Gene Pattern Analysis Using Expectation Maximization Algorithm

J.Sathia Parkavi[1], AP.Abinaya[2], C. Harini[3], S.Jamuna Lakshmi[4], R.Kiruthiga[5]
[1, 2, 3, 4, 5] Department of Computer Science and Engineering
[1, 2, 3, 4, 5] Saranathan College of Engineering,Tamil Nadu.

Abstract- *: Microarray technology is one of the important biotechnological means that allows recording the expression levels of thousands of genes simultaneously within a number of different samples. An important application of microarray gene expression data in functional genomics is to classify samples according to their gene expression profiles. When the number of genes is greater than the number of samples, it is possible to find biologically relevant correlations of gene behavior with the sample categories. Hence, one of the major tasks with the gene expression data is to find groups of co-regulated genes whose collective expression is strongly associated with the sample categories. So we implement reduce dimensionality, removing irrelevant data and increase diagnosis accuracy and presents learning method which is able to group genes based on their interdependence so as to mine meaningful patterns from the gene expression data using Spatial EM algorithm. It can be used to calculate spatial mean and rank based scatter matrix to extract relevant patterns and further implement KNN (K- nearest neighbor classification) approach to diagnosis the diseases. We have identified that semi supervised clustering algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability. The experimental results prove that Spatial EM based classification approach Provides improved accuracy rate in disease diagnosis.*

Keywords- Gene, Data, Clustering, EM Algorithm, KNN Algorithm, Data Mining

I. INTRODUCTION

In Today’s Era, Information plays a vital role. Initially when computers were invented, there is a need for mass digital storage, which leads to gathering and categorizing all sorts of data. Unfortunately these storages immediately became conquering, all of these factors gave rise to creation of structured database and database management system. In order to group various kinds of data into their categories, clustering is used. Here we cluster gene datasets to predict the disease of a person by using Expectation and Maximization algorithm and K-Nearest Neighbor algorithm. In the existing gene pattern clustering, clustering process is performed only based on a user defined manner, there is no classification to diagnosis the disease properly and it is also difficult to predict outliers in gene expression data. All these disadvantages cause improper clustering of data.

This Gene Pattern Analysis Using Expectation-Maximiation Algorithm provides proper clustering, predict the disease for each gene dataset and provides good accuracy.

DATA MINING

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large datasets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics.

i) DATA:

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- Operational or transactional data such as, sales, cost, inventory, payroll, and accounting.
- Nonoperational data, such as industry sales, forecast data, and macro economic data.
- Meta data - data about the data itself, such as logical database design or data dictionary definitions.

ii) INFORMATION:

The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

iii) KNOWLEDGE:

Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

iv) DATAMINING ELEMENTS:

Extract, transform, and load transaction data onto the data warehouse system. Store and manage the data in a multidimensional database system. Provide data access to business analysts and information technology professionals. Analyze the data by application software. Present the data in a useful format, such as a graph or table.

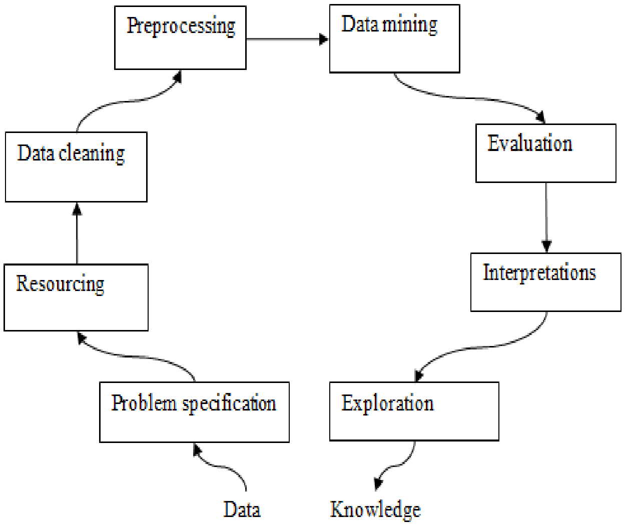


Figure 1. Levels of Data Mining

v) DATAMINING METHODS:

There are several major data mining technique have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns.

