Journal of Data Analysis and Information Processing, 2014, 2, 95-105
Published Online November 2014 in SciRes. http://www.scirp.org/journal/jdaip
http://dx.doi.org/10.4236/jdaip.2014.24012

Scientific Research

A Feature Subset Selection Technique for High Dimensional Data Using Symmetric Uncertainty

Bharat Singh, Nidhi Kushwaha, Om Prakash Vyas

Department of Information Technology, Indian Institute of Information Technology, Allahabad, India
Email: bharatbbd1@gmail.com, kushwaha.nidhi12@gmail.com, dropvyas@gmail.com

Received 18 September 2014; revised 20 October 2014; accepted 11 November 2014

Abstract

With the abundance of exceptionally High Dimensional data, feature selection has become an essential element in the Data Mining process. In this paper, we investigate the problem of efficient feature selection for classification on High Dimensional datasets. We present a novel filter based approach for feature selection that sorts out the features based on a score and then we measure the performance of four different Data Mining classification algorithms on the resulting data. In the proposed approach, we partition the sorted feature and search the important feature in forward manner as well as in reversed manner, while starting from first and last feature simultaneously in the sorted list. The proposed approach is highly scalable and effective as it parallelizes over both attribute and tuples simultaneously allowing us to evaluate many of potential features for High Dimensional datasets. The newly proposed framework for feature selection is experimentally shown to be very valuable with real and synthetic High Dimensional datasets which improve the precision of selected features. We have also tested it to measure classification accuracy against various feature selection process.

Keywords

High Dimensional Datasets, Feature Selection, Classification, Predominant Feature

1. Introduction

Data Mining is a multidisciplinary task to find out hidden nuggets of information from data. In recent years, as the technology advances in various fields, the data generated in these fields, have become increasingly larger in both number of instances and number of features in various field. The proliferation of High Dimensional data in

How to cite this paper: Singh, B., Kushwaha, N. and Vyas, O.P. (2014) A Feature Subset Selection Technique for High Dimensional Data Using Symmetric Uncertainty. *Journal of Data Analysis and Information Processing*, **2**, 95-105.
http://dx.doi.org/10.4236/jdaip.2014.24012

various applications poses challenges to Data Mining field. This enormity cause serious problems to many Data Mining and Machine Learning algorithms with respect to scalability and learning performance [1]. Feature selection is an active field of research and development since the 70's, in multidisciplinary field. It includes statistical pattern recognition [2] [3], machine learning [4]-[7], Data Mining [8]-[10] and it is extensively applied to various field such as text categorization [11] [12] image retrieval [13] [14], genomics analysis [7] [15] [16], CRM [17]. Due to these applications, not only the datasets get larger, but different and new kinds of data are also generated, for example, stream data, microarray in proteomics and genomics, social computing and system biology datasets.

Within this, High Dimensional datasets are flattering more and more copious in learning process. Relatively it has made traditional search algorithm too expensive in terms of time and memory storage resource. Thus, several modification or enhancement to local search algorithm can be found to deal with such problem. Therefore, feature selection is indispensable for the Data Mining and Machine Learning process while managing High Dimensional datasets. Various established search techniques have shown promising results in a number of feature selection problems, but there are only few techniques which deal with High Dimensional data. The central hypothesis is that the important attribute sets are strongly correlated with the target class, and uncorrelated attributes are less important. Further, strong correlation among attribute with other attributes makes strong only one of them and other can be removed. *If two or more attributes have the same importance to the target class values, it will be good to consider only one of them.* As the attributes of a particular application increases, the dimension of that dataset increases. Then feature selection algorithm becomes intractable for finding the best subset, so this problem, sometimes becomes the NP-hard.

Feature selection is a simple method that tries to find out a subset of original features that have the same information regarding the whole datasets, without the loss of generality. Here, the main goal is to identify a few features/genes from thousands of genes to identify a specific set features/gene for specific diseases. However, as the number of attributes becomes extremely larger, most of these presented techniques face the problem of unachievable time computation. In this context, the main problem with this type of data is due to less number of instances, within hundred, while the number of feature is in the order of thousands or even in order of millions. The major challenge in these types of applications is to haul out a set of impressive features, as small as possible, that accurately classifies the learning algorithms [18]. In various Data Mining tasks, the input is represented by a very large number of features, many of which are not required for classifying the class. Feature selection is the task of choosing a small subset of features that is sufficient to classify the target class effectively. The main reason to use feature selection is to reduce computational cost, improved accuracy, and problem understanding.

From the study, there is no feature selection method available for handling the all requirement presents in the inconsistent real world datasets. So the hybrid methods were also present for improving the efficiency of this method. Ranking of features is also applicable for managing the number of large set of feature. After ranking all the features we select only features that are above then some threshold value and then apply our traditional Data Mining approaches on the reduced features to check its correctness and accuracy of the trained model with the reduced set of features.

