ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 2, Issue 1, January 2014

MICROARRAY ANALYSIS USING FUZZY C-MEANS CLUSTERING ALGORITHM

Revathi T [1], Sathish A[2], Sumathi P[3]

Associate Professor & Research Scholar, Manonamniam Sundaranar University,Tirunelveli[1]

M.Phil Research Scholar, Department Of Computer Science, PSG College of Arts & Science, Coimbatore[2]

Assistant Professor, PG & Research Department of Computer Science, Govt.Arts College, Coimbatore[3]

Abstract: The technology of DNA microarrays has become the most sophisticated and the most widely used among other microarrays. This paper shows the feature of microarray analysis and the expanded information of DNA microarray analysis. The clustering technique is the process of finding a structured data from unlabeled data. It is a grouping process of dividing the data in groups of similar type and it contains different types of clusters like hierarchical, exclusive, overlapping and probabilistic. Each group is referred to as a cluster which contains objects of similar type. The data set for processing is taken from UCI Machine Learning Repository website and the analysis is done using the WEKA tool (Waikato Environment for Knowledge Analysis) which is an effective tool for machine learning. WEKA tool is Java-Based version which contains the collection of visualization tools and algorithms like clustering, classification, regression, preprocessing etc for data analysis. In this paper the microarray dataset is taken for predicting Breast cancer with the help of Fuzzy C-Means clustering technique.

Keywords: Microarray Analysis, DNA Microarray analysis, Cluster Analysis, Fuzzy C-Means Algorithm.

I. INTRODUCTION

A **microarray** is a 2D array on a solid substrate which is a glass slide or silicon thin-film cell that assays large amounts of biological material using high-throughput screening miniaturized, multiplexed and parallel processing and detection methods. The following are the types of microarrays:

- DNA microarrays, such as BAC microarrays and SNP microarrays
- MMChips, for surveillance of microRNA populations
- Protein microarrays
- Peptide microarrays, for detailed analyses or optimization of protein-protein interactions
- Tissue microarrays
- Cellular microarrays (also called transfection microarrays)
- Chemical compound microarrays
- Antibody microarrays
- Carbohydrate arrays (glycoarrays)
- Phenotype microarrays
- interferometric reflectance imaging sensor (IRIS)

II. DNA MICROARRAY

Scientists used to be able to perform genetic analysis of a few genes at once. DNA microarray allows us to evaluate thousands of genes in one experiment. A **DNA microarray** is also called as DNA chip or biochip. It is a collection of microscopic DNA spots attached to the solid surface[1]. Scientists use DNA microarrays to compute the expression levels of large numbers of genes

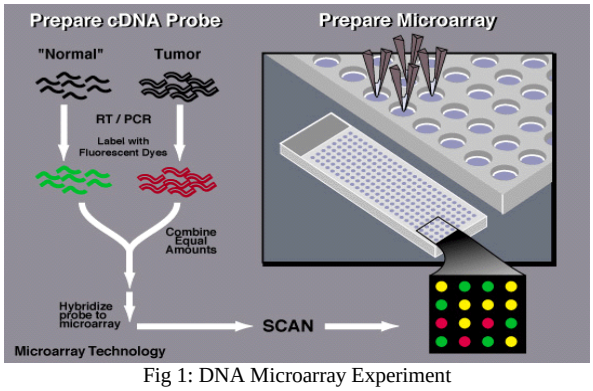

Fig 1: DNA Microarray Experiment

simultaneously or to genotype multiple regions of a genome. Each DNA spot contains picomoles of a specific DNA sequence, known as *probes*. These can be a short section of a gene or other DNA element that are used to hybridize a cDNA or cRNA also called anti-sense RNA sample which is referred as *target* under high-stringency conditions. Probe-target hybridization is mostly detected and quantified by discovery of fluorophore-, silver-, or chemiluminescence- labeled targets to determine relative abundance of nucleic acid sequences in the target.

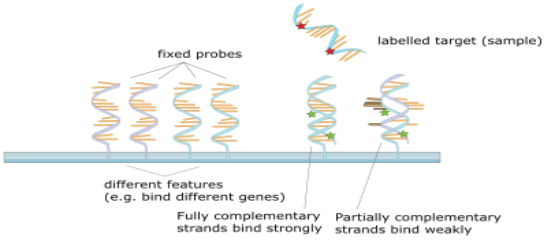

Fig 2: DNA Microarray Sample

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 2, Issue 1, January 2014

III. USES OF DNA MICROARRAY

1. To measure changes in gene expression levels – two samples gene expression can be compared from various samples, such as from cells of different stages of mitosis.
2. To observe genomic gains and losses. They are called as Microarray Comparative Genomic Hybridization (CGH)
3. To observe mutations in DNA.