1) ASSOCIATION

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in market basket analysis to identify what products that customers frequently purchase together. Based on this data businesses can have corresponding marketing campaign to sell more products to make more profit.

2) CLASSIFICATION

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics.

In classification, make the software that can learn how to classify the data items into groups. For example, can apply classification in application that “given all past records of employees who left the company, predict which current employees are probably to leave in the future.” In this case, divide the employee’s records into two groups that are “leave” and “stay”. And then can ask data mining software to classify the employees into each group.

3) CLUSTERING

Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes.

To make the concept clearer, we can take library as an example. In a library, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic without hassle. By using clustering technique, can keep books that have some kind of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in a topic, he or she would only go to that shelf instead of looking the whole in the whole library.

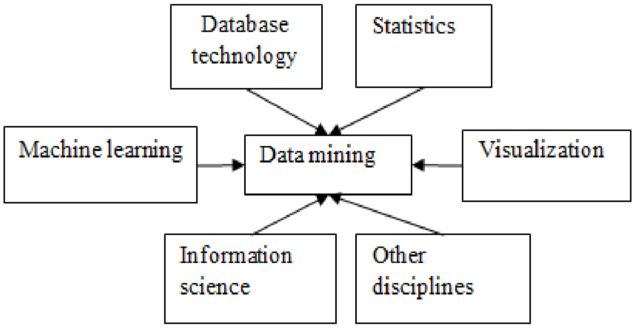


Figure 2. Techniques of Data Mining

II. OBJECTIVE

The overall objective of the data mining process is to provide information from a data set and organize them into an understandable structure for further use.

The following are some of the main objectives

1) MASSIVE DATA COLLECTION: Thousands and thousands of piece of gene datasets are clustered to diagnose the disease.
2) ANALYSING HIDDEN PATTERNS: By using EM and KNN algorithm, clustering and classification process will be done in an effective manner and hence the disease which can't be identified by the doctor, can be predicted easily in this proposed system.
3) RESOURCE REDUCTION: The genes are grouped into three's and they are converted into a matrix called RANK-BASED SCATTER MATRIX ,hence the resources required for storing the datasets will be low.
4) ALGORITHM'S BENEFIT: Outliers are predicted efficiently in gene expression data. Automatic clustering process is done. Efficiently diagnose the diseases using classification performance.
5) TIME CONSUMPTION: Since efficient algorithms has been used for clustering and classification process the time required will be less compared to that of the existing system.
6) ACCURACY: In KNN algorithm we use semi supervised method which will be more accurate than that of other methods. This semi supervised algorithm provide results in an accurate manner.

III. PROPOSED SYSTEM

In proposed system, we implement Spatial EM algorithm for analyzing microarray datasets. It is used to identify cluster location from group gene datasets by utilizing robust location and scatter estimators in each M-step. Able to represent arbitrarily complex structure of data. Another common technique for robust fitting of mixtures is to update the component estimates on the M-step of the EM algorithm by some robust location and scatter estimates. M estimator has been considered .

Minimum Covariance Determinant (MCD) estimator is used for cluster analysis. It is recommended use of S estimator. In order to diagnose the disease, spatial rank based location and scatter estimators is applied. They are highly robust and are computationally and statistically more efficient than the above robust estimators. We develop a Spatial-EM algorithm for robust finite mixture learning. Based on the Spatial-EM, supervised outlier detection and unsupervised clustering methods are illustrated and compared with other existing techniques.

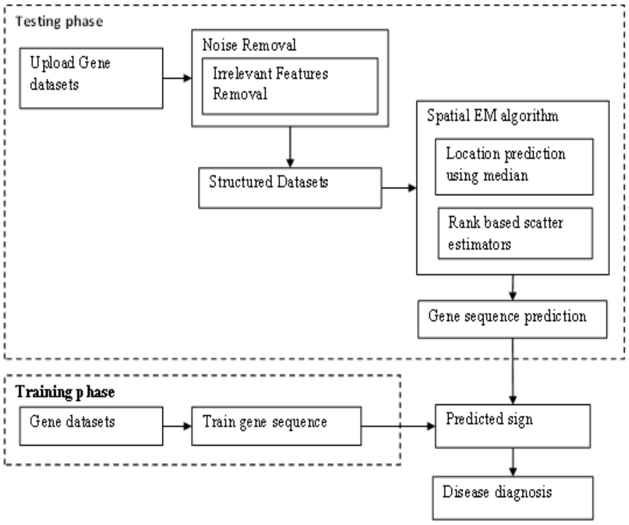


Figure 3. System Architecture

The following are 5 major modules in this proposed system

- Dataset Acquisition
- Median Estimation
- Rank Based Scatter
- Disease Prediction
- Evaluation Criteria

