1. **INTRODUCTION**

Data mining is a multidisciplinary effort to extract knowledge from data. In the present day scenario, organisations (and the government) have an enormous amount of data that is often stored in databases and data warehouses. The problem with this data is its sheer size – it might include thousands of features, a majority of which might be irrelevant or redundant or both and such data is said to be suffering from the ‘curse of dimensionality’ [1]. Several datasets with high dimensionality have now become publically available on the Internet. This fact has posed an interesting challenge to the research community since for machine learning methods it is difficult to deal with a large number of features as the presence of redundant*/*irrelevant genes may deteriorate the performance of a classifier significantly and lead to an increase in costs and learning time.

Microarray technology allows measurement and assessment of thousands of genes at a time, which has greatly affected the research that involves the molecular basis of disease diagnosis. Many data mining techniques have been extensively applied for classification of cancer microarray data. Cancer microarray classification refers to the prediction of the cancer in a set of given genes determining whether the tissue is cancerous and if it is, then determining the type of cancer. In general, this data is characterized by a large number of genes (in thousands) and a small number of available samples (fewer than hundred) and hence suffers from the limitation called the curse of dimensionality. This problem can be solved by identifying a smaller set of genes accountable for a given disease. Since such genes are only a handful, there is a need to identify them. All the remaining genes are either redundant or irrelevant. In addition, observed that the presence of redundant*/*irrelevant genes might deteriorate the performance of a classifier significantly and lead to an increase in costs and learning time[6]. Hence, dimensionality reduction is a crucial step towards the removal of irrelevant and redundant genes and in the identification of a set of relevant genes responsible for a particular disease. This also helps in cancer diagnosis at a premature stage and in catalysing drug discovery for early cure of cancer. The reduction in dimensionality can be achieved by two ways, feature extraction and feature selection, and each of them has its own merits. The feature extraction technique achieves a reduction in dimensionality by combining the original features thereby creating a set of new features. These features are usually more compact and have a strong discriminating power. This method is used in applications like image processing and signal analysis, where model accuracy is more desirable than the model interpretability.

In machine learning, **feature selection**, also known as **variable selection**, **attribute selection** or **variable subset selection**, is the process of selecting a subset of relevant features for use in model construction. Feature selection techniques are used for three reasons:

* They simplify the data models making them easier for interpretation by the research community.
* Reduce the time used in training.
* Enhance generalization by reducing over fitting of data.

Feature selection techniques have become the focus of much research in areas of application where there are many features and comparatively few samples (or data points). These areas include gene selection from microarray data, text categorization and combinatorial chemistry. Other areas include study of data related to various diseases, financial data, geospatial data, etc., where the number of features range from a hundred to thousands. Researchers are realizing that in order to achieve successful data mining, feature selection is an indispensable component[6]. Feature selection techniques do not alter the original representation of the variables, but merely select a subset of them. Thus, they preserve the original semantics of the variables and hence offer the advantage of interpretability to the person who uses them. There are many potential benefits of feature selection: they facilitate data visualization and data understanding, reduce the measurement and storage requirements, reduce the time used in training and defy the curse of dimensionality to improve prediction performance[5].

Gene ranking method aims to retain a certain number of genes, especially by ranking threshold, with scores determined according to a measure of relevance, discriminatory capability, information content or quality index. Some of the commonly used ranking methods for gene selection are Mutual Information (Shannon & Weaver, 1949), Cochran test statistics [3], Adjusted Welch test statistics (Welch, 1951) and Brown-Forsythe test statistics[2]. In literature, each method is evaluated only on a handful of microarray datasets.

In this study, we have investigated five popular gene-ranking methods to determine a simple and efficient gene ranking method that can determine a smaller set of discriminatory set of genes to provide high classification accuracy. For this, we have considered four publicly available cancer microarray datasets that are considered challenging and are used by research community for the evaluation of their learning model.

