ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 2, February 2013

Rough set with Effective Clustering Method

G.Ramani

Assistant Professor, Department of MCA, Padamasri Dr. B. V. Raju Institute of Technology, Narasapur, Medak Dist., India

Abstract: Rough set theory is a powerful mathematical tool that has been applied widely to extract knowledge from many databases .Rough set theory is proposed to mine rules from the Data warehouse. It constructs concise classification rules for each concept satisfying the given classification accuracy. Due to some drawbacks we suggest rough set with clustering methods to achieve more precision and Shows greater accuracy results in the proposed approach where groups a set of data so that the intra-cluster similarity is maximized and the inter-cluster similarity is minimized.

Key Words: Rough set, clustering, Precision.

I. INTRODUCTION

Rough set theory was first developed by Zdzislaw Pawlak in the early 1980 [1]. Mainly deals with the discretization of the data in a supervised learning (used to discriminate the dataset) and the ability to handle qualitative data. The knowledge discovery in Data mining with the rough set is a multi process which mainly consists of Discretization and Rules generation on training set and fits completely into the real life. It is a mathematical approach to imperfect knowledge. The applications of Rough set theory are in various domains such as engineering environment, banking, medicine, decision making and many more.

II. RELATED WORK

a. Rough Set definition:

Rough set approach [2] [3] is a classification to discover structural relationships within noisy data, applies on discrete valued attributes. A set of indiscernible objects is elementary set and has a basic knowledge about the set of objects. A union of these elementary sets is precise set or a rough set .Rough set has a boundary lines. The rough set is a set which is approximately described and the sets can be defined by using standard set operators. The set is associated with a pair of bounds; (For each object of the universe an information is associated and the objects are characterized by the same information. Set of similar objects is called elementary set and form a basic knowledge).

Lower Approximation: The lower approximation consists of all the data with out any ambiguity based on the knowledge of the attributes.

Upper Approximation: The objects are probably belong to the set, cannot be described as not belonging to the set based on the knowledge of the attributes.

The expanse of the set is the difference between the lower approximation and the upper approximation.

Rough sets are also used for attribute subset selection and relevance analysis.
Example:

I: Information System

U: universe (nonempty set)

R: equivalence relation on U, called indiscernibility relation, If x, y € U and x R y then x and y are indistinguishable in I.

Decries by an equation R € U×U on finite and non empty universe U, the relation '<' partitions U into a disjoint subsets $^{U}/_{R}$ called a quotient set of U. For two elements X, Y € U, if X < Y are indistinguishable. $^{U}/_{R}$ Are elementary sets and the union of more elementary sets are called definable sets.A and 2^U represents the distinct concept, used with elements of the quotient set $^{U}/_{R}$.The two elements.

I: information system

U: nonempty set of objects, called universe

The underlying system for constructing rough sets is the set algebra (2 , ¬,∩, ∪). An element A of 2U represents a non-vague concept. When such a non-vague concept is viewed with respect to elements of the quotient set U/<,i.e., the equivalence classes of <, it becomes vague and uncertain. Consider two elements x, y ∈ U and a subset A ⊆ U with x<y, x ∈ A, and y ∈ the non-vague concept to be vague and uncertain. For a subset A ⊆ U, one may describe it by a pair of lower and upper approximations:

apr(A)={x ∈ U | [x]< ⊆ A},= [{[x]< ∈ U/< | [x]< ⊆ A},apr(A)={x ∈ U | [x]< ∩ A = ∅},=[{[x]< ∈ U/< | [x]< ∩ A = ∅}, where

$[x]_R = \{y|xRy\}$, equivalence class containing x.

The lower approximation is the greatest definable set consists of A and upper approximation is the least definable set consists of A.

There are three types of rules, Association Rule mining, Rare Association Rule Mining and Multi-Objective Rule mining. The approach of these rules is incapable of handling inconsistency, generating minimal rule set and generating non-redundant rule set. Inconsistency is caused by the existence of indiscernbility relation of the data set. Data set is represented in the form of Decision table. Where each row and column represents object and attribute respectively. Consider a decision table with objects p1, p2, p3 along with condition attributes and decision attributes. The attributes of objects p1, p3 are equivalent but decision attributes are different and p1 and p3 are in discernable and the data set has inconsistency. The above mentioned methods cannot generate classification rules.

Table 1
.Inconsistent Data Set Table

Objects	Condition Attributes		Decision Attributes
	A	C	D
P1	Low	High	Yes
P2	Low	Low	No
P3	Low	High	No

b. Rough set theory in Data Mining:

A data mining system is designed with an objective to automatically discover the knowledge.

