Maximum Spanning Tree Based Redundancy Elimination for Feature Selection of High Dimensional Data

Bharat Singh and Om Prakash Vyas
Department of Information Technology, Indian Institute of Information Technology-Allahabad, India

Abstract: *Feature selection adheres to the phenomena of preprocessing step for High Dimensional data to obtain optimal results with reference of speed and time. It is a technique by which most prominent features can be selected from a set of features that are prone to contain redundant and relevant features. It also helps to lighten the burden on classification techniques, thus makes it faster and efficient.We introduce a novel two tiered architecture of feature selection that can able to filter relevant as well as redundant features. Our approach utilizes the peculiar advantage of identifying highly correlated nodes in a tree. More specifically, the reduced dataset comprises of these selected features. Finally, the reduced dataset is tested with various classification techniques to evaluate their performance. To prove its correctness we have used many basic algorithms of classification to highlight the benefits of our approach. In this journey of work we have used benchmark datasets to prove the worthiness of our approach.*

Keywords: *Data mining, feature selection, tree based approaches, maximum spanning tree, high dimensional data.*

Received February 15, 2015; accepted December 21, 2015

1. Introduction

The advancement in the current rapidly increasing world of digitally connected data, creates high dimensional data. To extract knowledge from this raw data, Data Mining proposed an automated solution which unfortunately proven unsuccessful for handling high dimensional data because of high computational cost while dealing with such type of data. One of the solution for aforementioned problem is feature selection [3, 7, 14, 21, 25], which is known as a pre-processing step before applying Data Mining process.

Data mining or knowledge discovery is the practice of investigating various data from numerous characteristics and encapsulating it into more useful and profitable information by summarizing its inter-relationships. It enhance decision making [20, 27], thereby cutting costs and incrementing revenue.

However, when the data in consideration is of high dimensionality, data mining algorithms incur an exorbitant cost of computation. This is where feature selection algorithms come into picture, as they remove irrelevant features, cutting down the processing time drastically.

Furthermore, Data Mining [5, 24] is a process of identifying hidden patterns of data by analyzing interlinks between features and its class labels or with various other perspectives. But as mentioned above, it can not handle High Dimensional data and thus needs special pre-processing step known as feature selection.

It makes this field more prominent for researchers to identify novel methods for feature solution that will be able to retrieving features that are not irrelevant, redundant this helps to reduce the noise in the overall dataset. For Example: If a book-selling company want to search cities where the book can be sold for profit, the predictive algorithms of Data Mining can be able to identify it through analysis of past books sold in that particular area. In that case information related to, where publisher of that book is residing now, will be less important than the genre of the book. However, in such cases a feature selection method should drop out the less important feature before mining the data to obtain hidden patterns in the dataset.

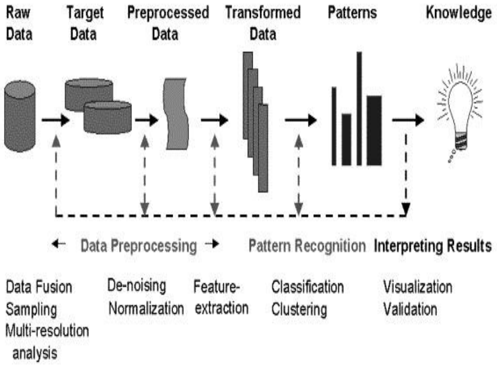

Figure 1. Knowledge discovery process.

There are numerous steps in knowledge discovery, which are shown in the Figure 1. Pre-processing, and pattern recognition are the important parts of the knowledge discovery process. Furthermore, in patterns recognition step, Clustering, association rule mining and classification [5, 6] are important ways of interpreting analysis results of data mining. Among this, classification methods used for building a predictive model from datasets that is the main method which we utilized for comparative analysis in this paper.

Most recently, various application utilized feature selection methods in the diverse field of study such as genomics, microarray, sensor , bioinformatics [3, 18, 27], Pattern Recognition [1], Image processing, big data and medical data analysis [11, 23].

Finding hidden patterns from complex data, while maintaining less complex and computationally expensive algorithmic process it the main motive of the Feature Selection method [3, 19]. Here, complex data means highly correlated w.r.t class label or other features in the dataset. Due to this high coupling, it is difficult to distinguish among the features. Thus, makes the task of prediction more complex and necessities appropriate method for preprocessing before the actual mining of the data.

To alleviate the above challenges, feature selection and feature reduction will be a possible solution. But, in this paper, we are going to tackle these challenges through the feature selection methods.

The intend of feature selection is similar to dimensionality reduction.More specifically, in comparison with the existing and most established technique of feature extraction techniques, like Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA) [23], that maintain the underlying properties of the original features [16].

To prove the worthiness of Maximum Spanning Tree based Feature Selcetion (MST-FS) method diverse benchmark datasets are analyzed, i.e., Mfeat-karhunen,Madelelon, Internet-ads, Ionosphere and Central Nervous System (CNS) [17].

In this paper, we proposed redundancy elimination techniques for microarray as well as high dimensional data classification. Here, three kinds of classifiers, namely IB1 instance-based, Naive Bayesian classifier and Decision Tree C 4.5 are used. These classifiers are utilized to classify original as well as reduced data individually, and a comparative analysis has been done and explained in section 6.3.

2. Related Works

The "Curse of Dimensionality" has been proven to the biggest challenge to mine the high dimensional data [3, 5, 19]. It's not only computationally expensive and time consuming, but also decreases the accuracy of the predictive techniques. To handle these issues various basic feature selection methods were proposed [6, 8, 13, 15], as mentioned in prestigious survey papers [17, 19, 21, 25] etc., in literature.

It has been an important and lively research area in the applications of machine learning and data mining with supervised, unsupervised and semi-supervised learning. Recently, many feature selection techniques have been devised by researchers to solve the concerns and challenges of high dimensionality. In most of the cases, the basis of feature selection is how itjoin the optimal feature subset search along with the creation of learning models. There are mainly three categories of feature selection methods, namely filter [1, 10, 13, 15, 20, 24], wrapper [7, 28], and embedded method [7, 11]. Filter methods were proven to be best from earlier methods which are independent from learning algorithms [8].