The motivation for investigating the feature subset selection algorithms came from the requirement to give support to application domain experts with very important quantified evidence that the selected features ultimately become more robust to variations in the training data. This requirement is particularly decisive in biological applications, e.g. DNA-microarrays, genomics, and proteomics, mass spectrometry. These applications are generally characterized by high dimensionality; the goal is to find a small output set of highly uncorrelated variables on which biomedical and Data Miner experts will subsequently invest considerable less time and research effort.

The remaining of this paper is organized as follows. In Section 2 we give the related work and background of feature selection techniques that are required for our proposed algorithm. Section 3 details the methodology and correlation based feature subset selection for High Dimensional data using SU. In Section 4 we have presented our framework and algorithm. In the Section 5 we have done complexity analysis of the proposed algorithm. Then, we have analyzed our algorithm's result, on synthetic data as well as on real world data in Section 6, and finally we conclude in Section 7.

2. Literature Review and Background

The recent problem in Machine Learning and Data Mining is to discovering representative set of attributes from

which to construct a model for classifying or clustering for a specific task. The Feature selection aims at selection a small subset of feature that meets certain criteria given by the user [15] [16] [18]-[20]. It reduces the number of attribute, separate out the irrelevant, noisy and redundant data, that improve and speed up the Data Mining techniques like prediction accuracy of classification. The central hypothesis is that the important attribute sets are strongly correlated with the class values, and uncorrelated attributes are less important. Further, strong correlation among attribute with other attributes makes strong only one of them and other can be removed. If two or more attributes having the same importance to the target class values then it will be good to consider only one of them.

In literature, a large number of feature selection algorithms have been already proposed and they were applied to different fields: bioinformatics [19], text categorization [11] [12], image processing [13] [14], etc. Various taxonomies can be found in the literature in order to classify feature selection algorithms [9]. These algorithms can be of three models: Filter model, Wrapper model, and Embedded model. Filter model does not consider any Data Mining algorithm. They are strongly relies on underlying characteristics of the data variable depends on certain criteria. For example feature selection using, information gain [7], fisher score, Laplacian score; these are methods to find features with the largest information gain, fisher score, Laplacian score respectively. These methods have some limitation as most of the methods are univariate. Consequently, each feature is considered individually without consideration of correlation among features. Wrapper model consider a learning process. Its aim to select a subset of feature that is used to predict or classify efficiently, that gives more discriminative power with that particular learning process; therefore, it consume more time compare to filter. The advantage of wrapper techniques is the suitably used the correlation among the features and the simultaneously interaction with the learning process. However, this type of algorithm has some limitations as it require very large calculation and computation. To overcome the wrapper model, we found very few algorithms in literature, which combine filter and wrapper search to benefit from the singular advantages of each methodology. But, in this paper, we will only consider filter-based approach. And, based on ranking of feature, we will select important attribute which are not redundant and whose SU value is greater than a particular threshold value.

In the process of feature selection, the most important and necessary key operation is, how the individual feature are clearly discriminated. For evaluation of discrimination power of attribute various methods have been proposed, in which information gain is the older and often used techniques [21]. Ding *et al.* [19] used mutual information gain for feature selection from biological dataset *i.e.* Microarray Gene Expression data.

A characteristic feature selection method consists of four fundamental steps as depicted in **Figure 1**, namely, generation of all possible subset, evaluation of generated subset, stopping criterion, and validation of result [10]. Generation of all possible subset is a brute force method that generates candidate feature subsets for estimation based on a particular search procedure. Each generated candidate feature subset is estimated and compared with the preceding most excellent one according to a certain estimating criterion [10]. If the criteria score of new subset come out to be better, it replaces the previous best subset. The process of generation of subset and estimation is recurring until a given stopping criterion is not satisfied [10]. Then, the most excellent subset usually needs to be tested by prior knowledge or many separate tests on synthetic datasets and/or real-world datasets. Feature selection can be found in various areas of Data Mining and Machine Learning such as classification, clustering, association rules, and regression with different applications in different domain [10].

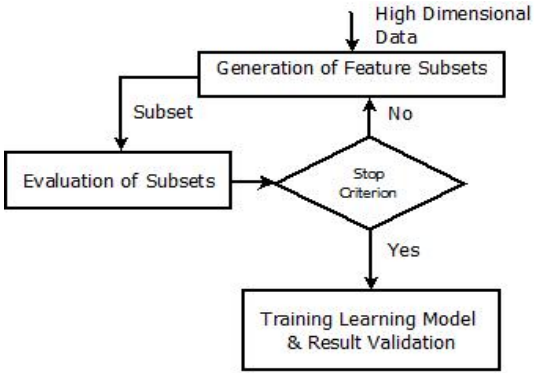

Figure 1. Four basic steps of feature selection.

