ISLAMIC UNIVERSITY OF TECHNOLOGY

A subset selection method using Filter and wrapper algorithms based on the nature of the expression values in microarray data sets for gene feature selection

Authors:

Tamzid Azad (104432)
Md. Mazharul Islam (104408)

Supervisor:

Prof. Dr. M. A. Mottalib
Head of the Department
Computer Science and Engineering

Co-Supervisor:

Md. Abid Hasan
Lecturer
Computer Science and Engineering Department.

A thesis submitted to the Department of CSE
In partial fulfillment of the requirements for the degree
B. Sc. Engineering in CSE
Academic Year: 2013-2014

A Subsidiary Organ of the Organization of Islamic Corporation
Dhaka, Bangladesh

Declaration of Authorship

This is to certify that the work presented in this thesis is the outcome of the Analysis and investigation carried out by Tamzid Azad and Md. Mazharul Islam under the supervision of Prof. Dr. M. A. Mottalib, Head, Computer Science and Engineering Department and Md. Abid Hasan, Lecturer Computer Science and Engineering Department, IUT, Dhaka, Bangladesh. It is also declared that neither of this thesis nor any part of this thesis has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Authors:

____

Tamzid Azad
Student ID – 104432

____

Md. Mazharul Islam
Student ID – 104408

____

Supervisor:
Prof. Dr. M. A. Mottalib
Head of the Department
Computer Science and Engineering
Islamic University of Technology (IUT)

Abstract

In the field of micro-array data analysis the crucial first step is gene selection. The process refers to selecting a subset consisting of a few genes which are of genetic significance out of thousands of genes to make the job of the classifier algorithm computationally easy and efficient at the same time.

Feature selection plays an important role in classification. The first set of data are gene expression profiles from Acute Lymphoblastic Leukemia (ALL) patients. In this paper an algorithm is proposed for feature subset selection (FFS) which is based on the nature of intensity values in microarray datasets. The proposed method is a combination of a filter and a wrapper algorithm which selects subsets. It is based on two assumptions. Our results demonstrate the importance of feature selection in accurately classifying new samples

Keywords: Gene expression profiles, Feature selection, Classification, Filter, Wrapper, Weka etc

Contents:

Chapter 1: Introduction

In the field of Bio-informatics gene feature selection holds a significant importance. It is because of the need of a balance between the computational ability and the aspiration for the perfect result.

1.1 Microarray data analysis:

In each type of cell, like a muscle cell or a skin cell, different genes are expressed (turned on) or silenced (turned off). If the cells that are turned on mutate, they could—depending on what role they play in the cell trigger the cell to become abnormal and divide uncontrollably, causing cancer. By identifying which genes in the cancer cells are working abnormally, doctors can better diagnose and treat cancer.

One way they do this is to use a DNA microarray to determine the expression levels of genes. When a gene is expressed in a cell, it generates messenger RNA (mRNA). Overexpressed genes generate more mRNA than under expressed genes This can be detected on the microarray. The first step in using a microarray is to collect healthy and cancerous tissue samples from the patient.

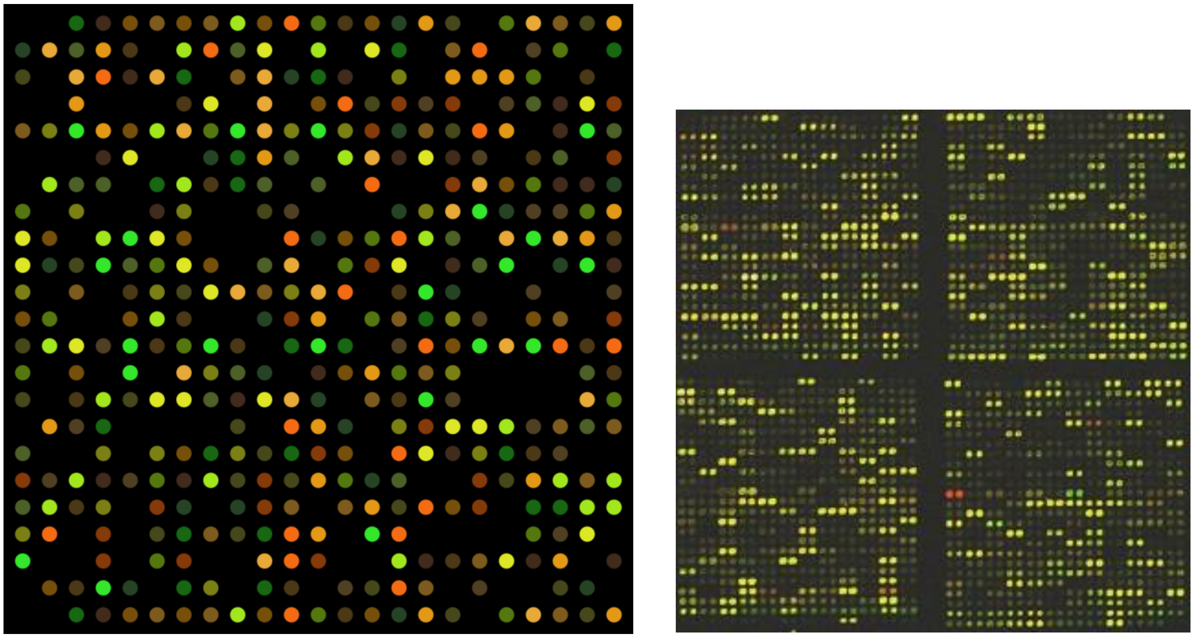


Figure-1.1: DNA microarrays

This way, doctors can look at what genes are turned on and off in the healthy cells compared to the cancerous cells. Once the tissues samples are obtained, the messenger RNA (mRNA) is isolated from the samples. The mRNA is color-coded with Fluorescent tags and used to make a DNA copy (the mRNA from the healthy cells is dyed green; the mRNA from the abnormal cells is dyed red.) The DNA copy that is made, called complementary DNA (cDNA), is then applied to the microarray. The cDNA binds to complementary base pairs in each of the spots on the array, a process known as hybridization. Based on how the DNA binds together, each spot will appear red, green, or yellow (a combination of red and green) when scanned with a laser.