1) DATASET ACQUISITION

In this module, upload the datasets. The dataset may be microarray dataset. A microarray database is a repository containing microarray gene expression data. The key uses of a microarray database are to store the measurement data, manage a searchable index, and make the data available to other applications for analysis and interpretation. Data pre-processing is an important step in the data mining process.

The phrase "garbage in, garbage out" is particularly applicable to data mining and machine projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values, missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult.

Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set. Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant, etc.

Parts of the data and then replacing, modifying, or deleting this dirty data or coarse data. After cleansing, a data set will be consistent with other similar data sets in the system.

The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleansing differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at entry time, rather than on batches of data. The actual process of data cleansing may involve removing typographical errors or validating and correcting values against a known list of entities.

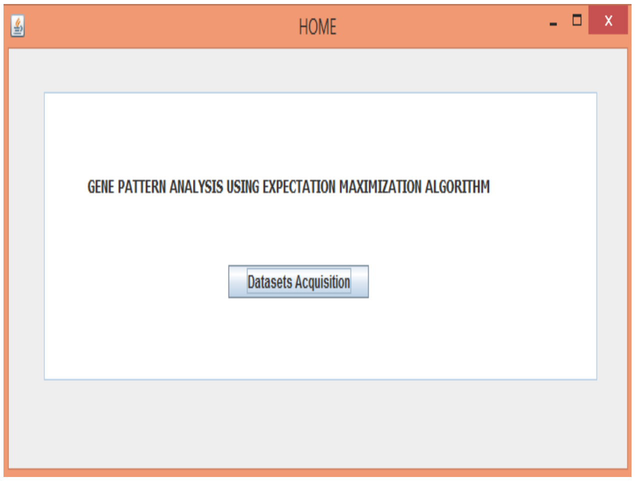
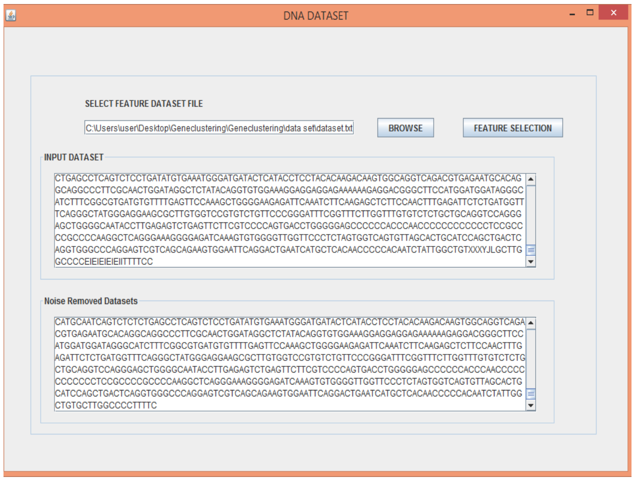

HOME
GENE PATTERN ANALYSIS USING EXPECTATION MAXIMIZATION ALGORITHM
Datasets Acquisition

Figure 4.