Almuallim and Dietterich proposed FOCUS [22], an exhaustive search algorithm, in this they showed that FOCUS can find the important and required features in quasi-polynomial time, but having some constraint like:

1) Limitation of difficulties in target class;

2) Data is free from the noise.

But the main problem in High Dimensional data is the computational complexity, that can be as large as $O(2^p)$, for example when all the features are relevant, it may be intractable. Devijver and Kittler in their paper review heuristic search algorithm. They have find Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) algorithms. These algorithms totally based on heuristic, "*find out the most important attribute to add in every step of the iteration is the attribute to be selected and find out the most important attribute to remove in every step of the iteration is the attribute rejected*". These techniques cause problem with High Dimensional data because they did not consider the interaction among attributes.

Relief [23] has been a fundamental and traditional technique for feature selection for normal datasets but when handling High Dimensional data, it failed due to infeasibility of computation. An optimization using supervised model construction has been proposed to improve starter selection. Relief is a basic technique which depends on near-hit, near-miss and some statistical techniques for an instance. It is also noise tolerance and could be untouched by interaction of feature.

2.1. Mutual Information

How to measure the correlation between two or more attributes based on label data? Mutual information (MI) is a basic technique to measures how much knowledge between two attributes are correlated. It is defined as the difference between the sum of the marginal entropies and their joint entropy. For two totally independent objects the mutual information is always zero. In [20], maximum dependency condition based on MI is used for feature selection, and various implementations for classification accuracies have been done. In this paper, we use mutual information, where Shannon's entropy is utilized [20].

Consider the High Dimensional data $D = N * M$, where M is the number of feature and N is the number of the instances. Let x and y be two random features or variables, p(x) and p(y) be their probability density functions and p(x, y) be their joint probability density function. Then their mutual information (MI) has been defined as follows [24]:

$$\mathrm{MI}(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x) * p(y)} \tag{1}$$

We can use the relation of entropy and mutual information to solve the problem in different ways, these are as follow. Let H(X) denote Shannon's entropy of X, then

$$H(X) = -\int p(x) \log\left(p(x)\right) dx \tag{2}$$

The entropy is related to mutual information as follows:

$$\mathrm{MI}(X,Y) = H(X) - H(X,Y) \tag{3}$$

$$\mathrm{MI}(X,Y) = H(X,Y) - H(X \mid Y) - H(Y \mid X) \tag{4}$$

The feature's values are considered to be discrete. Here, the marginal entropies is represented by H(X) and H(Y), $H(X \mid Y)$ and $H(Y \mid X)$ are the conditional entropies, and the joint entropy of X, Y is represented by H(X, Y). Mutual information is a symmetric estimation. That is the amount of mutual information with the relation with Y is equal to the amount of mutual information after observing X. We can say that the sequence of calculation for two variable X and Y (e.g., (x, y) or (y, x)) will not change the measurement.

As a feature selection criterion, the best feature will maximize the mutual information MI(X, Y), where X is the feature vector and Y is the class indicator. This is a nonlinear statistics of correlation between feature values and class values. The symmetric uncertainty (SU) [1] [7] [16] [19], is extended from MI with normalizing it to the entropy value of features and entropy value of features with class label. SU has been used to evaluate the effectiveness of features for classifying the data by various number of researchers [4] [13] [15] [16] [20]. Our approach also based on SU for correlation between either two features or a feature and a class value.

2.2. Symmetric Uncertainty

Symmetric uncertainty can be used to calculate the fitness of features for feature selection by calculating between feature and the target class. The feature which has high value of SU gets high importance. Symmetric uncertainty defined as

$$SU(X,Y) = \frac{2 \times MI(X,Y)}{H(X) + H(Y)} \tag{5}$$

where H(X) is the entropy of a discrete random variable X. If the prior probability of each element of X is p(x), then H(X) can be calculated by Equation (2).

Symmetric uncertainty, Equation (5), behave a couple of variables symmetrically, it compensates for mutual information's bias towards features having large number of different values and normalizes within range [0, 1]. A value 1 of SU(X, Y) indicate that knowledge of the object value strongly represent the values of other and the SU(X, Y) value 0 indicate the independence of X and Y. In this paper, we also deal with continuous features by normalized in proper discrete form.

2.3. Relevant Feature and F-Correlation

The definition of relevant feature is defined as: F_i is relevant to the target concept C if and only if there exists some D_i', f_i, and c, such that, for probability $p(D_I' = d_I', F_i = f)_i > 0$, $p(C = c \mid SU_{i,c} \geq \lambda)$ otherwise, feature F_i is an irrelevant feature [1]. Definition of relevant Feature indicates that there are two type of relevant features due to different S_i':

1) When $D_i' = D_i$, from the above definition, we can know that F_i is fundamentally relevant to the target class;

2) When D_i' is not proper subset of D_i, from the definition we may obtain that $p(C \mid D_i, F_i) = p(C \mid D_i)$. It can be concluded that F_i is irrelevant to the target class. It's a general concept that most of the information contained in redundant features is containing by some other features. As a result, redundant features do not have strong interpretability for the target class.