**2. RELATED WORK**

In literature, many feature selection methods have been proposed for the classification of microarray data by the data mining and the pattern recognition community. Since the classification involves a high-dimensional feature set, the selection of relevant genes is imperative to improve the performance of the learning system. These techniques, depending on how they combine the feature selection search with the construction of the classification model, are divided into three categories: Filter Methods such as the Brown Forsythe test, the Cochran test, minimum redundancy maximum relevance (mRmR), etc.; Wrapper Methods such as the Sequential Forward Selection (SFS), the Sequential Backward Elimination (SBE) and Genetic Algorithms (GA); and Embedded Methods like the Decision Tree method.[1]

Filter techniques assess the relevance of features by looking only at the intrinsic properties of the data. In most cases, a feature relevance score is calculated, and low scoring features are removed. Afterwards, this subset of features is presented as input to the classification algorithm. Advantages of filter techniques are that they are easily scalable, are computationally simple and fast, and are independent of the classification algorithm. As a result, feature selection needs to be performed only once, and then different classifiers can be evaluated. It is disadvantageous in the sense that it ignores the interaction with the classifier. This means that each feature is considered separately, thereby ignoring feature dependencies, which may lead to worse classification performance when compared to other types of feature selection techniques. Such is the general case of univariate filter model (which takes into consideration only one feature at a time). In order to overcome the problem of ignoring feature dependencies, a number of multivariate filter techniques are present, aiming at the incorporation of feature dependencies to some extent.

In literature, some multivariate methods have been suggested to reduce redundancy among the selected set of genes. A good gene selection method not only selects the relevant genes but also reduces the redundancy amongst the selected gene subset. While univariate feature filters evaluate (and usually rank) a single feature at a time, the multivariate approach evaluates dependency between a set of features and the corresponding class variable.

**3. BACKGROUND**

**3.1 Feature Ranking Techniques**

Feature ranking is one of the most widely explored feature selection techniques because of its simplicity and computational efficiency. Many feature ranking methods are proposed in this endeavor. Feature ranking approaches evaluate features using some statistical characteristics of the data. A scoring function is generally used to measure either correlation between individual feature and target class or scatter of a given feature among data of different class.Let us assume that there are *k* (≥ 2) distinct classes for the problem under consideration and there are *p f*eatures/genes and n samples. Suppose *Xfs* is the measurement of the feature *f* from sample *s* for *f =*1, 2, ...., *p* and *s =*1, 2, ..., *n*. Data can be represented in terms of a matrix *X* where columns and rows of the matrix *X* correspond to samples and features respectively. The matrix *X* is given by

We assume that the data matrix *X* is standardized so that the features have mean 0 and variance 1 across samples. Given a fixed feature, let *Zij* be the feature from the *jth* sample of the *ith* class where *Zij* is obtained from the corresponding row of *X*. We consider the following general model for *Zij* :

*Zij = µi +Ɛij* for *i* = 1, 2, ..., *k ; j=* 1, 2, 3, …., ni

with n1+ n2 +…..+ nk  = n. In the model, *µi* is a parameter representing the mean value of the feature in class *i*, *Ɛij* are the error terms such that *Ɛij* are independent normal random variables, with expectation (E) and variance(V) given by

*E*(ε*ij* ) = 0 , *V*(ε*ij* ) = < ∞

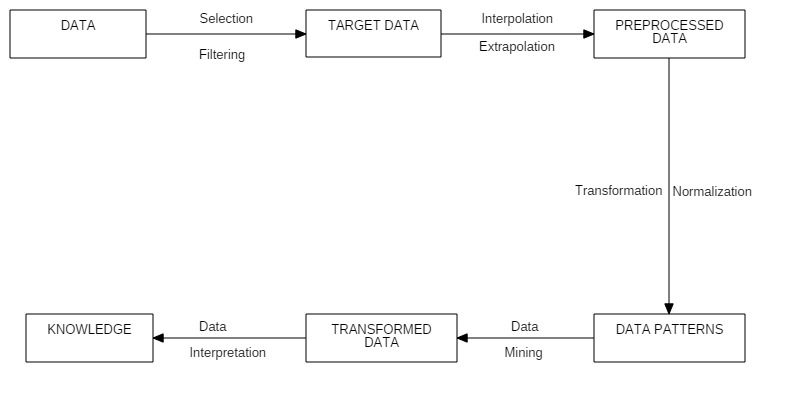
for j = 1, 2, ….., *ni i* = 1, 2, ……, *k*