C:\Users\user\Desktop\Geneclustering\Geneclustering\data set\dataset.txt

```
CTGAGCCTCAGTCTCCTGATATGTGAAATGGGATGATACTCATACCTCCTACACAAGACAAGTGGCAGGTCAGACGTGAGAATGCACAG
GCAGGCCCTTCGCAACTGGATAGGCTCTATACAGGTGTGGAAAGGAGGAGGAGAAAAAAGAGGACGGGCTTCCATGGATGGATAGGGC
ATCTTTCGGCGTGATGTGTTTTGAGTTCCAAAGCTGGGGAAGAGATTCAAATCTTCAAGAGCTCTTCCAACTTTGAGATTCTCTGATGGTT
TCAGGGCTATGGGAGGAAGCGCTTGTGGTCCGTGTCTGTTCCCGGGATTTCGGTTTCTTGGTTTGTGTCTCTGCTGCAGGTCCAGGG
AGCTGGGGCAATACCTTGAGAGTCTGAGTTCTTCGTCCCCAGTGACCTGGGGGAGCCCCCCACCCAACCCCCCCCCCCCTCCGCC
CCGCCCCAAGGCTCAGGGAAAGGGGAGATCAAAGTGTGGGGTTGGTTCCCTCTAGTGGTCAGTGTTAGCACTGCATCCAGCTGACTC
AGGTGGGCCCAGGAGTCGTCAGCAGAAGTGGAATTCAGGACTGAATCATGCTCACAACCCCCACAATCTATTGGCTGTXXXYJLGCTTG
GCCCCEIEIEIEIEIITTTTCC
```

```
CATGCAATCAGTCTCTCTGAGCCTCAGTCTCCTGATATGTGAAATGGGATGATACTCATACCTCCTACACAAGACAAGTGGCAGGTCAGA
CGTGAGAATGCACAGGCAGGCCCTTCGCAACTGGATAGGCTCTATACAGGTGTGGAAAGGAGGAGGAGAAAAAAGAGGACGGGCTTCC
ATGGATGGATAGGGCATCTTTCGGCGTGATGTGTTTTGAGTTCCAAAGCTGGGGAAGAGATTCAAATCTTCAAGAGCTCTTCCAACTTTG
AGATTCTCTGATGGTTTCAGGGCTATGGGAGGAAGCGCTTGTGGTCCGTGTCTGTTCCCGGGATTTCGGTTTCTTGGTTTGTGTCTCTG
CTGCAGGTCCAGGGAGCTGGGGCAATACCTTGAGAGTCTGAGTTCTTCGTCCCCAGTGACCTGGGGGAGCCCCCCACCCAACCCCC
CCCCCCCTCCGCCCCGCCCCAAGGCTCAGGGAAAGGGGAGATCAAAGTGTGGGGTTGGTTCCCTCTAGTGGTCAGTGTTAGCACTG
CATCCAGCTGACTCAGGTGGGCCCAGGAGTCGTCAGCAGAAGTGGAATTCAGGACTGAATCATGCTCACAACCCCCACAATCTATTGG
CTGTGCTTGGCCCCTTTTC
```

Figure 5.

2) MEDIAN ESTIMATION

To tackle the effect of outliers in cluster analysis to consider the Spatial EM clustering which replaces the squared Euclidean distances in the objective function of the k-means clustering with the absolute Euclidean distances. In spatial EM, can analyze coverage of the data before clustering begins and propose an algorithm, which modifies the nearest centroid sorting and the transfer algorithm, of the spatial medians clustering.

It has two distinct phases: one of transferring an object from one cluster to another and the other of amalgamating the single member cluster with it's the nearest cluster. Given a starting partition, each possible transfer is tested in turn to see if it would improve the value of clustering criterion. When no further transfers can improve the criterion value, each possible amalgamation of the single member cluster and other clusters is tested.

The amalgamation of the single member cluster should be executed with the detachment of an object which is far from its cluster centroid when it is found to be beneficial. When no further amalgamations give an improvement, the transfer phase is reentered and continued until no more transfers or amalgamations can improve the clustering criterion value.

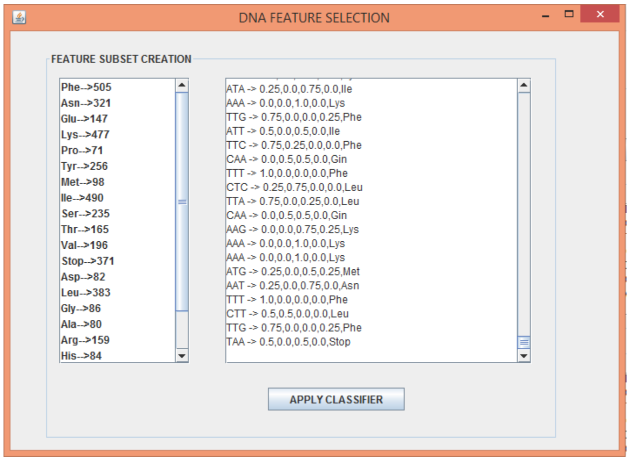
In this module, we can calculate the mean values for each gene features. These gene features listed as it is. Spatial-EM modifies the component estimates on each M-step by spatial median and rank covariance matrix to gain robustness at the cost of increasing computational burden and losing theoretical tractability. For single component elliptically symmetric models, consistency and efficiency of the rank covariance have been established in and with an amount of effort.