Given SU(X, Y) the symmetric uncertainty of features X & Y, the correlation between two attributes is refers as *F-correlation*. The correlation between any pair of attributes F_i and $F_j(F_i, F_j \in F, j \neq i)$ is refer *F-correlation* of F_i and F_i, and we are denoting it as $SU_{i,j}$.

3. A Correlation Based Feature Subset Selection Algorithm

In this section, we propose the framework of our feature subset selection techniques which can improve the classification and clustering technique. To select important feature for classification or clustering accuracy, we require some aspect *i.e.*

- How to decide which of the attributes is relevant for a particular class and which of the all attributes are not?
- How to decide among all relevant which attribute is redundant?
- How to decide whether two attribute are closely correlated?

Using the symmetric uncertainty (SU) as the fitness function, we are able to generate an algorithm and framework to select important features for Data Mining task. This framework and algorithm is totally based on the correlation analysis of attributes using supervised High Dimensional datasets.

The answer to these questions can be sorted out by applying appropriate approaches, like, for first question we can use a user defined threshold value, generally used with filter approach of feature selection. For example, let us consider a dataset D having M feature and N instances and set of C classes. Let $SU_{i,c}$ denote the SU value of a feature f_i and class C, then a subset D' of the important features can be decided by a user defined threshold value, which is the second step in our framework. It can be defined as: $F_i \in D'$, $SU_{i,c} >= \lambda$, for $1 \leq i \leq M$.

The answer to the next question is important because this is the main question on which we are focusing. For this, we have to analyze pair-wise correlations among all attributes, but if we calculate the pair wise correlation, the time complexity for this will be $O(M^2)$, where M is the number of attributes which are very high in High Dimensional data.

Correlation between attributes are also captures by symmetric uncertainty values, but to decide and differentiate between relevant and redundant attribute, we have a reason, why we are selecting a particular threshold value. We can say, need to define whether the value of correlation or symmetric uncertainty between two attributes in D' is very high to cause the redundancy, if it is true, then one of them may be deleted from D'. For a attribute $F_i \in S'$, as we discussed the value of $SU_{i,c}$ is shows the correlation of F_i to the class C. If we examine the value of $SU_{j,i}$ for every $F_j \in D'$, we will also calculate the extent to which F_i is correlated to the remaining important attribute in D'. Therefore, it is the main advantage of using this algorithm; in this we can find the strongly correlated attributes to a feature F_i in as usual way, after that we decide D' using a threshold in step 2 that is equal or similar to λ and reject the other attribute that have value less than λ. Similar things we can do for M' attributes in D'. However, when we are finding strong correlated features with one concept and not considering the another concept then this method is not logically good. In the context of a set of important feature D' which we sort out by comparing the user defined threshold value, when we try to calculate strong correlation features for a specific feature F_i within D', we have found that it is more important to use the class correlation value between F_i and the class label, while $SU_{i,c}$ as a reference. The logic behind this is lies in our hypothesis, that is: *an attribute that is correlated to one class at a particular level cab also be correlated to some other attributes at the equal or a higher level*. So, though the correlation between attribute and the class is higher than some threshold value λ and there for we are just considering this attribute is important to the class, but not considering this attribute correlation predominant. Lei Yu *et al.* [20] proposed and define the concept of predominant correlation which is as follow.

The correlation between a feature F_i and the class C is predominant iff $SU_{i,c} \geq \lambda$ and for each $F_j \in D'$, there should not be any F_i such that $SU_{j,i} \geq SU_{i,c}$ when $j \neq i$.

If there exists such F_i to a feature F_i, which follow the above condition then we can say it is a redundant attribute to F_i and use D_i, to represent the set of all redundant attribute for F_i. Given $F_i \in D'$, we divide the D' into two part D_h and D_l,

where $D_h = \left\{ F_j \middle| F_j \in D',\ \text{and}\ SU_{j,c} \geq SU_{avg,c} \right\}$ and $D_l = \left\{ F_j \middle| F_j \in D',\ SU_{j,c} \leq SU_{avg,c} \right\}$.

According to the above definitions, a attribute is good if it is predominant in predicting the class value, and feature selection, for classification, is a process that determine all predominant attributes to the class value and remove other attributes.