IV. STEPS INVOLVED IN DNA MICRO ARRAY PROCESSING

1) Collect Samples.
2) Isolate mRNA.
3) Create Labeled DNA.
4) Hybridization.
5) Microarray Scanner.
6) Analyze Data.

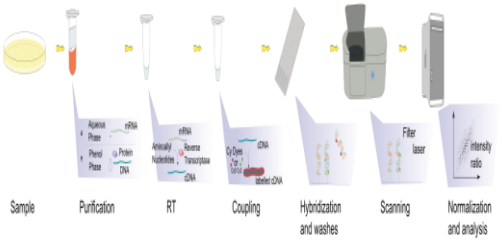

Fig 3: Steps in DNA Microarray Process

Step 1: Collect Samples.

- This can be from a variety of organisms. We'll use two samples – cancerous human skin tissue & healthy human skin tissue.

Step 2: Isolate mRNA.

- Extract the RNA from the samples. Using also a column, or a solvent such as phenol-chloroform.
- After isolating the RNA, we need to isolate the mRNA from the rRNA and tRNA. m RNA has a poly-A tail, so we can use a column containing beads with poly-T tails to bind the mRNA.
- Rinse with buffer to release the mRNA from beads.

Step 3: Create Labeled DNA.

- Add a labelling mix to the RNA. The labelling mix contains poly-T (oligo dT) primers, reverse transcriptase (to build cDNA), and fluorescently dyed nucleotides.
- We have to add cyanine 3 (fluoresces green) to the healthy cells and also cyanine 5 (fluoresces red) to the cancerous cells.
- The primer and RT bind to the mRNA first, and then add the fluorescently dyed nucleotides, generating a complementary strand of DNA

Step 4: Hybridization.

To apply the cDNA we have created to a microarray plate.
- When comparing two samples, apply both samples to the same plate.
- The ssDNA will attach to the cDNA already present on the plate.

Step 5: Microarray Scanner.

- The scanners have a laser, a computer, and a camera also.
- The laser contains the hybrid bonds to fluoresce.
- The camera records the images and it produced when the laser scans the plate.
- The computer allows us to instantly view our results and it also stores our data easily.

Step 6: Analyze Data.

- GREEN – the color indicates healthy sample hybridized more than the diseased sample.
- RED – the diseased/cancerous sample hybridized more than the nondiseased sample.
- YELLOW - both samples hybridized equally to the target DNA.
- BLACK - areas where neither sample hybridized to the target DNA.
- By comparing the differences in gene expression between the samples, we can know more about the genomics of a disease.

V. CLUSTER ANALYSIS

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of grouping data mining, and a common technique for statistical data analysis, used in many areas, including pattern recognition and image analysis, information retrieval, and also in bioinformatics. There are different types of clustering which can be used for analysis. They are:

(i) Connectivity based clustering:

Connectivity based clustering, also known as *hierarchical clustering*, is mainly based on the important ideas of objects being more related to nearby objects than to objects farther away. These algorithms include "objects" to form "clusters" based on their distance. Connectivity based clustering is a whole family of methods that differ by the way distances are computed.

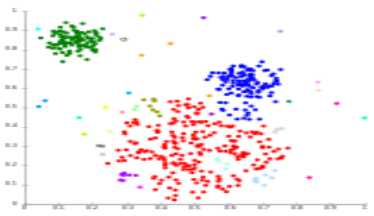
Fig 4: Connectivity-Based Clustering

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 2, Issue 1, January 2014

(ii) Centroid-based clustering:

In centroid-based clustering, clusters are represented by a central vector, which not necessarily is a member of the data set. The Centroid-based clustering is K-means clustering. Most k-means-type algorithms require the number of clusters - k- to be specified in advance, which is measured to be one of the biggest drawbacks of this algorithm. Furthermore, the algorithm prefers clusters of approximately similar size, as they will assign an object to the nearest centroid.

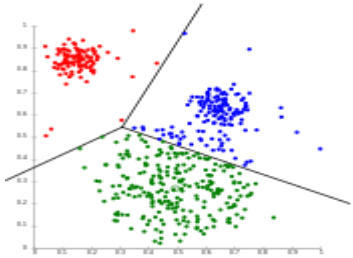

Fig 5: Centroid-Based Clustering

(iii) Distribution-based clustering:

The clustering model most closely related to statistics is based on the distribution models. Clusters can easily be defined as objects belonging most likely to the same distribution models. A nice property of this approach is that this closely resembles the way artificial data sets are generated: by sampling random objects from a distribution.

