BENCHMARKING ATTRIBUTE SELECTION TECHNIQUES FOR MICROARRAY DATA

S. DeepaLakshmi[1] and T. Velmurugan[2]
[1]Bharathiar University, Coimbatore, India
[2]Department of Computer Science, D. G. Vaishnav College, Chennai, India
E-Mail: deepa.dgvc@gmail.com

ABSTRACT

Feature selection helps to improve prediction quality, reduce the computation time, complexity of the model and build models that are easily understandable. Feature selection removes the irrelevant and redundant features and selects the relevant and useful features that provide an enhanced classification results as the original data. This research work analysis the performance of the clustering and genetic algorithm based feature selection (CLUST-GA-FS) algorithm. The proposed algorithm CLUST-GA-FS has three stages namely irrelevant feature removal, redundant feature removal, and optimal feature generation. The algorithm involves removing the irrelevant features, removing redundant features by constructing a minimum spanning tree, splitting the minimum spanning tree into cluster, finding the representative feature from each cluster and finally finding the optimal set of features using genetic algorithm. CLUST-GA-FS algorithm is compared with the existing filter feature selection methods Fast correlation based feature selection (FCBF), Correlation based feature selection (CFS), Information gain (Infogain) and ReliefF. The work uses three microarray dataset Leukemia, Colon and Arcene that are high dimensional.

Keywords: feature selection, feature clustering, genetic algorithm, filter methods, microarray data.

1. INTRODUCTION

Data Mining is the process of extracting nuggets of knowledge and interesting patterns from large volumes of data. Feature selection or extraction is a technique that transforms and simplifies the data to make data mining tasks easier [1]. Feature selection involves removing irrelevant features, redundant features and selecting the relevant and optimal features [2]. Feature selection improves the comprehensibility of the classifier models. This plays a vital role when a huge number of features are managed and the learning algorithm loses prediction capacity using all of them. Feature selection reduces the dimensionality of the data and the data mining algorithms can be effectively used [3]. With the rapid growth of internet, data are available in different formats-text, multimedia, spatial, spatiotemporal, data streams etc. from different sources. Various machine learning applications deal with big data of large volumes and ultrahigh dimensional data. The high dimensionality requires huge memory space and high computational cost is incurred in training [4]. Feature selection is a process of selecting a subset of features that retain enough information for obtaining good or better performance results [5].

Three general classes of feature selection algorithms are filter, wrapper and embedded methods. Feature selection methods rank features based on the score obtained by applying a statistical measure to each feature[5]. Filter methods use independent criteria to evaluate the subset of features without using a learning algorithm[6]. Filters can be univariate that considers one variable at a time or multivariate that considers more than one variable at a time [7]. FCBF, CFS, Infogain, ReliefF and Markov blanket feature selection are some of the filter methods. Wrapper feature selection methods select a subset of features using the learning algorithm as part of the evaluation function[8]. Wrapper methods can be either deterministic that is simple or randomized that is prone to be stuck in local optima. Sequential forward selection, sequential backward selection, hill climbing and genetic algorithms are some of the wrapper methods[9]. Embedded methods combine both filter and wrapper methods. It uses independent criteria to decide optimal subset and a learning algorithm to select the final optimal subset [10]. The proposed algorithm CLUST-GA-FS is an embedded feature selection method. The work presented in this paper compares performance of the proposed algorithm with existing filter methods and wrapper methods.

Mining High dimensional data has the challenge of the ‘curse of dimensionality’ making classification a difficult task. Many research works have been proposed in the literature for feature subset selection of high dimensional data. Some of the approaches involving filter feature selection methods are discussed in this chapter. Baris Senliol *et al.* proposed a feature selection method FCBF# which selects any given size of feature subset [11]. Microarray dataset of high dimensional data are used and it is shown that FCBF# accuracy values are higher than FCBF. Correlation based feature selection(CFS) was proposed by Mark A. Hall [[12] that identifies irrelevant, redundant and noisy attributes. It was evaluated on artificial and natural datasets. The accuracy obtained is better than the accuracy obtained using complete set of features. Xiaofei He *et al.* proposed a filter method using laplacian score. The laplacian score for each feature is computed to reflect its locality preserving power[13]. A new feature selection method for ultrahigh-dimensional data was proposed which is different from traditional gradient approach and solves a sequence of multiple kernel learning subproblems [14].

The paper is organized as follows: in chapter 2 the various filter methods are explained and in chapter 3 the proposed algorithm CLUST-GA-FS is discussed. Chapter 4 compares the various filter algorithms, and the proposed algorithm CLUST-GA-FS.