We are considering some assumptions in development of this framework, that is, if two attributes are seems redundant to each other and we have to remove one attribute, then we will remove the attributes that is less relevant to the class value and keeps more information to predict the class. The attribute with the highest $SU_{i,c}$ values is always a predominant attribute and removal of all other feature in the list is a very initial point. Similarly, the last attribute having smallest $SU_{i,c}$ will also considering as a start point and relevance of this attribute is checked with other attribute in the reverse order. Furthermore, we have proposed some assumption that can efficiently identify predominant features and avoid redundant features among all important attributes, without identify all important attributes for each attributes in D' and in this way, we may be refrain from pair-wise analysis of correlations between whole important attributes.

4. Proposed Framework and Algorithm

As we discussed the methodology so for, we are now going to propose a framework and algorithm. By using SU, that reimburse for the Information Gain's bias toward attributes with more values and normalize their values in the range of [0, 1] where the value 1 represent the knowledge of either one of the values totally classify the value of the another and value 0 represent that X and Y are independent. The main advantages of using symmetric uncertainty are that it treats a pair of feature symmetrically.

We are using the SU value for two main reasons: as we can see in step two, it can remove the attributes that have less SU value than predefined threshold λ because those attribute which are having high $SU_{i,c}$ value are having higher weight, and the attribute having lesser value of $SU_{i,c}$ is removed. After this, gets every features weight that can be used for sorting the attribute and make it easy to partition the attributes in D_l and D_h. See the **Figure 2**. Second reason is that it is symmetric in nature *i.e.* $SU_{x,y}$ is equal to $SU_{y,x}$ for any feature x and y. A feature having higher SU value have to more representative, or containing more information for a particular class. Symmetric characteristics of SU is used to make algorithm faster. To make faster, we parallely calculate

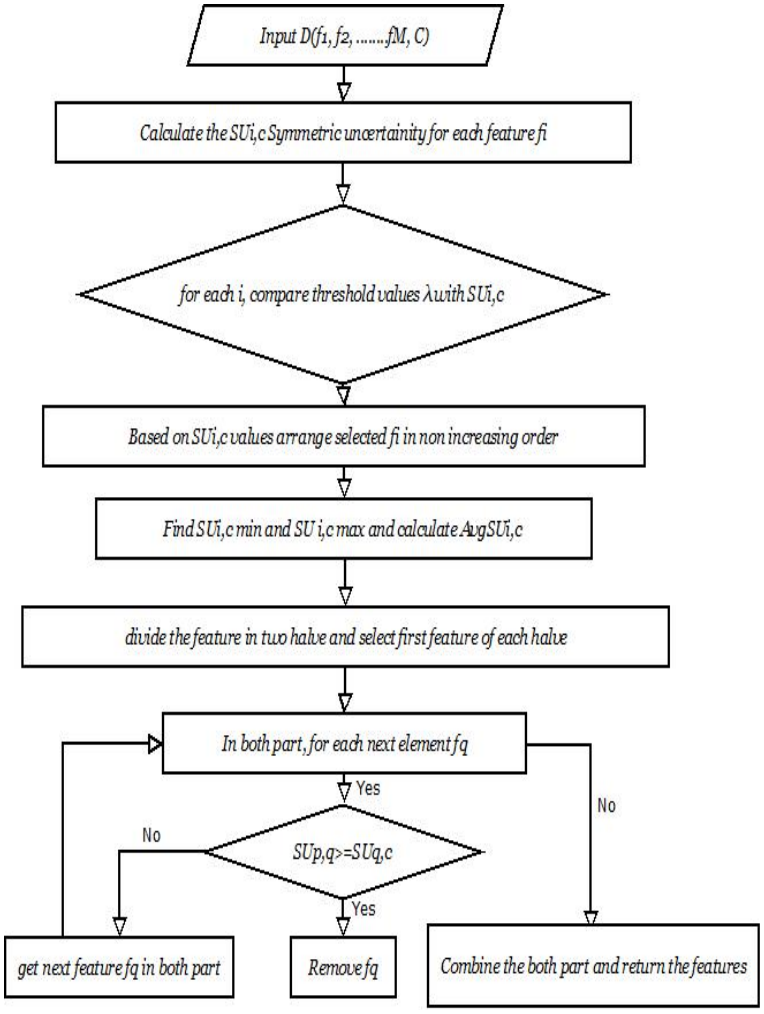

Figure 2. Proposed framework for feature subset selection.

the SU value for D'_h and D'_l in forward and reversed order respectively by using thread concept.

In this, we have given a High Dimensional dataset with M different attributes and a class label C, the approach finds a set of predominant attributes subset for the class values and reject all other attributes which are irrelevant. It can be divided in to two sections. In the earlier section, it calculates the $SU_{i,c}$ values for each attributes and arrange them in non increasing order according to their $SU_{i,c}$ values. Here the correlation among the attributes F_i and class C is represented by $SU_{i,c}$. Processing on the deletion of the non-relevant feature is done by making a ordered list of them and maintain the sustainability of the predominant features. An attribute f_p which is selected as predominant attribute based on SU value can be used for filter out the other attributes that have less SU values.