It assesses the significance of each attribute by exploiting just the inherent characteristics of the intended data set. In filter methods, a merit is intended for each attribute based on various evaluating standards. After this, those attributes that are qualifying greater than condition from some threshold value are regarded as the selection and the rest attributes are leaving as unselected. Embedded methods [7, 11], act as algorithm dependent methods and referred as a training part.

However, Wrapper methods [7, 15, 28] are working along with the classifier. In this model, a search technique in the whole feature set is defined. Then a variety of feature subsets are constructed and estimated by cross validation through a particular learning algorithm. The hybrid approach of filter and wrapper methods are known as embedded method [7].

Till now these approaches given better results in comparison to pure filter and wrapper methods.

It gives comparatively the best result with a specific learning algorithm while imposing equal complexity as filter methods. In general, though the embedded and wrapper techniques outperform filter techniques in term of accuracy, but filter techniques are mostly accepted for feature selection. Due to independency from learning technique, easy and quick computational cost, the filter based feature selection is suitable for high dimensional data. Feature selected by filter based techniques are more flexible for various learning tasks. Due to the above reason, the filter techniques for feature selection have greater emphasis.

Several filter methods have been introduced for feature selection. Space and ranking search are two classes of filter methods described by Guyon and Elisseeff [7].

In this paper feature selection method was considered as ranking problem. The rank of the features can be generated from some weighting functions. These ranking methods were able to set a threshold hold for selecting appropriate features. A lot of papers used this weighting technique with the Filter based methods [24]. For ex: Pearson Correlation Coefficient (PCC) [6], measure the similarity between two features based on their vector values. Instead of its simplicity, it can applied only on numerical or continuous variables value. To cope up such issue, a rank correlation coefficient was developed by Kendall

[13] that is used as feature weighting. An additional downside of PCC, merely linear dependency among the features and class/target feature is measurable. In addition, the information theory approach based techniques (such as Mutual Information (MI) [13] and Information Gain (IG)[15]) were devised to weight features to mitigate the aforementioned problem. But these methods fails when high dimensional data analyzed by them. Therefore, Kira and Rendell [13] demonstrated a further statistical feature weighting algorithm called as Relief. It considered the significance of each feature by calculating the association between an instance and its nearest neighbour from dissimilar class. Nearest hit and nearest miss are the parameters where nearest hit is the instances from the same class and nearest miss is from different class label. Due to limited applications of it only for two classes, variants' of its method ReliefF have been proposed [28]. The limitation of this technique is that it is applicable for only two classes within a data set. The new variant of Relief technique is ReliefF which is proposed in [28].

An improvement in text classification method Uguz [24] introduces a hybrid of Information Gain for ranking, the words and using Genetic Algorithm (GA) and Principal Component Analysis (PCA) applied feature selection. The main drawback of the paper is the negotiation of correlation words in documents.More recently, a Constraint Score technique was proposed for feature ranking [4] based on specifying the pair-wise constraints for feature selection with data instances.More details of the various feature weighting approaches are discussed in [3, 5, 17, 21, 23, 24].

We have validated our proposed work while considering the accuracy rate of the three classifiers i.e., Naive Bays, C4.5, and IB. The Naive Bayes classifier depends on the Bayesian theory and used for high dimensional data. It is based on probabilistic method that employs for classification by multiplying the individual probabilities of every pair of feature-value. For the data, with the hypothesis of independent variable between features, Bayes results are excellent. Decision Tree classifier i.e., Over ID3, C4.5 is an enhanced version which accounts for missing values, continuous attribute value, pruning of decision trees, rule derivation, and so on [24]. Instance-Based classifier (IB1) [24] is based on the idea of nearest-neighbour. It classifies instances by utilizing the class of the closest vectors in the data set via Euclidian distance metrics.

IB1 classifier is a lazy learner and the simplest among the algorithms used in this paper.

Yu and Liu [27], the Fisher Score method is utilized to access the feature weighting.That characterizes the differentiates ability of individual feature founded on the fisher criterion. The procedure for assigning weight in Fisher Score is a supervised feature weighting method. Variance [27] is the simplest feature weighting in an unsupervised manner. The interested user can find the detailed information regarding relevance of features in the Kohavi and Langley (1997) [16].

Let R_{fi} denote the f^th^ feature of the i^{th} instance, where f=1, 2,.....,n; xi, i=1, 2,....., m; represents the number of features and instances, respectively. Let the $\mu_f$depict the mean of the f^th^ feature, as given by Equation 1. Further, then the variance score of the f^th^ feature is V_f, which should be maximized, and calculated by Equation (2) [2]:

$$\mu_f = \frac{\sum_{i=1}^{m} R_{fi}}{m} \tag{1}$$

$$V_f = \frac{1}{m}\sum_{i=1}^{m}(R_{fi} - \mu_f)^2 \tag{2}$$

While using of variance, another unsupervised feature weighting method, i.e., He *et al.* [10] proposed Laplacian Score. Compared with the above unsupervised methods such as Variance [2], it not only prefers the variance of each feature which is having more representative power, but also prefers the locality preserving ability into account. The Laplacian score of the f^{th} features L_f, which should be minimized and calculated as follows [10]:

$$L_f = \frac{\sum_{i,j}\left(R_{fi} - R_{fj}\right)^2 S_{ij}}{\sum_i (R_{fi} - \mu_f)^2 D_{ii}} \tag{3}$$

Where D is the diagonal matrix with

$$D_{ij} = \sum_j S_{ij} \tag{4}$$

And S_{ij} is defined by the neighbourhood relationship between instances x_i, i=1,2,....., m. which is define as follows:

$$S_{ij} = \begin{cases} e^{\frac{-\|xi-xj\|^2}{t}} & If \quad x_i \quad neighbors \quad of \quad x_j \\ 0 & Otherwise \end{cases} \tag{5}$$

Where, t is constant to be set by the user, and x_i and x_j are neighbour's means, such that x_j is among k nearest neighbours of x_i.

