International Journal of Computer Architecture and Mobility
(ISSN 2319-9229) Volume 3 -Issue 3, May 2015

Rough Set based Approach to enhance the accuracy of Datasets

Kalyani Upadhyay[1], Swati Lodhi[2]

[1]*PG students of Sanghavi Innovative Academy Indore*

[2]*Assistant Professor of Sanghavi Innovative Academy Indore*

Abstract— **In this research a new technique is proposed to enhance the clustering accuracy of machinery datasets. In this approach to analyze the rough attributes in datasets a new mathematical tool is used known as Variable Precision Rough Set Theory. All the attributes are tested first and the attribute which provide redundant information will be removed. To check clustering accuracy of datasets FCM algorithm is used. Fuzzy C-Mean algorithm will cluster the decision attributes and thus information gain of datasets are improved effectively. Experiments show that accuracy of clustered data has improved at recognized level.**

Keywords— **Cross Validation Model, Variable Precession Rough Set Theory, FCM algorithm.**

I. INTRODUCTION

Clustering is widely used to group collection of objects in such a way that the objects whose property are approximately same will be consider in one group. There are many algorithms based on clustering like C-Mean, Fuzzy C-Mean (FCM). FCM is a data clustering technique in which a dataset is grouped into n clusters with every data point in the dataset belonging to every cluster to a certain degree. FCM have many applications like market research, pattern recognition, data analysis, and image processing. It can be used in the field of biology also like categorize genes with similar functionalities and gain insight into structures inherent to populations.

A new mathematical tool widely known as Rough set theory (RST), which is introduced by a Poland mathematician Pawlak in 1982 which is used to deal with vague and uncertain data in large datasets. Its main idea is that to classify objects into similar classes that are indiscernible with respect to attributes. Rough-set-based decision tree algorithms have been studied within resent years. However, it has some limitations like it can only do well in accurate classification where objects are strictly classified according to equivalence classes; hence the induced classifiers lack the ability to tolerate possible noises in real world datasets. In order to improve the shortcomings of rough set model, the classical rough set model is extended, and Ziarko proposed a new model named as variable precision rough set model which introduced the level of β at $\beta(0\leq\beta<0.5)$. Thus little bit of misclassification is allowed under VPRS.

In this approach we have partitioned the datasets into multiple fold by using Cross Validation Technique. Under this technique each attribute will be tested at least once. 5-Fold validation scheme will partition data into 5 folds and give train data and test data. Variable precession rough set model will analyze the attributes and remove rough attributes. FCM algorithm will cluster the datasets. This process will repeat for 5 times and each time accuracy of datasets are improved to recognized level. Thus mixture of VPRS-FCM will enhance the clustering accuracy of machinery datasets.

II. BASIC CONCEPTS

A. Variable Precision Rough Set Model

In data analysis, Variable Precision Rough Set (VPRS) is widely used to analyze the data and find the redundant attributes in datasets. This model is very useful in solving the problems where the datasets have lots of boundary objects. The main property of VPRS is shown below:

- Information System

An Variable precision rough sets (VPRS) [4] attempts to improve upon rough set theory by relaxing the subset operator. It was proposed to analyze and identify data patterns which represent statistical trends rather than functional. The main idea of VPRS is to allow objects to be classified with an error smaller than a certain predefined level. This approach is arguably easiest to be understood within the framework of classification. Let P, $Q \subseteq U$, the relative classification error is defined by $C(P,Q)=\{ 1-(|P\cap Q)/(|P|)@0)\}$
$|P|$
Where $|P|$ is the cardinality of that set

- Degree of inclusion

  We Let P, Q be any two sets, if $0 \leq\beta< 0.5$, the majority inclusion relation can be defined as:

  $P\subseteq \beta$ if $C(P, Q) \leq \beta$, $0 \leq\beta< 0.5$

- *β -lower and β - upper Approximation of Set*

  Let R be the indiscernible relation on the universe U. Suppose (U, R) is an approximation space.U/R = {P1, P2…Pn} where Pi is an equivalence class of R. For any subset $P\subseteq U$, lower approximation $R\beta$-P and upper approximation R-βP of P with precision level β respect to R is respectively defined as

  $R\beta P = U\{ Q\in U/R|(P\cap Q)/P\leq\beta\}$
  $\bar{R}\beta P = U\{Q\in U/R|(P\cap Q)/P<1\}$

