An Optimal Feature Selection Process Using Roughset Theory in High Dimensional Data Classification

Ms. N. Gayathri[1], Ms. K. Yemuna Rane M.Sc., M. Phil., M.Sc (App. Psy) [2]

[1]*M.Phil Research Scholar, Dept of Computer Science, Kongunadu Arts and Science College, Tamil Nadu, India*

[2]*Assistant Professor, Dept of Computer Applications, Kongunadu Arts and Science College, Tamil Nadu, India*

ABSTRACT

This research entitled "AN OPTIMAL FEATURE SELECTION PROCESS USING ROUGHSET THEORY IN HIGH DIMENSIONAL DATA CLASSIFICATION" incorporates information theory, which is the process of deriving the information from the feature selection from the unsupervised dataset. Feature Selection is the application of data mining techniques to discover patterns from the micro array datasets. Finding the best features that are similar to a test data is challenging task in current trend. To discover the significance features have more frequent change in the structural information, which involves feature dimensionality reduction, linked to one another and elimination of non-structural information. This research presents a framework for discovering best feature selection from unsupervised datasets. By aligning the relevant features from the datasets and by using the matching sequence or its frequency of match, the searching between the data features are determined.

The proposed research work presents a new approach to measure the features (attributes) in micro array datasets using the methodologies namely, data cleaning, Adaptive Relevance Roughset Feature Discovery, minimal-Redundancy-Maximal-Relevance (mRMR) and classification. Data feature selection and dimensionality reduction is characterized by a regularity analysis where the feature values correspond to the number times that term appears in the dataset. The relevance Roughset feature discovery method gives a useful measure is used to find the similarity features between data points are likely to be in terms of their features property. Despite the usefulness of searching measures in these applications, accurately measuring the similarity between the features or attributes remains a challenging task.

Some of the challenges faced in finding the best feature selection include positive, negative and inconsistency. This research proposes an enhanced relevance rough set based classification method to estimate the feature searching is measured using minimal redundancy optimization method corresponding micro array data. Each feature contains objective function and their own description which is used to identify the type of datasets. Initially, the total numbers of features are identified to enhanced feature selection of the datasets where the terms of match between the features are identified with help of classification algorithms.

Keywords: - Feature selection, Micro array dataset, Adaptive Relevance Roughset Feature Discovery, minimal-Redundancy-Maximal-Relevance (mRMR), classification, Dimensionality reduction and unsupervised dataset.

1. INTRODUCTION

Due to the rapid growth of digital data made available in recent year, knowledge discovery and data mining have attracted a great deal of attention with an imminent need for turning such data into useful information and knowledge. Many applications, such as market analysis and business management, can benefit by the use of the information and knowledge extracted from a large amount of data. Knowledge discovery can be viewed as the process of nontrivial extraction of information from large databases, information that is implicitly presented in the data, previously unknown and potentially useful for users. The presence of high dimensional data is becoming more

common in many practical applications such as data mining, machine learning and microarray gene expression data analysis. Feature selection is common in machine learning, where it may also be termed feature subset selection, variable selection, or attribute reduction. Finding relevant features simplifies learning process and increases prediction accuracy.The learning algorithms in the evaluation of subsets, some of which can encounter problems when dealing with large datasets. The main objective is to identify important and desirable theoretical properties of algorithms for feature selection and to identify suitable performance measures for evaluating algorithms for relevance classification in roughset feature selection data. The use of rough set method to solve a specific complex problem has attracted world-wide attention. Feature based methods include efficient computational performance as well as mature theories for feature weighting.To observe that most nonlinear techniques have major problems when faced with a dataset with a high intrinsic dimensionality. Finds minimal sets of data (data reduction).Predictive instances are instances that may produce predictive rules which hold true with a high probability.

1.1 Related Work and Drawbacks

Mutual Information (MI) approach is to first discretize the continuous features in the preprocessing step and use mutual information (MI) to select relevant features. Drawback of Mutual Information is a huge number of the features with continuous values using the definition of relevancy are quite a difficult task. **Fisher linear discriminant analysis** can be as poor as random guessing as the number of features gets larger. Drawbacks of Fisher linear discriminant analysis is the problem of *statistical variable selection* such as forward selection, backward elimination and their combination can be used for FS problems. **Bipartite Graphs is** to combine both the clicked and skipped URLs from users in the query-URL bipartite graphs in order to also consider rare query suggestions (using clicked URLs only favors popular queries).Drawback of Bipartite Graphs is to utilized forward selection method but not considered backward elimination method.

Problem Definition:

- Determining dimensionality reduction subset optimality is a challenging problem.
- A non-linear algebraic formulation of the high dimensional Classification problem.
- The learning algorithms in the evaluation of subsets, some of which can encounter problems when dealing with large datasets.

