M.Tech. (Computer Science) Dissertation Series

SYNOPSIS

Knowledge Discovery from Gene Expression Data in a Computational Intelligent Framework: Identifying Marker Genes and Cancer Subtypes

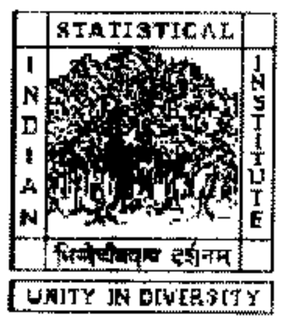
a dissertation submitted in partial fulfillment of the
requirements for the M.Tech. (Computer Science)
Degree of the Indian Statistical Institute

By

Shyam Sudhakar Chaturvedi

under the supervision of

Prof. Nikhil Ranjan Pal
ECSU



INDIAN STATISTICAL INSTITUTE
203, Barrackpore Trunk Road
Calcutta-700035

Motivation

Cancer treatment is highly dependent on proper diagnosis of the type of cancer, and that too at an early stage. Patients, hardly find symptoms to recognize these diseases(except in some cases like breast cancer, where swollen lumps appear on the breast). Thus, in most of the cases cancer is detected at advanced stages. Treatment of cancer at an advanced stage is far difficult. Also, different types cancer follow different courses of treatment. But, a greater challenge in cancer treatment is that tumors with similar histopathological appearance can follow different clinical processes. So proper diagnosis and an early detection of cancer are very important.
With the sequencing of human genome we have a flood of information associated with genes. Theoretically, it is believed that the genes have the information for all the metabolic activities in the body. Thus a gene or a combination of genes determine the response of an individual to a disease or disorder. We believe that the gene expression profiles have the information of cancer and specifically the type of cancer an individual is having. So, by intelligently fetching this information from gene expression profiles , we can help in the diagnosis of cancer at early stages. This can be done, just by analysing the expression profiles without going through costly and time consuming clinical tests at the screening phase.

Here, we have worked on Lung Cancer Data. It is to be noted that lung cancer is a major type of cancer which causes higher mortalities .

Problem Definition

In this thesis we address two problems: finding a set of marker genes and finding cancer subtypes, if present. The first problem can help to find genes responsible for cancer, while the cancer subtypes can help the treatment of cancer patients. Finding cancer subtypes involves clustering of gene expression data. The problem with gene expression data is that, they are having very high dimension and comparatively very few samples(in our case it is

203 samples and 12,600 genes) are available. Grouping(precisely, clustering) in high dimensional space is difficult , since distance based clustering methods fail due to "curse of dimensionality". So, the dimensionality of the data needs to be reduced.

Note that, the set of genes required for finding cancer subtypes may not just be the marker genes - it may require more genes to be considered.

Contributions

Our specific contributions in this thesis include:

1. We used a Neural Network based system for identification of marker genes, and found only 9 genes that can explain 5 types of cancer.
2. We modified the implementation of OFS scheme so that we can deal with high dimensional data easily.
3. In this work, we used 3 clustering algorithms viz. HCM, FCM and GKFCM.
4. We made a simple modification of GKFCM so that for illconditioned covariance matrices, the algorithm can continue.
5. We also used SOFM for class discovery - this can avoid the problem of finding the optimal number of clusters.
6. For class discovery first, we tried to use the original data with various normalizations but could not achieve much due to curse of dimensionality.
7. We then selected features using standard deviations criterion - an unsupervised method, but the result was not quite satisfactory.
8. We used a set of features selected by OFS for class discovery using HCM, FCM, GKFCM and SOFM.
9. Our results confirm the subtypes found by others. Also, we found that OFS selected features are better than that of standard deviation based method.

M.Tech. (Computer Science) Dissertation Series

Knowledge Discovery from Gene Expression Data in a Computational Intelligent Framework: Identifying Marker Genes and Cancer Subtypes

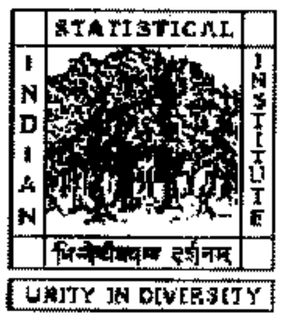
a dissertation submitted in partial fulfillment of the
requirements for the M.Tech. (Computer Science)
Degree of the Indian Statistical Institute

By

Shyam Sudhakar Chaturvedi

under the supervision of

Prof. Nikhil Ranjan Pal
ECSU



INDIAN STATISTICAL INSTITUTE
203, Barrackpore Trunk Road
Calcutta-700035

Contents

1 Introduction **7**

2 Gene expression data **8**

3 Normalization **11**
3.1 Normalization Types 11

4 Data Set **12**

5 Feature selection **12**
5.1 Standard Deviation Based Method 13
5.2 Online Feature Selection 13
5.3 Identification of Marker Genes 15
5.4 Clustering 16

6 Materials and Methods **18**
6.1 C-means(HCM) algorithm 18
6.2 The Fuzzy c-means algorithm 18
6.3 Gustafson-Kessel(GK) method 20
6.4 Self Organizing Feature Map(SOFM) 21

7 Results and Discussion **23**

8 Conclusions **28**

9 Programming environment **30**

List of Tables

1 Listing of marker genes. 16
2 List of genes which are capable to find subtypes of adenocarcinomas 24
3 comparison of disagreement indices out of 203 samples. 25
4 We have selected 15 features from standard deviation method and cluster by FCM and GK-Method 26
5 We have selected 15 features from OFS and cluster by HCM, FCM and GK-Method 27
6 OUTPUT OF SOFM 29
7 OUTPUT OF SOFM 29

Certificate

Indian Statistical Institute,
203 , B.T.Road,
Kolkata 700108.

Certificate of Approval

This is to certify that this thesis titled "Knowledge discovery (informative genes and cancer subtypes) from cancer gene expression data using a computational intelligent framework" submitted by Mr. Shyam Sudhakar Chaturvedi towards partial fulfillment of requirements for the degree of M.Tech in Computer Science at Indian Statistical Institute, Kolkata embodies the work done under my supervision.

Prof. N. R. Pal
ECSU,
Indian Statistical Institute,
Kolkata.

13.07.05

Acknowledgement

I take my immense pleasure in thanking Prof. N. R. PAL (Professor, ECSU, ISI) for his valuable guidance throughout the dissertation. His pleasant and encouraging words have always kept my spirits up. I would also like to express my sincere gratitude to Dr. Animesh Sharma (Visiting Scientist, ECSU, ISI)and Mr. Somitra Kumar Sanadhya (Research scholar ECSU, ISI) for helping me continuously throughout this work .

