Journal of Engineering Technology
Volume 3, July. 2015, Pages 135-144

ORIGINAL ARTICLE

Correlation and Clustering based Efficient Feature Subset Selection

Preeti Kumari[1], Prof. K.Rajeswari [2], Dr. V.Vaithiyanathan[3]
[1]ME scholor department of Computer Engg, Pimpri Chinchwad College of Engg, Pune University, India
[2] Associate professor in Pimpri Chinchwad College Of Engg, Pune University,India.
[3]Associate Dean Research, SASTRA university, Tanjore

Abstract - Feature subset selection is an effective and efficient way for reducing dimensionality, removing irrelevant and redundant data, increasing learning accuracy and thus improving the quality of results in less time. Feature selection composes of identifying a subset of the most important and useful features that produce compatible results as the original entire set of features produces. A feature selection algorithm can be measured from both the efficiency and effectiveness points. The efficiency composes of the time required to find a subset of features, whereas effectiveness is related to the quality of the subset of features finally selected. Based on the above criteria, a Correlation and clustering-based feature selection algorithm, is proposed. The algorithm works in two stages. In the first stage, irrelevant features are removed using correlation between feature and class by using a user defined threshold, then in the second stage with the help of these relevant feature, redundant features are removed by constructing a max heap from the feature set, after that calculation of correlation between the feature - feature set for the edges present in the tree is done. A tree is formed which is divided into clusters, and from each cluster a strong representative feature is selected to give a final subset of feature. The feature of each cluster differs and is relatively independent of each other. The clustering-based technique has a high probability of producing a subset of useful and independent features.

Keywords – Feature subset selection, correlation, tree, feature clusters, representative feature.

1. INTRODUCTION

Many factors contribute to the success of machine learning on any given task. The data quality is one such factor, if the data or information is redundant or irrelevant, or the data is noisy or unreliable, the knowledge discovery process is difficult and not up to mark. Machine learning provides us tools by which large quantities of data can be automatically analyzed. Feature selection is quite fundamental to machine learning. Feature subset selection is an efficient and effective way of increasing learning accuracy by reducing dimensionality and removing irrelevant data. The feature selection methods can be broadly classified into four categories: Embedded, Filter, wrapper and hybrid approach. The embedded method use feature selection as a part of training process which are specific to any given learning algorithms, and hence perform better than the other three categories [18]. An example of this approach is decision trees or artificial neural networks. Wrappers use feedback methods that incorporate machine learning algorithms in the feature selection process, so they depend on the performance of a specific classifier to measure the quality of feature set. Wrapper method try to search in the feature subsets and find out estimated accuracy of single learning algorithm for every feature that can be removed or added to or from feature subset .There are various strategies to search in feature space like forward, backward search, non-exhaustive, heuristic search, best fit, greedy or random search are often used.

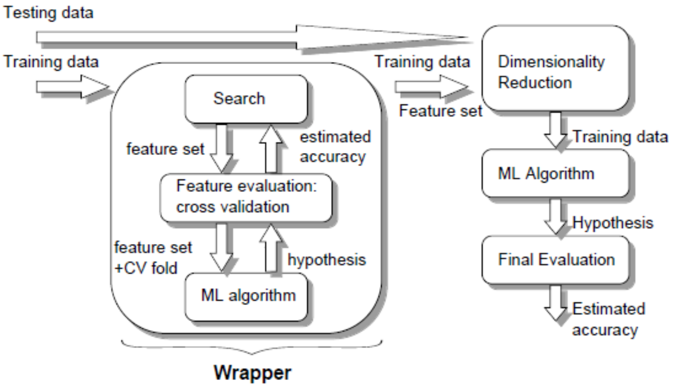

Figure. 1: Wrapper approach

The filter methods of feature selection are independent of learning algorithms and offer good generality. Their computational complexity is also less.