The extension to a mixture model loses mathematical tractability due to deletion of a portion of smallest values of the projected data in the update of covariance matrix. The whole procedure hybridizes soft and hard labels at each iteration, which makes the connection to maximum likelihood approximation extremely difficult to verify theoretically. In such a desperate situation, demonstrating empirical evidence seems to be the only thing we can do.

3) RANK BASED SCATTER:

In this module, can create scatter matrix based on median values that are derived by clustering algorithm. Then construct scatter matrix and reflecting as the within-cluster scatter, the between-cluster scatter and their summation — the total scatter matrix. The determinant of a scatter matrix roughly measures the square of the scattering volume. And minimizing this measure is equivalent to both minimizing the intra-cluster scatter and maximizing the inter-cluster scatter. Based on scatter matrix, classification is performed in following modules. Mixture model-based clustering is one of the most popular and successful unsupervised learning approaches. It provides a probabilistic (soft) clustering of the data in terms of the fitted posterior probabilities (Tji's) of membership of the mixture components with respect to the clusters. An outright (hard) clustering can be subsequently obtained by assigning each observation to the component to which it has the highest fitted posterior probability of belonging. That is, xi is assigned to the cluster argmaxjTji.

Model-based clustering approaches have a natural way to select the number of clusters based on some criteria, which have the common form of log-likelihood augmented by a model complexity penalty term. For example, Bayesian Inference Criterion (BIC), the Minimum Message Length (MML) , the Normalized Entropy Criterion (NEC) etc. have yielded good results for model choice in a range of applications. In this paper, we deal with robustness of model-based clustering. We assume that the number of clusters is known, otherwise, BIC is used. BIC is defined as twice of the log-likelihood minus p log N, where the likelihood is the Gaussian based, N is the sample size and p is the number of independent parameters. For a K component mixture model, with d being the dimension.



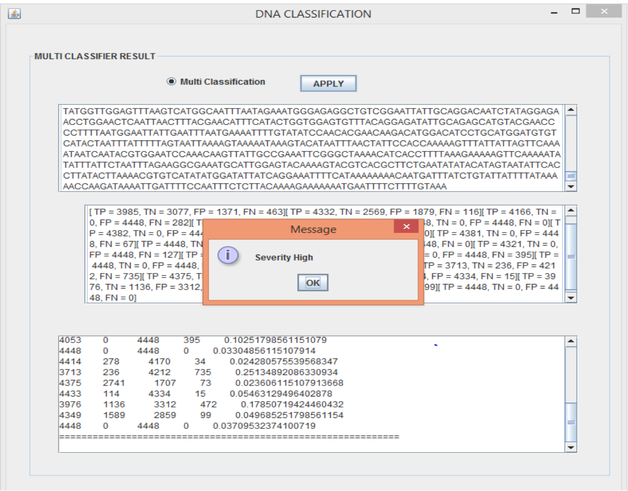
```
Phe-->505
Asn-->321
Glu-->147
Lys-->477
Pro-->71
Tyr-->256
Met-->98
Ile-->490
Ser-->235
Thr-->165
Val-->196
Stop-->371
Asp-->82
Leu-->383
Gly-->86
Ala-->80
Arg-->159
His-->84
```

```
ATA -> 0.25,0.0,0.75,0.0,Ile
AAA -> 0.0,0.0,1.0,0.0,Lys
TTG -> 0.75,0.0,0.0,0.25,Phe
ATT -> 0.5,0.0,0.5,0.0,Ile
TTC -> 0.75,0.25,0.0,0.0,Phe
CAA -> 0.0,0.5,0.5,0.0,Gin
TTT -> 1.0,0.0,0.0,0.0,Phe
CTC -> 0.25,0.75,0.0,0.0,Leu
TTA -> 0.75,0.0,0.25,0.0,Leu
CAA -> 0.0,0.5,0.5,0.0,Gin
AAG -> 0.0,0.0,0.75,0.25,Lys
AAA -> 0.0,0.0,1.0,0.0,Lys
AAA -> 0.0,0.0,1.0,0.0,Lys
ATG -> 0.25,0.0,0.5,0.25,Met
AAT -> 0.25,0.0,0.75,0.0,Asn
TTT -> 1.0,0.0,0.0,0.0,Phe
CTT -> 0.5,0.5,0.0,0.0,Leu
TTG -> 0.75,0.0,0.0,0.25,Phe
TAA -> 0.5,0.0,0.5,0.0,Stop
```