Shyam Sudhakar Chaturvedi

13.07.05

Abstract

There has been a substantial improvement in cancer classification over last decades.But, there is no general approach for identifying new cancer types(class discovery) or for assigning a tumor to known classes (class prediction).There is no well accepted method for identifying "marker genes". Traditional histological classification of cancer subtype is informative, but incomplete. Recent studies of gene expressions suggest that molecular classification can be used for effective diagnosis and prediction of the cancer type and treatment outcome. Here, we have made a study on microarray gene expression data of Lung Cancer with a view to discovering two types of knowledge : finding cancer subtypes(class discovery) and finding marker genes. In the context of the first problem the effect of various normalization schemes are studied in conjunction with different clustering algorithms. Experimentally, we found that because of the high dimensionality of the data c-means type clustering algorithms and their variants are not found to be very effective. So we applied a feature selection algorithm to reduce the dimension. Typically researchers use unsupervised feature selection for class discovery. Such features does not ensure class discriminating power. So we take a different route. We first find the marker genes. These reduces the dimensionality retaining the class discriminating power of the genes. These marker genes are then used to discover cancer subtypes. We could find just nine informative features(genes) to preserve the discriminating power. However, for class discovery we considered fifteen important features. Application of various clustering algorithms(C-means, Fuzzy c-means and Gustafson-Kessel method) and self organizing feature map (SOFM) finds clusters that are generally consistent with the histological classification. The analysis reveals previously defined types, subtypes and many additional details of Lung Cancer.

Our specific contributions are :

- We used a Neural Network based system for identification of marker genes, and found only 9 genes that can explain 5 types of cancer. For finding subtypes in the histologically known classes we have used a set of 15 genes, which includes the 9 marker genes also.
- We modified the implementation of OFS scheme so that we can deal with high dimensional data easily.

- In this work, we used 3 clustering algorithms viz. HCM, FCM and GKFCM.
- We made a simple modification of GKFCM so that for illconditioned covariance matrices, the algorithm can continue.
- We also used SOFM for class discovery - they can avoid the problem of finding the optimal number of clusters.
- First, we tried to use the original data with various normalizations but could not achieve much due to curse of dimensionality.
- We then selected features using standard deviation criterion - an unsupervised method, but the result was not quite satisfactory.
- We used a set of features selected by OFS for class discovery using HCM, FCM, GKFCM and SOFM.
- Our results confirm the subtypes found by others. Also, we found that OFS selected features are better than that by standard deviation based method.

1 Introduction

One of the challenges in cancer treatment is that tumor samples with similar histopathological appearance can follow significantly different clinical processes. Sub-classification of molecular profiles of these tumors may help explain why these tumors respond so differently to treatment. Thus, cancer treatment is highly dependent on proper diagnosis of cancer type(subtype), as different types(subtypes) of cancer follow different courses of treatment.

Lung Cancer is the most common causes of cancer related mortality [1] totaling more death than breast, colon and prostate cancer combined. The traditional classification based on histological methods divides all Lung Cancer types into -

small cell lung carcinoma(SCLC) and
non-small-cell lung carcinoma(NSCLC).

The latter, in turn is subcategorized as adenocarcinomas, squamous cell carcinomas, and large-cell carcinomas of which adenocarcinomas is the most common[2]. The histopathological subclassification of lung adenocarcinoma is challenging . In one study, independent lung pathologists agreed on lung adenocarcinoma subclassification in only 41% of cases[3]. However, researchers [4] argue for refining such distinction in case of Bronchioalveolar carcinoma(BAC) which is a histological subclass of lung adenocarcinoma. In addition, metastasis of nonlung origin can be difficult to distinguish from lung adenocarcinomas [5].

Cluster analysis of microarray data may have different objectives even if the same or similar algorithmic approaches are applied. On the other hand, the choice of an algorithm should be determined by the purpose of the cluster analysis : "class comparison", "class prediction", and "class discovery" [6]. The purpose of class comparison is to compare the gene expression in predefined groups of specimen, such as primary vs. metastatic tumors or diabetic vs. non-diabetic patients. Class prediction is also based on pre-determined classification. The purpose of class prediction is the correct identification of a new sample as a member of one of the known classes. Comparison of "molecular signatures" between specimens can reveal certain combination of genes that can predict chances of survival, effectiveness of treatment and other characteristics associated with a particular class. Class discovery is fundamentally different from class comparison or class prediction. The purpose of

class discovery is not testing but *generating hypothesis.*

DNA microarrays [21] have made possible the challenging task of monitoring expression levels of thousands of genes simultaneously. In the present study, we have two goals: discovering marker genes and looking for the presence of any sub-class within a cancer type. Also, we want to find how these subclasses, if present are related to each other. To accomplish this we analyze gene expression data using computational intelligence tools, such as neural networks, fuzzy clustering algorithms and self-organizing-feature-map(SOFM).

2 Gene expression data

Gene transcription in two or more different kinds of cells can be compared with comparative cDNA hybridization technique [20], [22]. Genetic disease is often caused by genes which are inappropriately transcribed – either too much or too little, or which are missing altogether. Such defects are especially common to cancers, which can occur when regulatory genes are deleted, inactivated, or become constitutively active. Unlike some genetic diseases (e.g. cystic fibrosis) in which a single defective gene is always responsible, cancers which appear clinically similar can be genetically heterogeneous. For example, prostate cancer (prostatic adenocarcinoma) may be caused by several different, independent regulatory gene defects even in a single patient. In a group of prostate cancer patients, every one may have a different set of missing or damaged genes, with differing implications for prognosis and treatment of the disease.

Comparative hybridization can serve two purposes in studying cancer : it can pinpoint the transcription differences responsible for the changes from normal to cancerous cells, and it can distinguish different patterns of abnormal transcription in heterogeneous cancers. Understanding the diverse basis of a cancer is crucial for inventing therapies targeted to the different varieties of the disease, so that each patient receives the most appropriate and effective treatment.

Cancers are common examples of genetically heterogeneous diseases, but they are by no means the only ones. Diabetes, heart disease, and multiple sclerosis are among the diseases for which genetic risk factors are known to be heterogeneous.

Microarrays [15] are made from a collection of purified DNA's. A drop of each type of DNA in solution is placed onto a specially-prepared glass microscope slide by an arraying machine. The arraying machine can quickly produce a regular grid of thousands of spots in a square about 2 cm on a side, small enough to fit under a standard slide coverslip. The DNA in the spots is bonded to the glass to keep it from washing off during the hybridization reaction.

Once the cDNA probes have been hybridized to the array and any loose probe has been washed off, the array must be scanned to determine how much of each probe is bound to each spot. The probes are tagged with fluorescent reporter molecules which emit detectable light when stimulated by a laser. The emitted light is captured by a detector, either a charge-coupled device (CCD) or a confocal microscope, which records its intensity. Spots with more bound probe will have more reporters and will therefore fluoresce more intensely.

Each of the two fluorescent reporters (fluors) used has a characteristic excitation wavelength; only light of this wavelength will cause the molecule to fluoresce. The emitted light has a characteristic emission wavelength which is different from the excitation wavelength. The detector for the emitted fluorescence from the array is sensitive to the emission wavelength but filters out the excitation wavelength; in this way, the fluorescent light of interest can be separated from the laser light scattered off the slide.

