ISSN:0975-9646



# Survey on Improvement of Efficiency of Classified Machinery Datasets Using VPRS

Rashmi Rani Upadhyay[1], Ritesh Shah[2]

[1] *PG student of Digital communication, SIMS Indore*
[2] *Associate Professor of Information Technology Department, SIMS Indore*

*Abstract*-**With this research a new technique is proposed in the field of classification, where dataset attribute reduction and efficiency of classification has achieved successfully. As the data grows on explosive rates, it has lead to several problems such as increase difficulty of finding relevant information from the huge amount of dataset's. Efforts are being used to overcome this problem and thus concept of classification was introduced. In this paper, an effective technique is proposed for classifying the rough data sets based on variable precision rough set theory which can deal with inconsistent, uncertain or vague knowledge. In this project, we have proposed a new technique in which we have used variable precision rough set theory (VPRS) and Artificial Neural Network(ANN) with K – fold technique. First of all, the knowledge dependence in variable precision rough set theory is used to reduce the test attribute set of data setse, that is, the test attribute space is optimized and hence the attributes which are not correlated with the decision information are deleted. We have introduced neural network based on back-propagation algorithm for classification of machinery dataset's. Experiments proved that the accuracy of dataset's have got increased by using the combination of K-fold technique. The data sets are applied to VPRS algorithm which reduces useless attribute and these reduced attributes will get applied to ANN classifier so that output comes in classified form. This is repeated for 5 times by using K-Fold Technique, and finally we have removed rough data from the machinery data set and increased classification accuracy.**

**Keywords:** ***ANN algorithm, degree of β-dependency, variable precision rough set (VPRS), enhanced information gain, K-Fold Technique.***

## 1. INTRODUCTION:

Machine learning, a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms. Machine learning is programming computers to optimize a performance criterion using example data or past experience. Classification is the prediction approach in data mining techniques. There are many algorithms based on classification that is Instance Based, Bayesian Networks, Support Vector Machine and Decision tree, Neural Networks. Artificial Neural Networks are a family of information processing techniques inspired by the way biological nervous systems process information. The fundamental concept of neural networks is the structure of the information processing system. Composed of a large number of highly interconnected processing elements or neurons, a neural network uses the human-like technique of learning by example to solve problems. A Neural network is often configured for a specific application, such as data classification or pattern recognition, through a learning process called training. Just as in biological systems, learning involves adjustments to the synaptic connections that exist between the neurons. Neural networks can differ based on the way their neurons are connected, the specific kinds of computations their neurons perform and the way they transmit patterns of activity throughout the network. Neural networks are being applied to an increasingly large number of real-world problems. Their primary advantage is that they can solve problems that are too complex for conventional technologies, problems that do not have an algorithmic solution, or for which an algorithmic solution is too complex to be defined.

Rough set theory (RST), proposed by Poland mathematician Pawlak in 1982, is a new mathematic tool to deal with vagueness and uncertainty. Its main idea is that to classify samples into similar classes containing objects that are indiscernible with respect to some attributes. RST can solve many problems occurred in data reduction, feature selection and pattern extraction so that we can get rid of redundant data even in the information system with null values or missing data. [2] Rough-set-based decision tree algorithms have been studied within resent years. However, these proposed approaches also have their limitations. They 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, Ziarko proposed a variable precision rough set model , which introduced the $\beta(0\leq\beta<0.5)$ [1] based on the basic rough set model, and allowed some degree of misclassification rate. Aijun also proposed a variable precision rough set model, which introduced $\beta(0.5\leq\beta<1))$ as the correct rate. The main concept of variable precision rough set theory is degree of dependency and significance of attributes which is used in the proposed algorithm to select splitting attribute, therefore this approach proposes a new attribute selection criterion, the enhanced information gain based on degree of β-dependency and significance of condition attributes on decision attribute is used as a heuristic for selecting the optimal splitting attribute to overcome problem of attribute

reduction and also extends variable precision rough set theory [6].

We have used some machinery dataset for classification. It this approach we have used Back Propagation Neural network to classify the input datasets. VPRS technique is used so that the main task of finding reduct and core will get more simplified and easy. We have also used K-Fold technique in this proposed work. Experiments proved that VPRSANN using K-Fold technique has reduces the complexity and of the datasets and increase the classification accuracy in every fold's.

## 2. BASIC CONCEPTS

### 2.1 VPRS

In data analysis, *Variable Precision Rough Set* (VPRS) is very useful for addressing problems where data sets have lots of boundary objects. In addition, this model allows identifying data patterns that otherwise would be lost. The standard definition of the set inclusion relation is too rigorous to represent any *almost* complete set inclusion.[1] [6]

#### *2.1.1 Information Systems*

An information system [5] is a pair $S= (U, W, V, f)$ where $U$ is a non-empty finite set of objects called universe. $W$ denotes the set of attributes, it is usually divided into two subsets $P$ and $Q$, which denote the set of condition attributes and the set of decision attribute, respectively. $f: U \times W \rightarrow V$ is an information function, where $V = \_a \in WVa$ is the domain of attribute.