The process start with the calculation of SU for each attributes, after that we select the first and last element and continues as follow. Calculate the middle index of the sorted element and divide the whole attributes into two parts. In the first part, we start from the first element to the middle index and in the second part from last element till the middle index. For all the remaining feature f_q, if f_p represents a redundant to a feature f_q, then f_q will be deleted from the list. Attribute f_q will be redundant feature to the f_p if the correlation between f_q and f_p is greater than the correlation between f_q and the class value C. similarly in the second part, start from the last element having minimum $SU_{i,c}$ and compare with all other feature in the reverse way till middle index. After finding the important feature from both parts we will combine them to get the complete important feature. When the entire attribute have been tested in both part individually, then process will terminate. And finally its return the optimal feature subsets. We have given proposed algorithm.

SU based Algorithm for feature selection:

Input: D(f_1, f_2, f_3,……., f_M, C) and threshold λ

Output: optimal Subset of features

```
1  Begin
2      For i=1to i<=M;
3       begin
4        Calculate SU_{i,c} for each f_i;
5         If(SU_{i,c} >λ)
6            Store f_i into D';
7    end;
8    Apply sorting and arrange D' in non-increasing order of SU_{i,c} value;
9    Find length of D'
10   Calculate the middle_index(D');
11   Partition the D' in D'_h and D'_l;
12   f_ph=firstElement(D'_h);
13   f_pl=lastElement(D'_l);
14   do begin
15       f_qh=getNextElement(D'_h,f_ph);
16          if(f_qh≠ null)
17                      f'_qh=f_qh;
18                      if(SU_ph,qh ≥ SU_qh,c)
19                      delete f_qh from D'_h;
20                      f_qh=findNextElement(D'_h, f'_qh);
21                      else f_qh=findNextElement(D'_h, f_qh);
22                      until (D'_h=Null);
23   end;
24   do begin
25       f_ql=getPreviousElement(D'_l,f_pl);
26            if(f_ql≠ null)
27                      f'_ql=f_ql;
28                      if(SU_pl,ql ≤ SU_ql,c)
29                      delete f_pl from D'_h;
30                      f_ql=findPreviousElement(D'_l, f'_ql);
31                      else f_ql=findPreviousElement(D'_l, f_ql);
32                      until (D'_l=Null);
33   end;
34  D_optimal=D'_h+ D'_l;
```

5. Computational Complexity of Proposed Approach

We analyze time complexity of the proposed algorithm. In the computation of symmetric uncertainty (SU) values of each feature have linear time complexity in terms of the number of feature M. Most of the time this number of feature also called dimensionality of datasets. Subsequently, this task is performed only once and stored in D', the computation is consider negligible in compared to the further consideration of important features. In the second part (14 - 23) and (24 - 33), in each round, the proposed algorithm can delete a large number of attributes that are redundant to the f_p in the same loop. In the best case, all of the remaining f_q will be redundant and so all of attributes are removed and time complexity will be of order O(M)). In the worst case, when all the f_q are stored in the D' the time complexity will $O(M^2)$. In the average case, we can assume that out of important attribute half of the attributes are deleted in the each iteration. So, the time complexity may be of order O(M logM) where M is the number of attributes. We divide the D' into two part and treat them individually. On average, Line (14 - 23) and (24 - 33) can be computed in O(M/2 logM/2). Since, in the line (1 - 7) we calculate a pair of

attribute's SU values in term of the number of instances N in the data, so the complete complexity of the above proposed algorithm O(N M logM).

6. Experimental Result and Discussion

In our experimental work, we experimentally evaluate the effectiveness of the proposed technique. The objective of our proposal is to evaluate the method in term of speed, number of selected attributes, and predictive accuracy for a particular classifier on selected feature. The algorithm compared against some already existing techniques: Information gain (IG), Chi square, ReleifF and FCBF on the 5 benchmarking high dimension datasets. Because our approach finding less number of features as compared to information gain, chi square, FCBC and ReliefF, results in reduction of time for the resultant mining algorithm. A list of datasets used in our approach is listed in the **Table 1**. This table contains 5 benchmarking High Dimensional datasets along with their characteristics, number of attribute, how many classes contained in the datasets. All of these datasets are taken from the UCI Repository [25]. A brief summary of datasets is described in **Table 1**.

For each dataset from the **Table 1** we will run our algorithm and note down the time required to run in **Table 2** and the number of selected features by the proposed algorithm **Table 4**. We are also analyze the same from some traditional algorithm like ReliefF, information Gain, Chi square, FCBF and record time required and number of selected feature for each algorithm in **Table 2** and **Table 3**.