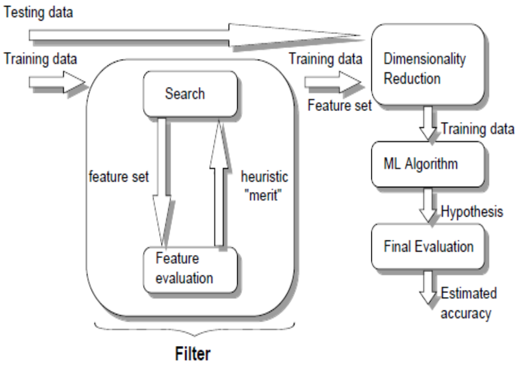

Figure.2: Filter approach

The hybrid method is combination of filter and wrapper methods, the filers can be used to reduce search space and then wrappers can be applied to that search space.

Here we are choosing the filter method because wrapper methods are computationally expensive and tend to over fit on the small training sets. The filter method in addition to their generality is a good choice when features are large. For the filter feature selection method a combination of correlation measure and clustering can be combined. In cluster analysis use of Graph theoretic method is well studied and used in many applications. Here we have used a tree based clustering algorithm because of their ability to not assume that data points are grouped across canters. Feature selection tasks can also be accomplished on the basis of correlation between features which is used to find relevant feature and remove redundant features.

The main reason behind using the correlation measures are:
1) The features which are irrelevant have no or weak correlation with the target concept.
2) Redundant features can be collected in a cluster and a representative feature can be selected out of the cluster to a select most representative and strong feature.

So a combined approach of correlation measures along with clustering is used to select feature subset of data. The rest material in this paper is organised as follows: In section II, we will describe literature survey.In section III , proposed approach and framework has been presented.Finally in section IV.We summarise the present work and draw conclusion along with future scope.

2. LITERATURE SURVEY

Within the filter model based feature selection, different feature selection algorithms can be further categorized into two groups namely, feature weighting algorithms and subset search algorithms, based on whether they compute the goodness of features individually or through feature subsets. Now we will discuss some of the advantages and shortcomings of representative algorithms in each group.

Feature weighting algorithms are based on the idea of assigning weights to features individually and then rank them based on their relevance to the target concept. There are a number of different definitions on feature relevance in machine learning literature (Blum & Langley, 1997; Kohavi & John, 1997). In this approach if a feature is good and thus will be selected if its weight of relevance is greater than a defined threshold value. A well known algorithm that relies on this concept of relevance evaluation is Relief (Kira & Rendell, 1992). The key idea behind Relief is to estimate the relevance of features in accordance to how well their values differentiate between the instances of the same and different classes that are near each other. Relief randomly samples a number (m) of instances from the training set and updates the relevance estimation of each feature based on the difference between the selected instance and the two nearest instances of the same and opposite classes. Time complexity of Relief for a data set with M instances and N features is O(mMN). where m being a constant, the time complexity becomes O(MN),which makes Relief very scalable to data sets having both a large number of instances and a very high dimensionality. However, it has disadvantage that Relief does not help with removing redundant features. As long as features are seem relevant to the class concept, they all be selected even when many of them are largely correlated to each other (Kira & Rendell, 1992). Many other algorithms belonging to this group have much similar problems as Relief does. They can only determine the relevance of features to the target concept, but fails to discover redundancy among features in the group. However, it has been seen that in feature selection along with irrelevant features, redundant features also majorly affect the speed and accuracy of learning algorithms and thus should be removed as well (Hall, 2000; Kohavi & John, 1997). Therefore, in the process of feature selection for high dimensional data where there may exist many irrelevant features along with redundant features, the purely relevance-based feature weighting algorithms do not meet all the need of feature selection that effectively.

Whereas in subset search algorithms the searching of candidate feature is guided by a certain evaluation measure (Liu & Motoda, 1998) , which determines the goodness of each subset. An optimal subset is selected when the search stops. Some of existing evaluation measures that are shown effective in removing both irrelevant and redundant features include the consistency measure (Dash et al., 2000) and the correlation measure (Hall, 1999; Hall, 2000). Consistency measure makes an attempts to find a minimum number of features that separate the classes as consistently as possible as the full set of features can do. An inconsistency is in data is defined as two instances having the same feature values but have different class labels. In Dash et al. (2000), different search strategies, namely, exhaustive, heuristic, and random search, are combined with this evaluation measure to form various algorithms. But the time complexity is exponential in terms of data dimensionality for exhaustive search algorithms and quadratic for heuristic search stategy. The complexity can be linear to the number of iterations in a random search, but experiments show that in process to find best feature subset, the number of iterations required is mostly at least quadratic to the number of features (Dash et al., 2000). In Hall (2000), a correlation measure can be applied to evaluate the goodness of feature subsets is based on the hypothesis that a good feature subset is the only one that contains features that are highly correlated to the class, but yet uncorrelated to each other. The underlying algorithm, named CFS, also exploits heuristic search. Therefore, with quadratic or higher time complexity in terms of their dimensionality, existing subset search algorithms have not strong scalability to deal with high dimensional data.