Furthermore, Information Gain based feature selection method is known as the Symmetric Uncertainty (SU). Various approaches used pure SU [1] as well as variants of it [11]. Ali and Shaahzad [1] used a hybrid of SU and Ant colony optimization method. It used Ripper and K-nearest algorithms. The main drawback of this approach is computationally expensive.

The mutual information between each variable and the class is:

$$MI(x_i, y) = \int_{x_i}\int_y p(x_i, y) log \frac{p(x_i, y)}{p(x_i)\,p(y)} dxdy \tag{6}$$

In Equation (6), the probability density of x_i and y is represented by $p(x_i)$ and $p(y)$, and $p(x_i, y)$ is the joint density. $MI(x_i, y)$ can be calculated by using the density of variable $x_{i,}$ and the density of class label y. The problem with the Equation (6) is that the densities $p(x_i)$, $p(y)$ and $p(x_i, y)$ are not known and hard to calculate from data. It might be good for nominal and discrete features as it uses sum operation instead of integration.

$$MI\left(x_{i,}y\right)=\sum_{x_i}\sum_{y}P\left(X=x_i,Y=y\right)log\frac{P\left(X=x_i,Y=y\right)}{P\left(X=x_i\right)P\left(Y=y\right)} \quad (7)$$

Besides this, the probability can be calculated from the frequency counts. But the estimation becomes harder with the large number of classes and huge numbers of feature values. While the estimation may be unreliable if the number of observations is not sufficient.

Continuous dataset can be handled by discrediting the variable or approximating their densities with a non-parametric technique [11, 20]. This was the simplest and easiest methods for feature's ranking.

Now we are discussing feature selection techniques used so far for analyzing data. Hall *et al.* [9] introduced an approach that considered only optimal feature subsets. More specifically, these features are highly correlated with a class label, however, less correlated with each others. Another work was proposed by Yu *et al.* [27, 28], as Fast Correlation based Filter (FCBF) which utilizes widely used Filter based algorithm. This algorithm used uncertainty for High Dimensional (HD) data and removed irrelevant features.

This algorithm uses C4.5 classifier. One of the limitations of this algorithm is that it still does not able to find the pairwise correlation between attributes. A GA based feature selection method is given by Valliammal and Subbarayan [25].

The proposed framework in this paper, selects the highly relevant feature subset through the maximum spanning tree and fisher score principle. Since the merits of each feature reflect its significance and the redundancy reflect through correlations among the attributes. We assume the highly correlated feature may represent the same information with respect to the class labels; therefore we have combined all such features into one feature. In Comparison with various other techniques, the algorithm MST-FS has shown some of the good estimations. First, the significance, representative ability and correlation of features are not limited to any particular measurement in MST-FS.

Consequently, the various prevailing feature selection and correlation techniques are embedded into the proposed framework. With this proposed framework, one can incorporate any ranking algorithm with any maximum spanning tree based methods.

3. Challenge: Feature Selection in Microarray Data

In this section, we will provide the challenges of feature selection imposed by micro-array data sets. The recent advent of microarray gene expression [3, 23, 25] data have made it possible to measure and analyse the high dimensional data. In most cases the unrefined data have noise or missing values. Again, due to technology modernizations, the generation of enormous size of the data set increases the difficulty for the researchers. For example, a number of genes are not significant to the related class labels which are irrelevant for bigger data size. Therefore, before applying a data mining techniques on micro-array data it must be a good practice to pre-process with some efficient techniques. There are many issues of feature selection for micro-array data. These are:

- How to enable data understanding.
- How could we reduce the storage requirement and computational time.
- How to defy the curse of dimensionality to increase the model performance with respect to the learning model.

One of its general applications in the micro-array analysis domain is to discover a set of drug leads, for this, a measurement is applied to find the genes that have maximum discriminative power between normal and diseased patient; this kind of genes may help to code for drug able proteins. Validation drug consists of a hard concentration problem in biology that could not come under the realm of the machine learning. For example, given a set for previously classified samples having different types of cancerous class, such as ALL and AML.A classifier will assign one of the above classes to a newly unlabeled sample. In medical diagnosis, classification and its pre-processing step, i.e., Feature Selection is very crucial. The described approach will save not only the costs associated with clinical testing,but increases the accuracy of the diagnosis also.

Due to lage number of features in comparision of less number of instances make high dimensional task more challenging. For example: In case of dealing with disease detection where genes informations are already known. Out of various genes, some are noisy and some are highly correlated. Therefore, it is effective to extract the representative genes from the actual data. Before learning process, identification of relevant genes is very important.

4. Feature Selection Techniques

Feature selection is a technique by which the most important features are selected from a given set of features; it ensures the alleviation of redundant and irrelevant features. It is also used as a pre-processing step [6, 24] in data mining and machine learning algorithms. Thus, pre-processing mainly change the layout of of data by eliminating irrelevant features. The classifiers performance and running time affected greatly due to the underlying presence of irrelevant and redundant attributes [6, 8]. In literature various methods for feature selection are present as discussed in section 2. Most of them are not considering redundant features that increase the computational cost. Therefore, the main motive in this paper is to handle redundant features [4].

4.1. Fisher Score

Fisher Score [5, 8] finds a good separator for classify multi-class data. The MST-FS method implemented in this paper is based on the merit of the fisher Score [5, 8]. In case of HD Data it gives better results. More specifically fisher score mainly provides a way of decreasing the dimension of the data sets. The fisher method obtains sets of attribute which are having maximum distance between inter class feature.

Moreover, the intra-class distance should be as minimum as possible.

For considerations, a set of 'm' features in the data sets X ϵ Rm×n projected to Z ϵ Rf×n , where 'f' is the set of top quality features. Here, the formula for finding Fisher Score is given as Equation (8):

$$F(Z) = tr\left\{\left(\hat{K}b\right)\left(\hat{K}t + \gamma I\right)^{-1}\right\} \tag{8}$$

In this equation, is a total scatter matrix, Ḱb is a scatter matrix between class and γ is a regularization terms.