1.2 Study of Algorithm

The main aim of feature selection (FS) is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features. In real world problems FS is a must due to the abundance of noisy, irrelevant or misleading features. For instance, by removing these factors, learning from data techniques can benefit greatly. Given a feature set size *n*, the task of FS can be seen as a search for an optimal feature subset through the competing 2*n* candidate subsets. The definition of what an optimal subset is may vary depending on the problem to be solved. Although an exhaustive method may be used for this purpose, this is quite impractical for most datasets. Usually FS algorithms involve heuristic or random search strategies in an attempt to avoid this prohibitive complexity. However, the degree of optimality of the final feature subset is often reduced. The usefulness of a feature or feature subset is determined by both its *relevancy* and *redundancy*. A feature is said to be relevant if it is predictive of the decision feature(s), otherwise it is irrelevant. A feature is considered to be redundant if it is highly correlated with other features. Hence, the search for a good feature subset involves finding those features that are highly correlated with the decision feature(s), but are uncorrelated with each other.

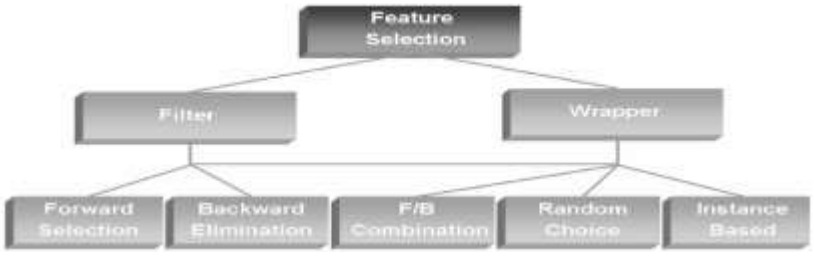

Fig -1: Aspects of feature selection

Determining subset optimality is a challenging problem. There is always a trade-off in non-exhaustive techniques between subset minimalist and subset suitability. The task is to decide which of these must suffer in order to benefit the other. For some domains (particularly where it is costly or impractical to monitor many features), it is much more desirable to have a smaller, less accurate feature subset. In other areas it may be the case that the modeling accuracy (e.g. the classification rate) using the selected features must be extremely high, at the expense of a non-minimal set of features.

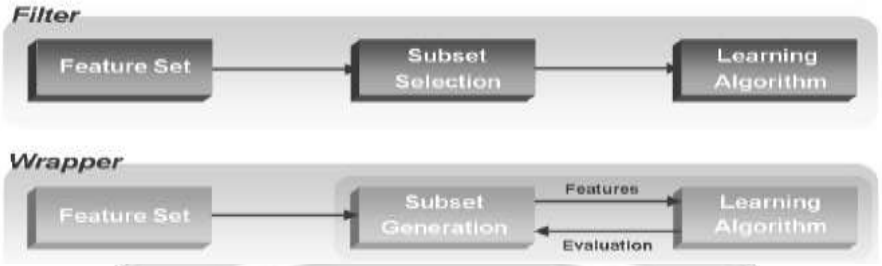

Fig -2: Filter and wrapper methods

Feature selection algorithms may be classified into two categories based on their evaluation procedure (Figure 2).If an algorithm performs FS independently of any learning algorithm (i.e. it is a completely separate preprocessor), then it is a filter approach. In effect, irrelevant attributes are filtered out before induction. Filters tend to be applicable to most domains as they are not tied to any particular induction algorithm. If the evaluation procedure is tied to the task (e.g. classification) of the learning algorithm, the FS algorithm employs the wrapper approach. This method searches through the feature subset space using the estimated accuracy from an induction algorithm as a measure of subset suitability. Although wrappers may produce better results, they are expensive to run and can break down with very large numbers of features. This is due to the use of learning algorithms in the evaluation of subsets, some of which can encounter problems when dealing with large datasets.

1.3 Dimensionality Reduction

The main distinction between techniques for dimensionality reduction is the distinction between linear and nonlinear techniques. Linear techniques assume that the data lie on or near a linear subspace of the high-dimensional space. Nonlinear techniques for dimensionality reduction do not rely on the linearity assumption as a result of which more complex embeddings of the data in the high-dimensional space can be identified. Figure 3 shows taxonomy of techniques for dimensionality reduction. Linear techniques perform dimensionality reduction by embedding the data into a subspace of lower dimensionality. Although there are various techniques exist to do so, PCA is by far the most popular (unsupervised) linear technique. Therefore, in this comparison, PCA only included as a benchmark.