2. FILTER METHODS

Filter methods are simple and fast and scale easily to high dimensional dataset. Filter methods rely on various measures of the data such as distance, information and consistency. Filter methods are usually a good choice for high dimensional dataset. Some of the filter methods are discussed below.

Fast correlation based feature selection (FCBF) is a filter method that identifies relevant features and redundant features [15]. A feature is strongly relevant if it is always necessary for an optimal subset and a feature is weakly relevant if it is not always necessary but becomes necessary at certain conditions. A feature is irrelevant if it is not necessary at all. FCBF consist of two steps. The first step selects a subset of relevant features and the second step selects predominant features from relevant features. Symmetrical uncertainty is the correlation measure used to find the relevant and redundancy features. Symmetrical uncertainty is defined as

$$\mathrm{SU(X,Y)} = \frac{2*\mathrm{MI(X,Y)}}{\mathrm{H(X)+\ H(Y)}} \qquad (1)$$

Where MI(X, Y) is the mutual information and X and Y are features. H(x) is the entropy of the feature X and H(Y) is the entropy of the feature Y.

CFS(Correlation based feature selection) algorithm identifies irrelevant, redundant and noisy features [12]. It also identifies features that are relevant as long as their relevance does not strongly depend on other features. The user need not specify the threshold or the number of features to be selected. It selects features that are highly correlated with the class but uncorrelated with each other. Pearson's correlation coefficient ρ is the correlation measure used when all the attributes are numeric.

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \qquad (2)$$

Where X and Y are the features or attributes, μ_x is the mean of X and μ_y is the mean of Y, σ_x is the standard deviation of X and σ_y is the standard deviation of Y and E is the expectation.

Information gain measures the amount of information in bits about the class prediction, where the presence of a feature and the class distribution is given [16]. The information gain is a measure of the reduction in entropy of the class variable after the value for the feature is observed. Information gain is a purely information-theoretic measure and it does not consider any actual classification algorithms. Information Gain IG is given by

$$H(X) = -\sum p(x)\log_2(p(x)) \qquad (3)$$

$$H(Y/X) = -\sum p(y)\sum p(x/y)\log_2(p(x/y)) \qquad (4)$$

$$IG = H(X) - H(Y/X) \qquad (5)$$

Where Entropy of a feature X is H(X), entropy of feature X after observing Y is H(Y/X).

ReliefF algorithm is a non-parametric feature weighting algorithm which is not limited to two class problems and it is more robust and can handle incomplete and noisy data [17]. The relevance of a feature is the average relevance of the feature across all training samples. It chooses the instances randomly and changes the weights of the feature relevance based on the nearest neighbor.

3. CLUST-GA-FS ALGORITHM-A PROPOSED METHOD

The proposed feature selection algorithm CLUST-GA-FS has three components: irrelevant feature removal, redundant feature removal, and optimal feature generation [18]. The first component removes the irrelevant features by using mutual information, a filter method. The second component removes the redundant features by choosing the representatives from each cluster. The genetic algorithm is used as the third component to find the optimal set of features. The irrelevant feature removal obtains features relevant to the class by eliminating the features which are irrelevant to the target class using mutual information. Redundant feature removal removes redundant features in 3 steps: Constructing a minimum spanning tree from the relevant features, grouping the features in the forest into clusters and selecting the representative feature from each cluster. The set of a representative feature from each cluster, the class variable and the number of features desired is provided as input to a genetic algorithm. Genetic algorithm (GA) is an adaptive heuristic search that uses optimization techniques to find true or approximate solutions based on the evolutionary ideas of natural selection and genetics [19]. GA begins with a set of chromosomes called the population. New populations are evolved by mutating solutions according to their fitness value. The fitness function used in this proposed method is based on the principle of min-redundancy and max-relevance. The mutual information is defined as follows:

$$MI(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \qquad (6)$$

Where X and Y are two features, p(x,y) is the joint probability distribution function of X and Y, p(x) and p(y) are the probability distribution functions of X and Y. The relevance of a feature set S is given as

$$D(S,C) = \frac{1}{|S|}\sum_{f_i \in S} I(f_i, C) \quad (7)$$

The redundancy of all features in S is given as

$$R(S) = \frac{1}{|S|^2}\sum_{f_i f_j \in S} I(f_i, f_j) \quad (8)$$

The minimum redundancy-maximum-relevance (mRMR) is given as

$$\text{Fitness Function:} mRMR = max_S\left[\frac{1}{|S|}\sum_{f_i \in S} I(f_i, C) - \frac{1}{|S|^2}\sum_{f_i f_j \in S} I(f_i, f_j)\right] \quad (9)$$