- **Mining**: It's a multi step process in the extraction of Knowledge Discovery Data.
- **Basic Concept:**

a. Discritization: Continuous data is transformed into nominal by splitting the range of the attribute values into a finite number of intervals with maximizing the interdependency between the attributes. The simple method of discretizing the data is by Boolean method and the technique is very efficient to return decision algorithms. There are different methods of discretizations, are, Chi-square based methods, Binning method, Histogram Analysis, Entropy based methods, Wrapper based methods, and Adaptive Discretization and Evolutionary based methods discritization is also done by Clusters, Decision Tree, correlation Analysis and Other methods.

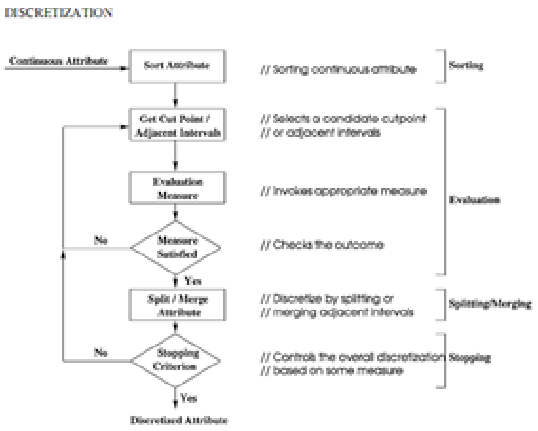

Fig.1.Discretization

b. Data Reduction: Generating Rules: The data is introduced to the reductant algorithms which is simple, fast and generates rules. Finding the reducts of attributes that can describe the concepts in the set is **NP-hard**. One of the method is discernibility matrix is used that stores the differences between the attribute values for each pair data tuples, rather than searching on the entire set of data. In order to understand the idea of reduct, let B subset of A and A€B in an information system I = (U, A) where U is the universe of objects, A is the set of attributes and R (B) is a binary relation.

1. A is dispensable in B if R (B) = R (B-{a}) otherwise a is indispensible in B.

2. Set B is independent if all its attributes are indispensible.

3.$\dot{B}$ Subset B is a reduct of B if $\dot{B}$ is independent and R$(\dot{B})$ = R(B)

The observed work is a classification rule generation method based on LEM2 algorithm used on datasets which has inconsistencies and was tested on the different data sets in rule generation. (They yield better results which we go for a mixed approach that local as global coverings, the working is towards LEM2 algorithm based minimal rule generation for the noisy data set. The inconsistency data set initially is introduced to the rough set theory to make the whole data set to a reduct data, on which we can induce various mining techniques to retrieve the knowledge information.

c. **Relevance Feature Selection**:

Relevance feature selection is the process of finding the subsets of features from the original data. In a given supervised data set A with N cases (I, target) consists of n-features i associated with the targets.

Let vector of features v = $(X_1\, X_{2\,,\ldots,}, Xi-1, Xi+1,\ldots.,Xn)$ which have been taken from the original data X. The feature Xi is relevant if any value exists to that feature a_{xi} and a predicator value a_y for which

p$(x_i = a_{xi})$ > 0.

The feature is strongly relevant if there exists a value of the featurea_{xi}, and a predicator value a_y and a value a_v (vector value) for which p (x_i = a_{xi}, v = a_v) > 0.Therefore strong relevance implies that the features are similar. The feature x_i is weakly relevant where their exists subset of features from the set of features.Value of the feature a_{xi} and a predicator value a_y and also a value a_{zi} for which p(x_i= a_{xi}, z_i = a_{zi}) > 0.This implies that the features are dissimilar. To make more effective the proposed method suggests different technique to retrieve information from the data.

III. PROPOSED APPROACH

Clustering Analysis is the important analysis in discovering of the Knowledge data (KDD).Data set is of numerical or categorical data. The clustering accuracy can be increased by fuzzy processing, the fuzzy K modes.