These two matrixes are mathematically presented by the Equations (9) and (10):

$$\hat{K}b = \sum_{k=1}^{c} n_k \left(\mu_k - \mu\right)\left(\mu_k - \mu\right)^T \tag{9}$$

$$\hat{K}t = \sum_{k-1}^{c}\left(z_k - \mu\right)\left(z_k - \mu\right)^T \tag{10}$$

μ_k is a mean vector of k^{th} class andn$_k$is the size of k^{th} class and μ is the overall mean vector of feature. Due to having a singular matrix problem in $\acute{K}b$, we are adding γI term for avoiding negative factors. Another challenge which we are alleviating is combinatorial problem, by utilizing a heuristic strategy [5, 10] because most of the feature selection methods are largely suffered by it. In a heuristic manner we are able to find the Fisher Score independently. The formula for calculating the Fisher Score is given as Equation (11):

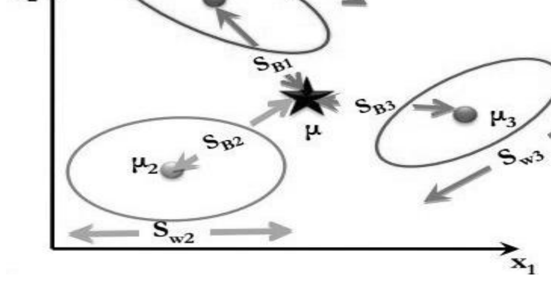

Figure 2. Random variable depiction in the spanned space system.

$$FS\left(X^j\right) = \frac{\sum_{k=1}^{c} n_k \left(\mu_k^j - \mu^j\right)^2}{\left(\sigma^j\right)^2} \tag{11}$$

where, n_k is size of k^{th} class, μ^j is the mean of whole j^{th} feature, $\mu^j{}_k$ is the mean of k^{th} class of j^{th} feature, σ^j is the standard deviation of whole j^{th} feature.

The Figure 2 shows that how a feature is co-related to its target class.

4.2. Correlation Analysis Measure: A Methodology

To measure the some of the dependency between two feature correlations are used. There are few methods such that Spearman and Pearson method are used to find the quantitative correlation. In these, correlation is measured between attribute and class, between two attributes. For example, Information Gain (IG), gain index [22], symmetric uncertainty [11] and many distance based methods. SU is derived by Written *et al*. [26], which is a new variant of information gain, while normalizing the IG of it. Due to its limitation to handle only nominal or categorical data SU is not always prefer for all kinds of data sets.

To eliminate the redundant features, we calculate pair wise measure of SU between attributes w.r.t. class. As mentioned in section 4.2., to measure correlation between class and attributes we apply Equation (11). In addition, we apply Equation (12) to calculate the correlation between two features. By considering the assumption that feature and a class label as one features, Face Centered Cubic (FCC) is calculated for finding the correlation.

- *Definition1*. Let us assume two features (|f|=2) F_1, F_2 and target class (C=2). Also assume that target class strongly correlated to both features. Therefore by taking consideration of Fisher Score (FS), we are measuring the correlation between the features f_i and f_j, FCC(f_i, f_j :C) that is defined as:

$$FCC\left(f_i, f_j : C\right) = \frac{\sum_{i,j}\left(f_i - f_j^{'}\right)\left(f_j - f_j^{'}\right)}{\sqrt{\sum_i\left(f_i - f_j^{'}\right)^2}\sqrt{\sum_j\left(f_j - f_j^{'}\right)^2}} \tag{12}$$

wheref'$_i$ andf'$_j$ are the mean of whole F_i feature and F_j feature, and $\sigma = \sqrt{\sum_i\left(f_i - f_j^{'}\right)^2}$ is the variance of whole F_i feature and F_j features and is the mean of k^{th} class of F_i feature and F_j feature.

- *Remark 1*. Variance of F_i and F_j features is higher than variance of F_i or F_j. Using such remarks, we can say FCC(f_i, f_j :C) is often less than FS(f_i, C) or FS(f_j, C).
- *Definition2*. Let us assume two features (|f|=2) x_1, x_2 and target class(C=2) where target class strongly correlated to both features, i.e. correlation between x_1, x_2 is often depend upon the target class. Therefore, we consider a correlation measure (Q(x_1,

x_2: C)) along with target class C whose equation is as follows:

$$Q(f_i, f_j : C) = \frac{\frac{FS(f_i, C)}{FS(f_j, C)} - FCC(f_i, f_j : C)}{FS(f_i, C) + FS(f_j, C)} \qquad (13)$$

- *Remark 2.* Here, FCC(f_i, f_j :C) = FCC(f_j, f_i:C), because variance of(f_i, f_j) =variance of (f_j, f_i).

4.3. Maximum Spanning Tree

Let V(G) and E(G) be the vertex and edge sets of a graph G respectively. A spanning tree T of G is a connected tree pertaining V(T)=V(G) and E(T) subset of E(G). Means, a spanning tree is a sub-graph of a graph G that contains all vertices without any cycle. In the case of MST, the total sum of the weight is larger than all other spanning trees. This is a well-known solution in various applications like Wireless Sensor Network (WSN), Traffic management. In literature there are three techniques to find the spanning tree. These are Prims, Kruskal and Boruvka algorithm [12]. Moreover, all these techniques are greedy algorithm in nature and run in polynomial time. Here is this work; we are applying Kruskal technique which was developed in 1956, and are used in many applications. It takes O(E log V) time complexity to generate the Maximum Spanning Tree (MST).

In the process of elimination of features, the notion of complete graph was desired to understand by us. In the complete graph each and every node is connected to other remaining nodes. By considering each feature represents a node, we will create a complete graph. The weights are assigned to each vertex and edge through the value of Fisher Score and correlation values, respectively as stated in Equations (11) and (13). Edges between two features are assigned a weight that is Fisher Score of two different distinct features. The value of V_i that is a set of nodes{ (FS(f_1, C), FS(f_2, C), FS(f_3, C) ....FS(f_m, C) } and E_{ij} is the set of edges {Q (F_i, F_j :C)} that is given by Equation (13).