To solve and overcome the problems of algorithms in both categories and meet the demand for feature selection for data, we develop a novel algorithm which can effectively identify both irrelevant and redundant features with less time complexity than subset search algorithms.

3.PROPOSED WORK

Feature subset selection problem involves removal of irrelevant feature along with redundant feature.The work is based on the concept that “good subsets of feature contain features that are highly correlated with the class, yet uncorrelated with each other.” In general a feature is considered to be good if it is relevant to the class concept but not redundant to any of the other relevant features in the dataset.

Here we are taking correlation between the two variables as for evaluating a goodness measure, if the correlation between a feature and the class is as high enough to make it relevant to the class and at the same time the correlation between it and any other relevant features does not reach a level so that it can be predicted by any of the other existing relevant features, it will be considered as a good feature for the classification task. In this process, the problem of feature selection comes and narrows down to finding a suitable measure of correlations between features and a good and effective ,sound method to select features which is based on this measure.

The proposed algorithm work in three steps:
a)In the first part relevant features are selected depending upon the threshold value. b)In the second step, we construct a heap (binary tree) of the relevant features.
c) In the third step depending upon the correlation between features set, the tree is partitioned into cluster and then most strong and representative feature that is strongly related to the target classes is selected from each cluster to form a subset of features.

3.1 Correlation Based Measures

Here we are considering correlation between the two variables as a goodness measure, if the correlation between a feature and the class is high enough to make it relevant to the class and at the same time the correlation between it and any other relevant features does not reach a level so that it can be predicted by any of the other present relevant features, it will be regarded as a good feature for the classification task.Broadly there are two methods to measure the correlation between any two random variables. One is based on classical linear correlation and the other is dependent on information theory concept. In the first approach, the most well known measure is the linear correlation coefficient. For a pair of variables (X,Y) the linear correlation coefficient r is computed by the formula:

$$r = \frac{\sum_i (x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_i (x_i - \overline{x_i})^2}\sqrt{\sum_i (y_i - \overline{y_i})^2}}$$

Where $\overline{x}$i is the mean of X, and $\overline{y}$i is the mean of Y .The value of r lies between range -1 and 1, inclusive. If X and Y are completely correlated with each other , r takes the value of 1 or -1; and if X and Y are totally independent of each other, r is zero. It is a symmetrical measure of evaluating correlation between two variables.
However, in real world data it is not always safe to assume linear correlation between two features in the data. Linear correlation measures may not be able to capture correlations between variables that are not linear in their nature . Another limitation of linear approach is that the calculation requires all features contain numerical values only, which restricts it to be applied on the data which are not of numerical values.

To overcome these mentioned limitations, in our approach we have adopt another approach and choose a correlation measure based on the information theoretical concept of entropy, a measure of the uncertainty of a random variable. The entropy of a variable X is defined as

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i))$$

and the entropy of X after observing values of another variable Y is defined as
where P(xi) is the prior probabilities for all values of X, and P(xi|yi) is the posterior probabilities of X given the values of Y . The amount by which the entropy of X decreases reflects additional information about X provided by Y and is called information gain (Quinlan, 1993), given by

$$IG(X|Y) = H(X) - H(X|Y)$$

According to this measure, a feature Y is regarded more correlated to feature X than to feature Z, if IG(X|Y) > IG(Z|Y).