In the proposed method the data is supported by the clusters defined by the previous defined clusters. The objects within a cluster are similar and the objects of different clusters are dissimilar. Regularly used partitioning methods are K-means and K-medians. K-means algorithm is the method of taking the mean value of the objects in a cluster and the K-medians method takes the centrally located object in the cluster. Database with noise and outliers the K-median is more strong .Each object in the database is clustered with the median clustering method, it iteratively replaces one of the mediods by one of the non-mediods till the cluster reaches to the high quality of similar objects.

a. First step: **K-Mediod Algorithm**: Number of clusters K and a database containing n objects. The K clusters minimizes the sum of the dissimilarities of all the objects to their nearest mediods. The following are the various steps involved;

1. Initially choose K objects as a mediod.

2. Assign each object to the nearest mediod in the cluster.

3. Select non-mediod object randomly (O random)

4. Compute the cost, S swapping with O random O_j.

5. Finish, else go to step 2.

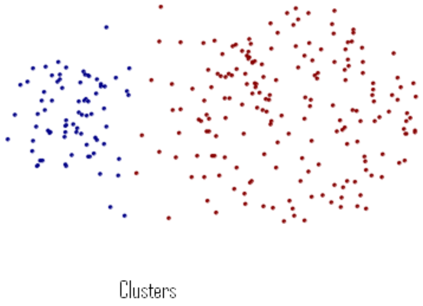

Fig.2. Data set in two clusters

b. Second step: **A multi –strategy** is implemented for the extraction of the precision from the data. The data analysis is done through the rough set approach leading to rule generation. The method has various steps, 1.Cluster Formation through unsupervised clustering algorithm (K-mediod-Clusters of similar groups are formed).2.**Attribute selection through data discritization**: Once the clusters are formed the attributes are selected through the data discrimination. 3. **Rule Extraction** through rough set theory. Which generates cluster defining rules from the continuous valued data so that the emergent rules can be directly applicable?

Extracting cluster defining rules – A hierarchical approach: Extracts rules from the collection of continuous valued data vectors for which the classification attribute is not defined. We propose a systematic transformation of the data set to the rules. The below is the description of the transformation.

1. Cluster Analysis: To the given data set, initially it is partition into K clusters. In which each cluster has the data of similar characteristics and how the data is partition. Here the data clustering is done without prior knowledge of the class and at the end we get distinct clusters.

2. Data Discretization: The continuous valued attributes are not suitable for the extraction of the rules. Discretization is an important aspect to discrete the attributes to their discrete intervals (ranges) and represented by a label. Various discretization methods are chosen based on the input data set. Discriminate function analysis is used to determine which variables discriminate between two or more groups.

3. Rule Discovery: To derive rules which explains the dependencies, structural characteristics and attribute significance of the clustered data? The steps involved in the rule defining are.

- Construction of reducts: Techniques are obtained to reduce the representation of the data with the minimizing the loss of the data. There are different methods of data reduction (Dimensionality reduction, Data Compression etc) the method which is used in the proposed method is Numerosity Reduction.

a. Numerosity reduction: In this method, Parametric and non-parametric models are used, where it stores only the model parameters instead of the actual data. Ex. Regression and log linear models etc.

In our system we are proposing the statistical theory ,support vector machines (SVM) is used to improve the ability of processing large-scale data of support vector machine, and through the simulation experiments to verify the superiority and adaptability of algorithm which is a supervised learning method used for classification and regression also called as Maximum Margin Classifiers.SVM maps the vectors to a higher dimensional space to reach maximal separating hyper planes are constructed. Parallel hyper planes are constructed on both the sides of the hyper plane that separate the data. The separating hyper plane is the hyper plane that maximize the distance between the two parallel hyper planes. The larger the distance the gen1995eralization will be better for the classifier.

$$w . x + b = o$$

Where; b is the scalar, w is the p-dimensional vector. Vector w points perpendicular to the separating hyper plane. Parallel hyper planes can be described as;

$$w.x + b = 1$$

$$w.x + b = -1$$

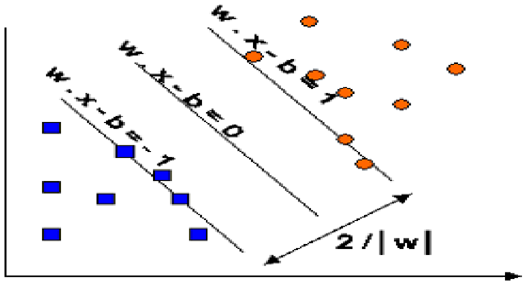

Fig.3.Hyperplane with support vectors

Therefore, samples along the plane are called the support vectors, hyper plane with the margin defined by M= $2/|w|$.

b. Generating rules: In generating the rules, Dynamic reducts are selected which has short length and satisfies the user defined accuracy. Involves the most popular measures of interest and significance. This approach aims to identify subsets of the most important attributes.

c. Rule filtering: Rule filtering involves in the step wise elimination of the less significant from the data set. It comprises of the following higher accuracy level, shortest length and maximum RHS support.