A good pair of fluors for a comparative hybridization experiment should have very different emission or excitation wavelengths. If the emission wavelengths are different, light emitted from the two fluors can be selectively filtered to measure the amount emitted by each fluor separately. If the excitation wavelengths are different, the two fluors can be stimulated and scanned one at a time. If one of these conditions is not met, the scanned intensities can be contaminated by crosstalk between the two fluorescent channels.

The end product of a comparative hybridization experiment is a scanned array image. The intensities provided by the array image can be quantified by measuring the average or integrated intensities of the spots. The ratio of fluorescent intensities for a spot is interpreted as the ratio of concentra-

tions for its corresponding mRNA in the two cell populations. Researchers [15] have Researchers [15] have demonstrated the ability to as a factor of two, with reasonable agreement between expression ratios measured on the array and ratios measured by an alternate form of RNA blotting. Numerous software packages, both free and commercial, exist for quantitating microarray data. In this work we have used Harvard Data Set [7] . Here the gene expression profiles have been produced using U95A oligonucleotide probe arrays (affymetrix , Santa Clara, CA).

Arindam Bhattacharjee, [7] has worked on this dataset They have done Hierarchical and Probabilistic clustering. They have emphasized the diagnostic potential of the gene expression profiling by its ability to discriminate primary lung adenocarcinomas from metastasis of extra-pulmonary origin. They have used a supervised approach to extract "marker genes"(genes that best define the proposed clusters), and have reported 44 such genes. They have put forward the existence of subtypes in the adenocarcinomas. They have concluded that :

1. Normal lung cancer samples form a distinct group, but are most similar to the adenocarcinomas.

2. Unlike other major lung tumor classes , adenocarcinomas were not defined by a unique set of marker genes.

3. Possibly there are 4 subtypes of adenocarcinomas viz.C1, C2, C3 and C4.

4. C4 were similar to normal lung, C3 were similar to normal lung and C2, C2 similar to squamous and pulmonary tumors, C1 were similar to SCLC and squamous tumors.

5. Squamous tumors have similarity with subtype of adenocarcinomas.

6. Discovered putative metastases of extra pulmonary origin with non-lung expression signatures among presumed lung adenocarcinomas.

Their results suggest that integration of expression profile data with clinical parameters could aid in diagnosis of lung cancer patients.

Andre Ptitsyn, [10] has reported the existence of subtypes in adenocarcinomas. Also, these subtypes show similarities with the other histologically known lung tumors and with normal lung specimens.

3 Normalization

There are many sources of systematic variation in microarray experiments which affect the measured gene expression levels. Normalization is the term used to describe the process of removing such variation. We need to try to remove systematic variation, to bring the data from different experiments onto a leveled playing field. Even when these effects are accounted for, it is interesting to standardize each vector to unit variance(dividing by SD)to avoid uninterpretable correlations resulting from one or two vectors having a negligible variation.However, dividing by the SD is necessary for many techniques, it amounts to giving equal prior importance to all genes(Frank and Friedman, 1993).Normalization of expression of a gene over samples captures the "shape" of the variation of the data over samples on an equal level(when compared to that of other genes). So, the properties related to shape of the expressions can be understood. Simultaneously, properties related to the magnitudes of individual expression levels are hardly captured. The following normalization techniques are typically used.

3.1 Normalization Types

0-1 normalization : In this technique the expression value of each gene is scaled down to [0, 1]. It is obtained by the min. and max. expression levels of a gene over all the samples. The normalized value(x_{new}) from x_{old} is computed as:

$$x_{new} = (x_{old} - min.)/(max. - min.)$$

log_{10} normalization: In this method the minimum of the expression value of a gene is found. If it is less than 1 then 10 is added to the expression level of the gene in all samples. After doing this we take log base 10 of the expression.

Thus,

$x_{new} = log_{10}(x_{old})\ if\ min \geq 10$

$= log_{10}(x_{old} + 10)\ if\ min < 10.$

Z-score: In this method minimum expression level of a gene is found and variance(var.) of that gene is computed . Then the new normalized value is computed as:
$x_{new} = (x - min.)/var.$

4 Data Set

We have used the Harvard Data Set in which a total of 203 samples were taken . The dataset consists of 186 lung tumors and 17 normal lung specimens. The specimens are known to be from histologically defined lung adenocarcinomas(n= 127), other adenocarcenomas(n= 12, which are suspected to be extrapulmonary metastases based on clinical history) implying 139(127+ 12) adenocarcinomas, squamous cell lung carcinomas(n= 21), pulmonary carcinoids(n= 20), SCLC(n= 6) cases, and normal lung(n= 17)specimens. U95A Oligonucleotide Affymetrix was used in the microarray experiments. The dataset is having expressions of 12,600 genes. Thus, there are 12,600 features and 203 samples.

5 Feature selection

Since the gene expression data are in huge dimension, the clustering of such data is a difficult task. The large dimensionality contains redundant or noisy features which causes difficulty in analyzing the data. In gene expression data the sample size is less (a few hundreds) whereas features(genes) are huge(several thousands). The sparsity of data in high dimensional space is known in the literature as "curse of dimensionality". Due to this Euclidean metric looses its significance. We shall see later, the clustering methods in this high dimensional data space using Euclidean metric fails. In light of the above discussion we need to select enough number of features(genes) to perform clustering .

Feature selection techniques aim to reduce the feature space to a highly predictive subset of the space, i.e., they aim to discard the bad/ irrelevant features from the available set of features without loosing the accuracy of

prediction. In addition, it is easier to interpret analysis, if the data can be reduced to a manageable number of features. The literature is quite rich in feature selection methodologies. Some of these methods use neural networks, neuro-fuzzy systems [18], while others use fuzzy logic or statistical techniques . Other approaches to dimensionality reduction involve replacing the given set of features by a new but smaller set of computed features.

Here features(genes) were selected using two methods: The first method is based on standard deviation of gene expressions and the other method does on-line feature selection(OFS) based on Neural Networks. The first one is *unsupervised*, so it does not guarantee whether the reduced dataset will have adequate discriminating power or not. For the class discovery problem, typically unsupervised method of feature selection is used. The OFS is *supervised* one that ensures a significantly high preservance of the discriminating power of the original dataset in the reduced dataset.

5.1 Standard Deviation Based Method

In the standard deviation based feature selection method we computed the standard deviation of the normalized expressions of each gene, then sorted them with respect to their numerical magnitude. Then, we took the most variant ones, say top 5 or top 100 or even more for use in clustering. Note that, such a method cannot be used for finding marker genes as it does not take into account the class information.

5.2 Online Feature Selection

In a standard multilayer perceptron (MLP) network, the effect of some features (inputs) can be eliminated by not allowing them into the network, i.e., by equipping each input node (hence each feature) with a gate and closing the gate. For good features the associated gates can be completely opened. On the other hand, if a feature is partially important, then the corresponding gate should be partially opened. Pal and Chintalapudi [17] suggested a mechanism for realizing such a gate so that *useful* features can be identified and attenuated according to their relative usefulness. In order to model the gates we associate a gate function to each node in the input layer of the MLP. A gate function should produce a value of 1 or nearly 1 for a good feature; while for a bad feature, it should be nearly 0. We call the network an Online

Feature Selection (OFS) network as it does both feature selection and system identification together.