  Where the domain of β is $0 \leq\beta<0.5$, Rβ-X is also called β-Positive region (POS (P, Q)). The β boundary of P with respect to R is defined as:

  $BND\beta P = U\{Q\in U/R|\beta<(P\cap Q)/P+<1-\beta\}$

When β= 0, Ziarko variable precision rough set model becomes Pawlak rough set model.[6]

5-Fold Cross Validation Scheme

Cross validation is a prototype estimation technique that is better than residuals. The problem with residual evaluations is that it do not give an indication of how well the learner will do when it is asked to make new predictions for data it has not already seen. One way to overcome this problem is to not use the entire data set when training a learner. Some of the data is removed before training begins. Then when training is done, the data that was removed can be used to test the concert of the learned model on not existing before data.

In 5 fold validation we divide the data into k repeated times. In this process we divide the data in different subset, one k subset are used as training set and other k-1 sets are used together for testing set, when this task completed Then the centre misconception over all k instance is computed . The main edge of cross validation method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set k-1 times. The variance of the resulting estimate is reduced as k is increased.

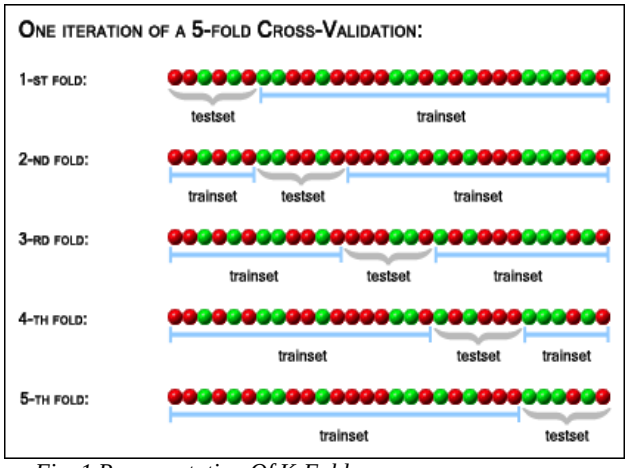


Fig. 1.Representation Of K-Fold

B. Fuzzy C-Mean Algorithm

Fuzzy c-means (FCM) is a data clustering technique in which a dataset is grouped into n clusters with every data point in the dataset belonging to every cluster to a certain degree. For example, a certain data point that lies close to the center of a cluster will have a high degree of belonging or membership to that cluster and another data point that lies far away from the center of a cluster will have a low degree of belonging or membership to that cluster. It starts with an initial guess for the cluster centers, which are intended to mark the mean location of each cluster. The initial guess for these cluster centers is most likely incorrect. Next, fcm assigns every data point a membership grade for each cluster. By iteratively updating the cluster centers and the membership grades for each data point, fcm iteratively moves the cluster centers to the right location within a data set. This iteration is based on minimizing an objective function that represents the distance from any given data point to a cluster center weighted by that data point's membership grade.

It also makes data description possible by means of clustering visualization, association and sequential analysis. Data clustering is primarily a method of data description which is used as a common technique for data analysis in various fields like machine learning, data mining, pattern recognization, image analysis and bio-informatics. Cluster analysis is also recognised as an important technique for classifying data, finding clusters of a dataset based on similarities in the same cluster and dissimilarities between different clusters [13]. Putting each point of the dataset to exactly one cluster is the basic of the conventional clustering method where as clustering algorithm actually partitions unlabeled set of data into different groups according to the similarity. As compare to data classification, data clustering is considered as an unsupervised learning process which does not require any labelled dataset as training data and the performance of data clustering algorithm is generally considered as much poorer. Although data classification is better performance oriented but it requires a labelled dataset as training data and practically classification of labelled data is generally very difficult as well as expensive. As such there are many algorithms that are proposed to improve the clustering performance. Clustering is basically considered as classification of similar objects or in other words, it is precisely partitioning of datasets into clusters so that data in each cluster shares some common trait. The hierarchical, partitioning and mixture model methods are the three major types of clustering processes that are applied for organising data. The choice of application of a particular method generally depends on the type of output desired, the known performance of the method with particular type of data, available hardware and software facilities and size of the dataset.

III. PROPOSED WORK

In this approach to improve the clustering accuracy of machinery datasets combination of multiple techniques are used that are VPRS, FCM,5-fold cross validation scheme We have explained our proposed work in following points shown below:

A.We have taken dataset from uci Machanery dataset.