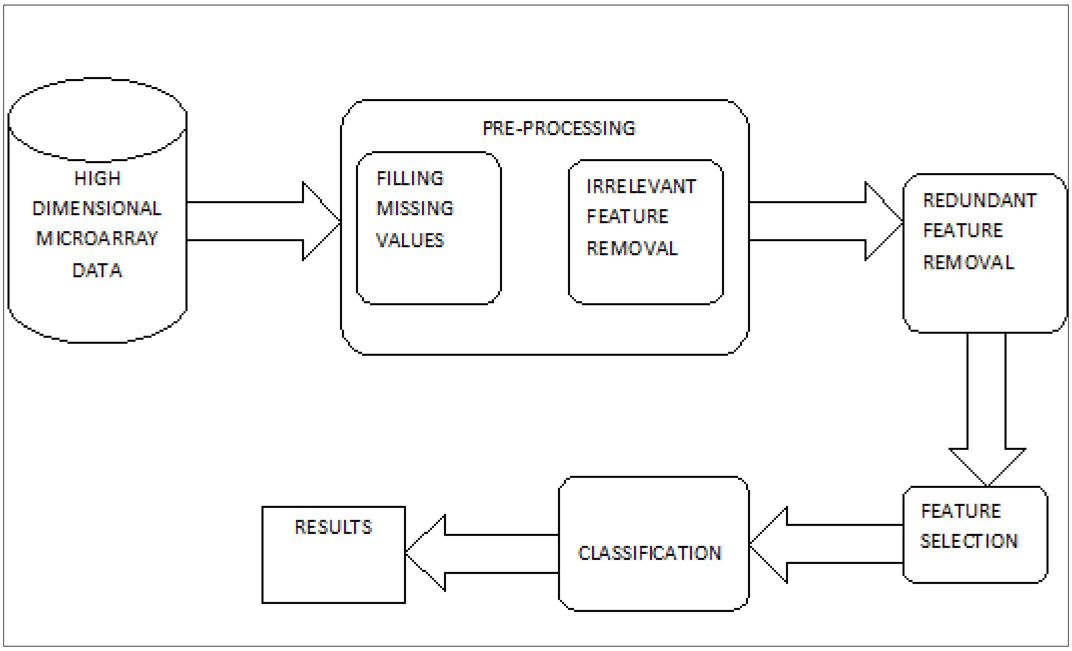
HIGH DIMENSIONAL MICROARRAY DATA
PRE-PROCESSING
FILLING MISSING VALUES
IRRELEVANT FEATURE REMOVAL
REDUNDANT FEATURE REMOVAL
FEATURE SELECTION
CLASSIFICATION
RESULTS

Figure-1. Architecture of CLUST-GA-FS Algorithm.

The first component removes the irrelevant features from the dataset by computing the mutual information of the features and removing the features whose mutual information value is less than the threshold value. A graph G is constructed using the features and mutual information between the features as the vertices and edges respectively. A minimum spanning tree of the graph is constructed in which the features are grouped into clusters. A forest is constructed from the minimum spanning tree by deleting the edge whose mutual information MI(f'_i,f'_j) are smaller than MI(f'_i,C) and MI(f'_j,C) where f'_i and f'_j are pair of features and C is the class variable. The feature which has the maximum mutual information value in each cluster is selected as the representative feature.

The set of representative features from each cluster in the forest is given as input to a genetic algorithm that generates an optimum set of features. The population is composed of the representative feature set which forms the chromosomes. The fitness function used is minimum redundancy maximum relevance. The chromosome generated during each generation is evaluated using joint conditional entropy. If the fitness function remains constant over the generations, the process is terminated and the optimal set of features is generated.