To use the gates we multiply each input feature value by its gate function value and the modulated feature value is passed into the network. The gate functions[19] attenuate the features before they propagate through the net, so we may call these gate functions *attenuating* functions. Useful gate functions $F_i : R \rightarrow [0, 1]$ should have a tunable parameter and should be differentiable with respect to the tunable parameter. It should be monotonic with respect to its tunable parameter. The sigmoidal function satisfies the above criteria and in this paper we have used it.

The basic philosophy of learning would be to keep all gates almost closed at the beginning of the learning (i.e. no feature is important) and then open the gates as required during the training. To complete the description in connection with MLP, let F_i be the gate or attenuation function associated with the i^{th} input feature. F_i has an argument m_i , $F_i'(m_i)$ be the value of derivative of the attenuation function at m_i. Let μ be the learning rate of the attenuation parameter; ν be the learning rate of the connection weights, x_i be the i^{th} input of an input vector $\mathbf{x}$; x_i' be the attenuated value of x_i, i.e., $x_i' = x_i F(m)$; w_{ij}^o be the weight connecting the j^{th} node of the first hidden layer to the i^{th} node of the input layer; and δ_j^1 be the error term for the j^{th} node of the first hidden layer [17].

It can be easily shown that except for w_{ij}^0, the update rules for all weights remain the same as that for an ordinary MLP trained with backpropagation. Assuming that the first hidden layer has q nodes, the update rules for w_{ij}^0 and m_i are :

$$w_{ij,new}^0 = w_{ij,old}^0 - \nu x_i \delta_j^1 F(m_i) \tag{1}$$

$$m_{i,new}^0 = m_{i,old}^0 - \mu x_i (\Sigma_{j=1}^q w_{ij}^0 \delta_j^1) F'(m_i) \tag{2}$$

As mentioned earlier, for the gate function, several choices are possible but we use here the sigmoidal function $F(w) = 1/(1 + e^{-m})$. The p gate parameters are so initialized that when the training starts $F(m)$ is practically zero for all gates, i.e., no feature is allowed to enter the network. As the gradient descent learning proceeds, gates for the features that can reduce the error faster are opened faster. The learning of the gate function continues along with other weights of the network. At the end of the training we can

pick up important features based on the values of the attenuation function. Typically, the training can be stopped when the training error is reduced to an acceptable level.

Note that, different initializations of the network may lead to different subsets of good features (genes). If this happens, this indicates that there are different sets of features that can do the classification job equally well. One may rank the features based on the extent the gates are opened and use a set of top ranked features. This is expected to do a good job because OFS looks at all the features at a time during the training process.

5.3 Identification of Marker Genes

Since the dimensionality of the data is very high(12,600), it poses a computational problem. Since training of OFS takes a lot of time, we proposed an approximate version of OFS. In this method, we split the features into sets of say, 1000 features each. Thus, we split the feature set into 12 parts. Therefore, instead of using 12,600 features to a single network we use only 1050 features. But we need to train 12 such networks. Now, for each set we run the OFS for 500 iterations , with the architecture 1050 neurons in the input layer, 400 neurons in the hidden layer and 5 neurons in the output layer(as histologically the samples are of 5 types). Note that, there is no specific reason behind the choice of the number 1050, any other choices can be done depending upon the available computing power. Any other partitioning of the features can also be used.

The misclassifications in all cases varied from 0 to 5. So, the training was adequate to select important features(genes). Note that, the purpose of OFS was just to discard the bad features(genes) and select a few good ones, so we did not train the net till error (misclassification) reduces to zero. From each of such division of features we selected only those features for which the "gate opening" was more than 0.1 in this way we got 121 features. Now, all these 121 features were taken as a new set of features and OFS was run on them for 1500 iterations, with the architecture 121 neuron in the input layer, 60 hidden layer neuron and 5 output layer neurons. We got 0 misclassification in several runs of OFS. Finally, we selected only 9 features based on the gate opening as shown in Table 1. We call these 9 features marker genes as they have adequate discriminating power to reduce the misclassifica-

Table 1: Listing of marker genes.

1.	31596_f_at immunoglobin lambda-like polypeptide 2
2.	34056_g_at activin A receptor, type IB
3.	34423_at DKFZP564J102 protein
4.	34666_at super oxide dismutase 2, mitochondrial
5.	34667_at nuclear transcription factor, X-box binding 1
6.	34668_at acetyl-Coenzyme A transporter
7.	34677_f_at Cluster Incl AJ012755 Homosapiens mRNA for TL132
8.	34680_s_at KIAA0107 gene product
9.	41201_at dynamin 1-like

tion to zero. Thus, ultimately 9 genes(features) were selected by this method.

For the knowledge discovery part we used the top 15 features (top in terms of gate opening) as listed in Table 2. For knowledge discovery we consider more genes because marker genes may not have adequate variation to find subtype in a class, although they can discriminate between the classes. We made several runs of Neural Network with 15 features. As expected, in this case we got zero misclassification. So, these 15 genes are having the capability to cluster the histologically known classes, and they may have discriminating power to represent the subtypes in a class, if present .

5.4 Clustering

Clustering[14] can be considered the most important unsupervised learning problem. It deals with finding substructures in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

The goal of clustering is to reduce the amount of data by categorizing or grouping similar data items together. In addition, we may sometimes discover new subgroups within a given group by analyzing the results of clustering. Fuzzy clustering algorithm could be useful here, because fuzzy membership

value might help to detect subgroups in a cluster.

Clustering methods can be divided into two basic types:-

1. hierarchial clustering

2. partitional clustering

Also neural network based clustering techniques viz. self-organizing-map(SOM) are available.

Within each of the types there exists a wealth of subtypes and different algorithms for finding the clusters. Hierarchical clustering proceeds successively by either merging smaller clusters into larger ones, or by splitting larger clusters. Partitional clustering, on the other hand, attempts to directly decompose the data set into a set of disjoint(or overlapped for fuzzy algorithms) clusters. Whereas, SOM gives the number of clusters as different groups of neurons on a 2-dimensional grid. Neurons, which are near to each other in terms of the weights learned by them are also topologically near to each other. Thus, looking at the map we may be able to find useful clusters. Since, one may not have any idea about the number of subgroups to look for and there is no reliable cluster validity measure, SOM may be very useful for finding subtypes.

A commonly used partitional clustering method is K-means clustering. The crisp or conventional k-Means clustering assigns each input vector to exactly one cluster. On the other hand, in fuzzy C-means clustering, a given pattern does not necessarily belong to only one cluster, but can have varying degrees of memberships to several clusters. In this work we have used C-means, fuzzy C-means(FCM), Gustafson-Kessel method(GK, a variant of FCM)[8] and SOFM [14] to group different cancer types.

Distance based clustering may fail to produce tight clusters for very high dimensional data because of the "Curse of dimensionality". Selecting important features before clustering the data may solve this problem. In this work first we report some result on the original dataset and then we shall select some important features and use these features for clustering the data.