B. After taking the data from uci dataset we have used 5 fold cross validation Scheme.here we apply different validations.

C. During performing validation we have divide the data set in two parts that is training data and testing data.

D.in training procisior we take the validated data and apply vprs for remove the redudent data.Vprs is basically used for remove the redudent attribute and overcome the redudent data set.

E.For increasing the accuracy of dataset we use Fuzzy C Mean algorithm Fuzzy c-means (FCM) is a data clustering technique in which a dataset is grouped into n clusters with every datapoint in the dataset belonging to every cluster to a certain degree.This algorithm match the attributes and if the attributes are same then it shows the attributes in different cluster and shows the accuracy of data.

F. In Testing process we again use the remaining validated data in our network.

G. certain datapoint that lies close to the center of a cluster will have a high degree of belonging or membership to that cluster and another datapoint that lies far away from the center of a cluster will have a low degree of belonging or membership to that cluster. It starts with an initial guess for the cluster centers, which are intended to mark the mean location of each cluster. . This iteration is based on minimizing an objective function that represents the distance from any given data point to a cluster center weighted by that data point's membership grade.

H.when In classification we classify the data sets by using different parameters, and check the accuracy of testing data sets.

I.In final phase we calculate the avarage of both training data and testing data if our average value is according to the desired average value then our network works properly.

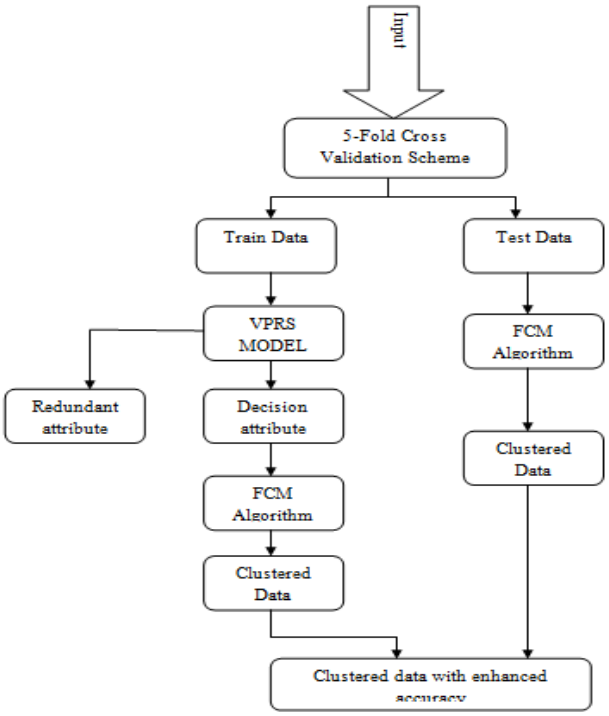
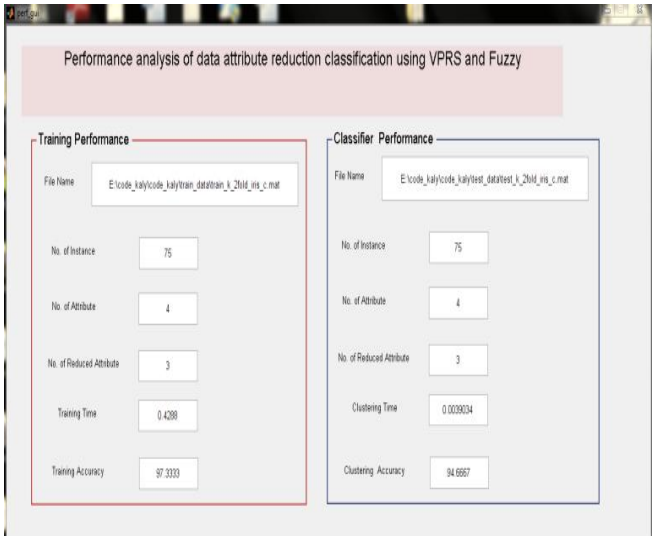
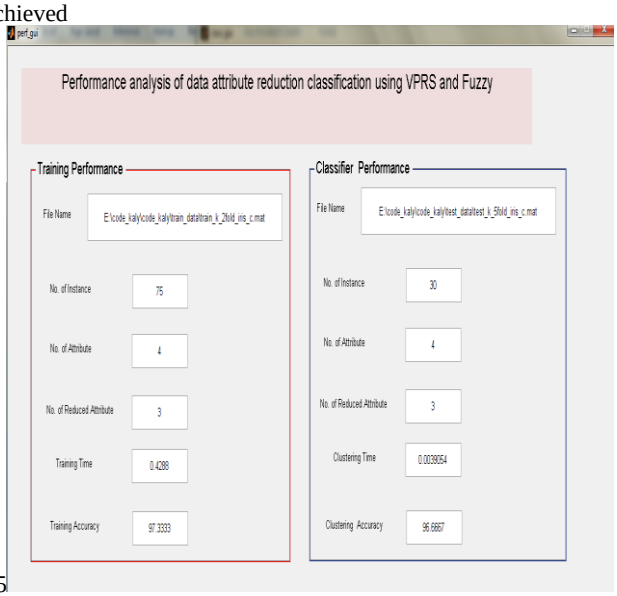