1) A red spot indicates that that gene was strongly expressed in cancer cells.

2) A green spot indicates that that gene was strongly repressed in cancer cells

3) If a spot turns yellow, it means that that gene was neither strongly expressed nor strongly repressed in cancer cells.

4) A black spot indicates that none of the patient’s cDNA has bonded to the DNA in the gene located in that spot. This indicates that the gene is inactive.

1.2 Importance of Feature Selection:

From a microarray data we have an immense amount of data to calculate. Here feature selection becomes useful. Basically what feature selection algorithms do is to reduce the number of features in a dataset. But at the same time it should be taken into account that no significant or useful data is being lost. To meet these criteria numerous approaches and algorithms have been proposed and used. In this document we introduce another approach which makes the feature selection closer to this criterion.

1.3 Differential Gene Expression:

If the genome is the same in all somatic cells within an organism (with the exception of the above-mentioned lymphocytes), how do the cells become different from one another? If every cell in the body contains the genes for hemoglobin and insulin proteins, how are the hemoglobin proteins made only in the red blood cells, the insulin proteins made only in certain pancreas cells, and neither made in the kidney or nervous system? Based on the embryological evidence for genomic equivalence (and on bacterial models of gene regulation), a consensus emerged in the 1960s that cells differentiate through differential gene expression. The three postulates of differential gene expression are as follows:

- Every cell nucleus contains the complete genome established in the fertilized egg. In molecular terms, the DNAs of all differentiated cells are identical.
- The unused genes in differentiated cells are not destroyed or mutated, and they retain the potential for being expressed.
- Only a small percentage of the genome is expressed in each cell, and a portion of the RNA synthesized in the cell is specific for that cell type.

1.4 Problem Statement:

One of the main problems of high dimensional data is the inclusion of noisy and irrelevant data in the information set. Space and time complexity increases due to large number of noisy, redundant and uninformative gene expression. Reducing the dimensionality is the goal in feature selection.

1.5 Research Challenges:

Execution of a brute force exhaustive search is not encouraged due to high dimensional feature space. Therefore an optimal method is to be devised to achieve an accurate and efficient outcome. The desired outcome of the method is minimizing the number of features and increasing the predictive power of the classifiers. To add more intensity to the problem domain this field of bioinformatics produces inadequate testing and training samples. With the removal of noisy, irrelevant and redundant information the proposed method must be able to handle the correlation factor existing between the features and thus utilize the combined predictive power. Our proposed method encompasses all these factors and theoretically expects to bring about better results handling noisy, redundant and correlated data.

1.6 Motivation:

This study aims at deriving a better method for feature selection using a hybrid approach. This approach has an upper hand on other approaches as it does consider the collective measure of genes and not only focus to individuals. The approach inherits both the merits from filter and wrapper approaches. The adaptive wrapper approach gives better result taking less runtime than the other traditional wrapper approaches.

1.7 Scopes:

In case of research, to improve the search criteria for feature selection, two probable approaches can be taken. Firstly, the improvements can come from existing approaches. Secondly, it may come from generating new approaches. In this study, we have worked with a hybrid approach which is the combination of two algorithms. This approach tends to reduce some shortcomings of the existing methods which we will discuss later. Computer vision, Pattern recognition, Artificial intelligence etc. are the fields where feature selection can be of great use.

1.8 Research contribution:

In terms of contribution our research can add on the classification accuracy. The comparison between the state-of-the-art approaches is given in chapter 4. If the features can be classified with a higher accuracy it can add a great value to further researches and treatment.

1.9 Thesis Outline:

Our thesis outline is limited to the feature selection. The work that will be explained in this document is concerned with reducing the abount of data in a microarray dataset, Applying filter, wrapper methods, Comparing the results with the state-of-the-art etc. We have introduced a new approach combining filter and wrapper method. The wrapper method has been implemented by using software

named ‘weka’. However our future work motivation is about the classification techniques.

Chapter 2: Literature Review

2.1 Feature Subset Selection:

This technique measures the goodness of each found feature subset. A great deal of work has also been done Feature subset selection (for example Guyon et al., 2004, Ma & Huang, 2005, Ooi & Tan 2003). Feature Subset Selection techniques are more effective than FR techniques. In our study we have emphasized on FSS. FSS follows three basic methodologies:

- **Filter methods:** Filter methods work with the intrinsic characteristics of data. It reduces the size of a dataset by filtering the data in terms of their mean, variance, standard deviation etc. Filters don't work in iterations. The advantage of using filter methods is that it possesses the capability to radically decrease the size of the dataset by applying thresholds. On the other hand the disadvantage is that the risk of losing relevant data while doing so is rather high compared to other methodologies.

- **Wrapper methods:** Wrapper methods have more reliability than the filter methods to have the most relevant dataset because its integration with a classifier. Wrapper works in iterations. In each iteration it takes a subset and the classifier integrated to it evaluates the relevance of the data. Comparing between the subsets in terms of their relevance wrapper chooses the best subset. While we have a great many advantages of the wrapper methods it should be taken into consideration that the wrapper methods consume higher rates of computational resources than other methodologies.

- **Embedded Methods:** In Embedded methods the feature space is taken to search in order to find optimal subsets of features. The search procedure is built into the classifier in this variation. The optimal subset is selected using different algorithms such as Bayesian decision trees, SVM etc.

2.2 Filter approaches:

Usually in feature selection, subset is selected from a microarray dataset using the following methods:

- Student’s t test
- Z test
- S test
- Hill climbing

2.2.1 T-test:

- data sets should be independent from each other except in the case of the paired-sample t-test
- where n<30 the t-tests should be used
- the distributions should be normal for the equal and unequal variance t-test
- compares between two means to suggest whether both samples come from the same population.
- In feature selection assigns a score(IS: importance score) to each intensity value.

Using a threshold on IS selects the subset.

2.2.2 Z-test:

- Very little difference with t-test

- compares between two means to suggest whether both samples come from the same population.
- z-test is preferable when n is greater than 30.
- the distributions should be normal if n is low, if however n>30 the distribution of the data does not have to be normal
- the variances of the samples should be the same
- sample sizes should be as equal as possible but some differences are allowed

Most of the feature selection methods focuses on the supervised learning and in this study it's not different. An effective learning model is constructed based on supervised learning.