6 Materials and Methods

As already discussed earlier, we use NN (Neural Network) based method called OFS [17] and standard deviation method to select a small set of features(genes). Features selected by both of these methods were used separately. We applied C-means (HCM), Fuzzy C-means(FCM), Gustafson-Kessel(GK) clustering method to cluster the data set in the reduced dimension to extract prototypes for each cluster. If the data vectors belonging to a particular class are part of a tight cluster then the prototype obtained for that class will be highly representative of data vectors of that class. On the other hand, if the prototype for a class is not very representative then we may expect that either the data for that class is noisy or there is a possibility of a sub-class within that class.

6.1 C-means(HCM) algorithm

In this method first, divide the dataset into 'C' almost equal groups(randomly), iter = 0 . Choose iter_max and ϵ (a small positive number).
1. Find the mean of each group, thus 'C' such means are obtained.
2. Find the distance of each sample point from each of the 'C' such means.
3. Assign each sample point a "label" describing the mean to which it is nearest.
4. Group the sample points with the same the "label" in one group and different label samples in different groups, thus 'C' new groups are formed.
5. Find the mean of each of these new 'C' groups.
6. Find the max_diff(maximum difference) between the mean points of step 1 and that in step 2 , also increment iter by 1 i.e iter= iter+ 1 .
7. If iter $>$ iter_max or max_diff $< \epsilon$ Then stop, else go to step 1 with groups in step 5 as the initial groups.

6.2 The Fuzzy c-means algorithm

The Fuzzy C-Means (FCM) clustering algorithm [14, 13] attempts to cluster data vectors into C groups based on the distances between them.

The FCM algorithm minimizes the objective function

$$J = \sum_{i=1}^{C} \sum_{k=1}^{N} u_{ik}^{m} \|\mathbf{x}_k - \mathbf{v}_i\|^2,$$

subject to

$$\sum_{i=1}^{C} u_{ik} = 1 \ \forall \ k = 1, 2, ..., N$$

and

$$0 < \sum_{k=1}^{N} u_{ik} < N \ \forall \ i = 1, 2, ..., C \ ,$$

where C is the number of clusters, $\mathbf{x}_k \in \mathbf{R}^p$ is the k^{th} data vector, N is the number of data vectors, $m > 1$ is the fuzzifier, u_{ik} denotes the membership of k^{th} data vector to i^{th} cluster and $\mathbf{v}_i \in \mathbf{R}^p$ is the centroid of the i^{th} cluster.

First order necessary conditions on $\mathbf{U}$ and $\mathbf{V}$ at a local minima of J are:

$$u_{ik} = \left[\sum_{j=1}^{C} \left(\frac{\|\mathbf{x}_k - \mathbf{v}_i\|}{\|\mathbf{x}_k - \mathbf{v}_j\|} \right)^{\frac{2}{m-1}} \right]^{-1}, \quad \forall i, k \tag{3}$$

and

$$\mathbf{v}_i = \frac{\sum_{k=1}^{N} u_{ik}^{m} \mathbf{x}_k}{\sum_{k=1}^{N} u_{ik}^{m}}, \quad \forall i. \tag{4}$$

The algorithm iterates between equations (4) and (3) in that order, as described below.

Algorithm:Fuzzy C-Means

1. Initialize a valid fuzzy c-partition $U = [u_{ik}]_{C \times N}$.
2. Compute a new set of prototypes using eq. (4).
3. Compute a new partition matrix using eq. (3) with these new prototypes.
4. Repeat this process (Steps 2 and 3 alternately) till the entries of the partition matrix stabilize.
5. Defuzzification : Use u_{ik} to find subtypes. Assign the data vector $\mathbf{x}_k$ to the cluster for which its membership value u_{ik} is largest.

The same procedure above can be carried out by initializing the prototypes instead of the partition matrix in which case the algorithm iterates between equations (3) and (4) in that order. The convergence properties remain the same under both schemes of initialization. As the value of m increases the algorithm produces more fuzzy partitions [13].

6.3 Gustafson-Kessel(GK) method

Gustafson-Kessel(GK) method is a variant of FCM in which the distance between a vector and a cluster centroid is obtained on the basis of a fuzzy covariance matrix of the cluster under consideration. They propped that the matrix A_i in the following expression (J function to be minimized) to be a positive definite matrix.

$$J = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^{m} \|\mathbf{x}_k - \mathbf{v}_i\|_{A_i}^2$$

where,

m > 1 is the fuzzifier.
p dimension of each sample vector.
$\mathbf{x}_k$ k: 1 to N are sample vectors .
$\mathbf{v}_i$ i: 1 to c are the mean vectors of c clusters.
u_{ik} membership of vector $\mathbf{x}_k$ in cluster C_i.
A_i i: 1 to c positive definite matrices of order p by p.

$A_i's$ are obtained as following :

$$A_i = [d_i]^{1/p} Cov_i^{-1}, 1 \leq i \leq c$$

where,

$Cov_i = \sum_{k=1}^{n} u_{ik}^{m} (\mathbf{x}_k - \mathbf{v}_i)(\mathbf{x}_k - \mathbf{v}_i)^T / \sum_{k=1}^{N} u_{ik}^{m}$, is the fuzzy covariance matrix of the cluster i,

d_i= determinant of Cov_i.

The algorithm follows the same steps as FCM , except in finding the distance between $\mathbf{x}_k$ (sample vector) and $\mathbf{v}_i$(mean of cluster i). Typically, FCM uses Euclidean metric. So, all the A_i matrices are I_p (identity matrix of order p) in FCM, although they are not written explicitly. Whereas in GK-Method A_i depends on the covariance matrix and its determinant. It is to be noted

that, while computing the inverse of the covariance matrix Cov_i it may not be invertible and determinant could be infinitesimally small (say less than 0.00001). In all such cases we proposed a modification of the usual GK-Method by choosing A_i as I_p. A good choice of the exponetial weight(m i.e.,the fuzzifier) is is found to be 1.50 .

6.4 Self Organizing Feature Map(SOFM)

We have used Kohonen's Model of SOFM. It describes a Feature map(ϕ) from set of data points($\mathbf{X}$) to a two-dimensional rectangular grid($\mathbf{A}$). This is a "competitive way" of learning. It consists of p(dimension of each sample vector) sensory nodes as input layer of neurons and a two-dimensional lattice of neurons as output layer.

The objective is to formulate a correspondence between a "pattern" or a sample vector to a node in the lattice so, that the topology of the input dataset is preserved on the two dimensional lattice. The output layer (i.e., lattice) consists of sufficient number of neurons (say, N) such that the objective is accomplished. All the sensory nodes are connected to the nodes in the lattice. so, there are N∗p links between the two layers of neurons. The links are assigned arbitrary(small random numbers) weights to begin with. Each neuron in the lattice is having a p-dimensional weight vector. The activation(i.e., firing strength) of a node j(in the lattice with weight vector $\mathbf{w}_j$), corresponding to a sample vector($\mathbf{x}_i$) is defined as the Euclidean distance between $\mathbf{w}_j$ and $\mathbf{x}_i$. The neuron which is having the minimum distance between the sample vector ($\mathbf{x}_i$) and it's weight vector ($\mathbf{w}_j$) is the "winner neuron ".
For every input, the network updates the weight vector associated with the winner as well as the weights associated with the nodes in the neighborhood of the winner.
Algorithm SOFM:

1. Initialize randomly, the weight vectors $\mathbf{w}_j$, j from 1 to N, choose ϵ, a small positive number, choose λ, a positive number .
2. Select randomly an input sample vector $\mathbf{x}$ from dataset.
3. Find the winning neuron. i.e., index($\mathbf{x}$) = j, such that $||\mathbf{x} - \mathbf{w_j}||$ is minimum compared to

that of all other $\mathbf{w}_k$, k from 1 to N.