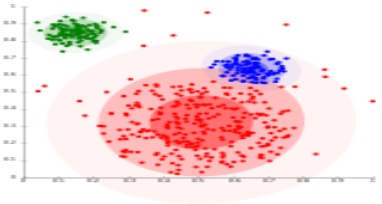

Fig 6: Distribution-Based Clustering

(iv) Density-based clustering:

In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Most of the objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points. The very popular density based clustering method is DBSCAN. In this contrast to many new methods, it features as well-defined cluster model called "density-reachability".

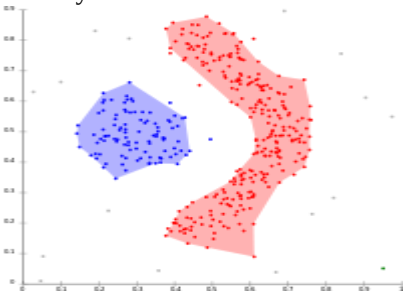

Fig 7: Density-Based Clustering

Clustering result evaluation:

There are two types of evaluation techniques in clustering analysis. They are as follows:

(i)Internal Evaluation:

When a clustering result is evaluated based on the data that was clustered itself, it was called as internal evaluation. These methods normally assign the best score to the algorithm that produces clusters with high similarity within a cluster and low similarity between clusters.

The subsequent methods can be used to assess the quality of clustering algorithms based on internal criterion:

- **Davies–Bouldin Index:** Davies–Bouldin index can be calculated by the formula they are:

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{i\neq j}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

where n is the number of clusters, $c_x$is the centroid of cluster x, $\sigma_x$is the average distance of all elements in cluster $x$to centroid c_x, and $d(c_i, c_j)$is the distance between centroid $c_i$and c_j. Since algorithms that produce clusters with low intra-cluster distances (high intra-cluster similarity) and high inter-cluster distances (low inter-cluster similarity) will have a low Davies–Bouldin index, this was an one of the clustering algorithm that produces a collection of clusters with the smallest Davies–Bouldin index is considered the best algorithm based on this criterion.

- **Dunn index**

The Dunn index aims to identify dense and well-separated the clusters. It is defined by the ratio between the minimal inter-cluster distances to maximal intra-cluster distance. Each cluster partition, the Dunn index can be calculated by the following formula:

$$D = \min_{1\leq i\leq n}\left\{\min_{1\leq j\leq n, i\neq j}\left\{\frac{d(i,j)}{\max_{1\leq k\leq n} d'(k)}\right\}\right\}$$

Where $d(i,j)$ represents the distance between clusters i and$j$, and $d'(k)$measures the intra-cluster distance of cluster$k$. The inter-cluster distance $d(i,j)$between two clusters may be any number of distance measures, which distance between the centroids of the clusters. Likewise, the intra-cluster distance $d'(k)$may be measured in a different ways, such as maximal distance between any pair of elements in cluster$k$. Since internal criterion seek clusters with high intra-cluster similarity and low inter-cluster similarity, these algorithms that produce clusters with high Dunn index are more desirable.

(ii)External Evaluation:

- **Rand measure**

The Rand index computes how similar the clusters (returned by the clustering algorithm) are to the benchmark classification. This can also view the Rand

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 2, Issue 1, January 2014

index as a measure of the percentage of correct decisions made by the algorithm. It can also compute using the following formula:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

Where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. There is an issue with the Rand index is that false positives and false negatives are have equal. This may be an undesirable characteristic for some clustering applications.

- **F-measure**

The F-measure can be used to balance the contribution of false negatives by weighting recall through a parameter $\beta \geq 0$. Let precision and recall be defined as follows:

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

Where P is precision rate and R is recall rate. Where we can calculate the F-measure by using the following formula:

$$F_\beta = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R}$$

Notice that when $\beta = 0$, $F_0 = P$. In other words we can say, recall has no impact on the F-measure when $\beta = 0$, and increasing β allocates an increasing amount of weight to recall in the final F-measure.

- **Jaccard index**

The Jaccard index is used to quantify the similarity between these datasets. The Jaccard index takes on a value between 0 and 1. Indexes of 1 means that are two dataset are identical and an index of 0 indicates that the datasets have no any common elements. The Jaccard index is calculated by this formula:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN}$$

This was very simple, the number of unique elements common to both sets divided by the total number of unique elements in both sets.