Proposed Algorithm: CLUST-GA-FS

```
Inputs: Data set S{$f_1$,$f_2$,….,$f_m$ C}, C-class, ϴ - threshold value
Output: representative feature subset
//=====Part 1: Irrelevant feature removal=======
Step 1:  for each feature $f_i$ in the data set
compute MI($f_i$,C)
if MI($f_i$,C)<ϴ
remove the feature $f_i$
end
end
relevant feature set F'={$f'_1$,$f'_2$,….,$f'_l$}(l≤m).

//===== Part 2: Removal of Redundant feature=====

Step 2: for each feature $f_i$ in F'
construct graph G=(V,E) with feature $f_i$ as vertices
and MI($f_i$,$f_j$) as edges
end
generate the minimum spanning tree of G
forest=minimum spanning tree of G
for each edge in forest
if MI($f_i$,$f_j$ )<MI($f_i$,C ) and MI($f_i$,$f_j$ )<MI($f_j$,C)
remove edge from the forest
end
```

```
end
for each tree in the forest
find the maximum MI($f_i$,C)
select $f_i$ as the representative feature of the cluster
end
representative feature set is F''={$f'_1$,$f'_2$,….,$f'_k$}(k≤l).
//======= Part 3: Generating Optimum set of features using Genetic Algorithm======

Step 3: Input: Representative feature set F''={$f'_1$,$f'_2$,….,$f'_k$}, Class C, desired no of features
Output: Optimal set of features, sel
Max_gen=no.of generations desired
find the entropy of F''- $H_f$ and C- $H_C$ and mutual information between the features - $MI_{ff}$
class C- $MI_{fC}$
generate the population consisting of the feature set
while generation is less than max_gen
find the fitness function=$max_S\left[\frac{1}{|F''|}\sum_{f_i \in F''} MI(f_{i,}c) - \frac{1}{|S|^2}\sum_{f_i f_j \in F''} MI(f_{i,}f_j)\right]$
rearrange the population according to their fitness values
create a new generation
if the chromosomes generated are identical
sel=population rearranged
end
end
sel=optimal set of features
```

Steps1 and 2 generate the relevant set of features and remove the redundant feature. The representative set of features is given as input to a genetic algorithm that determines the optimal set of features.

4. EXPERIMENTAL RESULTS

The Proposed algorithm CLUST-GA-FS is executed using MATLAB software and the results are verified by some of the classification algorithms. This verification is done by using WEKA software. Three classification algorithms are employed to classify datasets. The classifiers used are Naïve Bayes, C4.5, Jrip and Adaboost. CLUST-GA-FS and the existing filter methods FCBF, CFS, Infogain and ReliefF are implemented on publicly available three microarray datasets. The performance of the algorithms is compared based on the number of features selected and the accuracy of classification. The description of the data set and the performance metrics are discussed.

4.1 Data source

The datasets used for the evaluation of algorithms contains microarray data taken from the UCI repository. The number of features of the data sets varies from 2000 to 10000 and the number of classes is 2. The data sets used are Leukemia, Arcene and Colon dataset. Arcene contains a maximum of 10001 features with 200 instances and Colon has 2000 features with 62 instances. The data set used is given in the Table-1.

Table-1. Data set description.

S. No.	Data set	No. of features	No. of instances	No. of classes	Domain
1	Colon	2000	62	2	Microarray
2	Leukemia	7129	38	2	Microarray
3	Arcene	10001	200	2	Microarray

4.2 Minimum spanning tree and forest construction

The proposed algorithm consists of three steps - removing irrelevant features, removing redundant features and selecting the optimal features. On removing the irrelevant features using mutual information, the redundant features are removed by constructing a minimum spanning tree as shown in the Figure-2 and constructing a forest from the minimum spanning tree as shown in Figure-3. The minimum spanning tree and the forest constructed for the Arcene dataset is shown in the Figure-2 and Figure-3. The minimum spanning tree is constructed with 1067 nodes and 1045 edges. The forest is constructed with 1066 nodes and 609 edges.