We use the test statistics to determine the discriminating genes for microarray classification. Given a test statistic T, we define the discrimination power of a feature as the value of T evaluated at the n levels of the feature. This definition is based on the fact that with larger T the null hypothesis *H0 :* μ1 = μ2 = --- = μk will be more likely to be rejected. Therefore, the higher the discrimination power, the more powerful the feature is in discriminating between different sample classes. Finally, we choose those genes as salient features having a high power of discrimination. For the case of homogeneity of variances, the well-known ANOVA F-test is the optimal test to accomplish the task. However, with heterogeneity of the variances, the task is challenging. Therefore, some alternatives to the F-test are worthy of investigation.

**3.2 Embedded Technique**

In the embedded technique, the search for an optimal subset of features is built into the classifier construction, and can be seen as a search that includes the filtering algorithm to find the required feature subset along with the classifier. Just like the wrapper method, embedded approaches are thus specific to a given learning algorithm. They have the advantage that they include interaction with the classification model, while at the same time are far less computationally intensive than wrapper methods[5].

**4. MODEL ARCHITECTURE**

**4.1 Data Filtering Mechanism**

*Fig 4.1*

**5. SYSTEM REQUIREMENTS**

**5.1. System Requirement Specification**

**5.1.1 Software Requirements**

Operating System : Windows 8/10

User Interface : MATLAB

Programming Language : MATLAB

Technologies : MATHWORKS

**5.1.2 Hardware Requirements**

Processor : Pentium IV

Hard Disk : 40GB

RAM : 256MB

**6. PROBLEM DEFINITION**

**6.1. Problem Statement**

Datasets are usually of large size which includes many features, a majority of which are irrelevant or redundant, so the irrelevant features are needed to be filtered. Datasets of higher dimension are usually very complex.

**6.2. Main Objectives**

**6.2.1.** To apply feature selection techniques to maximize relevance and minimize redundancy

of abstracted datasets.

**6.3. Sub Objectives**

**6.3.1.** To use the test statistics and visualize the discriminating genes for microarray classification.

**6.3.2.** Investigate the performance of these ranking methods on microarray datasets regarding health care. The performance is evaluated in terms of classification accuracy and number of genes.

**6.4. Defining of Objectives**

**6.4.1 To prepare the programs executing the ranking methods.**

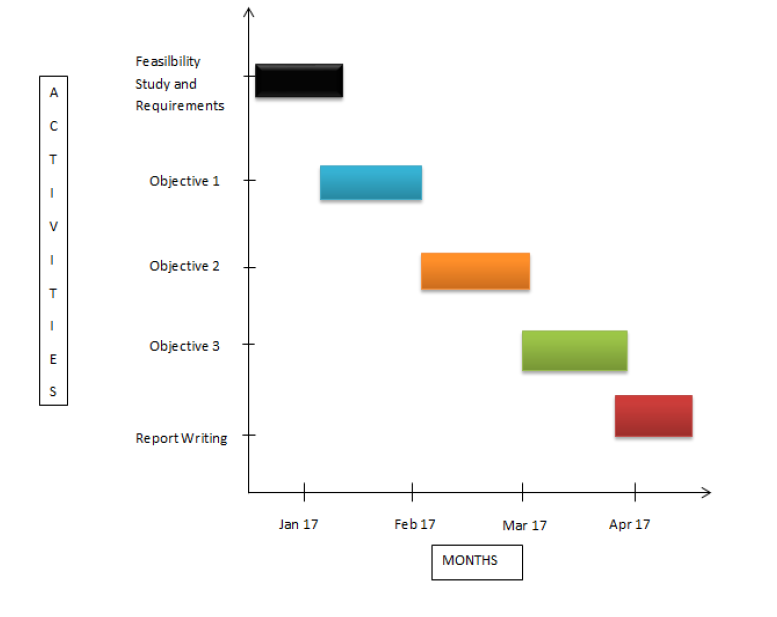
We have investigated five popular gene-ranking methods to determine a simple and efficient gene ranking method that can determine a smaller set of discriminatory set of genes to provide high classification accuracy.