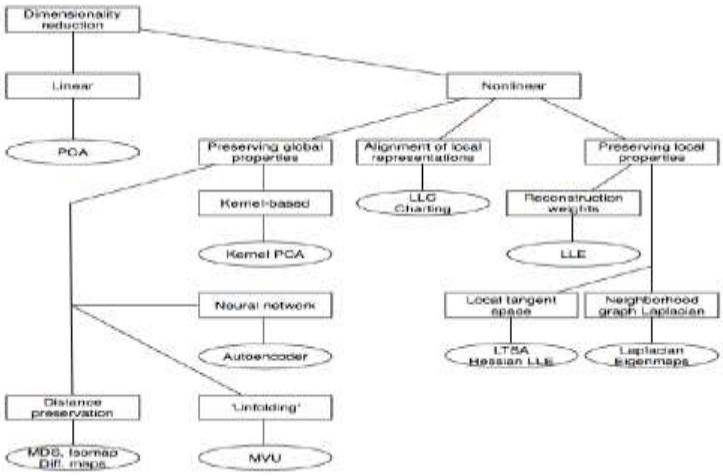

Fig -3: Taxonomy of dimensionality reduction techniques

1.4 Rough Set Theory

Rough set theory (RST) can be used as a tool to discover data dependencies and to reduce the number of attributes contained in a dataset using the data alone, requiring no additional information (Pawlak, 1991; Polkowski, 2002) in figure 4.Over the past ten years, RST has become a topic of great interest to researchers and has been applied to many domains. Given a dataset with discretized attribute values, it is possible to find a subset (termed a *reduct*) of the original attributes using RST that are the most informative. All other attributes can be removed from the dataset with minimal information loss. From the dimensionality reduction perspective, informative features are those that are most predictive of the class attribute. There are two main approaches to find rough set reducts. It considers the degree of dependency and with the discernibility matrix. This section describes the fundamental ideas behind both approaches. To illustrate the operation of these, an example dataset (Table 1) will be used.

Table -1: An example dataset

$x \in U$	A	B	C	D	E
0	1	0	2	2	0
1	0	1	1	1	2
2	2	0	0	1	1
3	1	1	0	2	2
4	1	0	2	0	1
5	2	2	0	1	1
6	2	1	1	1	2
7	0	1	1	0	1

Central to Rough Set Attribute Reduction (RSAR) is the concept of indiscernibility. Let $I = (U, A)$ be an information system, where U is a non-empty set of finite objects (the universe) and A is a non-empty finite set of attributes such that $a:U \rightarrow Va$ for every $a \in A$. Va is the set of values that attribute a may take. With any $P \in A$ there is an associated equivalence relation $IND(P)$.

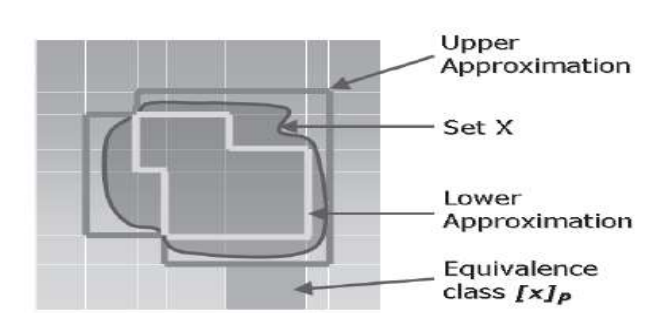

Fig -4: A Roughset Concept

Classification algorithms typically contain two phases,

- **Training Phase:** In this phase, a model is constructed from the training instances.
- **Testing Phase:** In this phase, the model is used to assign a label to an unlabeled test instance.

In some cases, such as lazy learning, the training phase is omitted entirely, and the classification is performed directly from the relationship of the training instances to the test instance. Instance-based methods such as the nearest neighbor classifiers are examples of such a scenario. Even in such cases, a pre-processing phase such as a nearest neighbor index construction may be performed in order to ensure efficiency during the testing phase.

2. METHODOLOGY

The proposed architecture accepts the data classification parameters as input which contains the MATLAB simulation where the novel boosting feature selection classification algorithm is applied to the Prostate Tumor micro array dataset. This overall architecture in figure 5 follows a high dimensional classification from the start to end state. The users initialize the dataset instances, attributes and classes as initial parameters in which the classification process is to be evaluated.