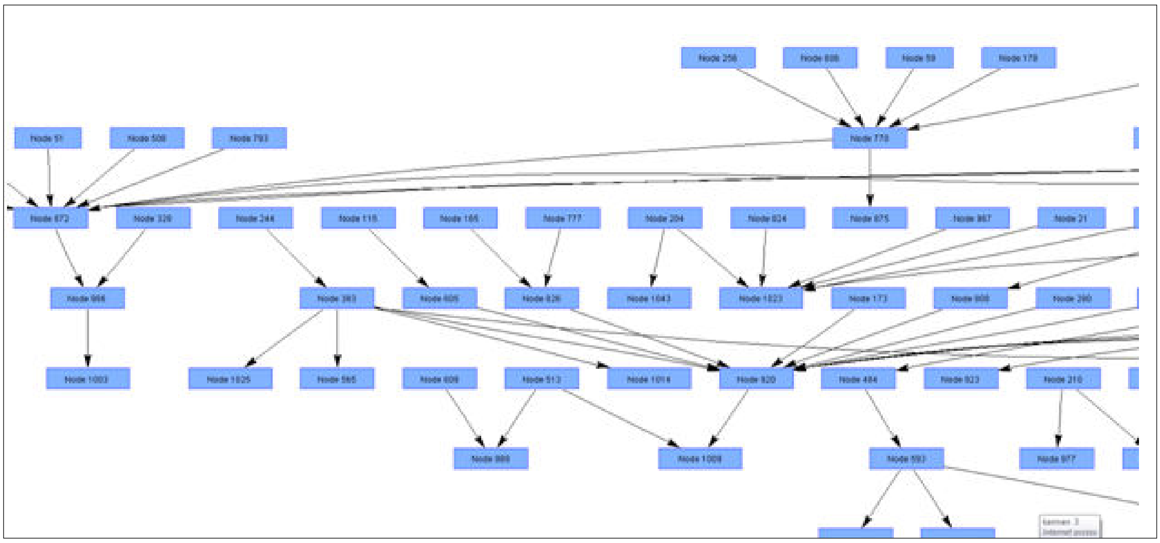

Figure-2. Minimum spanning tree of Arcene dataset.

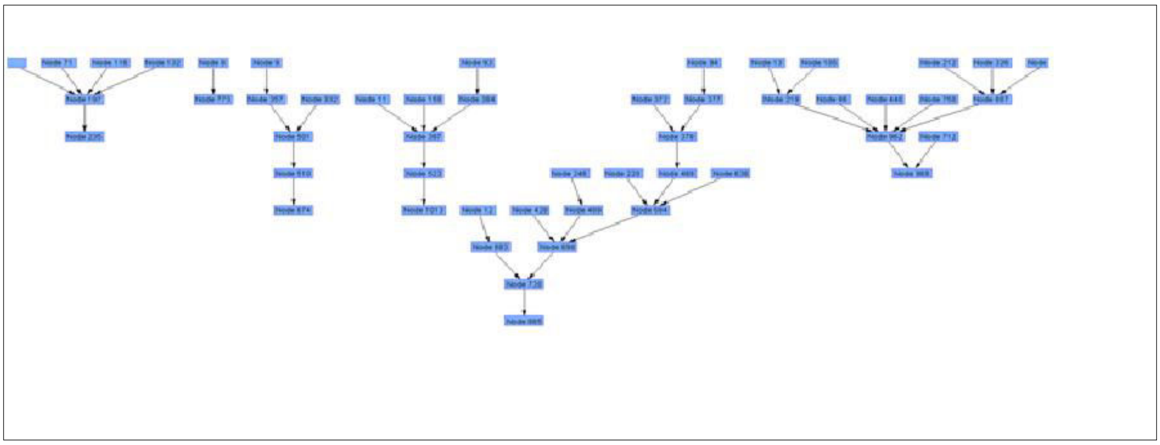

Figure-3. Forest of Arcene dataset.

4.3 Proportion of features

CLUST-GA-FS algorithm selects optimal features and minimum set of features from the chosen high dimensional data set. The proportion of the selected features is specified in the Table-2. The total number of features for each data set and the number of features selected by the filter methods FCBF, CFS, Infogain, ReliefF and the proposed algorithm CLUST-GA-FS is specified in the Table-2. It can be seen from Figure-2 that the number of features selected by the proposed algorithm and ReliefF for Leukemia dataset is less than other methods. For Arcene dataset, the number of features selected by the proposed algorithm is the least of all other filter methods. For Colon dataset, FCBF selects the least number of features.

Table-2. Proportion of features.

Dataset	CLUST-GA-FS	FCBF	CFS	InfoGain	ReliefF
LEUKEMIA	30	52	75	35	30
COLON	50	27	42	55	50
Arcene	32	34	35	35	36

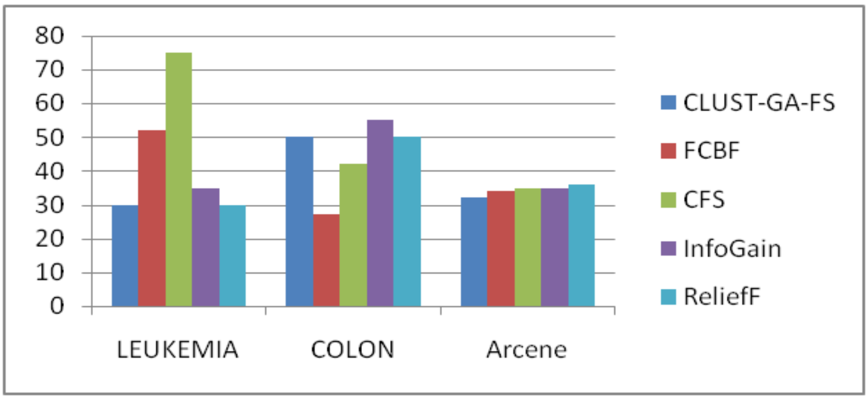

Figure-4. Proportion of the selected features.