2.3 Feature Selection:

A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimises the error rate. This is an exhaustive search of the space, and is computationally intractable for all but the smallest of feature sets. The choice of evaluation metric heavily influences the algorithm. They are categorized below:

- Supervised Learning: generates a function mapping input to the desired output. Output is predetermined here.
- Unsupervised Learning: models a set of input but there is no mapping to desired output.
- Semi-supervised Learning: combines both labeled and unlabeled examples to generate an appropriate function or classifier.
- Reinforcement Learning: an observation of real world is given. Every action has some impact on the environment. Environment provides feedback that guides the learning algorithm.
- Transduction: predicts new output based on training input, training output and test input.
- Learning to Learn: learns its own inductive bias based on previous experience.

2.4 Feature Selection Techniques:

Feature selection can be applied on a set of features which can be a better solution than choosing all possible subsets of features [3] - 14. It is impractical if a large number of sub-sets are available. Feature selection can be classified into two broad categories:

2.4.1. Feature Ranking:

This is also known as feature weighing which assesses individual features and assigns them weights according to their degree of relevance. Many researches have been done with feature ranking as the base method (for example Bekkerman et al., 2003, Caruana and de SA, 2003, Weston et al., 2003).

In order to perform feature selection the feature space needs to be traversed i.e. feature searching. Feature searching involves going through the feature space to select features to be used for classification. Many approaches to feature selection exist which can be broadly classified into the following:

2.4.2 Exhaustive:

Exhaustive search or brute-force is a general problem-solving technique that traverses all the possibilities for the solution checking whether each of the candidates satisfies the solution criteria. Exhaustive search is easy to implement and will guarantee a solution if it exists. The downside of this method is its cost. Cost is proportional to the number of candidate solutions which tends to grow very rapidly as the size of the problem increases. Thus this approach should avoided when sample size is very large.

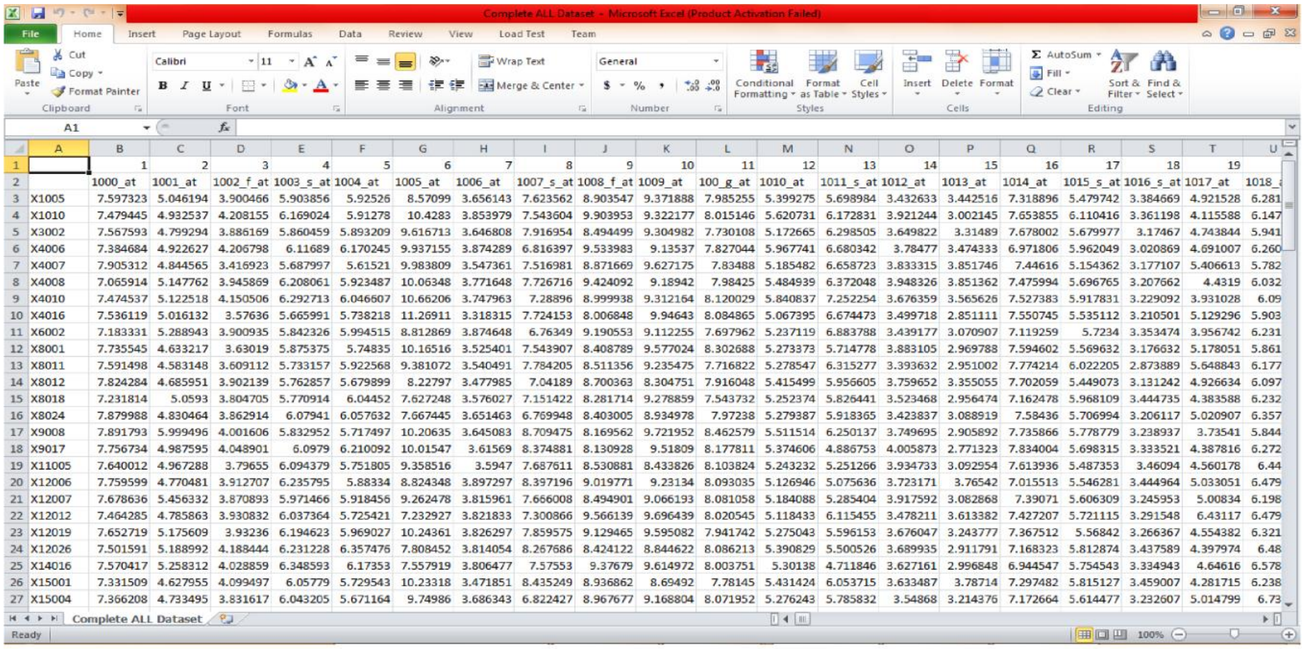
2.4.3 Best first:

Best first is a heuristic searching technique. It traverses through the candidate solution and selects that appears to be the best choice under the current situation and moves forward. But this approach does not ensure an optimum solution. An evaluation function defines the selection of a particular candidate

Chapter 3: Proposed Method

In the proposed method two algorithms have been applied a filter algorithm which Filters the Dataset and reduces the number of features based on the standard deviation values of the features and the students t-test. And then after reducing the number of features the broaden search wrapper algorithm has been applied to the reduced dataset.