Figure 6.

4) DISEASE PREDICTION

Classifiers based on gene expression are generally probabilistic, that is they only predict that a certain percentage of the individuals that have a given expression profile will also have the phenotype, or outcome, of interest. Therefore, statistical validation is necessary before models can be employed, especially in clinical settings. KNN approach matches each neighborhood genes to predict the diseases. In this module, implement classifier design in semi supervised format.K nearest neighbor classifier allowed to access and provides predicted sign for corresponding diseases such as diabetic, leukemia and so on. And calculate the severity of the disease.



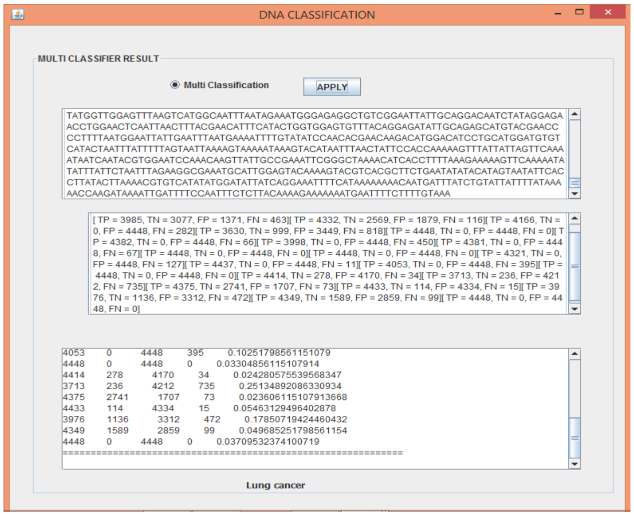
```
TATGGTTGGAGTTTAAGTCATGGCAATTTAATAGAAATGGGAGAGGCTGTCGGAATTATTGCAGGACAATCTATAGGAGA
ACCTGGAACTCAATTAACTTTACGAACATTTCATACTGGTGGAGTGTTTACAGGAGATATTGCAGAGCATGTACGAACC
CCTTTTAATGGAATTATTGAATTTAATGAAAATTTTGTATATCCAACACGAACAAGACATGGACATCCTGCATGGATGTGT
CATACTAATTTATTTTTAGTAATTAAAAGTAAAAATAAAGTACATAATTTAACTATTCCACCAAAAAGTTTATTATTAGTTCAAA
ATAATCAATACGTGGAATCCAAACAAGTTATTGCCGAAATTCGGGCTAAAACATCACCTTTTAAAGAAAAAGTTCAAAAATA
TATTTATTCTAATTTAGAAGGCGAAATGCATTGGAGTACAAAAGTACGTCACGCTTCTGAATATATACATAGTAATATTCAC
CTTATACTTAAAACGTGTCATATATGGATATTATCAGGAAATTTTCATAAAAAAAACAATGATTTATCTGTATTATTTTATAAA
AACCAAGATAAAATTGATTTTCCAATTTCTCTTACAAAAGAAAAAAATGAATTTTCTTTTGTAAA
```

```
[ TP = 3985, TN = 3077, FP = 1371, FN = 463][ TP = 4332, TN = 2569, FP = 1879, FN = 116][ TP = 4166, TN =
0, FP = 4448, FN = 282][ T                                              8, TN = 0, FP = 4448, FN = 0][ T
P = 4382, TN = 0, FP = 44                                               0][ TP = 4381, TN = 0, FP = 444
8, FN = 67][ TP = 4448, TN                                              48, FN = 0][ TP = 4321, TN = 0,
FP = 4448, FN = 127][ TP =                                              = 0, FP = 4448, FN = 395][ TP =
 4448, TN = 0, FP = 4448,                                               TP = 3713, TN = 236, FP = 421
2, FN = 735][ TP = 4375, T                                              , FP = 4334, FN = 15][ TP = 39
76, TN = 1136, FP = 3312,                                               99][ TP = 4448, TN = 0, FP = 44
48, FN = 0]
```