4. Update weights.
   $\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \eta(t)\pi_{ji}(t)[\mathbf{x} - \mathbf{w}_j(t)]$, if j is in the neighborhood(Nbd.) of the winning neuron index($\mathbf{x}$).
   $= w_j(t)$, otherwise.

5. Repeat the steps 2 to 4 typically, 500*N times.

Here, π_{ji} is the degree of influence of the winning neuron i on the j^{th} neuron. usually π_{ji} is considered a Gaussian function of the distance between the neuron j and i on the lattice. The Nbd. of a neuron k is defined as a set of neurons i whose distance on the lattice is less than some predefined value, say λ. The size of Nbd. decreases with time. The learning rate $\eta(t)$ also decreases with time.

After performing the above algorithm we achieve the following properties.

- Approximation of the input space. A large dataset X is represented by a smaller set of neurons(winner neurons) in the lattice. The weight vectors of the winner neurons are rerepresentation of the input pattern vectors. This property enables us to "Data Condensation" and "Vector Quantization". So, huge amount of data is represented by some prototypes.

- The feature map ϕ computed by SOFM algorithm is topologically ordered in the sense that the spatial location of a neuron in the lattice corresponds to a particular domain of input sample space. This implies, we can visualize "class structure" of the input data through the ordering of the neurons on the lattice.

- Density matching i.e., "clustering". The variation of distribution of input data implies variation of distribution of winner neurons in the lattice.

- It helps to avoid the difficult problem of choosing the right number of cluster-subtypes. We can always start with a large number of neurons in the output layer and detect the right number of cluster from the map.

7 Results and Discussion

When we apply the three clustering algorithms on the dataset without feature selection, we found all data points getting their membership values almost equally divided. We tried our experiments with different values of C(number of clusters) = 5, 6, 7, 8, 9, 10. The results are similar in the sense that the partitions are very fuzzy, the membership value for each data point to any cluster is nearly equal to 1/C. So, such clustering is not useful for detecting cancer subtypes.

Then, we use 15 features selected by standard deviation method and applied the clustering algorithms with C = 5, the confusion matrices generated by the algorithm are shown in Table 4. We repeat the experiment for different initializations and the confusion matrices obtained for two different runs are included in Table 4. Table 4 reveals that the confusion matrices obtained are not consistent between runs and between algorithms.

Using OFS also we selected 15 genes. We used these genes for clustering. The list of these 15 genes is given in Table 2. It is to be noted that we are using more number of genes viz. 15 than the marker genes (we are reporting 9) as the subtype properties of cancer may not necessarily be captured by the marker genes.

The results obtained after applying the clustering algorithms support the histologically known cancer types. There is significant evidence of having subtypes in Lung adenocarcinomas. Also, it is vivid that the subtypes have close resemblance with normal lung specimens. We also see the similarities of the subtypes of the Lung adenocarcinomas with pulmonary carcinoids and with SCLC.

The confusion matrices obtained by running the clustering algorithms on features selected by OFS method, are reported in Table 5. We made several runs and report in the Table 5 only two typical runs for each of the clustering methods used. We found similar distribution of data of the histologically known classes. On the basis of the "stability" of the distribution of the types, as shown in the confusion matrices we can claim for the "stability" of the clusters.

To demonstrate the stability of partition and hence that of the confusion matrices we computed a measure of disagreement between two confusion ma-

Table 2: List of genes which are capable to find subtypes of adenocarcinomas

1.	31596_f_at immunoglobin lambda-like polypeptide 2
2.	34056_g_at activin A receptor, type IB
3.	34423_at DKFZP564J102 protein
4.	34666_at super oxide dismutase 2, mitochondrial
5.	34667_at nuclear transcription factor, X-box binding 1
6.	34668_at acetyl-Coenzyme A transporter
7.	34677_f_at Cluster Incl AJ012755 Homosapiens mRNA for TL132
8.	34680_s_at KIAAO107 gene product
9.	41201_at dynamin 1-like
10.	35842_at Cluster Incl AL049265 Homosapiens mRNA
11.	35843_at Cluster Incl L40402 Homosapiens(clone Zap 2) mRNA fragment
12.	35848_at Cluster Incl L049432: Homosapiens mRNA
13.	39168_at AC-like transposable element
14.	1179_at Heat shock protein, 70 kda
15.	1142_at Fibroblast growth factor receptor k-Sam Alt.Splice 1

trices after establishing correspondence between them. In the Table 4 and Table 5. We observe the following:

1. The two confusion matrices(HCMSTD) of HCM in Table 4 show the following correspondence: Columns C1, C2, C3, C4 and C5 of first matrix correspond to C2, C1, C3, C4 and C5 of second matrix, respectively and results in a disagreement of (4 + 21 + 6 + 14 + 11)/203 = 56/203. Similarly in Table 5 Columns C1, C2, C3, C4 and C5 of first matrix correspond to C3, C2, C1, C4 and C5 of second matrix, respectively leading to a disagreement of (19 + 3 + 2 + 7 + 15)/203 = 46/203. Thus, clustering with OFS selected features gives lesser index of disagreement.

2. The two confusion matrices(FCMSTD) of FCM in Table 4 show the following correspondence: Columns C1, C2, C3, C4 and C5 of first matrix with C4, C3, C1, C2 and C5 of second matrix, respectively and results in a disagreement of (7 + 7 + 10 + 18 + 22)/203 = 64/203. Similarly in Table 5 Columns C1, C2, C3, C4 and C5 of first matrix correspond to C5, C3, C2, C1 and C4 of second matrix, respectively leading to a disagreement of (1 + 2 + 0 + 1 + 2)/203 = 6/203. Thus,

Table 3: comparison of disagreement indices out of 203 samples.

	STD Method	OFS Method
HCM	56	46
FCM	64	6
GK	98	47

clustering with OFS selected features gives lesser index of disagreement.

3. The two confusion matrices(GKSTD) of GK in Table 4 show the following correspondence: Columns C1, C2, C3, C4 and C5 of first matrix correspond to C2, C1, C3, C4 and C5 of second matrix, respectively and results in a disagreement of (13 + 23 + 33 + 7 + 22)/203 = 98/203. Similarly in Table 5 Columns C1, C2, C3, C4 and C5 of first matrix correspond to C2, C1, C3, C4 and C5 of second matrix, respectively leading to a disagreement of (14 + 6 + 14 + 4 + 9)/203 = 47/203. Thus, clustering with OFS selected features gives lesser index of disagreement.

