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

Rough set variant

Features selection refers to selecting representative attributes in high dimension data set. In the field of rough set, it is called attribute reduction. There exist many works in finding the reduct from the decision table or information system. Skowron and Rauszer [cited] used discernibility matrix and discernibility function to find reduct in rough set system. Qian and Shen[cite] used information entropy and heuristic function to calculate the reduct. Qian and Shen[cite] used distance measure to find reduct in variable precision rough set system. These works present different techniques in attribute reduction. However, their method could only work well on static information system. In the environment of growing data, these techniques need to run entirely whenever new data come in, which are not scalable and practical.

Incremental attribute reduction refers to dynamically updating reduct in the rough set system. In the field of decision rough set, new set of decision rules could be induced from the updated reduct. There are many research works on incremental attribute reduction: Qian et al in [cite] developed incremental attribute reduction based on information entropy; Hu et al in [cite] dynamically update reduct based on positive and negative region regard to decision attribute; Zheng and Wang[cite] introduce (RRIA), a tree based incremental features selection method in rough set system; Yang et al[cite] using discernibility matrix to incrementally update reduct in fuzzy rough set. Rather than updating reduct, there is also some works[cite] directly update decision rules. However, all these works only focused on updating selected features incrementally and neglected the fact that early data could lose the ability to represent the trend in recent time and fail to capture the phenomenon of concept drift. Since these early data still present in the system and get considered when updating new reduct, they could affect accuracy of the induced decision rule when the trend of the data is changing drifting. On the other hand, since it has been observed that rough set system is sensitive to outliers[cited], the existence of outlier could significantly affect the result of attribute reduction, thus induction of decision rules. All works above did not treat the outliers properly and require outlier detection as a preprocessing job. Our proposed model could handle the situation of concept drift well since data will be divided by batches based on arrival and could reduce the effect of outliers by using a voting mechanism as outlier could only affect the rules, thus vote, in its batch.

Beside the works above, the most related works to our project is Leung’s 3 models for association rule mining in data stream: landmark stream mining model[cite], sliding window stream mining model[cite] and time fading stream mining model[cite]. The proposed models are the combinations of Leung’s stream models with the variants to the discernibility matrix reduct updating model in [cite]

3.preliminaries I

In this section, we briefly review some preliminaries in crisp rough set theory[cite] that are related to our project, including lower and upper approximation, reduct, core and rule template; and the LEM2 algorithm[cite] to induce rule in crisp rough set.

3.1 rough set theory

In this project, we focus on the problem in the decision making based on rough set. Let DT = (U, A∪D) be a decision table with nominal conditional attribute A, decision value set D = {d}, and universe of objects U.

Definition 1: An equivalence relation defined by attribute subset B ⊆ A is call B-indiscernibility relation denoted by IND(B) is defined by:

IND(B) = {(x, y) ∈ U × U: a(x) = a(y), ∀ a ∈ B}

Definition 2: for any object x ∈ U, [x]B = {y ∈ U: a(x) = a(y), ∀ a ∈ B}

Definition 3: given an attribute subset B ⊆ A, the lower approximation of B, B↓, and upper approximation of B, B↑, regarding to concept set X is define as:

B↓(X) = {x: [x]B ∈ X}, those x such that all the elements in [x]B is in X

B↑ (X) = {x: [x]B ∩ X ≠ Ø}, those x such that at least one element in [x]B is in X

Definition 4: Suppose U = {x1, x2, x3..., xn}, a n × n matrix M (), is defined as a discernibility matrix of DT = (U, A∪D) if:

and

Definition 5: The core in decision table is defined as:

Definition 6: The discernibility of attribute a is defined as: , and the discernibility of A is:

Definition 7: The reduct Red of a decision table is defined as:

3.2 LEM2 algorithm

4. our 3 crisp models

This section will cover the detail of the 3 proposed crisp models. In subsection 4.1 will cover the algorithm we used for incremental attribute reduction, finding approximate and decision rules induction, and subsection 4.2 will introduce the detail of our three model.

4.1 Incremental Attribute reduction

We use discernibility matrix and discernibility relation as the framework and combining with heuristic function to finding one reduct in the rough set system.

Algorithm 1: Attribute reduction by discernibility matrix and hill climbing:

5. preliminaries II

Additional to crisp rough set, in this project we also proposed 3 membership rough set models based on the concept of membership function in fuzzy rough set and similarity measures between real values. This section will give some reviews about relevant definition in fuzzy rough set, fuzzy discernibility matrix and similarity measure.

5.1 fuzzy rough set

With the same definition of decision table, is called a fuzzy decision table. For each condition attribute a ∈ A, one could define a fuzzy binary relation , which is called a fuzzy binary relation if is reflexive, (R(x, x) = 1), symmetric, (R(x, y) = R(y, x)) and sup-min transitive .

Definition 1: the lower and upper approximate of attribute set B, B ⊆ A, regarding the concept set X is define as:

Here, X(u) denote the membership degree of u to a fuzzy set X and RB (x, u) denote the fuzzy equivariance relation between x and u regarding to attribute set B. And RB (x, u) = , where T is a t-norm aggregation function in fuzzy set theory.

One would notice that the definitions of lower and upper approximate is different from the one in the crisp rough set. These definitions are representing the membership degree of x to such lower and upper approximate instead of representing a set of elements.

In the contest of decision table, we mostly concern able the lower and upper approximate membership degree of each x, x ∈ U, regard to its decision class. That is:

In the assumption of our models, decision attribute is always crisp (nominal), then the membership of u to [x]D becomes: . And the definition could be refined as:

The refined lower approximate membership function will be used in our 3 membership rough set models to detect the membership degree of object belong to its decision class’s lower approximate and to making prediction, whereas the binary fuzzy relation is replaced by similarity measures between attributes, which will be covered in next sub-section.

Definition 2: According to [cite], A fuzzy discernibility matrix M () with size n × n is defined as:

5.2 similarity relation for real value attributes

Given a decision table , and R is a similarity relation define for real value attribute a, a ∈ A, if and only if R satisfy reflexivity and symmetry. We could define many similarity relations for real value attributes, such as:

1. Min-max scale similarity: .
2. Gaussian similarity:

Here , , denote the maximum and minimum value for attribute , and denote the variance of attribute .

Our membership models will use the Gaussian similarity to calculate the similarity relation between real value attributes. For nominal attributes, we use equivariance relation as defines in crisp rough set.

For attribute subset , would be the aggregation function of . And such aggregation function could be defined as:

3.Minimum T-norm in fuzzy relations:

4.Product T-norm in fuzzy relation:

5.

Equation 3 is suitable for decision table that is a mixture of nominal and real value attribute since it could preserve some tolerance even when the objects disagree in nominal attribute and the two T-norm version is suitable in the case that all attributes are real value or ordinal.

6. Incremental membership rough set models

This section will cover the detail of the proposed: IVMRS, ISwMRS and TFMRS. All these three models find and update new reduct by discernibility relation techniques in fuzzy rough set from [cite]. So, section 6.1 will introduce these discernibility relation techniques and the 3 proposed model will be covered in section 6.2.

6.1 Discernibility relation in fuzzy rough set for incremental attribute reduction

The

7. experiment

Comparison of other datasets

Performance of our model

8.

Conclusion and future work