4.4 Performance evaluation of CLUST-GA-FS algorithm

To analyze the performance of the proposed algorithm, classifiers Naïve Bayes, C4.5, Jrip and Adaboost were used to classify the dataset with the selected proportion of features. The performance of Filter methods FCBF, CFS, Infogain and ReliefF for the dataset specified was analyzed using WEKA tool. The performance of classifier naïve bayes is given in the Table-3. It can be seen that the accuracy of the classifier using the features selected by the proposed algorithm CLUST-GA-FS is comparable to the accuracy obtained using the features selected by the filter methods.

Table-3. Result of Naïve Bayes for accuracy.

Dataset	CLUST-GA-FS	FCBF	CFS	InfoGain	ReliefF
LEUKEMIA	91.18	91.17	85.29	91	91.17
COLON	87.1	90.32	87.09	85.48	88.7
Arcene	70.59	68	74.5	67.5	66.5

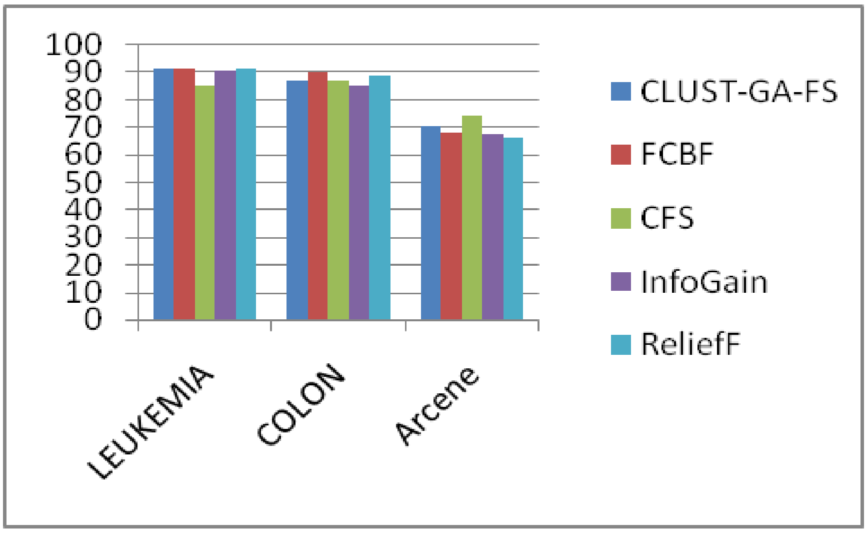

Figure-5. Accuracy of microarray dataset using Naïve Bayes.

Table-4 shows that the accuracy obtained using C4.5 classifier for the leukemia dataset using the proposed and the filter algorithms are the same. For colon dataset, the classifier has classified with more accuracy using the features selected by the proposed algorithm. For arcene dataset, accuracy using the proposed algorithm is better than the filter algorithms except infogain.

Table-4. Result of C4.5 for accuracy.

Dataset	CLUST-GA-FS	FCBF	CFS	InfoGain	ReliefF
LEUKEMIA	91.17	91.17	91.17	91.17	91.17
COLON	85.48	75.8	77.41	77.41	74.19
Arcene	75	73	66.5	81	72

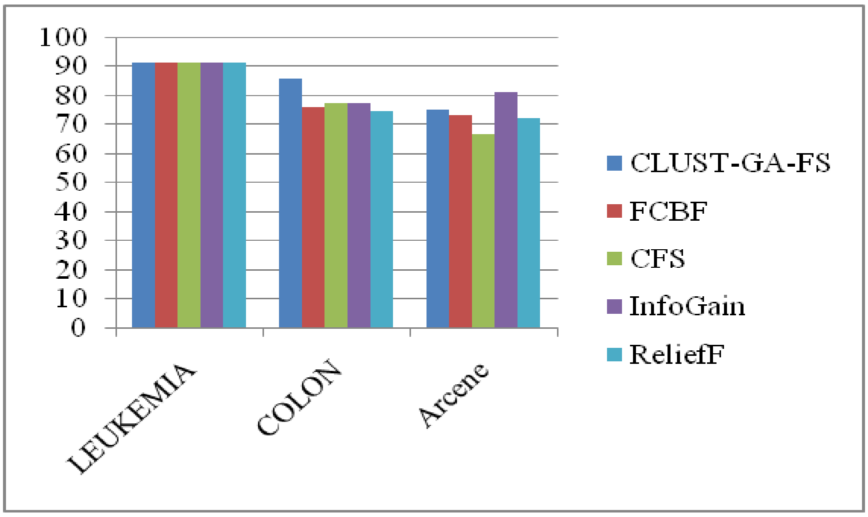

Figure-6. Accuracy of microarray dataset using C4.5.