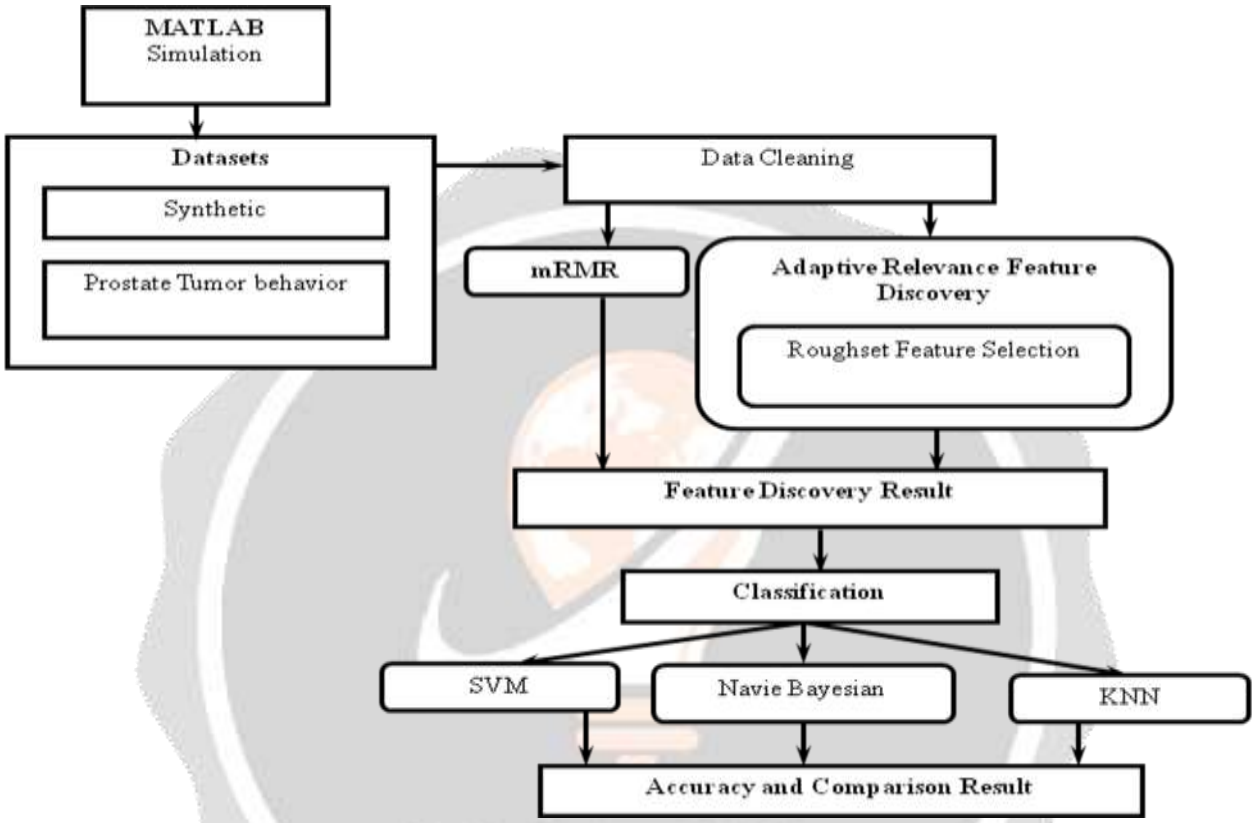

Fig -5: Architecture of Proposed System

Classification problems often have a large number of features, but not all of them are useful for classification. Irrelevant and redundant features may even reduce the classification accuracy. Feature selection is a process of selecting a subset of relevant features, which can decrease the dimensionality, shorten the running time, and/or improve the classification accuracy. Feature selection (FS) refers to the problem of selecting those input attributes that are most predictive of a given outcome; a problem encountered in many areas such as machine learning, pattern recognition and signal processing. Unlike other dimensionality reduction methods, feature selectors preserve the original meaning of the features after reduction. The proposed system attempts to use the uncertain information to improve the performance of rough sets and extensions thereof for the task of FS. These approaches are applied to two applications domain problems where the reduction of features is of high importance; microarray gene expression data analysis and complex systems monitoring. The utility of the approaches is demonstrated and compared empirically with several other dimensionality reduction techniques. In several experimental evaluation sections, the approaches are shown to equal or improve classification accuracy when compared to results obtained from unreduced data. Based on the new Roughset feature selection approaches and techniques are also presented in this thesis. The first of these is the application of a nearest neighbor classifier for the classification of real-valued data. This technique is evaluated within the microarray gene expression data analysis. The evidence that the clinical phenotypes and behavior of prostate cancer can be anticipated by the analysis of the gene expression profiles. Also, a novel unsupervised feature selection approach is proposed which reduces features by eliminating those which are considered redundant. The Adaptive Relevance Roughset Feature Discovery classifier mentioned are employed and evaluated for the complex systems monitoring application.

The Roughset feature selection process flow diagram is in figure 5 follows a completed feature selection from the start to end state.