Fig. 2 Proposed Model



IV EXPERIMENTS AND RESULTS

Our experiments are carried out on an Intel (R) Pentium(R) CPU B940@ 2.00 GHz, 2GB RAM, 32 bit Windows 7 Operating System. All procedures were implemented on MATLAB System. We have used datasets from the UCI Machine Learning Repository. In the experiments self test validation was conducted on all data sets to calculate the classification accuracy .The performance chart shows that the classifier accuracy has increased from 94.66 % in 2-fold to 97.33% in 5-fold. First of all we have focuses on dividing the machinery data sets in multiple-folds so to partitioned the datasets into train data and test data. K-fold cross-validation technique partitioned, the original sample randomly k subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the classifier, and the remaining k − 1 subsamples are used as training data. In (figure 3) 2-fold comparison result of train and test datasets is shown below. In 2-fold 75 instances are in training and 75 instances are in testing. VPRS has reduced 1 attribute remaining attribute are 3. FCM has clustered data and clustering accuracy has achieved as



95

I. When we apply 5 fold in this network the performance will increased. In this below figure we show the increasing performance of our network,here the training time is 0.4288ns and clustering time is 0.0039054ns,the training accuracy is 97.33 and clustering accuracy has achieved as 96.6667. .

V. CONCLUSION:

We proposed the concept of the enhanced information gain based on VPRS Model. The analysis of the results has been carried out by means of the proper statistical study, which shows the goodness of this approach for dealing with classified machinery dataset's this approach improved the classification rate and proposes new attribute selection criteria. K-Fold cross validation technique has evaluated the performance of classifier. A new technique is proposed dataset shift has potentially introduced which result in inaccurate performance estimation. This paper analyzes the prevalence and impact of partition in the field of classification. This model has produced stable performance in the era of classification. Thus Datasets are properly classified

References.

[1] Khaled Shabam, "A Cascade of Artificial Neural Networks to Predict Transformers Oil Parameters," Prediction approach, vol.16,pp.517-518,2009.

[2] Rajkumar Sharma, "An Optimize Decision Tree Algorithm Based on Variable Precision Rough Set Theory Using Degree of β - dependency and Significance of Attributes ,"Machine Learning, vol. 2,pp.3942-3947,2012 .

[3] Wojciech Ziarko, "Variable Precision Rough Set Model," Classification Approach,pp.40-41,1993.

[4] Bahram G. Kermani, "Performance of the Levenberg–Marquardt neural network training method in electronic nose applications" Machine learning,vol.110,pp.15-16,2005.

[5] Z. Pawlak, "Information systems theoretical foundations," Information Systems, vol. 6, pp. 205–218, 1981.

[6] Manolis I.A.," Is Levenberg-Marquardt the Most Efficient Optimization Algorithm for Implementing Bundle Adjustment?"Optimization algorithm,pp. 1-3.

[7] Jose García Moreno-Torres,"Study on the Impact of Partition-Induced Dataset Shift on k-fold Cross-Validation,"vol.23,pp.1305-1306,2012.

[8] Djaffar Ould Abdeslam," A unified artificial neural network archtecture for active power filters", vol.54,pp.61-62,2007.

[9] Suyun Zhao,"The model of fuzzy varaible precesion rough sets",vol.17,pp.451-452,2009.

[10] Jose García Moreno-Torres,"Study on impact of partition-induced dataset shift on k-fold cross validation",vol.23,pp.1304,2012.

. Thus Datasets are properly classified