Let's we have a data set D containing feature sets f={ F_1, F_2, F_3,….F_f} and having labeled data C for each instance. Our objective is to select a feature subset F′= {F1, F2, F3,….Fm} such that each feature is having greater information. For this, we apply a formula of equation (11), to find the merit of each feature. Then, filtered those feature having greater value while comparing with the pre-set values β. Let consider, V_i={F'} having the Fisher Score of each feature sets i.e., {FS(f_1,C), FS(f_2,C), FS (f_3, C)….. FS (f_m, C)} and E_{ij} = {Q(f_i, f_j :C)} for each i, j = 0 to m. In addition to this, those edges are removed from the MST whose Q weights are smaller than both Fisher values of features.

Such that FS(f_i, C) > Q (f_i, f_j: C) < FS (f_j, C). Then Q (f_i, f_j: C) edge will have to be discarded, otherwise not.

After this step, we get sets of forest. Then from each forest, the most relevant features are selected while utilizing their Fisher Score value.

5. MST-FS: Maximum Spanning Tree (MST) based Feature Selection

In this proposed approach, we utilize the MST along with Fisher Score [5] in to obtain the effective features. As we have already described Fisher Score in section 4.1. to find the merit of each feature. After finding its values, those features are removed whose merit is less than pre-set value (β). In such way the proposed approach produces considerably better results. We have designed our method MST-FS in four steps:

Algorithm 1: proposed MST-FS procedure

```
Input: D ($F_0$, $F_1$……$F_n$, C);
Output: F ($F_0$, $F_1$…..$F_f$);
• Step 1: Removal of Irrelevant Feature.
      A: Score = FS ($f_i$, C)
      B: if (Score>Threshold)
            Selected feature)
• Step 2: MSTCreation
      A: (G) = Null;
      B: Co-Relation = Q($f_i$, $f_j$:C);
      C: to draw an edge $f_i$, $f_j$ to Complete Graph (G)
      With Fisher Score as the weight of the
      analogousnodes;
      D: Apply KrusKal's (G) to Construct MST.
• Step 3: Tree Screening and Cluster Realization
      For each node($E_i$) {
            Select the node with max(FS($f_i$, C))
            Remove Edge (Q($f_i$, $f_j$: C)) and
      associatednode ;
            F'=$f_i$;
      reiterate step 3;
• Step 4: Select Feature from Each Cluster and get a set of
  feature: F′ = {($F_0$, $F_1$…..$F_f$)};
```

A pictorial representation of the overall process of the proposed method is given as a flow diagram in Figure 4. In this, our main objective of removing redundant features comes in last but two steps of the Figure 4.

The process starts with the calculation of Fisher Score of each feature. After that, it comparisons with β is performed for removing irrelevant features. Next, we make the complete graph which, considering each feature as a node. Then we apply Kruskal technique to obtain the MST.

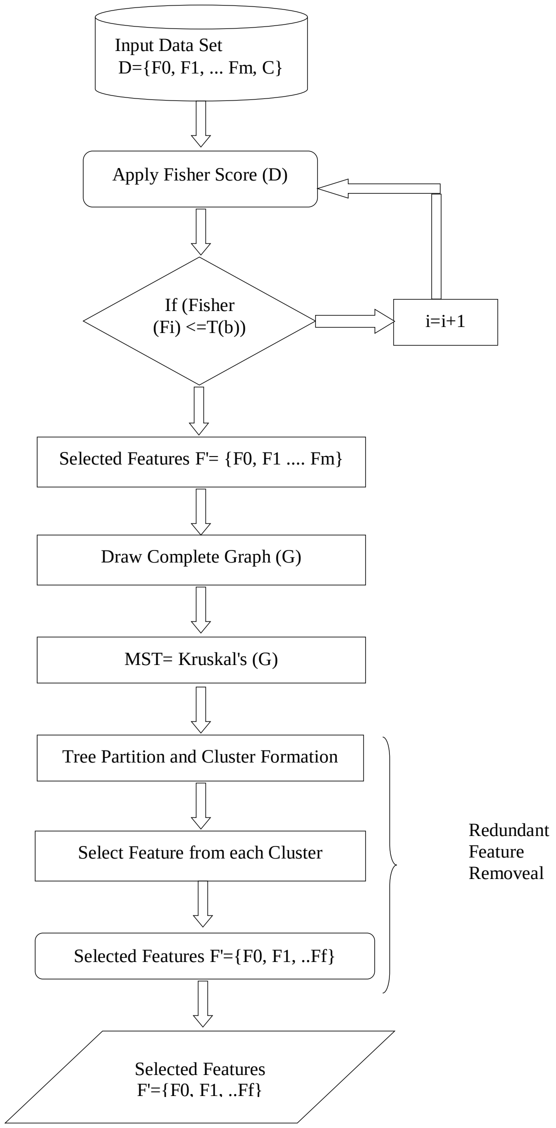

Figure 4. MST-FS algorithm's process flow diagram.

Kruskal technique presents a set of forest, in which each individual forest considered as a cluster. Last but not least, a cluster is constructed of each forest to find the set of selected features. For the justification of MST-FS technique, we have performed the classification task to achieve the accuracy provided by decision tree, Naive Bayes and Instance based classifier. While also considering three existing feature selection algorithm like CFS [9], Fisher Score [5, 8] and ConsSF [7], used to classify 5 datasets. The step by step functionality of MST-FS algorithm is described in the section 5.

6. Performance Evaluation

In this section, we apply the classification models IB1, C4.5, and naive bayes as well as other benchmark techinques to evaluate with the existing feature selection approaches, i.e., fisher score, CFS, and ConsSF on various data sets, in comparision with our proposed MST-FS approach.

The exploited data sets are demostrated in section 6.1. Subsequently, we describe our experimental methodology and model selection procedure. Section 6.2. presents the results we have obtained. The analysis of the different parameteres are discussed in 6.3. Lastly, we analyze running times of our proposed method implementation.

6.1. Datasets

In this paper, a proposed feature selection approach is applied on five high dimensional data sets that are derived from UCI Machine Learning Repository [18].

The descriptions of these data sets are given in Table 1 below. Although all data sets are taken from UCI repository [18], but the source of each data set is given as a footnote.