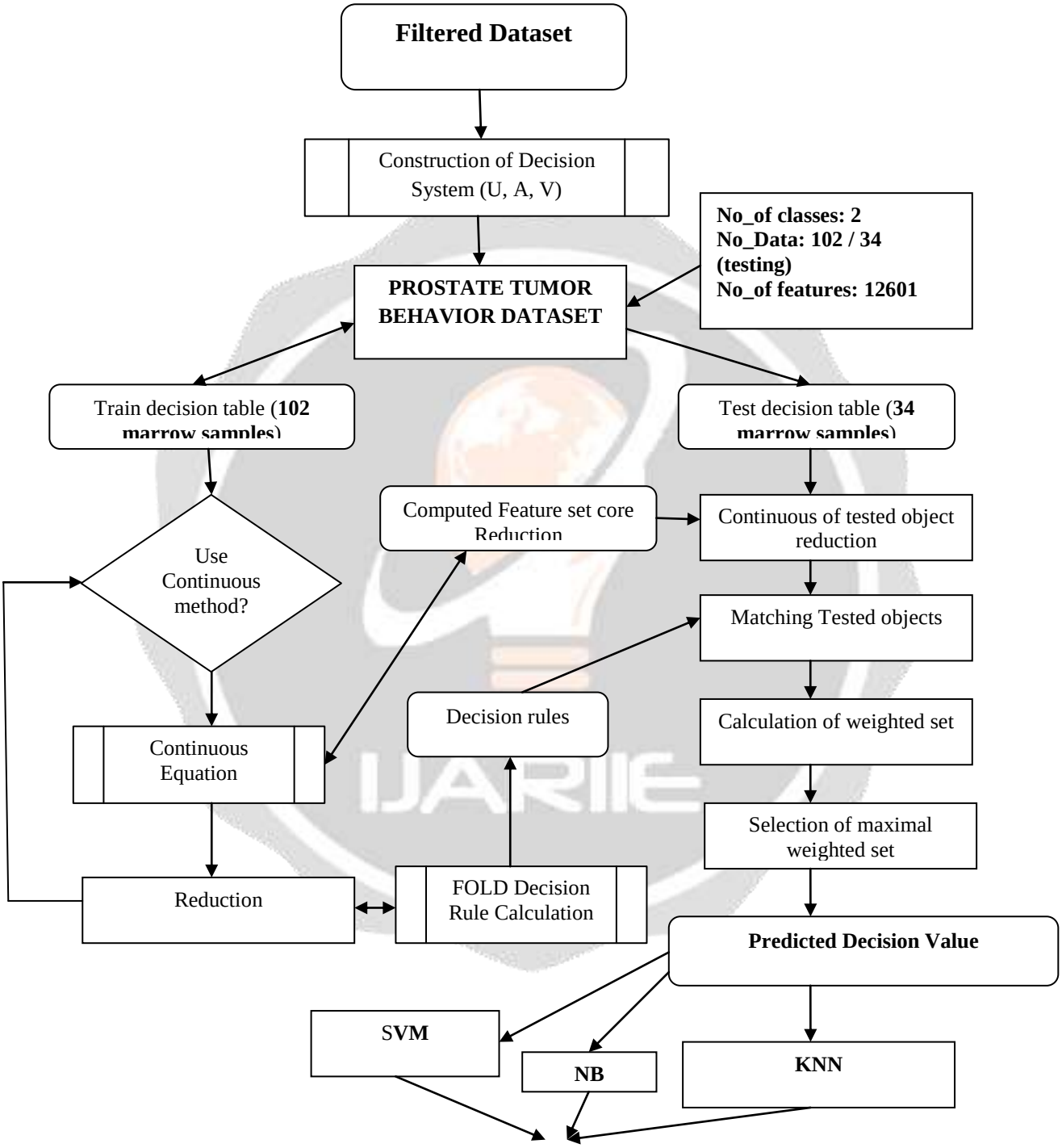

Fig -6: Roughset feature selection process flow

The following methodology is listed below:

- Data Cleaning
- Adaptive Relevance Roughset Feature Discovery
- minimal-redundancy-maximal-relevance (mRMR)
- Classification

2.1 Data Cleaning

Data cleaning method is kind of preprocessing technique it plays a very important role in data classification techniques and applications. The three key steps of data cleaning are ***Training set extraction***, ***Feature Attribute selection*** and ***filtering methods***. **Training set Extraction:** To compute the *cross validation classification* error for a large number of features and find a relatively stable range of small error. **Feature Attribute Selection:** It is a statistical technique that can reduce the dimensionality of data as a by-product of transforming the original attribute space. **Filtering Approach:** It has much ***lower complexity*** than wrappers; the features thus selected often yield comparable classification errors in different classifiers. The unsupervised raw dataset is first partitioned into three groups: (1) a ***finite set of objects***, (2) the ***set of attributes*** (features, variables) and (3) the ***domain of attribute***. For each groups in the dataset, a decision system is constructed. Each decision system is subsequently split into two parts: the ***training dataset*** and the ***testing dataset***. Each training dataset uses the corresponding input features and fall into two classes: ***normal*** (+1) and ***abnormal*** (−1).