Symmetry is a desired property for a measure of correlations between features. However, information gain is biased in favor of features with more values. Furthermore, the values have to be normalized to ensure they are comparable and have the same affect. Therefore, we choose symmetrical uncertainty (Press et al., 1988), defined as follows.

$$SU(X,Y) = 2\left[\frac{IG(X|Y)}{H(X)+H(Y)}\right]$$

It compensates for information gain's bias toward features with more values and normalizes its values to the range [0-1] with the value 1 indicating that knowledge of the value of either one completely predicts the value of the other and the value 0 indicating that X and Y are independent. In addition, it still treats a pair of features symmetrically. Entropy-based measures require nominal features, but they can be applied to measure correlations between continuous features as well, if the values are discretized properly in advance (Fayyad & Irani, 1993; Liu et al., 2002). Therefore, we use symmetrical uncertainty in this work.

Definition:

Feature Relevance: The relevance between the feature Fi ∈ F and the target concept C is referred as feature relevance of Fi and C and is denoted by SU (Fi,C). If SU (Fi,C) is greater than a predetermined threshold θ, we say that Fi is a strong F-Relevance feature.

Definition:

F-Correlation: The correlation between any pair of features Fi and Fj ,(Fi ,Fj ∈ F and i≠j) is called the F-Correlation of Fi and Fj and is denoted by SU(Fi, Fj).

3.2 Algorithm

```
Inputs:  D(F1, F2,…….. ,Fm,C)- the given data set.
         : θ –F-Relevance threshold.
Output: S- Selected feature subset.
//-----Part 1:Irrelevant feature removal----
1.For  i = 1 to m do
2.          F-Relevance=SU(Fi,C)
3.           If F-Relevance> θ  then
4.                        S=S∪{ Fi};
5.Sort relevant feature
//-----Part 2: Heap binary tree construction
6.G=NULL
7.For each pair og feature in G do
8.        F-Correlation=SU(Fi,Fj)
9.Obtain a graph with F-correlation as the weight of corresponding edge.
10.For each edge in the forest do
    If SU(Fi, Fj)< SU(Fi,C) and SU(Fi,Fj)< SU(Fj,C)     then
    Forest=forest-edge;
11.For each tree in the forest do
     Fs=max SU(F,C);
S=S∪ {F_s}
12. return S.
```

3.3 Framework of Proposed system

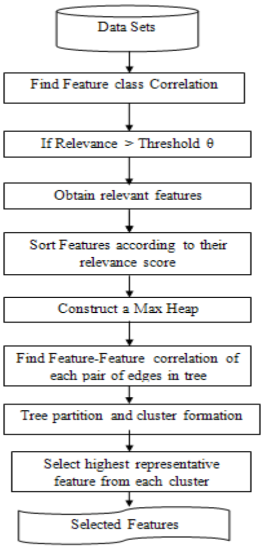
Data Sets
Find Feature class Correlation
If Relevance > Threshold θ
Obtain relevant features
Sort Features according to their relevance score
Construct a Max Heap
Find Feature-Feature correlation of each pair of edges in tree
Tree partition and cluster formation
Select highest representative feature from each cluster
Selected Features

Figure.3. : Framework of proposed model

3.4Time Complexity analysis

In Dash (2000), different search strategies like exhaustive, heuristic and random search are applied, their time complexity is exponential in terms of data dimensionality for exhaustive search and quadratic for heuristic search.
The algorithm improves the efficiency of earlier work by not calculating the pair wise correlation of all features and by removing redundant feature with the help of clustering.
The major amount of work in the algorithm is to compute the SU values for feature relevance and feature correlation, which has linear complexity in terms of number of instances. So the first part has O(m) in terms of number of features.
In second part ,if number of selected feature in first part is one(f=1) then there is no need to continue rest of algorithm, then the algorithm has a complexity O(m). It constructs a binary tree heap which has a complexity of O(log n).So the algorithm has a linear time complexity of O(m+log n).

4. EMPIRICAL STUDY

4.1 Data Sources

1)Congressional Voting Records Data Set

Data Set Characteristics: Multivariate
Number of Instances:435
Attribute Characteristics: Categorical
Number of Attributes:16
Data Set Information:

This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA. The CQA lists nine different types of votes: voted for, paired for, and announced for (these three simplified to yea), voted against, paired against, and announced against (these three simplified to nay), voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an unknown disposition).