- *Ionosphere.* This dataset have 2 class labels, namely good and bad radar manully classified 351 data points with 34 attributes.
- *Mfeat-Karhunen.* Donated by Robert P.W. Duin.This dataset consists of features of handwritten numerals (0-9) extracted from a collection of Dutch utility maps. 200 patterns per class (for a total of 2000 pattern) have been digitized in binary image.
- *Medalon.* The Medalon data set have 500 attributes and 2000 instances; it is a two-class data having continuous input variables. It was a part of feature selection challenge of Neural Information Processing System (NIPS) in 2013.
- *Internet_ads.* The dataset represents a set of possible advertisements on Internet pages. There are two class labels: advertisement ("ad") and not advertisement ("nonad"). It contains 3279 examples, in which 458 are from class "ads" and 2821 are "nonad" labelled class. It is having 1558 features in which 3 are continuous and remaining are binary values.

- *CNS.* The data set contains 60 patient samples, 21 are survivors (labelled as “Class1”) and 39 are failures (labelled as “Class0”). There are 7129 genes in the dataset.

We have described the properties of data sets in the above section 6.1. Now the experiments are conducted on these 5 high dimensional data sets, in order to examine the performance of the proposed method for feature selection and classification task under the constraints of high dimensionality. Pre-processing feature selection method, fisher score, and an MST algorithm Krushkal's and C4.5, NB and IB1 are implemented in the Netbeans IDE using java programming language. A 10-fold cross validation technique is used at the time of classification training and testing stage. All implementations are carried out on a machine with Intel (R) Core (TM) i7, 3.40GHz CPU, 12 GB of RAM, 1 TB HDD, and with Window 7 operating system.

6.2. Model Selection and Experimental Methodology

The training and test subsets were generated using 10-fold cross validation and the average accuracy was computed. Feature selection and classification were then performed on the training set and the classification performance was finally calculated from the test data set after the experiment. The MST-FS method first applied on the training data to find the representative features from the actual data. Afer this, the same features are filter from the test data to check the performance. Here, all the classifiers are applied on the actual data as well as on selected data and corresponding results are shown in the Tables 5, 6,and 7. We had performed various steps to decide the value of threshold for each data set individually.

The algorithm initially removes the irrelevant features in the dataset from Fisher score followed by removal of redundant features by MST algorithm. Here, MST algorithm is utilized for identifying most correlated features of the dataset based on the property of MST, as shown in the Figure 3 Weka API has been exploited for the implementation purpose.

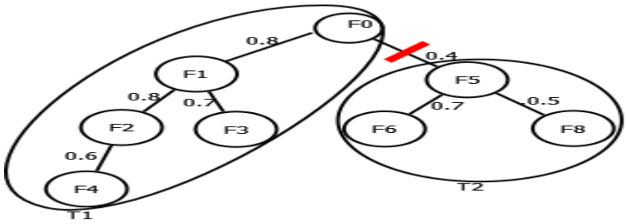

Figure 3. Representation of MST.

6.3. Result and Analysis

We have included microarray datasets as well as some other high dimensional data from multi-disciplinary areas in the Data Mining fields. Recently, we have found that Microarray classification is a typical area where feature selection methods [17, 24] are applied. So, we will check the performance of our proposed method, MST-FS, for high dimensional datasets classification problems. The proposed method has been experimented with five datasets, as shown in Table 1.

Table 1. Characteristics of data sets.

S. No.	Data set	#Feature	#Instances	#Class
1	Ionosphere[1]	35	351	2
2	Mfeat-karhunen[2]	65	2000	10
3	Madelon[3]	501	824	2
4	Internet_ads[4]	1559	3279	2
5	CNS[5]	7130	60	2

Table 2. Validation result for IB1 with other features selected by three feature selection methods and proposed method.

S.No	Data set	MST-FS	Fisher Score	CFS	ConsSF	Full set
1	Ionosphere	90.37	90.25	89.26	88.17	86.53
2	Mfeat-karhunen	96.87	96.42	95.97	93.26	95.30
3	Madelon	72.38	64.41	71.78	71.83	51.67
4	Internet ads	95.28	96.67	95.02	97.73	96.14
5	CNS	82.16	81.57	79.95	68.84	57.92
6	Average	87.01	84.99	86.18	83.55	77.37

Table 3. Validation result for C4.5 with other features selected by three feature selection methods and proposed method.

S.No	Data set	MST-FS	Fisher Score	CFS	ConsSF	Full Set
1	Ionosphere	94.12	92.98	91.37	87.69	90.27
2	Mfeat-karhunen	83.78	82.69	81.96	83.93	81.43
3	Madelon	73.71	65.55	71.10	67.84	57.35
4	Internet ads	97.92	96.20	97.05	96.27	97.15
5	CNS	65.34	57.12	65.25	79.41	58.87
6	Average	82.97	78.90	81.34	83.02	77.01

Table 4. Validation result for NB with other features selected by three feature selection methods and proposed method.

S.No	Data set	MST-FS	Fisher Score	CFS	ConsSF	Full Set
1	Ionosphere	86.57	83.69	92.56	87.58	83.15
2	Mfeat-karhunen	94.45	93.56	94.32	85.92	94.57
3	Madelon	59.82	62.15	61.17	61.17	57.28
4	Internet ads	95.91	95.13	95.27	95.24	96.73
5	CNS	80.81	78.84	75.51	57.19	62.18
6	Average	83.51	82.67	83.76	77.42	78.75

MST-FS provides less computationally expensive result than mentioned feature selection algorithm.

Tables 2, 3, and 4 gives the predictive performance on selected data by three existing feature selection methods along with full datasets, as well as selected data set by feature selection approach through various classifiers. In case of identical accuracy, the results with fewer variables are considered best.

