**Abstract**

Recent genomic studies and molecular biology imbibe microarray as a key component for expressing thousands of genes. A few challenges begot are high input dimensionality, heterogeneity in the data with a very low sample size. Ensemble feature selection method is known to allay this big data problem, improving computational time of a prediction model by selecting a subset of the genes from the entire set. This article presents a study on ensemble learning where efficacy of three individual feature selection methods is tested against their coalesced subset of recurring features for the training data. The results support the said ensemble method.

**Keywords**: Feature Selection, Ensemble method, Filter method, Microarray Data

**1. Introduction:**

The study of expression of cells and tissue is one of the major ways for discovery in medicines. Microarray is an efficient way to analyse gene expression data. Microarray gives a basic to genotype, thousands of different loci at a time. In general, there is an inordinate number of variables in gene expression data with unknown correlation structures [3]. This high dimensionality challenges class prediction done with typical statistical analysis methods. Here, the main problem with microarray data is it is not possible to research on large number of genes using traditional methods.To overcome this, gene subset selection has becomes prime importance to the researchers since long. Feature selection is a way to select this gene subset. Three different approaches namely filter, wrapper and embedded or hybrid are used for feature selection. Hybrid methods like ensemble filtering methods are also popular where different feature selection methods of same or different categories are combined to generate a subset with most relevant features [4].

In this article, three different filter methods namely, T-test [5], F\_classif [6] and Mutual Information (MI) [7] are tested on some microarray datasets individually, involving all the features to generate feature subsets. Another subset is then generated by combining the outcomes of the individual feature subset obtained in the preceding step. Finally this subset of features are used to classify the microarray data. Classification is done with the help of three different classifiers: K-Nearest Neighbours (KNN) [8], Support Vector Machine (SVM) [9] and Gaussian Naive Bayes (GNB) [10], of which KNN provides the best results. For evaluating the proposed ensemble approach, four different microarray datasets are used. The results corroborate that this system is proficient in achieving a better accuracy than the individual subsets generated from the said filter methods.

**2. Background**

**2.1 Classifiers**

Classification is the method of categorizing new sample into a group, based on a set of training data which trains the classifier. Following three popular classifiers used here.

2.1.1 KNN

This is a widely used non-parametric classifier used for both classification and regression predictive problems where the input consists of k-closest training examples. KNN [7][11] is particularly referred to as ‘lazy algorithm’ due to the fact that it does not resort to any generalization and retains all details while comparing every training sample to predict.

2.1.2 SVM

SVM [9] is a discriminative classifier formally defined by a separating hyperplane, selected to best segregate the points in the input variable space by their class. The supervised learning algorithm outputs an optimal hyperplane which categorizes new examples with help of labelled training data. In two-dimensional space, i.e. a plane surface this a discriminating line which crops the plane in two parts. This classifier tends to do well even with a limited amount of samples but is not always the best.

2.1.3 GNB

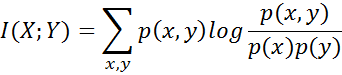
A form of naïve Bayes classifier (which are based on probabilistic model), this classifier works on continuous feature values under the assumption that all features are following a Gaussian distribution [10]. This operates primarily on Mean and Standard Deviation of training data i.e. the z-score of the dataset. Mostly Bayes classifiers are outperformed by other approaches, such as boosted trees or random forests. In addition that GNB requires only a small number of training data.

**2.2 Filter Methods**

Among the three feature selection approaches (Filter, Wrapper and Embedded), filter methods generally use different statistical approach to obtain the optimum subset of features with highest predictive powers not relying on any learning algorithm. Following three filter methods are used here.

2.2.1 Mutual Information (MI): This score measures the contribution of a variable towards reducing uncertainty about the value of another variable, much like a standard correlation analysis. Intuitively, MI [7] measures the information that variables share. For example, if two features are independent, then knowing one does not give any information about the other and vice versa, their MI score would be zero. But this method does not work so well with a low sample size.

Formally, MI of two discrete random variables X and Y can be defined as:{\displaystyle I(X;Y)=\sum \_{y\in Y}\sum \_{x\in X}p(x,y)\log {\left({\frac {p(x,y)}{p(x)\,p(y)}}\right)},}



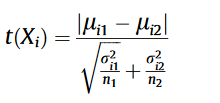
Where is the joint probability function of X and Y, and , and are the marginal probability distribution functions of X and Y respectively.