The Jrip classifier performance has been improved by the proposed and the filter algorithms. For Colon dataset, the Jrip classifier has obtained more accuracy using the proposed algorithm. For Arcene dataset, the accuracy obtained using the features selected by Infogain is higher than all the other algorithms. The performance of the Jrip classifier is given in Table-5.

Table-5. Result of JRIP for accuracy.

Dataset	CLUST-GA-FS	FCBF	CFS	InfoGain	ReliefF
LEUKEMIA	91.17	91.17	73.52	91.17	91.17
COLON	80.64	72.58	70.96	77.41	66.12
Arcene	69.11	70	69	71.5	69

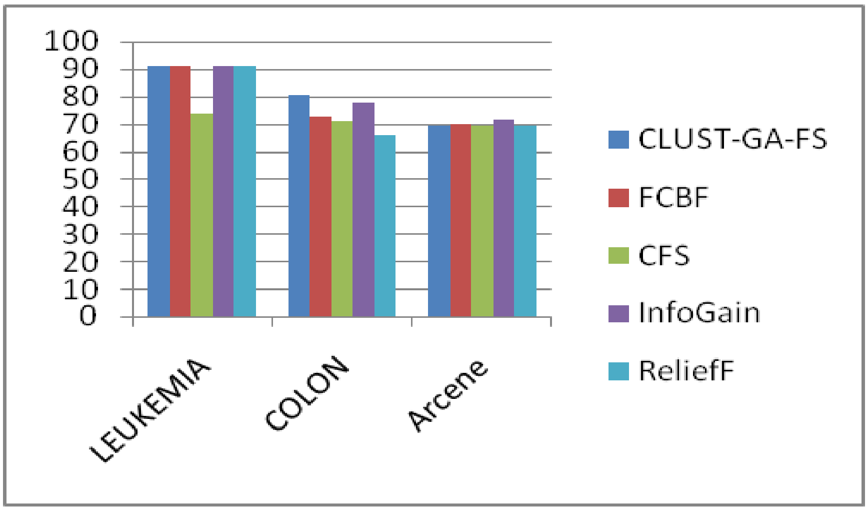

Figure-7. Accuracy of microarray dataset using JRip.

Adaboost classifier performance has been improved by the proposed algorithm and it has obtained a higher accuracy for Leukemia, Colon and Arcene dataset as shown in Table-6. The classifiers Naïve Bayes, C4.5, Jrip and Adaboost performance has been increased using the features obtained using the proposed CLUST-GA-FS algorithm and specifically classifier Adaboost has had the best performance for all the microarray dataset specified.

Table-6. Result of Adaboost for accuracy.

Dataset	CLUST-GA-FS	FCBF	CFS	InfoGain	ReliefF
LEUKEMIA	91.18	91.17	91.17	91.17	91.17
COLON	83.87	82.25	82.25	77.41	75.8
Arcene	80.89	71	74	71	68.5

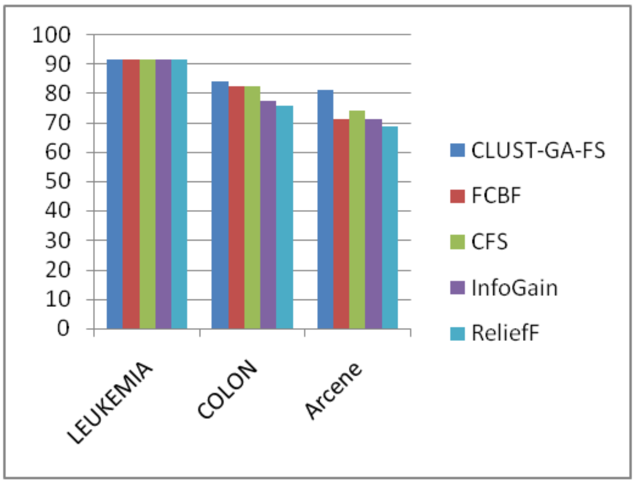

Figure-8. Accuracy of microarray dataset using adaboost.