3.1 The Dataset: The dataset that has been used for our implementation is an ALL Colon cancer dataset. The screenshot of the dataset-



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
2		1000_at	1001_at	1002_f_at	1003_s_at	1004_at	1005_at	1006_at	1007_s_at	1008_f_at	1009_at	100_g_at	1010_at	1011_s_at	1012_at	1013_at	1014_at	1015_s_at	1016_s_at	1017_at	1018_
3	X1005	7.597323	5.046194	3.900466	5.903856	5.92526	8.57099	3.656143	7.623562	8.903547	9.371888	7.985255	5.399275	5.698984	3.432633	3.442516	7.318896	5.479742	3.384669	4.921528	6.281
4	X1010	7.479445	4.932537	4.208155	6.169024	5.91278	10.4283	3.853979	7.543604	9.903953	9.322177	8.015146	5.620731	6.172831	3.921244	3.002145	7.653855	6.110416	3.361198	4.115588	6.147
5	X3002	7.567593	4.799294	3.886169	5.860459	5.893209	9.616713	3.646808	7.916954	8.494499	9.304982	7.730108	5.172665	6.298505	3.649822	3.31489	7.678002	5.679977	3.17467	4.743844	5.941
6	X4006	7.384684	4.922627	4.206798	6.11689	6.170245	9.937155	3.874289	6.816397	9.533983	9.13537	7.827044	5.967741	6.680342	3.78477	3.474333	6.971806	5.962049	3.020869	4.691007	6.260
7	X4007	7.905312	4.844565	3.416923	5.687997	5.61521	9.983809	3.547361	7.516981	8.871669	9.627175	7.83488	5.185482	6.658723	3.833315	3.851746	7.44616	5.154362	3.177107	5.406613	5.782
8	X4008	7.065914	5.147762	3.945869	6.208061	5.923487	10.06348	3.771648	7.726716	9.424092	9.18942	7.98425	5.484939	6.372048	3.948326	3.851362	7.475994	5.696765	3.207662	4.4319	6.032
9	X4010	7.474537	5.122518	4.150506	6.292713	6.046607	10.66206	3.747963	7.28896	8.999938	9.312164	8.120029	5.840837	7.252254	3.676359	3.565626	7.527383	5.917831	3.229092	3.931028	6.09
10	X4016	7.536119	5.016132	3.57636	5.665991	5.738218	11.26911	3.318315	7.724153	8.006848	9.94643	8.084865	5.067395	6.674473	3.499718	2.851111	7.550745	5.535112	3.210501	5.129296	5.903
11	X6002	7.183331	5.288943	3.900935	5.842326	5.994515	8.812869	3.874648	6.76349	9.190553	9.112255	7.697962	5.237119	6.883788	3.439177	3.070907	7.119259	5.7234	3.353474	3.956742	6.231
12	X8001	7.735545	4.633217	3.63019	5.875375	5.74835	10.16516	3.525401	7.543907	8.408789	9.577024	8.302688	5.273373	5.714778	3.883105	2.969788	7.594602	5.569632	3.176632	5.178051	5.861
13	X8011	7.591498	4.583148	3.609112	5.733157	5.922568	9.381072	3.540491	7.784205	8.511356	9.235475	7.716822	5.278547	6.315277	3.393632	2.951002	7.774214	6.022205	2.873889	5.648843	6.177
14	X8012	7.824284	4.685951	3.902139	5.762857	5.679899	8.22797	3.477985	7.04189	8.700363	8.304751	7.916048	5.415499	5.956605	3.759652	3.355055	7.702059	5.449073	3.131242	4.926634	6.097
15	X8018	7.231814	5.0593	3.804705	5.770914	6.04452	7.627248	3.576027	7.151422	8.281714	9.278859	7.543732	5.252374	5.826441	3.523468	2.956474	7.162478	5.968109	3.444735	4.383588	6.232
16	X8024	7.879988	4.830464	3.862914	6.07941	6.057632	7.667445	3.651463	6.769948	8.403005	8.934978	7.97238	5.279387	5.918365	3.423837	3.088919	7.58436	5.706994	3.206117	5.020907	6.357
17	X9008	7.891793	5.999496	4.001606	5.832952	5.717497	10.20635	3.645083	8.709475	8.169562	9.721952	8.462579	5.511514	6.250137	3.749695	2.905892	7.735866	5.778779	3.238937	3.73541	5.844
18	X9017	7.756734	4.987595	4.048901	6.0979	6.210092	10.01547	3.61569	8.374881	8.130928	9.51809	8.177811	5.374606	4.886753	4.005873	2.771323	7.834004	5.698315	3.333521	4.387816	6.272
19	X11005	7.640012	4.967288	3.79655	6.094379	5.751805	9.358516	3.5947	7.687611	8.530881	8.433826	8.103824	5.243232	5.251266	3.934733	3.092954	7.613936	5.487353	3.46094	4.560178	6.44
20	X12006	7.759599	4.770481	3.912707	6.235795	5.88334	8.824348	3.897297	8.397196	9.019771	9.23134	8.093035	5.126946	5.075636	3.723171	3.76542	7.015513	5.546281	3.444964	5.033051	6.479
21	X12007	7.678636	5.456332	3.870893	5.971466	5.918456	9.262478	3.815961	7.666008	8.494901	9.066193	8.081058	5.184088	5.285404	3.917592	3.082868	7.39071	5.606309	3.245953	5.00834	6.198
22	X12012	7.464285	4.785863	3.930832	6.037364	5.725421	7.232927	3.821833	7.300866	9.566139	9.696439	8.020545	5.118433	6.115455	3.478211	3.613382	7.427207	5.721115	3.291548	6.43117	6.479
23	X12019	7.652719	5.175609	3.93236	6.194623	5.969027	10.24361	3.826297	7.859575	9.129465	9.595082	7.941742	5.275043	5.596153	3.676047	3.243777	7.367512	5.56842	3.266367	4.554382	6.321
24	X12026	7.501591	5.188992	4.188444	6.231228	6.357476	7.808452	3.814054	8.267686	8.424122	8.844622	8.086213	5.390829	5.500526	3.689935	2.911791	7.168323	5.812874	3.437589	4.397974	6.48
25	X14016	7.570417	5.258312	4.028859	6.348593	6.17353	7.557919	3.806477	7.57553	9.37679	9.614972	8.003751	5.30138	4.711846	3.627161	2.996848	6.944547	5.754543	3.334943	4.64616	6.578
26	X15001	7.331509	4.627955	4.099497	6.05779	5.729543	10.23318	3.471851	8.435249	8.936862	8.69492	7.78145	5.431424	6.053715	3.633487	3.78714	7.297482	5.815127	3.459007	4.281715	6.238
27	X15004	7.366208	4.733495	3.831617	6.043205	5.671164	9.74986	3.686343	6.822427	8.967677	9.168804	8.071952	5.276243	5.785832	3.54868	3.214376	7.172664	5.614477	3.232607	5.014799	6.73

Figure 3.1- Screenshot of ALL Colon cancer dataset

3.2 Filter:

The filter that has been applied in our approach is a simplistic approach. It calculates the standard deviation of the features and applies students t-test on the features. The students t-test is a popular method for subset selection in data mining. It works in the following way-

T-test:

To apply the t-test the following rules has to be abided.