From Tables 2, 3 and 4, we analysed that the good accuracy is achieved with the proposed MST-FS method in mostly all of the cases. For example, for Ionosphere data, the best predictive accuracy is obtained by IB1 and C4.5, but the Naive Bayesian classifier has achieved better accuracy result with conventional techniques Correlation-based Feature Selection(CFS). Figures 5, 6, 7, and 8, we have analysed that the performance of MST-FS is better than Fisher Score, CFS and Consistency based Fetaure Selection (ConsSF) techniques. Therefore, we are able to achieve the stated objective with MST-FS techniques. From the results, we havealso analysed that in all cases MST-FS did not perform better, as we can see in Table 5 for Madelon.WhereasInternet_ads data MST-FS gives the minimum number of features as compared to other feature selection techniques.We

[1] https://archive.ics.uci.edu/ml/datasets/Ionosphere

[2] https://archive.ics.uci.edu/ml/datasets/Multiple+Features

[3] http://archive.ics.uci.edu/ml/datasets/Madelon

[4] http://archive.ics.uci.edu/ml/datasets/Internet+Advertisements

[5] http://www-genome.wi.mit.edu/mpr/CNS/

can say proposed approach work efficiently, but not always.

Table 5 represents the number of features selected after applying feature selection approach on the above five data sets. We have analysed that the feature selected by ConsSF algorithm gives the best results for Ionosphere and Mfeat-karhunen data as compared to Fisher Score, CFS, as well as proposed approach MST-FS. Due to the scarcity of data, ConsSF work fine on such data. But, the proposed approach works effectively on Madelon and Internet_ads data sets, due to its underlying structure. We found that MST-FS method's performance is enhanced than CFS technique for CNS data sets.

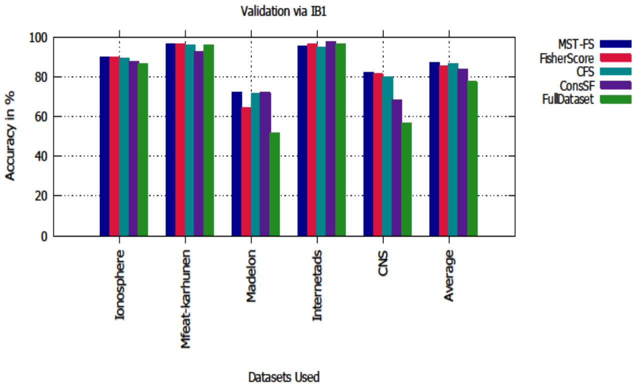

Figure 5. IB1 classification accuracy on 5 data sets with 4 filter based approach.

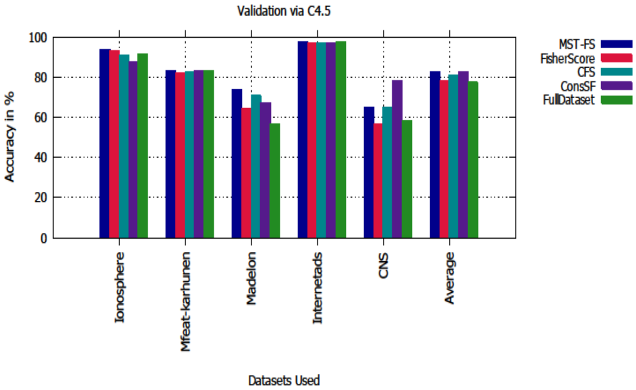

Figure 6. C4.5 classification accuracy on 5 data sets with 4 filter based approach.

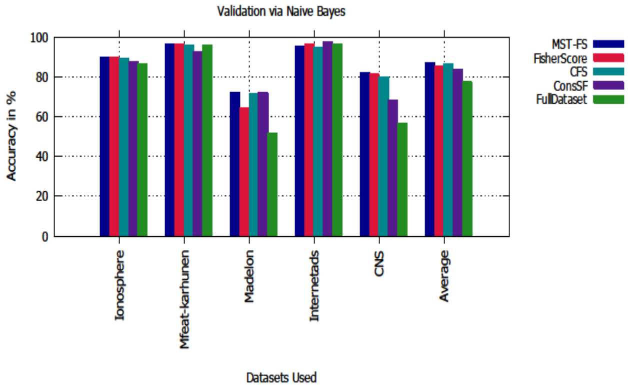

Figure 7. NB classification accuracy on 5 data sets with 4 filter based approach.

Table 5. A comparison of number of features/attributes selected by proposed methods and existing methods.

S.No	Data set	MST-FS	Fisher Score	CFS	ReliefF	ConsSF
1	Ionosphere	9	19	14	10	8
2	Mfeat-karhunen	38	43	56	17	9
3	Madelon	6	50	10	13	14
4	Internet_ads	30	90	31	35	37
5	CNS	31	37	30	45	36

7. Conclusions

In this paper, we proposed one efficient supervised algorithm, called MST-FS for feature extraction of microarray data. It gives better results in most of the cases. Furthermore, it is able to efficiently determine the redundant features that improve the accuracy of classifiers. This feature selection method can also be used to decrease the dimensionality of a feature space comprising of a huge number of genes.It also removes irrelevant and redundant features and thus improving the performance of classifiers. Maximum spanning tree based approach is proved to be prominent for finding the pair wise correlation between features.

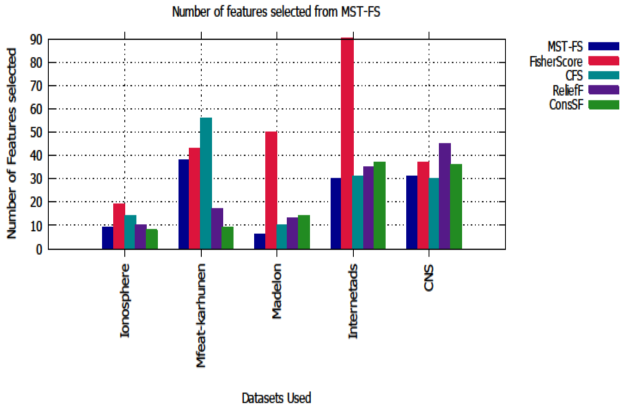

Figure 8. Result in terms of number of selected features by proposed MST-FS and three filter based approach suh as Fisher Score, CFS. As the MST-FS select the minimum number of features in most of the cased.

We compared our method with some of the existing filter and wrapper methods, and on different well known datasets. The empirical results proved that MST-FS works equally well onto both types of feature selection methods.It selects less number of features, reduces time for classification and provides better classification accuracy. So, the proposed method can be seen towards the first step for utilizing Maximum Spanning tree based method for feature selection technique. Also, it can be ensemble with wrapper based or embedded based feature selection techniques to minimize the computational cost.