We have reported these results in Table 3. Based on these indices also we found that the feature selected by OFS are more useful.

In the output of various clustering algorithm we noticed the that adenocarcinomas are always divided into 5 subtypes thus "supporting" the claim of existence of subtypes of cancer in the histologically known class "adenocarcinomas". Also, the pulmonary carcinoid(other histologically known type of cancer) is seen to cluster in one group most of the times and some time into two subgroups. They show resemblance with subtypes of adenocarcinomas. Normal lung specimens are seen to cluster in one or two groups and also resembling with subtypes of adenocarcinoma. We get a cluster of size 12 in the subtypes of adenocarcinomas in resemblance with normal specimens. These might be the metastasis specimens. The squamous tumor specimens also cluster into groups by all clustering methods. They are seen to cluster in one group and with resemblance with a subtype of adenocarcinomas. The SCLC specimens are seldom seen to cluster together by all the clustering algorithms used, even though they are pretty less(only 6) in number. They are seen to split into two to three groups most of the times of the runs of the algorithms.

Table 4: We have selected 15 features from standard deviation method and cluster by FCM and GK-Method

HCMSTD iterations 155 initialization 6772271 max_delta : 0.000091	C1	C2	C3	C4	C4
Adeno	11	22	27	29	50
Normal	0	6	0	9	2
Squamous	3	4	10	0	4
Pulmonary	17	1	2	0	0
SCLC	1	0	3	0	2

HCMSTD iterations 52 initialization 98650971 max_delta : 0.000098	C1	C2	C3	C4	C5
Adeno	37	13	25	19	45
Normal	5	0	0	11	1
Squamous	0	5	9	1	6
Pulmonary	0	17	2	1	0
SCLC	0	1	0	0	5

FCMSTD initialization 123812813 num of iterations 16 max_delta : 0.000009 exp_wt 1.500000	C1	C2	C3	C4	C5
Adeno	37	16	24	39	23
Normal	1	0	11	5	0
Squamous	4	5	1	1	10
Pulmonary	0	17	2	0	1
SCLC	5	1	0	0	0

FCMSTD initialization 290072813 num of iterations 144 max_delta : 0.000010 exp_wt 1.500000	C1	C2	C3	C4	C5
Adeno	25	40	11	35	28
Normal	6	1	0	1	9
Squamous	5	11	3	2	0
Pulmonary	2	0	17	0	1
SCLC	0	3	1	2	0

GKSTD initialization : 416292813 num of iterations : 45 max_delta : 0.000010 exp_wt 1.500000	C1	C2	C3	C4	C5
Adeno	35	10	52	21	21
Normal	0	1	5	1	10
Squamous	9	2	4	5	1
Pulmonary	0	1	0	19	0
SCLC	4	1	0	1	0

GKSTD initialization : 123812813 num of iterations : 500 max_delta : 0.000029 exp_wt 1.500000	C1	C2	C3	C4	C5
Adeno	23	28	26	26	36
Normal	0	0	1	0	16
Squamous	8	6	1	4	2
Pulmonary	1	0	0	19	0
SCLC	4	1	0	1	0

Table 5: We have selected 15 features from OFS and cluster by HCM, FCM and GK-Method

HCM iterations 195 initialization 1: 1297233271 max_delta : 0.000098	C1	C2	C3	C4	C4
Adeno	44	27	5	30	33
Normal	3	11	0	1	2
Squamous	15	4	0	2	0
Pulmonary	1	0	7	11	1
SCLC	4	0	2	0	0

HCM iterations 55 initialization 2: 670172271 max_delta : 0.000087	C1	C2	C3	C4	C5
Adeno	3	30	32	34	40
Normal	0	11	2	3	1
Squamous	0	4	10	2	5
Pulmonary	7	0	1	12	0
SCLC	2	0	3	0	1

FCM exp_wt 2.000000 iterations 88 initialization 1: 123316271 max_delta : 0.000100	C1	C2	C3	C4	C5
Adeno	12	38	46	25	18
Normal	0	1	12	2	2
Squamous	4	1	11	2	3
Pulmonary	0	3	1	15	1
SCLC	0	2	0	2	2

FCM exp_wt 2.000000 iterations 68 initialization 2: 2147483647 max_delta : 0.000097	C1	C2	C3	C4	C5
Adeno	25	46	37	19	12
Normal	2	12	1	2	0
Squamous	2	11	1	4	3
Pulmonary	16	1	2	1	0
SCLC	2	0	2	2	0

GK exp_wt 1.500000 iterations 289 initialization 1: 923296873 max_delta : 0.000010	C1	C2	C3	C4	C5
Adeno	25	48	23	11	32
Normal	2	0	12	0	3
Squamous	2	4	4	7	4
Pulmonary	6	1	0	0	13
SCLC	2	0	0	0	4

GK exp_wt 1.500000 iterations 213 initialization 2: 71877713 max_delta : 0.000009	C1	C2	C3	C4	C5
Adeno	45	23	29	12	30
Normal	0	2	10	1	4
Squamous	3	8	2	6	2
Pulmonary	2	2	1	0	14
SCLC	1	0	3	1	1

Thus, showing resemblance with the subtypes of adenocarcinomas.
In the above exercise we clustered the data into 5 groups, as there are 5 histologically known classes. Since, the biggest class has subtypes, we could discover this with C = 5. For the very special structure of the dataset, it was possible, but in most applications to look for subtypes, one will need to look for more clusters than the number of classes.

The results obtained from SOFM shown in Table 6 and Table 7 supports existence of 5 subtypes in the whole cancer dataset. We have experimented with lattices of size $5 \times 5, 7 \times 7$ *and* 9×9. On the map we show the number of sample points of each class(histologically known classes) coming to each node, in the order : first entry is for class 1 samples(adenocarcinoma), second entry for class 2 samples(normal), third entry for class 3 samples(squamous), fourth entry for class 4 samples(pulmonary) and fifth entry for class 5 samples(SCLC). An asterisk(*) in the map indicates that none of the samples has converged to that node in the grid. The results obtained from HCM, FCM and GK-Method are in concordance with the clusters obtained by SOFM.

Our results are consistent with the results obtained by Bhattacharjee, [7]. The results show the intricacy of the cancer dataset, probably the subtypes of cancer are interwoven but can't be said concretely. It is to be noted that the partition matrix obtained by FCM shows almost uniformly equal membership of samples to each cluster. But in that of GK-Method the partition matrix is showing membership varying significantly from cluster to cluster.

8 Conclusions

We attempted two problems : identification of marker genes and finding of cancer subtypes. We achieved the following :

1. We used a Neural Network based system for identification of marker genes, and found only 9 genes that can explain 5 types of cancer. For finding subtypes in the histologically known classes we have used a set of 15 genes, which includes the 9 marker genes also.