**6.4.2. To plot the graphs and investigate the performance.**

We need to create a program to plot the graphs and observe the variations for the different cancer datasets.

.

**6.5 Pert chart**

****

*Fig 6.1*

**7. IMPLEMENTATION**

**7.1.** **The implementation of our project is as follows:**

In this study, we have investigated five popular gene-ranking methods to determine a simple and efficient gene ranking method that can determine a smaller set of discriminatory set of genes to provide high classification accuracy. For this, we have considered four publicly available cancer microarray datasets that are considered challenging and are used by research community for the evaluation of their learning model. We considered the following parametric test statistics.

**7.1.1. Brown- Forsythe Test Statistic (BF-Test)[2]**:

This statistic, given by Brown & Forsythe (1974), is computed as

where , and

Under H0, B is distributed approximately as Fk-1,y where

**7.1.2. Adjusted Welch Test Statistic (AW-Test):**

It is a variant of the Welch test statistics. Welch test statistics is given by

Where, and Here H0, Whas an approximate distribution of *F*k-1,yw, where

The Adjusted Welch statistics described on similar basis is defined by the formula

Where with ϕi chosen such that 1 ≤≤ (ni – 1)/(ni – 3) and  Under *H*0, *W\** has an approximate distribution of *Fk-1, vw\**, where

In this paper, we choose, since this choice provides reliable results for small sample sizes and a large number (*k*) of populations.

**7.1.3. Cochran Test Statistics (C-Test)[3]:**

This is the quantity appearing as numerator of the Welch test statistics W and is given by

Under H0, C has an approximate distribution of (Cochran, 1937).

**7.1.4. Mutual Information (MI):**

It is another important ranking method based on information theoretic approach and measures dependency among variables. For ranking of genes, mutual information is calculated between each gene and class label.

The mutual information for a gene vector *Xi* and the class vector *c* is given by

Where *P* (*Xi*) and *P*(*c*) are marginal probability distribution functions for random variables *Xi* and *c* respectively and *P( Xi , c)* is joint probability distribution. For maximum information relevance, the selected features *Xi* should have largest mutual information *I ( Xi , c)* for target class *c*, which indicates the largest dependency on the target class. The advantage of mutual information is that it can capture even nonlinear relationship between the gene and the corresponding class label *c*.

To check the efficiency and accuracy of the univariate techniques we have discussed, we did extensive experiments on four well-known and publicly available cancer microarray datasets and used these techniques to find the relevant and the number of such genes in the microarray data. A brief description of datasets is given in Table 7.1. The performance of the methods is evaluated in terms of the classification accuracy and the number of relevant genes and the method with the highest accuracy is determined for a particular dataset. While comparing two methods, the one with the higher classification accuracy is termed as the better performer while in the case of the same classification accuracy, the method with lesser number of selected genes is considered better. Two popular classifiers viz. K-nearest neighbour (KNN) and Support Vector Machine (SVM), commonly used by machine learning and data mining communities, have been used for evaluation.

**7.1.4.1 Description of datasets**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S. No. | Dataset | Original Genes | Samples | Classes |
| 1 | Colon | 2000 | 62 | 2 |
| 2 | Leukemia | 7128 | 72 | 3 |
| 3 | Pomeroy | 2277 | 34 | 2 |
| 4 | Prostate | 5966 | 102 | 2 |

*Table 7.1*

These data sets have different number of features, classes and instances. Heterogeneity of the data is important for checking the strength and weakness of the different techniques in the analysis of different microarray data. The classification accuracy of the datasets is given in terms of Leave One Out Cross Validation (LOOCV). Researchers have been working on a large number of datasets they’ve found that using the 60 top ranked genes to calculate the classification accuracy gives an optimal result. Hence each ranking method is applied to a dataset to obtain 60 top ranked genes. These top ranked genes are incrementally included one by one to develop the decision model. At every stage, classification accuracy of the test data is determined.