Severity High

```
4053    0       4448    395     0.10251798561151079
4448    0       4448    0       0.033048561151079l4
4414    278       4170    34      0.024280575539568347
3713    236       4212    735       0.25134892086330934
4375    2741        1707    73       0.023606115107913668
4433    114       4334    15      0.05463129496402878
3976    1136        3312    472       0.17850719424460432
4349    1589        2859    99       0.049685251798561154
4448    0       4448    0       0.03709532374100719
==============================================================
```

Figure 7.



```
TATGGTTGGAGTTTAAGTCATGGCAATTTAATAGAAATGGGAGAGGCTGTCGGAATTATTGCAGGACAATCTATAGGAGA
ACCTGGAACTCAATTAACTTTACGAACATTTCATACTGGTGGAGTGTTTACAGGAGATATTGCAGAGCATGTACGAACC
CCTTTTAATGGAATTATTGAATTTAATGAAAATTTTGTATATCCAACACGAACAAGACATGGACATCCTGCATGGATGTGT
CATACTAATTTATTTTTAGTAATTAAAAGTAAAAATAAAGTACATAATTTAACTATTCCACCAAAAAGTTTATTATTAGTTCAAA
ATAATCAATACGTGGAATCCAAACAAGTTATTGCCGAAATTCGGGCTAAAACATCACCTTTTAAAGAAAAAGTTCAAAAATA
TATTTATTCTAATTTAGAAGGCGAAATGCATTGGAGTACAAAAGTACGTCACGCTTCTGAATATATACATAGTAATATTCAC
CTTATACTTAAAACGTGTCATATATGGATATTATCAGGAAATTTTCATAAAAAAAACAATGATTTATCTGTATTATTTTATAAA
AACCAAGATAAAATTGATTTTCCAATTTCTCTTACAAAAGAAAAAAATGAATTTTCTTTTGTAAA
```

```
[ TP = 3985, TN = 3077, FP = 1371, FN = 463][ TP = 4332, TN = 2569, FP = 1879, FN = 116][ TP = 4166, TN =
0, FP = 4448, FN = 282][ TP = 3630, TN = 999, FP = 3449, FN = 818][ TP = 4448, TN = 0, FP = 4448, FN = 0][ T
P = 4382, TN = 0, FP = 4448, FN = 66][ TP = 3998, TN = 0, FP = 4448, FN = 450][ TP = 4381, TN = 0, FP = 444
8, FN = 67][ TP = 4448, TN = 0, FP = 4448, FN = 0][ TP = 4448, TN = 0, FP = 4448, FN = 0][ TP = 4321, TN = 0,
FP = 4448, FN = 127][ TP = 4437, TN = 0, FP = 4448, FN = 11][ TP = 4053, TN = 0, FP = 4448, FN = 395][ TP =
 4448, TN = 0, FP = 4448, FN = 0][ TP = 4414, TN = 278, FP = 4170, FN = 34][ TP = 3713, TN = 236, FP = 421
2, FN = 735][ TP = 4375, TN = 2741, FP = 1707, FN = 73][ TP = 4433, TN = 114, FP = 4334, FN = 15][ TP = 39
76, TN = 1136, FP = 3312, FN = 472][ TP = 4349, TN = 1589, FP = 2859, FN = 99][ TP = 4448, TN = 0, FP = 44
48, FN = 0]
```