2.2 Adaptive Relevance Roughset Feature Discovery

The adaptive relevance feature discovery process considers the *mutual-information-based feature selection* for both supervised and unsupervised data. For discrete feature variables, the integral operation in eqn.(1) reduces to summation. Given two random variables x and y, their mutual information is defined in terms of their probabilistic density functions $p(x)$, $p(y)$, and $p(x, y)$:

$$MI(x,y) = \int\int p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy \quad (1)$$

In Maximum Relevance discovery, the selected features x_i are required, individually, to have the ***largest mutual information*** $MI(x_i, c)$ with the target class c, reflecting the largest dependency on the target class. Given N samples of a variable x, the approximate similarity function $Simm(x)$ has the following form:

$$Simm(x) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - x^i, h) \quad (2)$$

Where $\delta(.)$ is the sampling window function as explained below, $x^{(i)}$ is the *ith* sample, and h is the window length.

Rough set theory is a new mathematical approach to imprecision, vagueness and uncertainty. The Roughset feature selection as the process of finding a subset of features, from the original set of pattern features, optimally according to the defined criterion. Rough sets theory is based on the concept of an ***upper*** and a ***lower approximation*** of a set, the approximation space and models of sets. An information system can be represented as,

$$S = (U, A, V, f) \quad (3)$$

where U is the universe, a finite set of N objects $(x_1, x_2, \ldots, x_N)$ (a nonempty set),
A is a finite set of attributes, $V = U_{a \in A} V_a$ (where V_a is a domain of the attribute a).

The straightforward feature selection procedures are based on an evaluation of the predictive (Entropy) power of individual features, followed by a ranking.

Algorithm1: Feature selection based on Rough sets

Input : Set of conditional and decisional features C, D.
Output: A subset of features
Process
Step 1: Initialize the best subset of features as the empty set.

Step 2: For *i* in 1: number of conditional features Apply some evaluation measure based on dependency of Roughsets.

End for

Step 3: Order the features according to dependency measure

Step 4: Select only the features with high dependency measure.

2.3 minimal-Redundancy-Maximal-Relevance (mRMR)

The purpose of feature selection is to find a feature set S with m features {xi}, which jointly have the largest dependency on the target class c. This scheme, called Max-Dependency, has the following form:

$$\max D(S, c),\ D = I(\{x_i, i = 1, 2, \ldots, m\}; c) \qquad (4)$$

The Max-Dependency criterion is hard to implement, an alternative is to select features based on maximal relevance criterion (Max-Relevance). Max-Relevance is to search features which approximates with the mean value of all mutual information values between individual feature and class. The following minimal redundancy (Min-Redundancy) condition can be added to select mutually exclusive features:

$$\min R(S), \qquad R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j)$$

3. IMPLEMENTATION

The implementation for developing the boosting feature selection classification algorithm uses MATLAB R2010a simulator. The classification will provide a boosting approach in the feature selection method and prevent the Classification problems in high dimensional data in distributed manner. The Prostate cancer behavior microarray data set are considered for implementation process. It is a high dimensional data set with small sample sizes and large number of features. It contains 102 samples for training, 34 samples for testing and number of class size is 2. The prostate cancers of identical grade can have widely variable clinical courses, from indolence over decades to explosive growth causing rapid patient death. Cancer classification has been difficult in part because it has historically relied on specific biological insights, rather than systematic and unbiased approaches for recognizing tumor subtype. To divided cancer classification into two challenges: class discovery and class prediction. Class discovery refers to defining previously unrecognized tumor subtypes. Class prediction refers to the assignment of particular tumor samples to already-defined classes. The class prediction can be applied to any measurable distinction among tumors. Importantly, such distinctions could concern a future clinical outcome—such as whether a prostate cancer turns out to be indolent or a breast cancer responds to a given chemotherapy. Class discovery involves two issues: (I) developing algorithms to cluster tumors by gene expression and (II) determining whether putative classes produced by such clustering algorithms are meaningful—that is, whether they reflect true structure in the data rather than simply random aggregation.

4. CLASSIFICATION

The research work performed the experiments on a prostate gene expression micro array dataset. The performance evaluation considers the effect of b (reasonable choice as was the case with the random permutation) on the change of the accuracy and the *Roughset feature selection classification (RFSC)* statistic for all the combinations of the four

FS algorithms and the three classifiers.Four FS algorithms considered in this research are Roughset feature selection classification (RFSC), minimal-redundancy-maximal-relevance (mRMR), Fast Correlation-Based Filter (FCBF), and Fast clustering bAased feature Selection algoriThm (FAST).