2. We modified the implementation of OFS scheme so that we can deal with high dimensional data easily.

Table 6: OUTPUT OF SOFM

*	8 6 0 0 0	3 0 0 0 0	15 1 2 0 0	33 6 10 1 0
26 0 3 0 1	*	0 0 1 0 0	2 3 0 0 0	*
2 0 0 0 0	1 0 0 0 0	*	*	*
*	*	11 1 0 1 1	*	34 0 3 2 3
0 0 1 0 0	2 0 0 5 1	2 0 1 10 0	*	0 0 0 1 0

3 0 0 0 0	*	*	*	2 0 0 4 1	*	*
16 0 2 0 0	*	*	*	34 2 11 1	*	*
*	5 11 0 0 0	*	*	*	*	*
*	*	13 0 2 9 4	*	*	*	*
0 0 0 1 0	1 0 0 0 0	*	*	*	*	*
*	21 1 0 5 1	*	*	*	*	*
0 0 1 0 0	44 3 5 0 0	*	*	*	*	*

*	*	0 0 0 1 0	2 0 0 4 1	*	*	*
*	*	*	*	*	*	*
*	*	0 0 1 0 0	*	13 0 2 0 0	*	*
*	*	*	23 1 5 0 1	*	27 2 10 1 0	*
*	*	*	*	*	20 1 0 1 1	3 0 0 0 0
6 11 0 0 0	*	*	*	*	*	*
6 0 1 12 1	*	*	*	*	*	39 2 2 1 2

Table 7: OUTPUT OF SOFM

*	*	*	*	*	2 0 0 4 1	*	*	*
*	*	0 0 1 0 0	*	*	*	*	*	*
*	*	*	*	*	3 6 0 0 0	*	17 1 5 0 1	*
13 6 2 0 0	*	*	*	*	*	*	37 2 2 1 2	3 0 0 0 0
*	*	*	*	*	*	*	0 0 0 1 0	*
10 1 7 1 0	*	*	*	*	*	*	*	*
*	*	*	4 0 1 11 0	19 0 0 0 0	*	*	*	*
3 1 0 0 0	*	*	1 0 0 0 0	1 0 0 0 0	*	15 0 2 0 0	*	*
0 0 1 0 0	*	*	*	*	*	*	11 0 0 2 2	*

3. In this work, we used 3 clustering algorithms viz. HCM, FCM and GKFCM.

4. We made a simple modification of GKFCM so that for illconditioned covariance matrices, the algorithm can continue.

5. We also used SOFM for class discovery - they can avoid the problem of finding the optimal number of clusters.

6. First, we tried to use the original data with various normalizations but could not achieve much due to curse of dimensionality.

7. We then selected features using standard deviation criterion - an unsupervised method, but the result was not quite satisfactory.

8. We used a set of features selected by OFS for class discovery using HCM, FCM, GKFCM and SOFM.

9. Our results confirm the subtypes found by others. Also, we found that OFS selected features are better than that of standard deviation based method.

9 Programming environment

We have implemented the algorithms using C language. We have used "ruby" for scripting data files and used Matlab also. We have used Sun-Solaris and Pentium II systems to carry out the computations.

References

[1] Greenlee ,R.T.,Hill-harmon,M.B.,Murray,T. & Thun,M.(2001) CA Cancer J. Clin. 51,15 - 36.

[2] Travis, W.D., Travis, L.B. & Devesa, S.S.(1995) Cancer 75, 191-202.

[3] Sorensen, J.B., Hirsch, F.R., Gazdar, A. & Olsen, J.E. (1993) Cancer 71, 2971-2976.

[4] Breathnac, O.S., Kwiatkowski, D.J., Finkelstein, D.M., Godleski, J., Sugarbaker, D.J., JOhnson, B.E. & Mentzer, S. (2001) J. Thorac. Cardiovasc. Surg. 121, 42-47.

[5] Shirakusa Shirakusa T., Tsutsui, M., Motonaga, R., Ando, K. & Kusano, T. (1988) Am. Surg. 54, 655-658 and Flint A. & Lloyd, R.V., (1992) Arch. Pathol. Lab. Med. 116, 39-42.

[6] Golub, T.R., Slonim, D.K. Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M. L., Downing, J.R., Caligiuri, M.A., et al. (1999) Science 286, 531-537. and Simon et al., 2003.

[7] Arindam Bhattacharjee, William G. Richards, Jane Staunton, Cheng Li et al., Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, PNAS Vol. 98, 2001.

[8] Kim, D.W. et al., Detecting Clusters of Different Geometrical Shapes in Microarray Gene Expression Data, Vol. 21,2005. Bioinformatics.

[9] Baird D. et al., Normalization of Microarray Data Using a Spatial Mixed Model Analysis Which Includes Splines. Vol. 20 , 2004. Bioinformatics.

[10] Ptitsyn A., Pennington Biomedical Research Center.

[11] Eisen MB, Spellman PT, Brown PO, Botstein D. (1998) Cluster analysis and display of genome-wide expression patterns Proc Natl Acad Sci U S A 95:14863-14868.

[12] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531-537.

[13] NR Pal, JC Bezdek ,On Cluster Validity for the Fuzzy c-Means Model,1997,IEEE Trans. Fuzzy Systems.

[14] James c. Bezdek ,James Keller, Raghu Krishnapuram, Nikhil R Pal,1999,Kluwer Academic Publishers,p 14-23 and p 241-253.

[15] Schena, M., Shalon, D., Davis, RW & Brown,Use of a cDNA microarray to analyse gene expression patterns in human cancer,1999,N Genetics - Nature Genetics

[16] Richard O. Duda, Peter E. Hart, David G. Stork,Pattern Classification (2nd Edition),2000 Wiley-Interscience.

[17] Pal NR, Chintalapudi KK. (1997) A connectionist system for feature selection Neural, Parallel and Scientific Computations (5) 359-382

[18] D. Chakraborty and N. R. Pal, A neuro-fuzzy scheme for simultaneous feature selection and fuzzy rule-based classification, *IEEE Trans. Neural Networks*, Vol. 15, No. 1, pp 110-123, 2004.

[19] Nikhil Ranjan Pal, Animesh Sharma, Somitra Kumar Sanadhya and Karmeshu,On Identifying Marker Genes from Gene Expression Data in a Neural Framework through Online Feature Analysis,2005 communicated.

[20] Brown P, Bostein D. (1999) Exploring the new world of the genome with DNA microarrays Nature Genetics 21(1 suppl):33-37.

[21] B Lemieux, A Aharoni, M Schena,Overview of DNA chip technology-1998,Molecular Breeding

[22] KM Kurian, CJ Watson, AH Wyllie,DNA Chip Technology–1999,The Journal of Pathology

[23] HC Liu, Z He, Z Rosenwaks,Application of complementary DNA microarray (DNA chip) technology in the study of gene expression profiles during folliculogenesis. - 2001,Fertile Steric

[24] JC Bezdek, NR Pal ,Some new indexes of cluster validity,1998,IEEE Transactions on Systems, Man, and Cybernetics

[25] C. E. Forkner, Leukemia and Allied Disorders (Macmillan, New York, 1938); E. Frei et al., Blood 18, 431(1961); Medical Research Council, Br. Med. J. 1, 7