Attribute Information:
1. Class Name: 2 (democrat, republican)
2. handicapped-infants: 2 (y,n)
3. water-project-cost-sharing: 2 (y,n)
4. adoption-of-the-budget-resolution: 2 (y,n)

5. physician-fee-freeze: 2 (y,n)
6. el-salvador-aid: 2 (y,n)
7. religious-groups-in-schools: 2 (y,n)
8. anti-satellite-test-ban: 2 (y,n)
9. aid-to-nicaraguan-contras: 2 (y,n)
10. mx-missile: 2 (y,n)
11. immigration: 2 (y,n)
12. synfuels-corporation-cutback: 2 (y,n)
13. education-spending: 2 (y,n)
14. superfund-right-to-sue: 2 (y,n)
15. crime: 2 (y,n)
16. duty-free-exports: 2 (y,n)
17. export-administration-act-south-africa: 2 (y,n)

2) Spect Heart Data Set

Data Set Characteristics: Multivariate
Number of Instances:267
Attribute Characteristics: Categorical
Number of Attributes:22
Data Set Information:
The dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images. SPECT is a good data set for testing ML algorithms; it has 267 instances that are described by 23 binary attributes.

4.2 Experimental Setup

Configuration of system:
1)Software Requirement:

- Windows 2000, Windows XP, or Windows 7 as the operating system.
- Net beans(Java platform).
- WEKA 3.6.3 data mining tool.

2)Hardware Requirement:

- Minimum of 256MB RAM
- Pentium IV or Higher version
- Display Option: VGA/Flat Screen

To evaluate the performance of proposed algorithm, we have used different type of classification algorithms to classify data sets before and after feature selection. They are 1) probability based Naïve Bayes(NB) , 2)tree based C4.5,3)the instance based lazy learning algorithm IB1 and 4)rule-based ripper.
Naïve Bayes uses a probabilistic method to classify by multiplying the individual probabilities of every feature pair. This algorithm takes independence among the features and provides excellent classification results. Decision tree learning algorithm C4.5 is an extension of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees, rule derivation, and so on. The tree comprises of nodes (features) that are selected by information entropy. Instance-based learner IB1 is a single-nearest neighbour algorithm, and it classifies entities taking the class of the closest associated vectors in the training set via distance metrics. It is the simplest among the algorithms used in our study. Inductive rule learner RIPPER (Repeated Incremental Pruning to Produce Error Reduction) is a propositional rule learner that defines a rule based detection model and seeks to improve it iteratively by using different heuristic techniques. The constructed rule set is then used to classify new instances.

4.3Experimental procedure

In order to get good and stable results we have used 10 fold cross validation strategy is used. Each data set classification algorithm is applied before and after feature selection and its classification accuracy is calculated. If the accuracy of classification increases or remains same on reduced dataset our purpose is achieved. Because dimensionality of data set is reduced its efficiency successfully increases.

Procedure:

```
Data={D1, D2,……., D1}
Learners={NB/C4.5/DT/IB1/RIPPER/BayesNet/Part}
for each dataset
    For each fold ∈ [1,N] do
        TestData=bin[fold]
        TrainingData=data-TestData
          For each learner ∈ Learners do
              Classifier=learner(TrainingData)
              Accuracy=apply classifier to TestData
                    End for
          End for
End for
```

4.4 Results and Analysis

For evaluating the performance of the proposed algorithm we have used accuracy measure of different classifiers on original feature set and reduced feature set.The proposed algorithm when applied to the congressional voting records produces a subset of useful features, which increases the accuracy of classifiers as shown in table 4.1.