- Data sets should be independent from each other except in the case of the paired-sample t-test.
- Where n<30 the t-tests should be used.
- The distributions should be normal for the equal and unequal variance t-test.
- compares between two means to suggest whether both samples come from the same population.
- In feature selection assigns a score (IS: importance score) to each intensity value.

Using a threshold on IS selects the subset.

In the proposed model we see that the datasets are independent. The number of datasets is not more than 30. Basically we applied the t-test to only one dataset. The distribution in the dataset is normal in our model so it can be justified if we use the t-test in the proposed approach. Now the t-test actually calculates the importance scores for each of the features. And reduces or filters the data according to it.

3.3 Standard Deviation:

The standard deviation is a popular and simplistic tool to filter features where the variance of the values plays a significant role. In the given dataset we can see that the variance of sample values in a feature plays a big role as we are looking for those features which have differentially expressed genes. As in microarray dataset a single spot contains a whole feature the expression values in the feature will be significant when they contain values which are not alike. The nature of the expression values can be understood from the nature of the differentially expressed genes-

3.4 Differential Gene Expression:

If the genome is the same in all somatic cells within an organism (with the exception of the above-mentioned lymphocytes), how do the cells become different from one another? If every cell in the body contains the genes for hemoglobin and insulin proteins, how are the hemoglobin proteins made only in the red blood cells, the insulin proteins made only in certain pancreas cells, and neither made in the kidney or nervous system? Based on the embryological evidence for genomic equivalence (and on bacterial models of gene regulation), a consensus emerged in the 1960s that cells differentiate through differential gene expression. The three postulates of differential gene expression are as follows:

- Every cell nucleus contains the complete genome established in the fertilized egg. In molecular terms, the DNAs of all differentiated cells are identical.
- The unused genes in differentiated cells are not destroyed or mutated, and they retain the potential for being expressed.
- Only a small percentage of the genome is expressed in each cell, and a portion of the RNA synthesized in the cell is specific for that cell type.

The code where the standard deviation has been calculated is given in the appendix.

The students t-test calculation has been done using the built in function in MATLAB. By using both the students t-test and standard deviation a score has been obtained for each of the features. Then by using a threshold in that score the reduced number of features.

After applying the filter a reduced dataset has been obtained with 100 features.

3.5 Wrapper:

We have used the broaden search wrapper algorithm in the next step. The algorithm is based on the following ideas-

3.5.1 Problems associated with simple wrapper algorithms:

- The simple wrapper algorithms work in iterations. From the dataset they take subsets and check their classification accuracy. And then takes another subset. If the former subset has higher accuracy then the former subset is discarded. We felt this is problematic because the features' classification accuracy depends on each and every feature in the subset. So discarding the whole subset is unfair to the features which if combined with other features could've given better result.

- The best approach would be to calculate classification accuracy for all possible combinations. But it is a common knowledge that that is not a good option as the amount of computational time and resources it would take is immense.

The objective was to create a balance where the features were taken as much as possible and also taking the best combination out of the greater number of features.

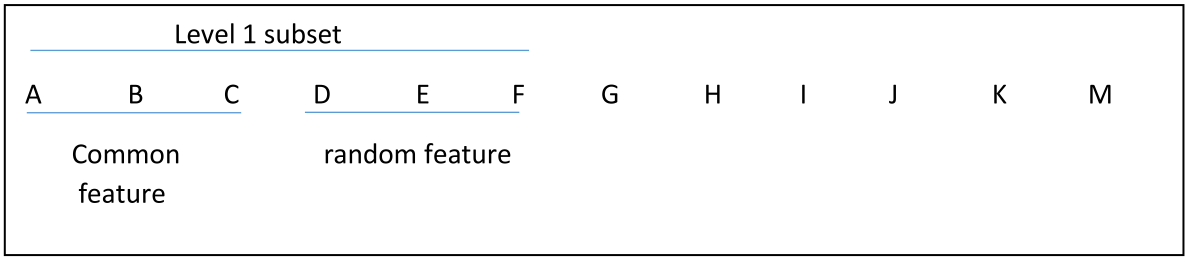
3.5.2 The Algorithm:

Our algorithm takes the subset from the greater set randomly at first. Each subset has two kinds of features in it.

Common features: The common features are those which will be used in the next subset of has been used in previous subsets.

Random features: The rest of the features are called random features.

In the given dataset the number of random features and common features in a subset is equally taken. The size of the subset is 10 and it is constant. In the 10 features there are 5 random features and 5 common features. To understand the concept the following figure helps-



The first subset is taken randomly then the next subset is taken in the following way-