2.2.2 F\_Classif: This feature selection method employs Analysis Of Variance (ANOVA) using f-tests [6] to statistically test the equality of means of three or more groups, where f-test is based on f-statistic which is nothing but a ratio of variances.



Where, MSA is mean square among groups and MSW is mean square within groups.

2.2.3 T-test: T-test is another statistical based filter method that uses t-score to rank features. The t-score [4] gives a ratio of the difference between two groups and the difference within the groups. A large t-score implies the groups are different, whereas a small t-score implies the groups are more similar. T-score is given as:



where µi denotes mean of ith feature (Xi) and 𝜎I denotes standard deviation of ith feature (Xi) for some class C in a binary classification problem.

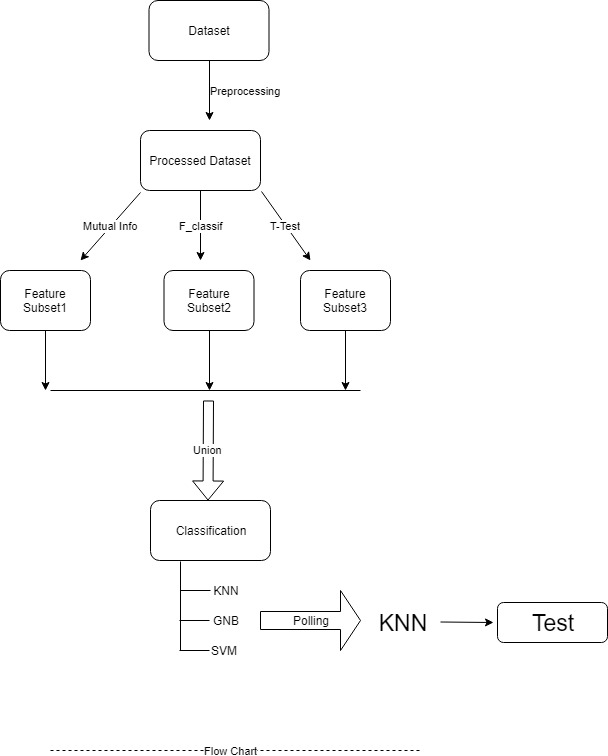
**3. Material and Methods**

An initial task is carried out with the datasets in which some values are missing. The missing values are filled with average of that attribute, so that it does not mislead the accuracy and class. Next for feature selection, three different filter methods (i.e. MI, F\_Classif, and T-test) are used and top-100 features from each filter method is selected. These features are used for ensemble method, so that final subset of features is significantly small compare to the original feature set.

Three subsets obtained by using the said methods are tested for common attributes and if the result is significantly large, then only these features are taken in consideration for classifying the microarray data. But in most of the cases we have found that number of common attributes is very small, hence, we combine these feature subsets to get the final subset. This subset is then used to generate the accuracy of the predicted model. Hence, union of these three subsets are used to decide the class of samples of microarray data. We have observed that these minimal set of features can predict the classes almost correctly.

Fig. 1 displays the flowchart of the proposed ensemble feature selection method. After preproceesing of datasets, univariate filters i.e. MI (Enropy based filter), Guassian Naive Bayes(GNB)[10] and T-score (Statistical based filter) are used for feature ranking. Feature ranking shows the importance of that attributes for determining class lables. We generally reduce the feature dimension so that processing times become minimal. But there is always a trade-off between class prediction efficiency and number of features.

We have inspected the result with the three different classifiers and KNN has seemed to be best among said three classifiers. KNN is a type of instance-based learning based on Euclidean distance between a test sample and the specified training samples [4][10].



**Fig 1: Flowchart of the proposed Ensemble Feature Selection method.**

**Results and Discussion:** The said ensemble feature selection method is applied on four different datasets taken from FRI Bio-lab (<http://www.biolab.si/supp/bi-cancer/projections/index.html>). The classification accuracy of the model with combined feature subset gives better accuracy. Also KNN is the selected classifier among the three as it outperforms others. Table 1 depicts dimension of dataset and Table 2 shows the comparison between the three classifiers on four datasets. Table 3 shows accuracy with different size training data with varying train to split ratio for classification.