IV. EXPERIMENTS AND RESULTS

The data set is taken from the super market. Initially we use the LIBSVM with different kernel linear, polynomial etc (also n fold cross validation is conducted to determine the best value of different parameter) for the given data set, the hyper plane differentiates the classes from the data set into clusters. Then the rules are generated with the help of the user specified threshold. where the generated rules can be utilized for the decision making. And finally the rule filtering process is done only to eliminate the unnecessary data from the system. The use of rough set theory with clustering provides an alternative to the traditional statistical approaches. The below mentioned table shows the results of the rough set rule generation based on clustering method.

Table2.
The two subsets of attributes Eyes and Height the rule which is extracted from the table has been discussed.

Object	Height	Hair	Eyes	Class
o_1	Short	blond	blue	c_1
o_2	Short	blond	brown	c_2
o_3	Tall	red	blue	c_1
o_4	Tall	dark	blue	c_1
o_5	Tall	dark	blue	c_1
o_6	Tall	blond	blue	c_1
o_7	Tall	dark	brown	c_2

Sample rules for the information Retrieval: The two rules which have been drawn from the attributes which is shown in the table are given below, The composition of the classes provides useful structural information about rough set approximations for rule induction and the rules are also generalized.

(Height = short, Eyes = brown) then $Class_2$

(Height = tall, Eyes = blue) then $Class_1$

V. CONCLUSION

An improved clustering algorithm based on rough sets has been put forward, and the application of the method of calculating equivalence class in rough sets has been studied in clustering. The improved clustering algorithm resolves the problems that the number of clusters cannot be set exactly and can only find clusters with spherical shape, making the partitioning method be able to discover clusters with arbitrary shape. The feasibility of the algorithm also is represented in the paper. In fact, the feasibility can be proved theoretically. The algorithm given in this paper illuminates that clustering method and rough sets can be syncretized, and the united method has the important significance to synthesis analyses methods of data mining. Efficiently finds hidden patterns and granular clusters which are modeled by the rules are mapped to different cases ,represented by fuzzy membership functions .since rough set theory is used to obtain cases through the rules and the generation time is very much reduced. The main issues which are discussed in the paper are;

- Data Reduction
- Data Significance
- Generates Rules from the data

The main concept of the clustering the unsupervised data from the huge data through the rough set theory is simple and accurate in the information retrieval.

REFERENCES

[1.] Agrawal R., Gehrke J., Gunopulos D., Raghavan P.: Automatic subspace clustering of high dimensional data for data mining applications. Proc. ACM-SIGMOD Int. Conf. Management of Data, Seattle, Washington (1998).

[2] Pawlak, Z.: Rough Sets. In: Lin T.Y., Cercone N. (eds.): Rough Sets and Data Mining: Analysis of Imprecise Data. Kluwer Academic Publishers, Dordrecht pp. 3-7. (1997)

[3] P.K. Agarwal and C.M. Procopiuc, Exact and ApproximationAlgorithms for Clustering,º Proc. Ninth Ann. ACM-SIAM Symp. Discrete Algorithms, pp. 658-667, Jan. 1998.

[4] Boser, B. E., I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers .InProceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144 -152. ACM Press 1992.

[5] P. Langley and H. A. Simon, "Applications of machine learning and rule induction," Commun. ACM, vol. 38, no. 11, pp. 55–64, 1995.

[6] Beynon, M., Curry, B. & Morgan, 'Knowledge discovery in marketing: An approach through rough set theory', European Journal of Marketing, vol. 35, no. 7/8, pp. 915-935. P. 2001

[7] Dougherty, J., Kohavi,R., and Sahami,M. Supervised and unsupervised discretization of continuous features.In Proc. Twelfth International Conference on Machine Learning. Los Altos, CA: Morgan Kaufmann, pp. 194–202. 1995

[8] Dubois, D. and Prade, H. Rough fuzzy sets and fuzzy rough sets, InternationalJournal of General Systems, 17, pp.191-209, 1990.

[9] Liu, H., Setiono, R.: Chi2: Feature Selection and Discretization of Numeric Attributes. Proc. 7th Int. Conf. on Tools with Artificial Intelligence, Washington D.C (1995).

[10] K. Alsabti, S. Ranka, and V. Singh, ªAn Efficient k-means Clustering Algorithm,º Proc. First Workshop High Performance Data Mining, Mar. 1998.

[11] Ankerst M., Breunig M., Kriegel H., Sander J., OPTICS: Ordering Points to Identify the Clustering Structure, Proc. ACM SIGMOD'99 Int. Conf. on Management of Data.