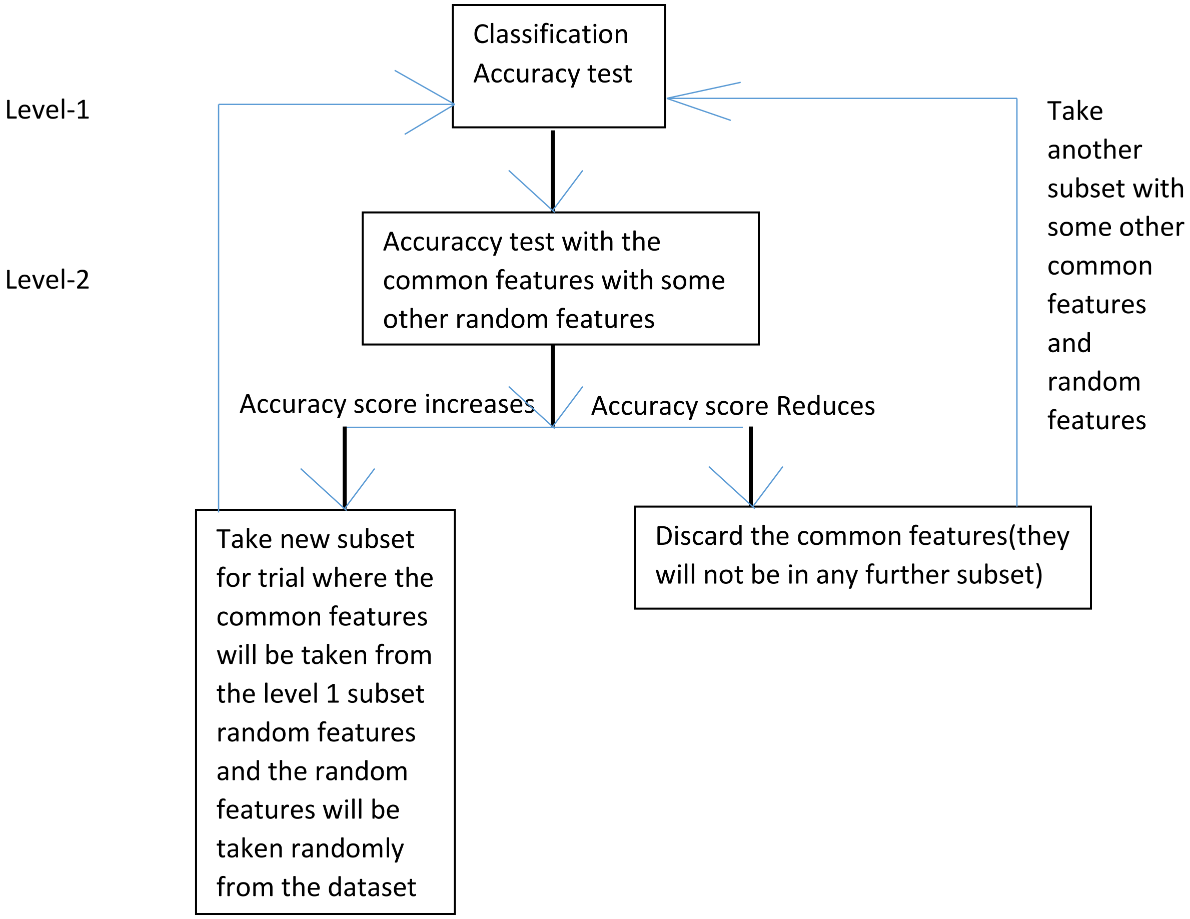

Classification Accuracy test
Level-1
Take another subset with some other common features and random features
Level-2
Accuraccy test with the common features with some other random features
Accuracy score increases
Accuracy score Reduces
Take new subset for trial where the common features will be taken from the level 1 subset random features and the random features will be taken randomly from the dataset
Discard the common features(they will not be in any further subset)

Figure 3.2- The algorithm

Step-1: Take a random subset and calculate the classification accuracy.

Step-2: Take another subset where the common features from level-1 remains and with them some other random features different from the ones previously used.

Step-3: If the classification accuracy in the step-2 is greater than step-1 then in the next subset the common features are taken from the random features used in the first step. And if the classification accuracy reduces then the common features are discarded from the dataset. And never used in any further subsets. However if the accuracy stays the same, a new subset is taken with new features.

3.5.3 Logical assumptions:

- The assumption for which we are discarding the common features is that the common features are present in both the trials. And therefore is responsible for the drop in the classification accuracy.

- In the case of the increase in the classification accuracy we are taking new subset with common features from level-1 random features.

- This implies that the common feature was not responsible for the low accuracy in the first trial. So it leaves only the random features of the first trial that could be responsible for the low accuracy.

- We take them in the next subset for the hope that we can have a drop in the next accuracy so that they can be discarded.

Chapter 4: Performance evaluations

For the implementation of the the filter we have used MATLAB. And for the implementation of the wrapper algorithm we have used weka.

4.1 Weka:

Weka (Waikato Environment for Knowledge Analysis) is a popular machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is a free software available under the GNU General Public License.

We have used it for procuring the classification accuracy for each of our selected subsets in the wrapper approach.

Advantages of Weka include:

- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- It is freely available under the GNU General Public License
- A collection of data preprocessing and modeling techniques.
- It’s Graphical User Interface is very user friendly.

Weka's main user interface is the Explorer, but essentially the same functionality can be accessed through the component-based Knowledge Flow interface and from the command line. There is also the Experimenter, which allows the systematic comparison of the predictive performance of Weka's machine learning algorithms on a collection of datasets.

The Explorer interface features several panels providing access to the main components of the workbench:

- The Preprocess panel has facilities for importing data from a database, a CSV file, etc. And for preprocessing this data using a so-called filtering algorithm. These filters can be used to transform the data and make it possible to delete instances and attributes according to specific criteria.

- The classify panel enables the users to apply classification and regression algorithms. These are called classifiers in weka. One of the classifiers is selected to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, ROC curves, etc.