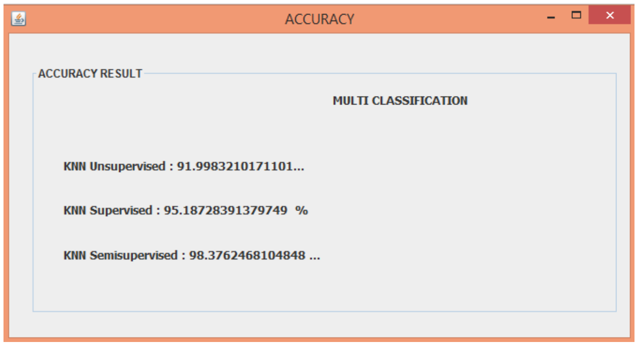
```
4053    0       4448    395     0.10251798561151079
4448    0       4448    0       0.033048561151079l4
4414    278       4170    34      0.024280575539568347
3713    236       4212    735       0.25134892086330934
4375    2741        1707    73       0.023606115107913668
4433    114       4334    15      0.05463129496402878
3976    1136        3312    472       0.17850719424460432
4349    1589        2859    99       0.049685251798561154
4448    0       4448    0       0.03709532374100719
==============================================================
```

Lung cancer

Figure 8.

5) EVALUATION CRITERIA

In this module, the performance of the proposed semi-supervised algorithm is extensively compared with that of some existing supervised and unsupervised gene clustering and gene selection algorithms. To analyze the performance of different algorithms, the experimentation is done on microarray gene expression data sets. The major metrics for evaluating the performance of different algorithms are the class separability index and classification accuracy of K-nearest neighbor rule. The proposed system provide improved accuracy rate in gene classification.



ACCURACY RESULT

MULTI CLASSIFICATION

KNN Unsupervised : 91.9983210171101...

KNN Supervised : 95.18728391379749 %

KNN Semisupervised : 98.3762468104848 ...

Figure 9.

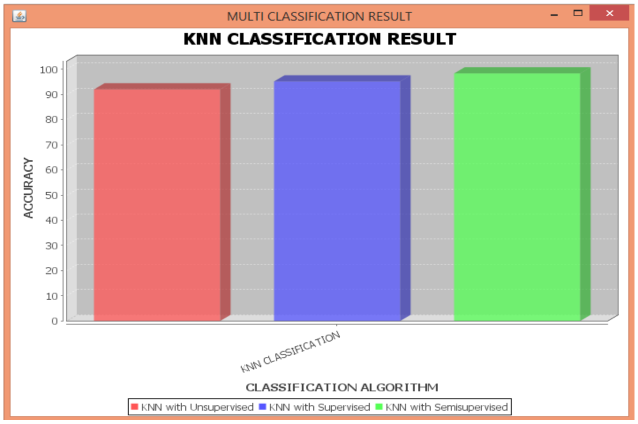


Figure 10.

IV. CONCLUSION

Recent DNA microarray technologies have made it possible to monitor transcription levels of tens of thousands of genes in parallel. Gene expression data generated by microarray experiments offer tremendous potential for advances in molecular biology and functional genomics. Here both classical and recently developed clustering algorithms are reviewed, which have been applied to gene expression data, with promising results. The proposed semi-supervised spatial EM clustering algorithm is based on measuring mean values and scatter matrix using the new quantitative measure, whereby redundancy among the attributes is removed. The clusters are then refined incrementally based on sample categories. The performance of the proposed algorithm is compared with that of existing supervised EM gene selection algorithm with its severity level whether it is medium, high or low.

REFERENCES

[1] Bashir and E. M. Carter, “High breakdown mixture discriminant analysis,” J. Multivariate Anal., Vol. 93, No. 1, pp. 102–111.

[2] Y. Chen, Bart H. Jr, X. Dang, and H. Peng, “Depth-based novelty detection and its application to taxonomic research,” in Proc. 7th IEEE Int. Conf. Data Mining, Omaha, Nebraska, 2007, pp. 113–122.

[3] Y. Chen, X. Dang, H. Peng, and H. Bart Jr., “Outlier detection with the kernelized spatial depth function,” IEEE Trans. Pattern Anal. Mach. Intell., Vol. 31, No. 2, pp. 288–305.

[4] Y. Chueng, “Maximum weighted likelihood via rival penalized EM for density mixture clustering with automatic model selection,” IEEE Trans. Knowl. Data Eng., Vol. 17, No. 6, pp. 750–761.

[5] X. Dang and R. Serfling, “Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties,” J. Stat. Inference Planning, Vol. 140, pp. 198–213.