1.Dataset: Congressional Voting Records

Table 4.1: Result of voting records

Name of the classifier	Accuracy of Original Data Set(%)	Accuracy of Reduced Data set(%)
Naive Bayes	92.05	97.07
C4.5	96.65	97.07
Decision Table	95.39	97.07
IB1	92.05	96.23
Bayes Net	92.39	97.07
Part	95.39	97.07

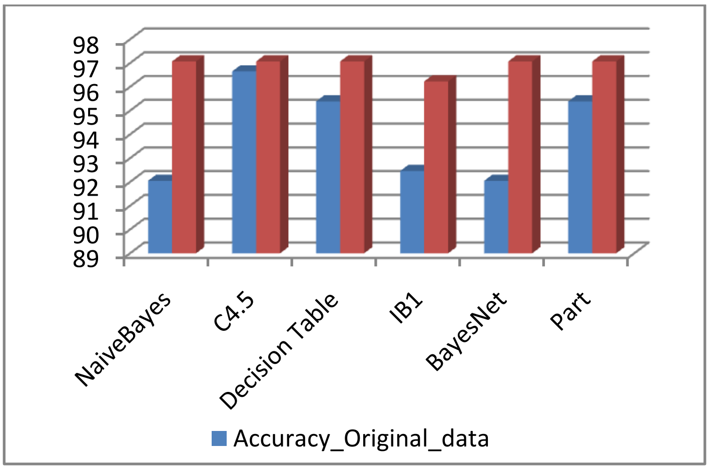

Figure. 4.1: Result for Congressional Voting Records Data Set

It is observed that percentage of accuracy increased by c4.5,Naive Bayes, Decision Tree,IB1,Part,Bayes Net are 0.43,5.43,1.76,4.54,1.76 and 5.06 respectively. Hence accuracy of classifier increases with reduced feature selected which increases the efficiency of dataset.

2.Dataset: Spect Heart Records

The increase in the accuracy of different classifiers in original and reduced data set for Spect Heart records is shown in table 4.2.

Table 4.2: Result of Heart records

Name of the classifier	Accuracy of Original Data Set(%)	Accuracy of Reduced Data set(%)
Naive Bayes	73.75	80
C4.5	68.75	70
Decision Table	72.5	75
IB1	53.75	70
Bayes Net	73.75	81.25
Part	70	75

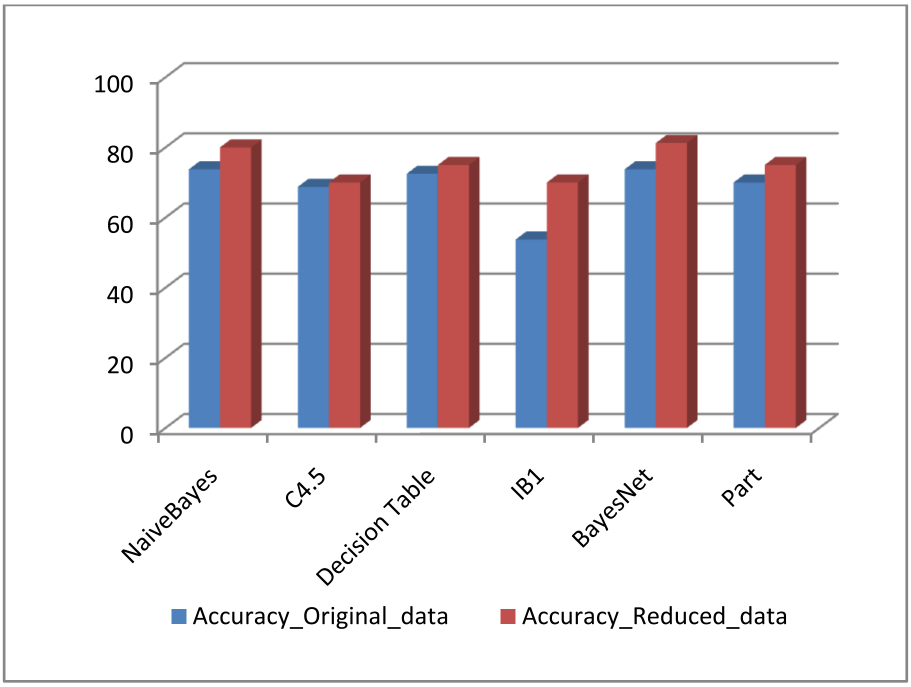

Fig 4.2: Result for Spect Heart Data Set

The graph in fig. 4.1 and 4.2 shows the result of applying the algorithm over congressional data set and heart data set. It is observed that percentage of accuracy increased by c4.5,Naive Bayes, Decision Tree,IB1,Part,Bayes Net are 1.81,8.47,3.44,30.23,7.14 and 10.16 respectively.