For the validation of our proposed algorithm we have tested the classification accuracy against to different classifier. Mainly, decision tree, SVM and NB classifier are used to check the classification accuracy with all 5 previous feature selection. **Table 4** shows the accuracy by J48 classifier on our techniques and four traditional features selection on reduced dataset. The full data column represent the accuracy on whole data. Similarly, **Table 5** and **Table 6** show the accuracy obtained by NB and SVM classifier respectively, while considering the four dataset obtained by four feature selection techniques and our technique. From **Tables 4-6**, we can observed that our proposed methods works more accurately as compared to the given feature selection techniques, that was the main goal of this work. We get the better result in compare to IG, ReliefF, Chi square, but similar to

Table 1. Dataset and their description.

Datasets	Number of attributes	Number of instances	Number of classes
Lung-cancer	57	32	3
Chemical	151	936	3
Isolat	618	1560	26
Leukemia	7129	72	2
Overian	15,154	253	2

Table 2. Time required (in ms) for relevent feature by different feature selection techniques and out proposed techniques.

Datasets	IG	Chi square	ReliefF	FCBF	Our techniques
Lung-cancer	238	325	62	25	20
Chemical	2766	2432	2622	130	103
Isolat	19,930	19,851	18,085	3098	2830
Leukemia	29,987	27,883	21,090	4143	3716
Overian	-	-	36,561	7613	7207

Table 3. A comparision of number of selected feature through various techniques and our tech.

Datasets	IG	Chi square	ReliefF	FCBF	Our techniques
Lung-cancer	16	15	9	7	8
Chemical	23	21	11	10	9
Isolat	37	39	22	21	25
Leukemia	52	62	36	33	33
Overian	-	-	107	96	100

Table 4. Classification accuracy by J48 classifier.

Datasets	Full data	IG	Chi square	ReliefF	FCBS	Our techniques
Lung-cancer	81.26	89.32	88.35	84.50	93.73	92.24
Chemical	94.13	92.13	92.78	93.27	95.36	95.83
Isolat	79.54	78.21	75.53	75.02	77.32	78.37
Leukemia	74.25	76.86	82.47	83.89	86.64	85.61
Overian	72.87	-	-	77.31	78.27	79.34

Table 5. Classification accuracy by NB classifier.

Datasets	Full data	IG	Chi square	ReliefF	FCBF	Our techniques
Lung-cancer	88.53	92.75	92.17	86.22	94.73	94.33
Chemical	95.56	93.80	95.53	93.53	96.65	95.30
Isolat	82.37	83.93	77.73	77.05	81.62	80.18
Leukemia	78.84	81.23	85.42	81.47	88.77	90.50
Overian	77.29	-	-	74.32	82.55	84.83

Table 6. Classification accuracy by SVM classifier.

Datasets	Full data	IG	Chi square	ReliefF	FCFS	Our techniques
Lung-cancer	83.42	90.75	90.97	84.32	93.18	94.63
Chemical	95.32	92.50	93.16	92.61	96.42	96.19
Isolat	80.21	80.26	77.49	76.56	80.86	78.77
Leukemia	74.23	79.47	84.35	81.28	88.27	88.50
Overian	72.94	-	-	73.98	80.75	83.29

FCFS method in respect to classification accuracy. But when we talk about the time consumption, our approach outperform then all other methods.

7. Conclusion

In this paper, we have proposed an algorithm for feature subset selection for High Dimensional datasets. We are using correlation based feature ranking method, symmetric uncertainty (SU), which forms the basis of our approach. Our future plan is to extend this approach on very High Dimensional data (it is proposed that the current approach be explored on very High Dimensional data (*i.e.* ovarian dataset)). We have noticed that this algorithm generally works fine with numerical data; we can also try to extend this approach to working with mixed type of data (containing both nominal and categorical) without normalizing them in discrete values. This may also solve the problem of feature selection for High Dimensional data and biological datasets with millions of features using this approach. Since, for example the next generation sequencing techniques in biological analysis can produce data with several millions features in a single computation. Existing approaches make it hard to access data of this dimensionality, which creates the challenges of computational power, algorithm stability and accuracy of algorithm in parallel.

References

[1] Song, Q., Ni, J. and Wang, G. (2013) A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data. *IEEE Transactions on Knowledge and Data Engineering*, **25**, 1-14. http://dx.doi.org/10.1109/TKDE.2011.181

[2] Ben-Bassat, M. (1982) Pattern Recognition and Reduction of Dimensionality. In: Krishnaiah, P.R. and Kanal, L.N., Eds., *Handbook of Statistics-II*, Vol. 1, North Holland, Amsterdam, 773-791. http://dx.doi.org/10.1016/S0169-7161(82)02038-0

[3] Mitra, P., Murthy, C.A. and Pal, S.K. (2002) Unsupervised Feature Selection Using Feature Similarity. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **24**, 301-312. http://dx.doi.org/10.1109/34.990133

[4] Blum, A.L. and Langley, P. (1997) Selection of Relevant Features and Examples in Machine Learning. *Artificial Intel-*

ligence, **97**, 245-271. http://dx.doi.org/10.1016/S0004-3702(97)00063-5

[5] Kohavi, R. and John, G.H. (1997) Wrappers for Feature Subset Selection. *Artificial Intelligence*, **97**, 273-324. http://dx.doi.org/10.1016/S0004-3702(97)00043-X

[6] John, G.H., Kohavi, R. and Pfleger, K. (1994) Irrelevant Feature and the Subset Selection Problem. *Proceedings of* 11*th International Conference on Machine Learning*, New Brunswick, 10-13 July 1994, 121-129.