#### *2.1.2 Relative Misclassification Rate*

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

$$C(P,Q) = \begin{cases} 1 - \frac{|P \cap Q}{|P|} & |P| \\ 0 \end{cases}$$

Where |P| is the cardinality of that set.

#### 2.1.3 Degree of inclusion

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

***P, ⊆β if C (P, Q) ≤ β ,, 0 ≤β< 0.5***

#### *2.1.4 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 $R^{-}{}_{\beta}P$ of $P$ with precision level $\beta$ respect to $R$ is respectively defined as

$$R^{\beta}\ P = U\{\ Q \in U/R | \frac{P \cap Q}{P} \leq \beta\}$$

$$\overline{R}_{\beta} P = U\{Q \in U/R | \frac{P \cap Q}{P} < 1\}$$

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

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

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

### 2.2 ANN

Artificial Neural Networks are a family of information processing techniques inspired by the way biological nervous systems process information. The fundamental concept of neural networks is the structure of the information processing system. Composed of a large number of highly interconnected processing elements or neurons, a neural network uses the human-like technique of learning by example to solve problems. A Neural network is often configured for a specific application, such as data classification or pattern recognition, through a learning process called training. Just as in biological systems, learning involves adjustments to the synaptic connections that exist between the neurons. Neural networks can differ based on the way their neurons are connected, the specific kinds of computations their neurons perform and the way they transmit patterns of activity throughout the network. Neural networks are being applied to an increasingly large number of real-world problems. Their primary advantage is that they can solve problems that are too complex for conventional technologies, problems that do not have an algorithmic solution, or for which an algorithmic solution is too complex to be defined. [3]

#### 2.2.1 *Back-Propagation Learning Algorithm*

A number of learning rules are available to train neural networks. The Back Propagation (BP) learning algorithm is used in this study to train the multi-layer feed-forward neural network. Signals are received at the input layer, pass through the hidden layer, and reach to the output layer, and then fed to the input layer again for learning. The learning process primarily involves the determining of connection weights and patterns of connections. The BP neural network approximates the non-linear relationship between the input and the output by adjusting the weight values internally instead of giving the function expression explicitly. [5] Further, the BP neural network can be generalized for the input that is not included in the training patterns. The BP algorithm looks for minimum of error function in weight space using the method of gradient descent. The combination of weights that minimizes the error function is considered to be a solution to the learning problem.

### 2.3 K-Fold Cross Validation Technique

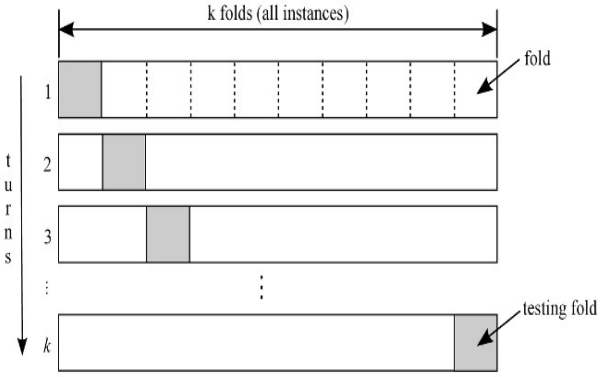
This paper analyzes how different partitioning methods can introduce dataset shift (or, more specifically, covariate shift) and the effect it has over both the reliability of the estimation of a classifier performance based on a low number of iterations of *k*-fold cross-validation, and the number of iterations needed to reach a stable classifier performance estimation. [4]

Cross-validation is a technique used for assessing how a classifier will perform when classifying new instances of

the task at hand. One iteration of cross-validation involves partitioning a sample of data into two complementary subsets: training the classifier on one subset (called the training set) and testing its performance on the other subset (test set).

In *k*-fold cross-validation, the original sample is randomly partitioned into *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. The cross-validation process is then repeated *k* times, with each of the *k* subsamples used exactly once as the test data. The *k* results from the folds are then averaged to produce a single performance estimation.

When we have one dataset with the answers (the "class" for each data point), we can split this dataset into training and testing portion. The training portion is used to build a model of the dataset, and the testing version is used to test that model. We'll want to split the dataset multiple times at random places and then average the results. Most common is 10-fold cross validation. This means we choose 90% of the data to be the training set, and 10% to be the testing set. We evaluate the precision/recall/etc. with this split, then choose a different 90/10 split and do it again. Because there are 10 possible splits, we do it 10 times and average 10 results




In oreder to evaluate the expected performance of a classifier over a dataset, *k*-fold cross-validation schemes are commonly used in the classification literature. Also, when comparing classifiers, it is common to compare them according to their performances averaged over a number of iterations of cross-validation. Even though it has been proved that these schemes asymptotically converge to a stable value, which allows realistic comparisons between classifiers in practice a very low number of iterations are often used. The most common variations are 2×5, 5×2, and 10×1, with this notation meaning 2-folds iterated five times, 5-folds iterated two times, and 10-folds iterated once, respectively. Note that when more than one iteration takes place, the partitions are assumed to be constructed independently