- **Fowlkes–Mallows index** :

The Fowlkes-Mallows index computes the similarity between the clusters returned by the clustering algorithm and the benchmark classifications. This higher value of the Fowlkes-Mallows index the more similar the clusters and the benchmark classifications. It can be computed by using these formula:

$$FM = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}$$

Where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives. The FM index is the geometric mean of the precision and recall P and R, while the F-measure is their harmonic mean. Moreover, precision and recall are also known as Wallace's indices B^I and B^{II}.

VI. FUZZY C-MEANS ALGORITHM

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or many clusters. These methods are frequently used in pattern recognition. This technique is mainly based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \left\| x_i - c_j \right\|^2 \quad , \quad 1 \leq m < \infty$$

where *m* is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster *j*, x_i is the *i*th of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}} \quad , \quad c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m}$$

This iteration will stop when $\max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas *k* is the given steps. This procedure converges to a minimum or a saddle point of J_m.

The algorithm is composed of the following steps:

1. Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$
2. At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m}$$

3. Update $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}}$$

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 2, Issue 1, January 2014

4. If $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$ then STOP; otherwise return to step 2.

PROCESSIN STEP OF GENE ANALYSIS USING DNA MICRO ANALYSIS WITH CLUSTERING

The following steps define the processing step of gene analysis using DNA microarray analysis with K-Mean algorithm in clustering.

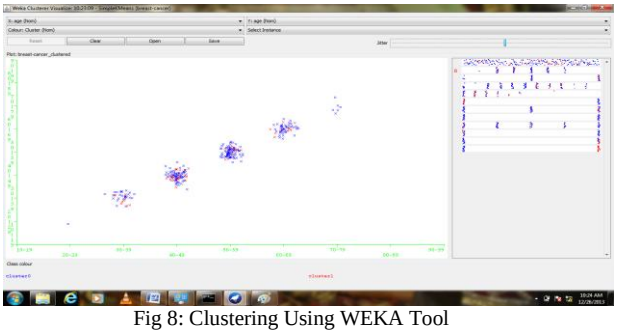

Fig 8: Clustering Using WEKA Tool

(i) The breast cancer data set is used to analyze the gene in microarray analysis.
(ii) Using DNA microarray analysis genes are analyzed from the dataset.
(iii) After analyzing the similar gene data are grouped using K-Mean algorithm in clustering.
(iv) The clustered data are stored in the database
(v) The analysis and clustering are performed using WEKA tool of data mining.

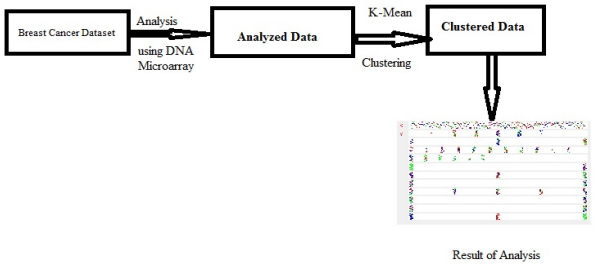

Fig 9: Process Step of this Dissertation

VII. CONCLUSION

The DNA microarray analysis as the emerging technique in genes analysis can analyze thousands of genes at a time. K-Mean the clustering algorithm on the other hand provides the effective grouping of similar data on analyzed dataset.
Thus the above paper provides the detailed information about the microarray analysis of breast cancer data set using DNA microarray analysis for genes and grouping the clustered data with effective K-Mean algorithm.

BIBLIOGRAPHY

[1]. Jesus Angulo and Jean Serra, "Automatic Analysis of DNA Microarray Images using Morphology", Bioinformatics. 2003 Mar 22;19(5):553-62.
[2]. Kathleen Kerr, Mitchell Martin and Gray A. Churchill, "Analysis of variance of Gene Expression Microarray Data", Journal of Computational Biology, vol. 7, number 6, 2000.
[3]. Buhler.J and Ideker.T, "Improved Techniques or finding spots on DNA Microarrays", UW CSE Technical Report UWTR 2000.
[4]. Eisen.M.B and Brown.P.O, "DNA arrays for analysis of Gene Expression", Meth.Enzymol, 303.
[5]. Osama Abu Abbas, " Comparison Between Data clustering Algorithms", The International Arab Journal of Information Technology, Vol. 5, No. 3, July 2008.
[6]. Han .J and Kamber .M, "Data Mining Concepts and Techniques", Morgan Kaufmann publishers, 2001.