[7] Chow, T.W.S. and Huang, D. (2005) Effective Feature Selection Scheme Using Mutual Information. *Neurocomputing*, **63**, 325-343. http://dx.doi.org/10.1016/j.neucom.2004.01.194

[8] Kim, Y., Street, W. and Menczer, F. (2000) Feature Selection for Unsupervised Learning via Evolutionary Search. *Proceedings of* 6*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington DC, August, 365-369.

[9] Dash, M., Choi, K., Scheuermann, P. and Liu, H. (2002) Feature Selection for Clustering a Filter Solution. *Proceedings of Second International Conference on Data Mining*, Florida, 19-22 November, 115-122.

[10] Liu, H. and Yu, L. (2005) Toward Integrating Feature Selection Algorithms for Classification and Clustering. *IEEE Transactions on Knowledge and Data Engineering*, **17**, 491-502. http://dx.doi.org/10.1109/TKDE.2005.66

[11] Yang, Y. and Pederson, J.O. (1997) A Comparative Study on Feature Selection in Text Categorization. *Proceedings of* 14*th International Conference on Machine Learning*, Nashville, 8-12 July 1997, 412-420.

[12] Nigam, K., Mccallum, A.K., Thrun, S. and Mitchell, T. (2000) Text Classification from Labeled and Unlabeled Documents Using EM. *Journal of Machine Learning*, **39**, 103-134. http://dx.doi.org/10.1023/A:1007692713085

[13] Guldogan, E. and Gabbouj, M. (2008) Feature Selection for Content-Based Image Retrieval. *Signal, Image and Video Processing*, **2**, 241-250. http://dx.doi.org/10.1007/s11760-007-0049-9

[14] Vasconcelos, M. and Vasconcelos, N. (2009) Natural Image Statistics and Low-Complexity Feature Selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**, 228-244. http://dx.doi.org/10.1109/TPAMI.2008.77

[15] Oveisi, F., Oveisi, S., Efranian, A. and Patras, I. (2012) Tree-Structured Feature Extraction Using Mutual Information. *IEEE Transactions on Neural Networks and Learning Systems*, **23**, 127-137. http://dx.doi.org/10.1109/TNNLS.2011.2178447

[16] Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1988) Numerical Recipes in C. Cambridge University Press, Cambridge.

[17] Ng, K.S. and Liu, H. (2000) Customer Retention via Data Mining. *Artificial Intelligence Review*, **14**, 569-590. http://dx.doi.org/10.1023/A:1006676015154

[18] Liu, H. and Motoda, H. (2001) Feature Extraction, Construction and Selection: A Data Mining Perspective. Second Printing, Kluwer Academic, Boston.

[19] Ding, C. and Peng, H. (2003) Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *Proceedings of the IEEE Computer Society Conference on Bioinformatics*, Berkeley, 11-14 August 2003, 523-528.

[20] Yu, L. and Liu, H. (2003) Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. 20*th International Conference on Machine Learning* (*ICML*-03), Washington DC, 21-24 August 2003, 856-863.

[21] Hariri, S., Yousif, M. and Qu, G. (2005) A New Dependency and Correlation Analysis for Features. *IEEE Transactions on Knowledge and Data Engineering*, **17**, 1199-1207. http://dx.doi.org/10.1109/TKDE.2005.136

[22] Almuallim, H. and Dietterich, T.G. (1991) Learning with Many Irrelevant Features. *Proceeding of* 9*th National Conference on Artificial Intelligence* (*AAAI*-91), Anaheim, 14-19 July 1991, 547-552.

[23] Kononenko, I. (1994) Estimating Attributes: Analysis and Extensions of RELIEF. *Machine Learning*: *ECML*-94, *European Conference on Machine Learning*, Secaucus, 6-8 April 1994, 171-182.

[24] Kannan, S.S. and Ramraj N. (2010) A Novel Hybrid Feature Selection via Symmetrical Uncertainty Ranking Based Local Memetic Search Algorithm. *Knowledge-Based Systems*, **23**, 580-585. http://dx.doi.org/10.1016/j.knosys.2010.03.016

[25] Blake, C.L. and Merz, C.J. (2010) UCI Repository of Machine Learning Database. Department of Information and Computer Sciences, University of California, Irvine.