**7.2.** **Files included in project are**:

**7.2.1** accuracy.m

**7.2.2** Crunfileranks.m

**7.2.3** getAccuracy.m

**7.2.4** getFeatRanks.m

**7.2.5** runfileranks.m

**7.2.6** MI.m

**8. RESULTS AND OUTPUTS**

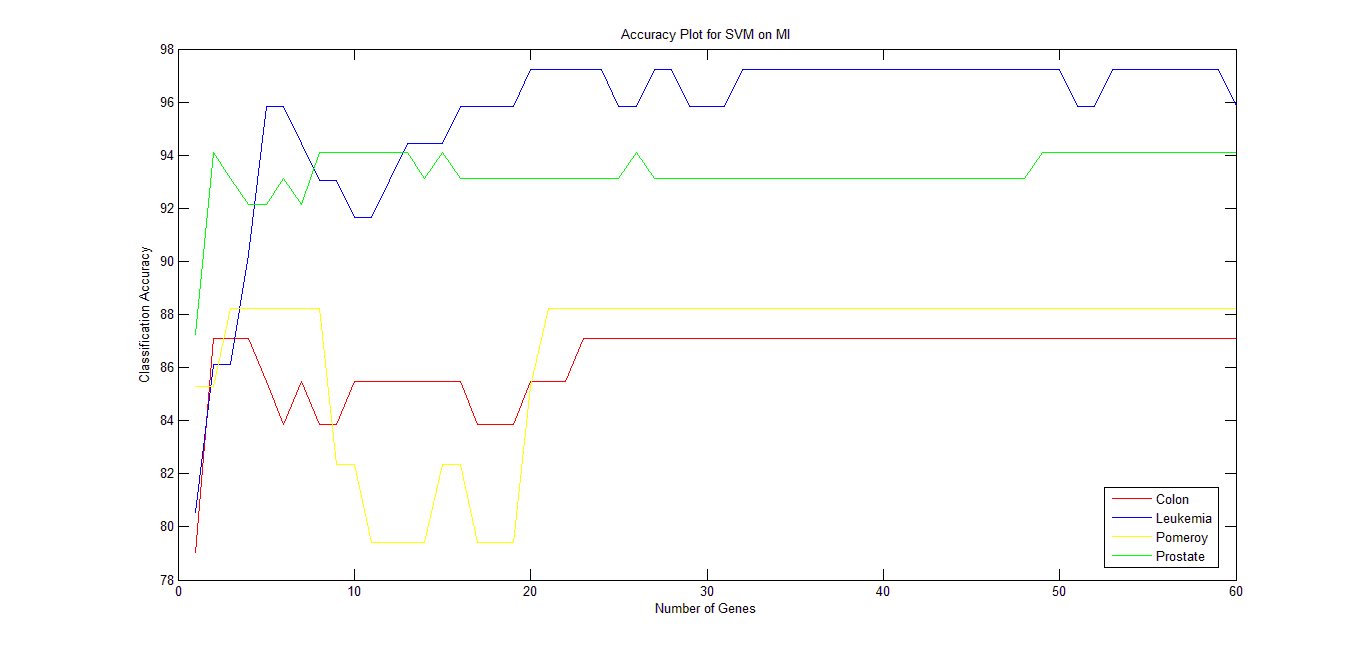
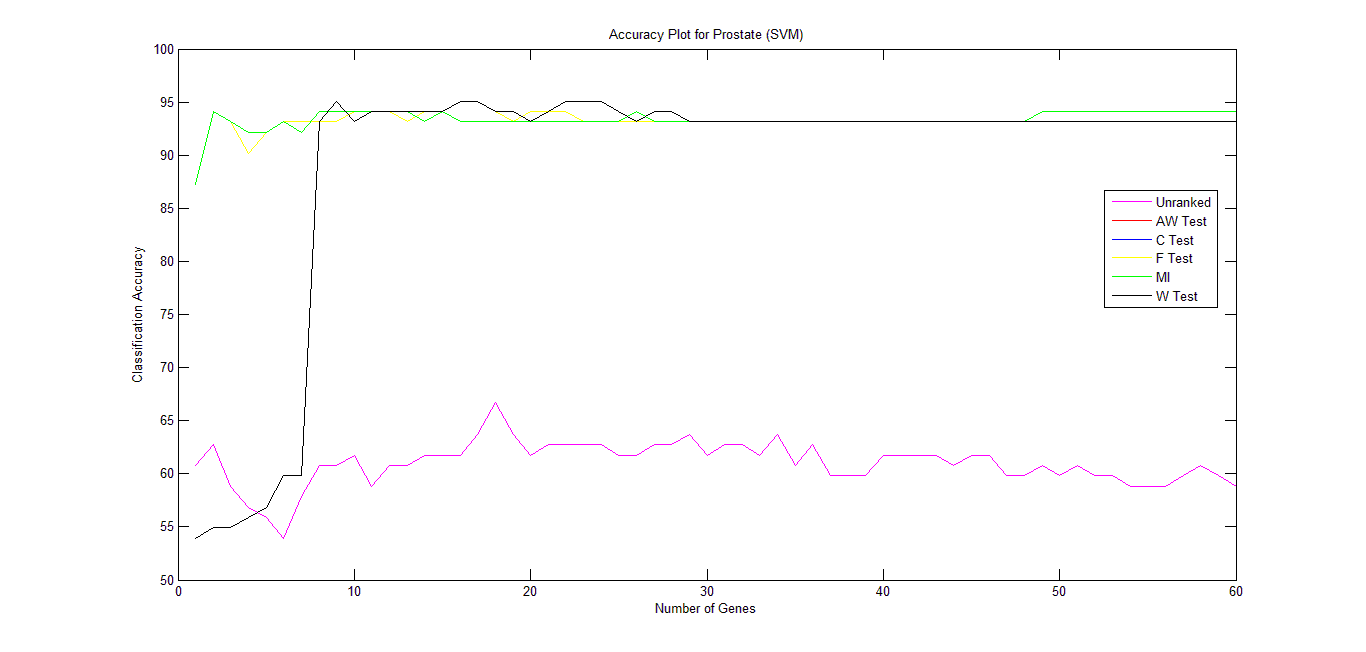
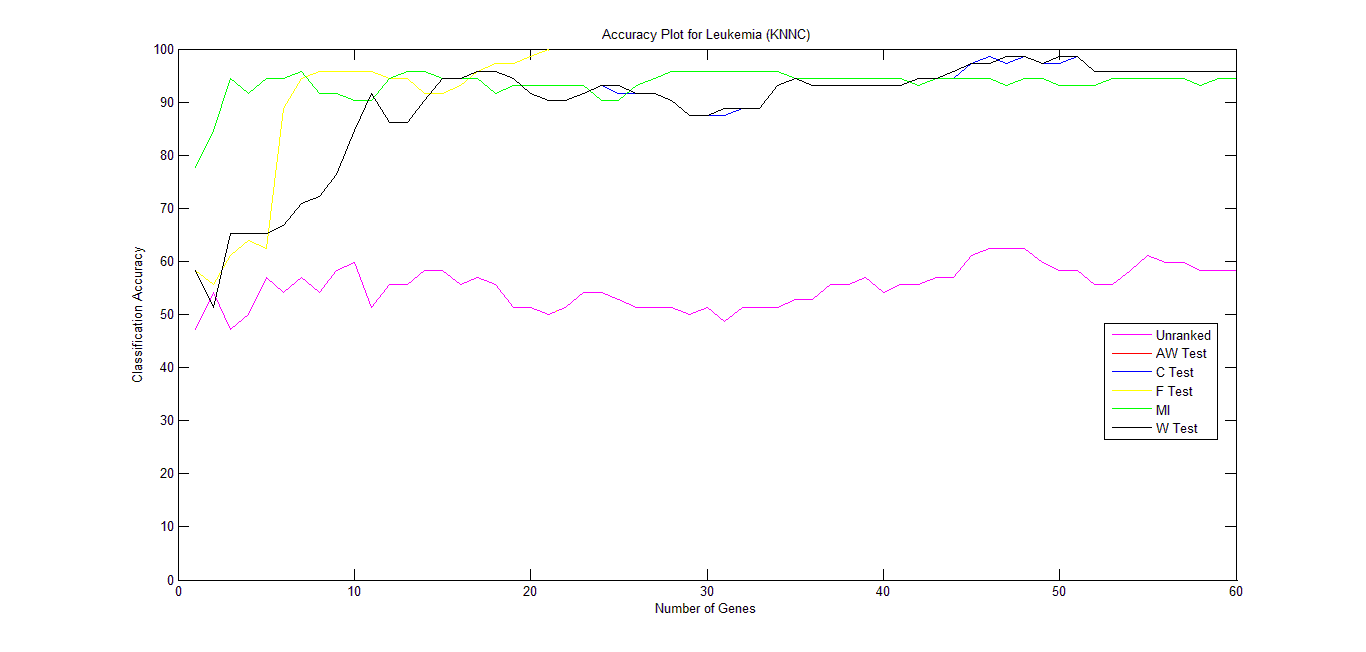
The classification accuracy of the datasets is given in terms of Leave One Out Cross Validation (LOOCV). Researchers have been working on a large number of datasets they’ve found that using the 60 top ranked genes to calculate the classification accuracy gives an optimal result. Hence each ranking method is applied to a dataset to obtain 60 top ranked genes. These top ranked genes are incrementally included one by one to develop the decision model. At every stage, classification accuracy of the test data is determined. Maximum classification accuracy obtained for each combination of ranking method and a classifier for a dataset is shown in Table 2. The number within parenthesis represents the number of genes corresponding to maximum classification accuracy for a given ranking method and a classifier. For each of the four datasets, combination of ranking and classification methods resulting in the maximum classification accuracy is shown in bold red in the same table.