## III. IMPROVED VPRS WITH ANN CLASSIFIER BY K-FOLD TECHNIQUE

In this approach we have proposed a new technique to classify the dataset's with improved accuracy. In order to achieve accuracy in classification standard dataset's are used. K-Fold cross validation partitioning is used to partition the dataset's into two set's i.e., training set and test set. We performed three independent experiments using the same procedure, where the only difference was the type of cross-validation scheme used. We tested the 1×2, 1×3,1×4,1×5 cross-validation schemes shown below.

```
k=[1/2 ,1/3, ¼, 1/5 ];
len1=round(n*k);
len2=n-len1;
data=standata1(1:len2,:);
data1=standata1(len2+1:end,:);

save(strcat(cd,'\train_data\train_k_',num2str(1/k),'fold_',fn
m),'data')
save(strcat(cd,'\test_data\test_k_',num2str(1/k),'fold_',fnm),'
data1')
```

**Now we have implemented this technique in following ways:**

**Input**: An information system $S=(U,P \cup Q,V,f)$, the training parameter and the thresh hold parameter $\beta, 0 \leq \beta < 0.5$

**Output**: Classified dataset's with improved classification accuracy.

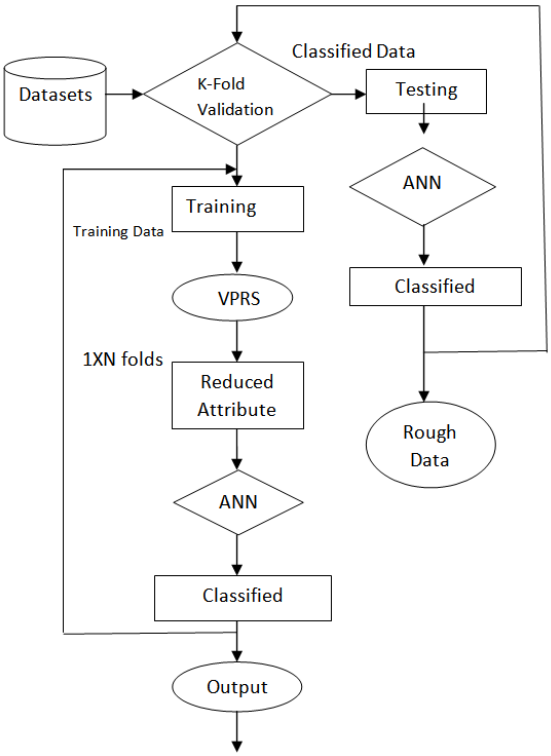
**Step 1**: Create the list of standard datasets based on machinery learning schemas.

**Step 2**: For each dataset's apply K-fold cross validation technique to partition the data sets in train dataset's in such a way that training the classifier on one subset (called the training set) and testing its performance on the other subset (test set).

**Step 3:** Apply the VPRS technique on training dataset's and classify the same by ANN neural Network. On the other hand Classify the test dataset's also. Then compare the classifier result of training datasets and test datasets.

**Step 4:** Now holdout method is repeated 5 time's. Each time one of the *k* subsets is used as *Test data* set and the other *k-1* subsets are put together to form the *New training* one. Finally, the *k* accuracy estimates are averaged to provide the accuracy of the final model trained on *new training dataset's*. The variance term of the resulting estimate reduces as *k* increases.

**Step 5:** After completion of 5-fold technique no more training samples will be there to classify, finally the datasets will be classified with more information gain.




## IV. CONCLUSION:

We proposed the concept of the enhanced information gain based on VPRS Model. 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 where dataset shift has potentially introduced, which result in accurate 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. Rajkumar Sharma, Pranita Jain, Shailendra K.Shrivastava "An Optimize Decision Tree Algorithm Based on Variable Precision Rough Set Theory Using Degree of β-dependency and Significance of Attributes." International Journal of Computer Science and Information Technologies, Vol. 3 (3) 2012,3942-3947
2. D. Calvo-Dmgz, J. F. G´alvez "Using Variable Precision Rough Set for Selection and Classification of Biological Knowledge Integrated in DNA Gene Expression" Journal of Integrative Bioinformatics, 9(3):199, 2012
3. Khaled Shaban, Ayman El-Hag "A Cascade of Artificial Neural Networks to Predict Transformers Oil Parameters".
4. Jose García, Moreno-Torres, José A. Sáez "Study on the Impact of Partition-Induced Dataset Shift on *k*-fold Cross-Validation" IEEE Transactions on Neural Networks and Learning Systems, vol. 23, no. 8, August 2012
5. Jason L. Wright, Milos Manic "Neural Network Architecture Selection Analysis With Application to Cryptography Location"
6. Suyun Zhao, Eric C. C. Tsang "The Model of Fuzzy Variable Precision Rough Sets" IEEE Transactions on Fuzzy Systems, vol. 17, no. 2, April 2009.
7. www.ics.uci.edu/mlearnlMLRepository.html