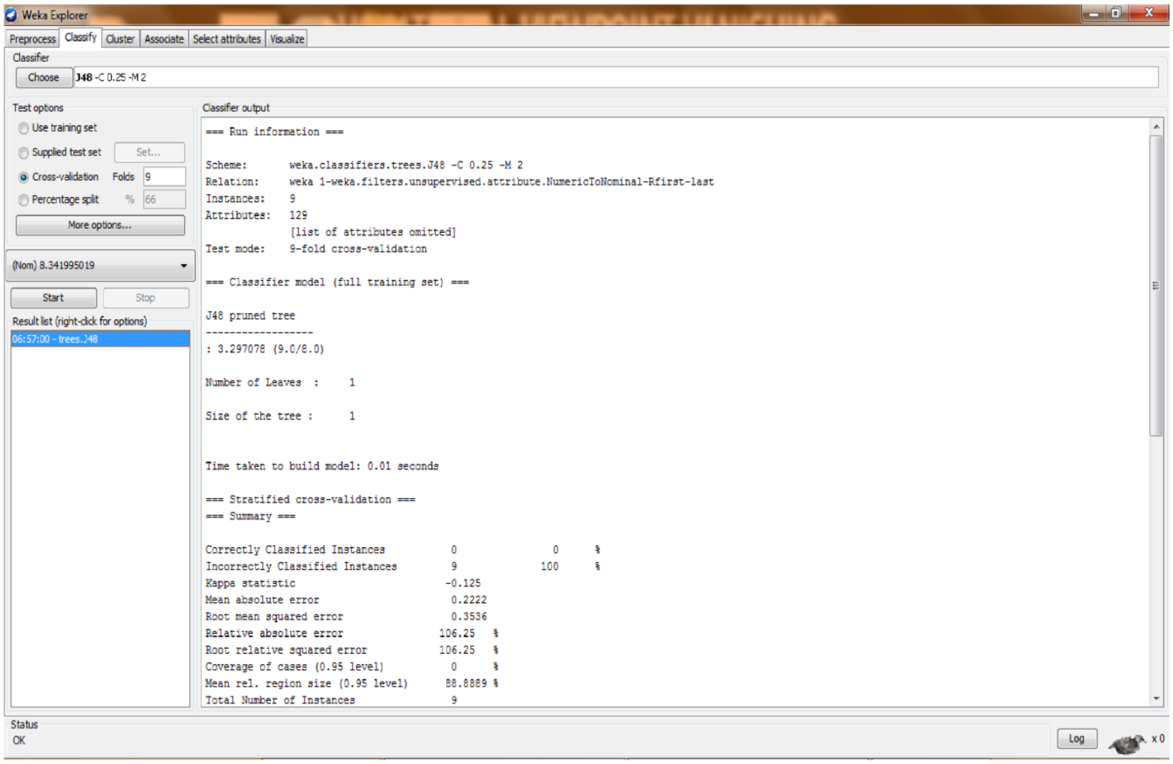
- The Associate panel provides access to association rule learners that attempt to identify all important interrelationships between attributes in the data.

- The Cluster panel gives access to the clustering techniques in Weka. For example the simple k-means algorithm. There is also an implementation of the expectation maximization algorithm for learning a mixture of normal distributions.

- The Select attributes panel provides algorithms for identifying the most predictive attributes in a dataset.

- The Visualize panel shows a scatter plot matrix, where individual scatter plots can be selected and enlarged, and analyzed further using various selection operators.

We have gone through the steps to calculate the classification accuracy scores which is very crucial for coming to our conclusion.



Classifier: J48 -C 0.25 -M 2

Test options

○ Use training set
○ Supplied test set
● Cross-validation Folds 9
○ Percentage split % 66

(Nom) 8.341995019

Result list (right-click for options)

06:57:00 - trees.J48

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     weka 1-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last
Instances:    9
Attributes:   129
              [list of attributes omitted]
Test mode:    9-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------
: 3.297078 (9.0/8.0)

Number of Leaves  :     1

Size of the tree :      1


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           0                0      %
Incorrectly Classified Instances         9              100      %
Kappa statistic                         -0.125
Mean absolute error                      0.2222
Root mean squared error                  0.3536
Relative absolute error                106.25   %
Root relative squared error            106.25   %
Coverage of cases (0.95 level)           0      %
Mean rel. region size (0.95 level)      88.8889 %
Total Number of Instances                9
```

Status

OK

Figure 4.1- The interface of weka explorer:

4.2 Result:

We have taken each subset manually and put it in weka as input. The results were saved and then the next subset would go as input.

The table in the following page shows the result of the classification accuracy-

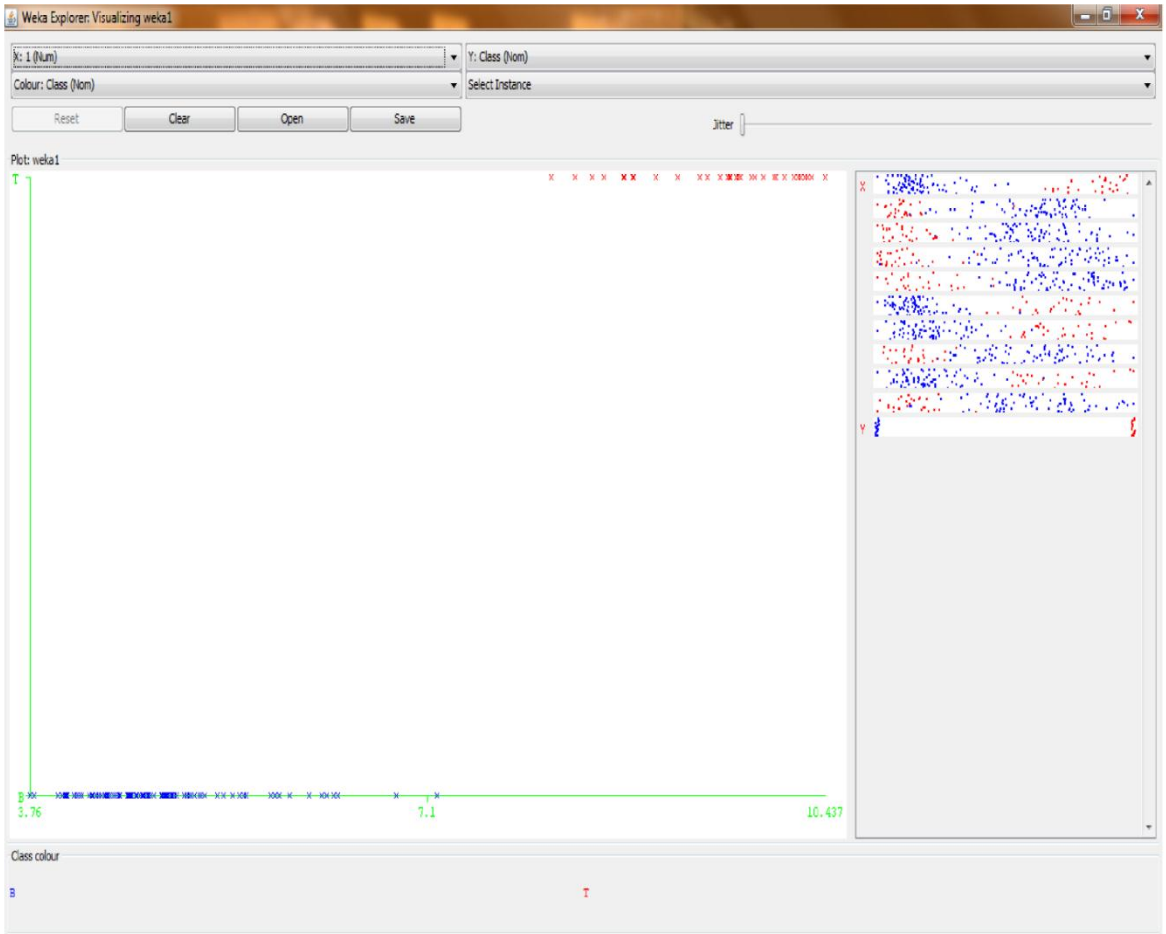
	correctly classified instances %	random features	common features	
weka1	99.2188	6,7,8,9,10	1,2,3,4,5	
weka2	99.2188	11,12,13,14,15	1,2,3,4,5	
weka3	94.5313	16,17,18,19,20	11,12,13,14,15	11,12,13,14,15 discarded
weka4	96.0938	21,22,23,24,25	16,17,18,19,20	
weka5	96.875	16,17,18,19,20	26,27,28,29,30	
weka6	98.4375	31,32,33,34,35	21,22,23,24,25	
weka7	96.0938	36,37,38,39,40	26,27,28,29,30	26,27,28,29,30 discarded
weka8	95.4375	36,37,38,39,40	16,17,18,19,20	16,17,18,19,20 discarded
weka9	93.75	36,37,38,39,40	41,42,43,44,45	36,37,38,39,40 discarded
weka10	90.625	41,42,43,44,45	46,47,48,49,50	41,42,43,44,45 discarded
weka11	99.2188	31,32,33,34,35	1,2,3,4,5	
weka12	90.625	31,32,33,34,35	46,47,48,49,50	46,47,48,49,50 discarded
weka13	99.2188	1,2,3,4,5	21,22,23,24,25	
weka14	96.875	6,7,8,9,10	21,22,23,24,25	21,22,23,24,25 discarded
weka15	99.2188	1,2,3,4,5	6,7,8,9,10	
weka16	97.6563	31,32,33,34,35	6,7,8,9,10	6,7,8,9,10 discarded
	features remaining:	10	feature IDs	1,2,3,4,5,31,32,33,34,35

Table 4.1- Result analysis

Here at last the 10 features were obtained which shows the maximum classification accuracy.

The feature IDs are 1, 2, 3,4,5,31,32,33,34,35.

The classification plotting is visualized in this image-



Weka Explorer: Visualizing weka1

X: 1 (Num)

Y: Class (Nom)

Colour: Class (Nom)

Select Instance

Figure 4.2-Red dots denote the cancerous samples and the blue dots denote the healthy samples

The classification accuracy using the obtained 10 features are 99.2188%. This accuracy is compared with other algorithms and with different datasets-

Dataset:	TSVM-RFE(60 genes 72 samples)	SVM-RFE(60 genes 72 samples)	GLAD(60 genes 72 samples)	Broaden search(100 genes 128 samples)
ALL-AML	96.03%	96.32	75.49%	99.2188%

Chapter 5: Conclusion

The approach to feature selection that has been described in this paper has scopes of improvement. Yet it can be useful for further research on gene feature selection. The wrapper method implemented here is works in a complex manner but reduces the number of the features in an efficient way.

Our future work motivation is for classification techniques. We will also try to improving the method implemented here. In terms of the loss of significant data our algorithm plays quite a smart role to select the most relevant features.

Appendix:

```
clc
clear all
load dataset.csv
x=dataset(1:128,1:12625);


L=12625;
a=0; b=0; m=0; Std=0;

for i=1:L

    for j=1:128
        a= a + x(j,i);
    end
    b(i) = a;
    m(i) = mean(b(i));

end

for i=1:L
    for j=1:128
        Std(i)= abs(m(i) - x(j,i));
    end
end

Std

M=max(Std);
```

References:

1. Guangtao Wang & Qinbao Song, “A Feature Subset Selection Algorithm Automatic Recommendation Method”, Journal of Artificial Intelligence Research 47 (2013) 1-34, Xi'an Jiaotong University, 2013.

2. Md. Abid Hasan, “Linear Regression Based Feature Selection for Microarray Data Classification”,M.Sc Thesis Report, Islamic University of Technology, 2012.

3. Noelia Sánchez-Maroño, Amparo Alonso-Betanzos & María Tombilla-Sanromán, “Filter Methods for Feature Selection – A Comparative Study”, Lecture Notes in Computer Science, Volume 4881, 2007, pp 178-187

4. Isabelle Guyon & Andre Elisseeff, “An Introduction to Variable and Feature Selection”, Journal of Machine Learning Research 3 (2003) 1157-1182, 2003.

5. Laetitia Jourdan et al. “A Genetic Algorithm for Feature Selection in Data-Mining for Genetics”, University of Lille, MIC’2001 - 4th Metaheuristics International Conference, 2001.

6. Jigang Wang et al. “Neighborhood size selection in the *k*-nearest-neighbor rule using statistical confidence”, Journal Pattern Recognition Volume 39 Issue 3, Pages 417-423, Elsevier Science Inc. New York, NY, USA March, 2006, Brown University.

7. Chris Ding and Hanchuan Peng, “Minimum Redundancy Feature Selection from Microarray Gene Expression Data”, NERSC Division, Lawrence Berkeley National Laboratory, University of California, Berkeley, CA, 94720, USA.

8. Hala Helmi, Jonathan M. Garibaldi & Uwe Aickelin, “Examining the Classification Accuracy of TSVMs with Feature Selectionin Comparison with the GLAD Algorithm”, 2005.

9. L. S. Oliveira, n. Benahmed, r. Sabourin, f. Bortolozzi & c. Y. Suen, “Feature Subset Selection Using Genetic Algorithms for Handwritten Digit Recognition”, 2008.

10. Gianluca Bontempi & Patrick E. Meyer, “Causal filter selection in microarray data”, Machine Learning Group, Computer Science Department, Faculty of Sciences ULB, Universit´e Libre de Bruxelles, Brussels, Belgium, 2002.

11. Ron Kohavi and George H. John, “Wrappers for feature subset selection”, Data Mining and Visualization, Silicon Graphics, Inc., 2011 N. Shoreline Boulevard, Mountain view, CA 94043, USA, 1997.