**8.1 Percentage accuracy of different filter techniques with different classifiers**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Datasets** | **Without Feature Selection** | | **C Test** | | **F Test** | | **W Test** | | **AW Test** | | **MI** | | |
|  | **KNNC** | **SVM** | **KNNC** | **SVM** | **KNNC** | **SVM** | **KNNC** | **SVM** | **KNNC** | **SVM** | **KNNC** | **SVM** | |
| **Colon** | 77.41 (58) | 87.09 (40) | 83.87  (15) | **88.70**  **(9)** | 83.87  (15) | **88.70**  **(9)** | 83.87  (15) | **88.70**  **(9)** | 83.87  (15) | **88.70**  **(9)** | 85.48  (15) | 87.09  (2) |
| **Leukemia** | 62.50 (46) | 70.83  (52) | 98.61  (46) | 98.61  (45) | **100.0**  **(21)** | 98.61  (18) | 98.61  (47) | 98.61  (44) | 98.61  (47) | 98.61  (44) | 95.83  (28) | 97.22  (20) |
| **Pomeroy** | 73.52 (59) | 73.52  (44) | 94.11  (16) | 94.11  (19) | 94.11  (6) | 94.11  (9) | 94.11  (16) | 94.11  (19) | 94.11  (16) | 94.11  (19) | **97.05**  **(31)** | 88.23  (3) |
| **Prostate** | 67.64  (24) | 66.66 (18) | 93.13  (21) | **95.09**  **(9)** | 93.13  (35) | 95.09  (16) | 93.13  (21) | **95.09**  **(9)** | 93.13  (21) | **95.09**  **(9)** | 95.09  (15) | 94.11  (2) |