It can be seen that the classification accuracy obtained on full data set is less than the classification accuracy obtained on reduced data set.So the algorithm successfully removes the relevant and redundant features from dataset and hence successfully reduces its dimensionality and achieves performance gain.

5. CONCLUSION AND FUTURE SCOPE

In the proposed solution new algorithm has being designed based on correlation and clustering which effectively removes the irrelevant and redundant features of dataset which results in increasing accuracy of classifiers. For Congressional Voting Records dataset the reduced percent of feature set is 87.5 % with a average increase of 3.16 %.Similarly for Spect Heart Data Set the reduced percent of feature set is 72.72 with an average increase of 10.21 accuracy measure.

In future scope, different correlation measures along with fuzzy logic can be incorporated in the present algorithm to improve performance of a system.

REFERENCES

1. N.Hoque,D.K Bhattacharyya,J.K Kalita."MIFS-ND : A mutual information-based feature selection method",Elssevier-Expert System with Application 41(2014) 6371-6385.

2.Qinbao Song,Jingiie Ni and Guangtao Wang,"A Fast Clustering- Based Subset Selection Algorithm For High Dimentional Data", IEEE Transaction on knowledge and data engineering vol. 25 No:1 (2013).

3. Lei Yu,Huan Liu,"Efficient Feature Selection via Analysis of Relevance and Redundancy",Journal of Machine Learning Research 1205-1224,2005.

4.Lei Yu,Huan Liu,"Efficiently Handling Feature Redundancy in High- Dimentional Data",ACM-2003.

5. Andreas G.K.Janecek,Wilfried N.Gansterer,"on the Relationship Between Feature Selection and Classification Accuracy",JMLR,2008.

6. Yu L. and Liu H., Feature selection for high-dimensional data: a fast correlation-based filter solution, in Proceedings of 20th International Conference on Machine Leaning, 20(2), pp 856-863, 2003.

7. Yu L. and Liu H, Efficiently handling feature redundancy in highdimensional

data, in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '03). ACM, New York, NY, USA, pp 685-690, 2003.

8. Yu L. and Liu H, Redundancy based feature selection for microarray data, In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 737-742, 2004.

9. M.Dash, H.Liu, and H.Motoda, "Consistency Based Feature Selection", Proc. Fourth Pacific Asia Conf. Knowledge Discovery and Data Mining, pp. 98-109, 2000.

10. M. Dash and H. Liu, "Consistency-Based Search in Feature Selection", Artificial Intelligence, vol. 151, nos. 1/2, pp. 155-176, 2003.

11. M.Dash,H.Liu,"Feature Selection for Classification",Elsevier-Intelligent Data Analysis,Vol. 1,no. 3,1997.

12. A. Arauzo-Azofra, J.M. Benitez, and J.L. Castro, "A Feature Set Measure Based on Relief", Proc. Fifth Int Conf. Recent Advances in Soft Computing, pp. 104-109, 2004.

13.H. Liu, H. Motoda, and L. Yu, "Selective Sampling Approach to Active Feature Selection", Artificial Intelligence, vol. 159, nos. 1/2, pp. 49-74, 2004.

14.Battiti,"Using Mutual Information for Selecting Features in Supervised Neural Net Learning", IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537-550, July 1994.

15.C.Krier, D.Francois, F. Rossi, and M. Verleysen, "Feature Clustering and Mutual Information for the Selection of Variables in Spectral Data", Proc. European Symp. Artificial Neural Networks Advances in Computational Intelligence and Learning, pp. 157-162, 2007.

16.Z. Zhao and H. Liu, "Searching for Interacting Features", Proc. 20th Int Joint Conf. Artificial Intelligence, 2007

17.R. Butterworth, G. Piatetsky-Shapiro, and D.A. Simovici, "On Feature Selection through Clustering," Proc. IEEE Fifth Conf. Data Mining, pp. 581-584, 2005.

18.Guyon I. and Elisseeff A., An introduction to variable and feature selection,Journal of Machine Learning Research, 3, pp 1157-1182, 2003.

19.UCI machine learning repository https://archive.ics.uci.edu