Reference

[1] Ali S. and Shaahzad W., "Feature Subset Selection Method Based on Symmetric Uncertainty and Ant Colony Optimization," *in*

Proceedings of International Conference on Emerging Technologies, Islamabad, pp. 1-6, 2012.

[2] Bishop C., *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[3] Cabrera W., Oronez C., Matusevich D., and Baladandayuthapani V., "Bayesian Variable Selection for Linear Regression in High Dimensional Microarray Data," *in Proceedings of the 7th International Workshop on Data and Text Mining in Biomedical Infortics*, San Francisco, pp. 17-18, 2013.

[4] Dash M., Liu H., and Motoda H., "Consistency Based Feature Selection," *in Proceedings 4th Pacific Asia Confrence Knowledge Discovery and Data Mining*, Kyoto, pp. 98-109, 2000.

[5] Duda R., Hart P., and Stork D., *Pattern Classification*, Wiley-Interscience Publication, 2001.

[6] Forman G., "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289-1305, 2003.

[7] Guyon I. and Elisseeff A., "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.

[8] Gu Q., Zhenhui L., and Han J., "Generalized Fisher Score for Feature Selection," *in Proceedings of 27th Conference on Uncertanity in Artifical Intelligence*, Barcelona, pp. 266-273, 2011.

[9] Hall M., "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning," *in Proceedings of 17th International Confernce on Machine Learning*, San Francisco, pp. 359-366, 2000.

[10] He X., Cai D., and Niyogi P., "Laplacian Score for Feature Selection," *in Proceedings of the 18th International Conference on Neural Information Processing Systems*, Vancouver, pp. 507-514, 2005.

[11] Jiang B., Ding X., Ma L., He Y., Wang T., and Xie W., "A Hybrid Feature Selection Algorithm:Combination of Symmetrical Uncertainty and Genetic Algorithms," *in Proceedings of the 2nd International Symposium on Optimization and Systems Biology*, Lijiang, pp. 152-157, 2008.

[12] Karumanchi N., *Data Structure and Algorithm Made Easy*, CareerMonk Plublications, 2011.

[13] Kira K. and Rendell L., "The Feature Selection Problem: Traditional Methods and a New Algorithm," *in Proceedings of 10th National Conference Artificial Intelligence*, San Jose, pp. 129-134, 1992.

[14] Kononenko I., "Estimating Attributes: Analysis and Extension of Relief," *in Proceedings of European Conference of Machine Learning*, Catania, pp. 171-182, 1994.

[15] Kononenko I., "On Biases in Estimating Multi-Valued Attributes," *in Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, pp. 1034-1040, 1995.

[16] Kohavi R., LangleyP., and Yun Y., "The Utility of Feature Weighting in Nearest-Neighbor Algorithms," *in Proceedings of 9th European Conference on Machine Learning*, Prague, pp. 85-92, 1997.

[17] Liu H., "Advancing Feature Selection Research - ASU Feature Selection Repository," Arizona: Technical Report, 2011.

[18] Lichman M., "UCI Machine Learning Repository,"http://archive.ics.uci.edu/ml,University of California, Last Visited, 2012.

[19] Murugappan I. and Vasudev M., "PCFA: Mining of Projected Clusters in High Dimensional Data Using Modified FCM Algorithm," *The International Arab Journal of Information Technology*, vol. 11, no. 2, pp. 168-177, 2014.

[20] Peng H., Long F., and Ding C., "Feature Selection Based On Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, 2005.

[21] Phiwma N. and Sanguansat P., "An Improved Feature Extraction and Combination of Multiple Classifiers for Query-by-Humming," *The International Arab Journal of Information Technology*, vol. 11, no. 1, pp. 103-110, 2014.

[22] Quinlan J., *C4.5: Program for Machine Learning*, Morgan Kaufmann, 1995.

[23] Sharma A., Paliwal K., Imoto S., and Miyano S., "A Feature Selection Method using Improved Regularized Linear Discriminant Analysis," *Machine Vision and Applications*, vol. 25, no. 3, pp. 775-786, 2013.

[24] Uguz H., "A Two Stage Feature Selection Method for Text Categorization by using Information Gain, Principal Component Analysis and Genetic Algorith," *Knowledge-Based Systems*, vol. 24, no. 7, pp. 1024-1032, 2011.

[25] Valliammal N. and Subbarayan G., "An Optimal Feature Subset Selection using GA for Leaf Classification," *The International Arab Journal of Information Technology*, vol. 11, no. 5, pp. 447-451, 2014.

[26] Witten I., Frank E., and Hall M., *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2011.

[27] Yu L. and Liu H., "Efficient Feature Selection Via Analysis of Relevance and Redundancy," *Journal of Machine Learning Research*, vol. 10, no. 5, pp. 1205-1224, 2004.

[28] Yu L. and Liu H., "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," *in Proceedings 20th International Conference on Machine Learning*, Washington, pp. 856-863, 2003.

Bharat Singh is Senior Research Scholar in Department of Software Engg. Lab at Indian Institute of Information Technology (IIIT-A), Allahabad, India. He completed his masters in computer Science from NIT-Durgapur, India.Currently, he is an active member of ACM, IEEE, IDES and IET. He has published various research papers in International Journals as well as in National/International Conferences into his credit. His research areas are Data Mining, Machine Learning, Feature Engineering and High Dimensional Data.

Om Prakash Vyas is currently working as Professor and Dean of (Research and Development) in Indian Institute of Information Technology-Allahabad (Govt. of India's Center of Excellence in I.T.). Prof. Vyas has done M.Tech.(Computer Science) from IIT Kharagpur and has done Ph.D. work in joint collaboration with Technical University of Kaiserslautern (Germany) and I.I.T. Kharagpur. With more than 25 years of academic experience Prof. Vyas has guided Four Scholars for the successful award of Ph.D. degree and has more than 80 research publications with two books to his credit. His current research interests are High Dimensional Data Mining and Service Oriented Architectures.