FCBF, CFS, Infogain and ReliefF are some of the popular filter methods used for feature selection. The proposed algorithm CLUST-GA-FS has selected minimum set of features from high dimensional datasets. The number of features selected is less than the features selected by the filter methods for the Leukemia and Arcene dataset. The accuracy obtained by the classifiers Naïve Bayes, C4.5, Jrip and Adaboost using the features selected by the proposed algorithm is comparable to the number of features selected by the filter methods specified.

5. CONCLUSIONS

This research work is carried out to analyze the performance of the proposed CLUST-GA-FS algorithm based on the number of features selected and the accuracy of the classifiers. The algorithm removes the irrelevant features using mutual information, removes redundant features by constructing a minimum spanning tree, splitting the generated minimum spanning tree into clusters and selecting a representative feature from each cluster. The set of representative features is given as input to a genetic algorithm to find the optimal set of features. The proposed CLUST-GA-FS algorithm has selected an optimal set of features from large dimensional microarray dataset. The performance of the classifiers Naïve Bayes, C4.5, Jrip and Adaboost using the features obtained using the popular filter methods FCBF, CFS, Infogain and ReliefF and the proposed algorithms are analyzed. The classifiers Naïve Bayes, C4.5 and Jrip have obtained good accuracy for the proposed algorithm and it is comparable with the accuracy obtained for the filter methods specified. It is also observed that the classifier Adaboost has outperformed and has obtained best classification accuracy for the features generated by the proposed algorithm than the other classifiers.

REFERENCES

[1] H. Liu and H. Motoda. 1998. Feature extraction, construction and selection: A data mining perspective. Vol. 453. Springer Science & Business Media.

[2] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. Journal of machine learning research. 5: 1205-1224.

[3] H. Liu and H. Motoda. 2012. Feature selection for knowledge discovery and data mining. Springer Science & Business Media. Vol. 454.

[4] E. Schadt, M. Linderman, and J. Sorenson. 2010. Computational solutions to large-scale data management and analysis. Nature reviews. Genetics. 11(9): 647.

[5] I. Guyon and A. Elisseeff. 2006. An introduction to feature extraction. Feature extraction. pp. 1-25.

[6] H. Liu and L. Yu. 2005. Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on knowledge and data engineering. 17(4): 491-502.

[7] S. DeepaLakshmi and T. Velmurugan. 2016. Empirical study of feature selection methods for high dimensional data. Indian Journal of Science and Technology. 9(39).

[8] G. John, R. Kohavi, and K. Pfleger. 1994. Irrelevant features and the subset selection problem. In: Machine

learning: proceedings of the eleventh international conference. pp. 121-129.

[9] M. Dash and H. Liu. 1997. Feature selection for classification. Intelligent data analysis. 1(1-4): 131-156.

[10] V. Kumar and S. Minz. 2014. Feature Selection. Smart CR. 4(3): 211-229.

[11] B. Senliol, G. Gulgezen and L. Yu. 2008. Fast Correlation Based Filter (FCBF) with a different search strategy. In Computer and Information Sciences, ISCIS'08. 23rd International Symposium on (pp. 1-4). IEEE.

[12] M. Hall. 1999. Correlation-based feature selection for machine learning. PhD Thesis, New Zealand Department of Computer Science, Waikato University.

[13] X. He, D. Cai and P. Niyogi. 2006. Laplacian score for feature selection. In Advances in neural information processing systems. pp. 507-514.

[14] M. Tan, I. Tsang, and L. Wang. 2014. Towards ultrahigh dimensional feature selection for big data. Journal of Machine Learning Research. 15: 1371-1429.

[15] L. Yu and H. Liu. 2003. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th international conference on machine learning (ICML-03). pp. 856-863.

[16] D. Roobaert, G. Karakoulas, and N. Chawla. 2006. Information gain, correlation and support vector machines. Feature extraction. pp. 463-470.

[17] W. Duch. 2006. Filter methods. Feature Extraction. pp. 89-117.

[18] S. Deepalakshmi and T. Velmurugan. 2017. A Clustering and Genetic Algorithm based Feature Selection (CLUST-GA-FS) for High Dimensional Data. International Journal of control theory and Applications. 10(23): 63-76.

[19] D. Fogel. 1994. An introduction to simulated evolutionary optimization. IEEE transactions on neural networks. 5(1): 3-14.