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Number of instances** | **Number of features** |
| Leukemia | 72 | 5147 |
| Prostate | 102 | 12533 |
| DLBCL | 77 | 7070 |
| Lung Cancer | 203 | 12600 |

Table 1: Details of microarray datasets used for evaluating the proposed feature selection model.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Accuracy of Classifiers(%) | | |
| KNN | GNB | SVM |
| Prostate | 95.10 | 87.34 | 70.01 |
| DLBCL | 92.60 | 63.33 | 90.34 |
| Leukaemia | 92.40 | 95.20 | 65.44 |
| Lung Cancer | 92.40 | 87.64 | 63.66 |

Table 2: Accuracy of classifiers using multiple datasets. [Train to Test split ratio:80/20]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset** | **No of Features** | **Train to Test split ratio** | **Number of samples in train set** | **Number of samples in test set** | **Accuracy (%)** |
| Lung Cancer | 12600 | 70 / 30 | 142 | 61 | 85.67 |
| 80 / 20 | 162 | 41 | 88.60 |
| 90 / 10 | 182 | 21 | 92.47 |
| Leukemia | 5147 | 70 / 30 | 50 | 22 | 90.23 |
| 80 / 20 | 58 | 14 | 93.2 |
| 90 / 10 | 65 | 7 | **96.33** |
| Prostate | 12533 | 70 / 30 | 70 | 32 | 91.46 |
| 80 / 20 | 80 | 22 | 93.44 |
| 90 / 10 | 90 | 12 | 95.20 |
| DLBCL | 7070 | 70 / 30 | 54 | 23 | 91.66 |
| 80 / 20 | 62 | 15 | 93.75 |
| 90 / 10 | 70 | 7 | 94.75 |

Table 3: Prediction accuracy of the KNN classifier with varying number of train and test samples

Table 4 shows comparative study of accuracy of the model with ensemble method and without it. From the table it can observed that the number of features has reduced exponentially, which would in turn reduce the computation time of the prediction model as well. You have to analyze the results. At least add few lines saying the detail of the results you have obtained.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **No of Features** | **Accuracy without any feature selection method (%)** | **Accuracy with proposed feature selection method(%)** |
| Leukemia | 5147 | 86.60 | 93.33 |
| **Prostate** | **12533** | **82.20** | **95.20** |
| DLBCL | 7070 | 81.50 | 93.70 |
| Lung Cancer | 12600 | 82.90 | 92.40 |

Table 4: Comparison of efficiency of the selected feature vector obtained by the proposed method with the efficiency of the entire feature vector for classifying the microarray data.

**Conclusion:** In this work, we have proposed an ensemble feature selection method based three popularly used filter methods namely MI, F\_Classif and T-test. The efficacy of the proposed method is established by enhanced the prediction accuracy obtained for four microarray datasets. The results provided by this method are better than individual filter method when we consider all the features. We have obtained only some key features (about 2-3%) for each dataset using our method. Therefore, we conclude that even with low number of features (here, genes), significant accuracy can be obtained with this ensemble filtering technique. This ensemble technique is limited by the filter methods used to generate the subset. Also, a high degree of intersection among the subsets ensures that important features are selected. In future, the work can be improved by considering the Similarity based and Sparse learning based filter methods. Also, different approaches may be tested where features selected from one filter method can be fed to another filter method, i.e. nested filtering. This work still needs to be tested on different datasets.

**References:**

1: Yvan Saeys, Thomas Abeel, Yves Van de Peer: Robust Feature Selection Using Ensemble Feature Selection Techniques.

2.Adi L. Tarca, Roberto Romero and Sorin Draghici: Analysis of microarray experiments of gene expression profiling

3. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H,

Olson JA, Marks J, J N: Predicting the Clinical Status of Human Breast

4. Sajid Nagi, Dhruba Kr. Bhattacharyya: Classification of microarray cancer data using ensemble approach

5. Mohamad Morovvat and Alireza Osareh: An Ensemble of filters and wrappers for microarray data classification

6. A new Feature Selection method on classification of medical datasets: Kernel F score Feature Selection, Kemol Polat, Salih Gunes

7. Mutual Information, Peter E. Latham and Yasser Roudi (2009), Scholarpedia, 4(1):1658.

8.ML-KNN Lazy Learning Approach to multi-label Learning- Min-Ling Zhang, Zhi-hua Zhou, https://doi.org/10.1016/j.patcog.2006.12.019

9.Support Vector Machines- M.A.Hearest, S.T.Dumains, E.Osuna,IEEE Computer Society

10. Rajeev Raizada, Yune Sang Lee: Smoothness without Smoothing: Why Gaussian Naive Bayes Is Not Naive for Multi-Subject Searchlight Studies.

11. K-nearest neighbour, Leif E. Peterson (2009), Scholarpedia, 4(2):1883.