4710 project report

1.Introduction

As the development of internet technologies, it has been observed that collected data is growing in terms of volume, variety, velocity, veracity, and value. This leads to the development of various new data analysis technique. Among these technique, incremental data mining is often used to discover previous unknow and potential useful knowledge from existing and growing information. Decision making based on extracted rule from growing data is one of such important application in incremental data mining. In real life, data usually with high dimension of attributes, it is more practical to select representative features before making decision. In our project, we are focusing on incremental feature selection and make decision based on selected features.

Pawlak’s Rough[cite] set model is one of such theoretical frameworks in feature selection as well as decision making based of selected feature. Feature selection in rough set is also called attribute reduction[cite]. The selected feature is called reduct[cite]. Attribute reduction is a process to reduce the number of attribute and preserve the discernibility of the original data while minimize the class separability[cite]. In the recent decades, several techniques in attribute reduction based on rough set were proposed[cite], but most of them is focus on static data. In the environment of growing data, they lack the ability to update the selected features and must recompute them every time, which is not scalable and efficient in the century of big data. To deal with dynamic data stream, there exist some research on finding reduct incrementally based on rough set[cite]. However, most of them are focusing on update the reduct regarding to the whole database without considering concept drifting. Furthermore, it has been overserved that rough set model is very sensitive to outliers[cite], most of the work listed above also does dealing outlier properly and maybe need more preprocess work before running their techniques.

In terms of decision making, Pawlak’s Rough set also have its drawbacks when dealing with real value attributes, which are usually the dominant attributes in real life. One must implement discretization or categorization before analyzing through rough set. Even with the help of discretization or categorization, traditional rough set theory tends to treat discretized real value as nominal data and ignore some important intrarelationship, such as similarity among them. Fuzzy rough set[cite] is a variant of rough set which treat the real value as a member in fuzzy set to preserve the relationship among real value. Some of our models will also adopt the concept of membership function in fuzzy rough set.

For convenience for the future discussion, here is the description of our assumption, contributions and main idea. This project assumes that the decision attribute in the decision table(dataset) is nominal (crisp), while the conditional attribute could be nominal or numeric. In this project we proposed 3 incremental feature selection and decision rule induction models for decision-making based on crisp rough set (discretize real value data before analyzing) and 3 corresponding feature selection for decision-making models based on our modified rough set model (using membership function in fuzzy rough set concept and using similarity to comparing real value attribute). The 3 proposed incremental crisp rough set models named: incremental voting rough set (IVRS), incremental sliding window rough set (ISwRS), and time fading rough set (TFRS). The 3 proposed modified rough set model called: incremental voting membership rough set (IVMRS), incremental sliding window membership rough set (ISwMRS), and time fading membership rough set (TFMRS). In fact, the 3 crisp models and 3 membership models are correspondingly built on the same base concept with different way to deal with real data value and different techniques in making decision. Inspired by Random Forest[cite] and association rule mining in stream data[cite], our models, unlike most of works, separate the growing dataset into different batches, and preform attribute reduction within each batch then make decision based on the combination of each batch. We perform attribute reduction by implementing the technique of discernibility matrix [cite] and make decision based on the rules induced by LEM2 algorithm[cite] in crisp rough set and based on membership function in membership rough set. These proposed models are robust to outliers since outliers would only affect the processing of attribute reduction they belong to, and outliers are usually rare in the comparison to normal data. 4 of our proposed models, ISwRS, TFRS, ISwMRS, TFMRS, could also handling the problem of concept drifting, which is a popular phenomenon in real life, by putting more attention in the recent data. And the proposed models also tend to be more efficient than other works in the process of attribute reduction since only the reduct in the latest batch will be updated.

The structure of this report is as follow: section 2 will list some relate work, section 3 will introduce some background information in rough set and LEM2 algorithm. section 4 will be the detail of our 3 crisp models. section 5 will introduce some background in fuzzy rough set. section 6 will contain the detail of our 3 membership-based models. Section 7 will be experiment and comparison to other existing models. Section 8 will conclude our paper and outline some futures works.

2.related work

This paragraph contains some variant of rough set model (variable precision rough set, tolerance rough set, decision, theoretical rough set)

This paragraph contains traditional rough set attribute reduction (search tree, discernibility matrix, heuristic)

This paragraph contains some existing work in incremental attribute reduction

3.preliminaries I

Definition of rough set, reduct, attribute dependency, discernibility matrix, LEM2

4. our 3 crisp models

IVRS

ISwRS

TFRS

Walk through example

5. preliminaries II

Definition of rough set, technique in comparing similarity, combining similarity in different attribute.

6. our 3 model in membership rough set

IVMRS

ISwMRS

TFMRS

Walk through example

7. experiment

Comparison of other datasets

Performance of our model

8.

Conclusion and future work