4.1 SVM Classification

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges.In this algorithm, the research work get accuracy or correct rate of each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.In matlab svmtrain function uses an optimization method to identify support vectors si, weights αi, and bias b that are used to classify vectors x according to the following equation:

$$c = \sum_{i} \alpha_i k(\alpha_i, x) + b$$

Table -2: SVM Accuracy for the Four FS Algorithms

Random value (b)	FAST	FCBF	mRMR	RFSC
1	87.9	94.3	90.2	98
2	89.8	93.7	93.5	95.45
3	90.6	94.4	94.2	96.21
5	90.9	94.6	94.9	95.11
10	90.8	94.9	94.1	96.78
20	91.5	95.1	93.9	96.21

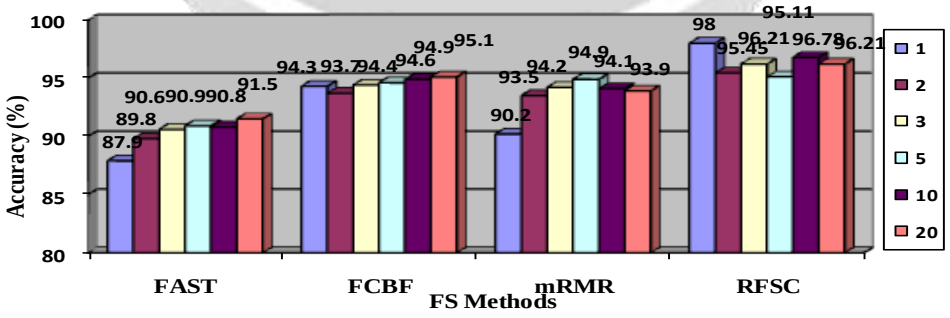

Chart -1: SVM Accuracy for the Four FS Algorithms

4.2 KNN Classification

In matlab *knnclassify* function classifies the rows of the data matrix Sample into groups, based on the grouping of the rows of Training. Classification *KNN* predicts the classification of a point X_{new} using a procedure equivalent to this:

- Find the Number of Neighbours points in the training set X that are nearest to X_{new}.
- Find the Num of Neighbours response values Y to those nearest points.
- Assign the classification label Y_{new} that has smallest expected misclassification cost among the values in Y.

Table -3: KNN Accuracy for the Four FS Algorithms

Random value (b)	FAST	FCBF	mRMR	RFSC
1	87.9	92.8	89.8	92.89
2	88.6	92.4	92.1	93.05
3	89.9	93.9	93.1	94.66
5	89.8	93.3	92.9	93.54
10	90.3	93.5	92.9	94.02
20	90.7	93.4	93.3	95.80

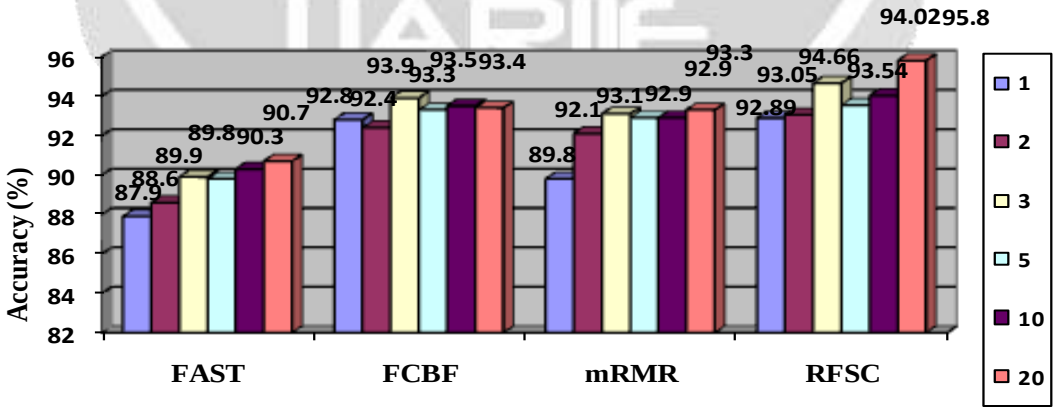

Chart -2: KNN Accuracy for the Four FS Algorithms

4.3 NB Classification

The Navie Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems.Navie Bayesian classifiers use Bayes theorem, which says

$$p(c_j|d) = \frac{p(d|c_j)p(c_j)}{p(d)}$$

$p(c_j|d)$ = probability of instance d being in class c_j
$p(d|c_j)$ = probability of generating instance d given class c_j,
$p(c_j)$ = probability of occurrence of class c_j,
$p(d)$ = probability of instance d occurring

Table -4: NB Accuracy for the Four FS Algorithms

Random value (b)	FAST	FCBF	mRMR	RFSC
1	88.8	94.9	90.6	**95.11**
2	91.0	93.5	93.5	**94**
3	91.8	94	94.1	**96.31**
5	91.5	95.3	94.4	**96.19**
10	91.8	94.7	94.2	**95.70**
20	91.8	94.8	93.8	**96.95**

