules extracted from a dynamic dat set need to be updated in time. Therefore, it is necessary to devise rules extraction algorithms for a dynamic decision table.

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, Mehran Sahami, "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.