*Table 8.1*

**8.2 Classification accuracy graphs**



*Fig 8.1*

The following is observed from Figure 8.1 and Table 8.1:

1. For each dataset, there is a significant variation in classification accuracy, which depends on the choice of the ranking method and the classifier.
2. The classification accuracy achieved using any filter feature selection technique is always better than the case where no such technique is applied.
3. There is no clear winner amongst the five ranking methods used. However, the best performance for most of the datasets was achieved by the F-test and MI.
4. The classification accuracy for Colon and Pomeroy is almost similar by the C-test, F-test, W-test and the AW-test. The accuracy for Prostate is similar for the C-test, W-test and the AW-test.
5. All the tests used show stability in terms of performance i.e. once a high accuracy is reached, it is maintained.
6. The performance of MI is better with only a few number of genes in most of the cases, while other ranking methods show improvement as the number of genes increases.
7. SVM as a classifier performs much better than the KNNC in most of the cases.
8. The highest accuracy recorded while conducting the experiment was 100.0% when F-test is used on the Leukemia dataset with KNNC as the classifier.
9. KNNC gives better classification accuracy when used with MI than with the other tests, in most of the cases.
10. SVM shows a comparatively lower accuracy when used with MI when compared to the other four tests.
11. The tests show that as the number of samples of data increase, a better classification accuracy is achieved by each of the techniques used. Colon, with the least number of samples, has a lower classification accuracy when the feature selection techniques are used onto them in comparison to the case when the same techniques are used on the other datasets.
12. A higher accuracy is also achieved when the number of features in the dataset is high, like in the case of Leukemia and Prostate, when compared to Colon and Pomeroy.

**9. CONCLUSION**

**9.1 Work we have carried out**

Microarray gene expression data are often characterized by a large number of features and a small sample size. Hence, microarray cancer classification suffers from the curse of dimensionality. To overcome this problem, the research community has devised some feature selection methods in literature. These methods determine a minimal set of relevant features that help in improving classification accuracy and computational cost. In the previous section, we discussed and analysed various univariate filter techniques and carried out extensive experiments on four publically available cancer microarray datasets. Our experiments have demonstrated that for each dataset there is significant variation in classification accuracy that depends on the choice of the ranking method and the classifier. The best performance is achieved with the F-test and the Mutual Information followed closely by the other three tests. The number of features and the number of samples present in the dataset have a direct impact on the classification accuracy when the feature selection techniques are applied on them.

**9.2 Future**

The analysis done can help to find the maximum accuracy in any new sample of cancer microarray dataset.

**10. REFERENCES**

[1] Bellman, R.: Adaptive Control Processes. In: A Guided Tour, Princeton University Press Princeton (1961).

[2] Brown, M.B., Forsythe, A.B.: The small sample behavior of some statistics which test the equality of several means. Technometrics 16, 129–132 (1974).

[3] Cochran, W.: Problems arising in the analysis of a series of similar experiments. J. R. Stat. Soc. Ser. C. Appl. Stat. 4, 102-118

[4] H. Liu and H. Motoda: Feature Selection for Knowledge Discovery and Data Mining, London: Kluwer Academic Publishers

[5] Yvan Saeys, Inaki Inza and Pedro Larran: A Review of Feature Selection Techniques in Bioinformatics, Department of Plant Systems Biology, VIB, B-9052 Ghent, Belgium and Bioinformatics and Evolutionary Genomics group (2007).

[6] Isabelle Guyon and Andre Elisseeff: An Introduction to Variable and Feature Selection, Shoreline Boulevard, Mountain View, CA 94043, USA (2003).

[7] Chris Ding and Hanchuan Peng: Minimum Redundancy feature selection from Microarray Gene Expression Data, Computational Research Division and Life Sciences/Genomics Division, Lawrence Berkeley National Laboratory, University of California, Berkeley, CA, 94720, USA, (2003).