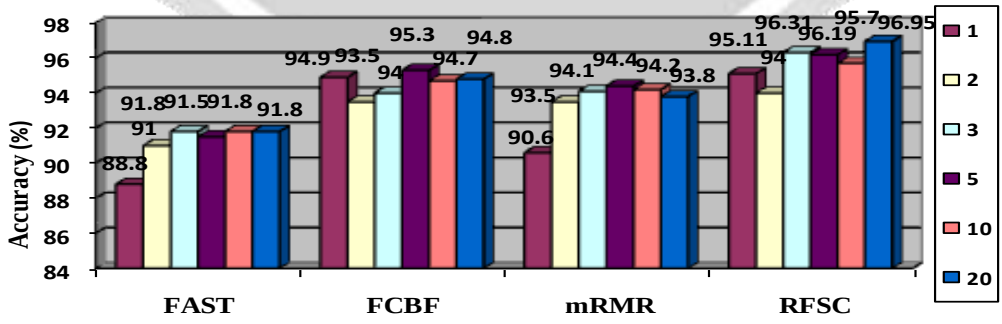

Chart -3: NB Accuracy for the Four FS Algorithms

5. CONCLUSION

The research work presents an enhanced method such as Randomized Feature Selection Classification using Adaptive Relevance Feature Discovery (ARFD) based Roughset method which combines classifications of SVM, KNN and NB to solve the problem of high dimensional classification. In the ARFD model, some of the new training documents will be selected using the knowledge currently held by the system. The theoretical analysis of the proposed methods is based on the fact that dimensionality reduction for roughest method has deep connections with boosting approximations to the data matrix that contains the points one wants to cluster. This research focuses on those connections in the text and employed modern fast algorithms to compute such low rank approximations and designed fast algorithms for dimensionality reduction in k-means.The proposed methodologies performance is analyzed with synthetic datasets and Micro array datasets those are downloaded from machine learning repository. The values are compared with several constrains such as number of dimensions versus objective, running time, accuracy. Based on the results generated this research concludes that accuracy increases compared to the previous method of mRMR algorithm.

6. FUTURE WORK

A further challenge is to identify an important future direction is to develop a computationally efficient method of determining the distance metric of the embedding space, manifold finding and dynamic/streaming data. Evolving some dimensional reduction methods like canon pies can be used for high dimensional datasets is suggested as future work.

7. REFERENCES

[1] T. Abeel, T. Helleputte, Y. V. de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," Bioinformatics, vol. 26, no. 3, pp. 392–398, 2010.

[2] D. Aha and D. Kibler, "Instance-based learning algorithms," Mach. Learn., vol. 6, no. 1, pp. 37–66, 1991.

[3] F. Alonso-Atienza, J. L. Rojo-Alvare, A. Rosado-Mu~noz, J. J. Vinagre, A. Garcia-Alberola, and G. Camps-Valls, "Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection," Expert Syst. Appl., vol. 39, no. 2, pp. 1956–1967, 2012.

[4] D. Dernoncourt, B. Hanczar, and J. D. Zucker, "Analysis of feature selection stability on high dimension and small sample data," Comput. Statist. Data Anal., vol. 71, pp. 681–693, 2014.

[5] J. Fan and Y. Fan, "High dimensional classification using features annealed independence rules," Ann. Statist., vol. 36, no. 6, pp. 2605–2637, 2008.

[6] A.J. Ferreira and M. A. T. Figueiredo, "Efficient feature selection filters for high dimensional data," Pattern Recog. Lett., vol. 33, no. 13, pp. 1794–1804, 2012.

[7] Y. Han and L. Yu, "A variance reduction framework for stable feature selection," Statist. Anal. Data Mining, vol. 5, no. 5, pp. 428–445, 2012.

[8] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," The J. Mach. Learn. Res., vol. 5, no. 2, pp. 1205–1224, 2004.

[9] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.

[10] Y. Li, C. Zhang, and J.R. Swan, "An Information Filtering Model on the Web and its Application in Jobagent," Knowledge-Based Systems, vol. 13, no. 5, pp. 285-296, 20

[11] S. Robertson and I. Soboroff, "The Trec 2002 Filtering Track Report," TREC,2002, trec.nist.gov/pubs/trec11/papers/OVER.FILTERING.ps.gz.

[12] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and minredundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[13] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," IEEE Trans. Knowl. Data Eng., vol. 25, no. 1, pp. 1–14, Jan. 2013.

[14] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," the J. Mach. Learn. Res., vol. 5, no. 2, pp. 1205–1224, 2004.