ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

SURVEY: A COMBINE APPROACH OF FEATURE SELECTION AND DIFFERENT CLUSTERING TECHNIQUE IN BREAST CANCER GENE DATA

Divya Kharbanda*, Geetika Munjal
[*] Department of Computer Science and Engineering Northcap University, Gurgaon-122001

DOI: 10.5281/zenodo.260068

ABSTRACT

DNA microarray datasets have large number of genes however only a small number of genes are required to detect a particular type of disease. So gene selection plays an important role in removing irrelevant features which improves accuracy. In this paper we discussed about feature extraction techniques like I-relief, PCA. As large microarray datasets have the issue of dimensionality, PCA technique can be also used to reduce dimensions. After feature selection cluster analysis is performed to identify interesting patterns. Different types of clustering algorithms have been discussed like k-means, hierarchical, partitioning, model-based clustering and DB-Scan having their advantages and disadvantages in the result. Clustering validation techniques are discussed which can be used to calculate the exact number of clusters.

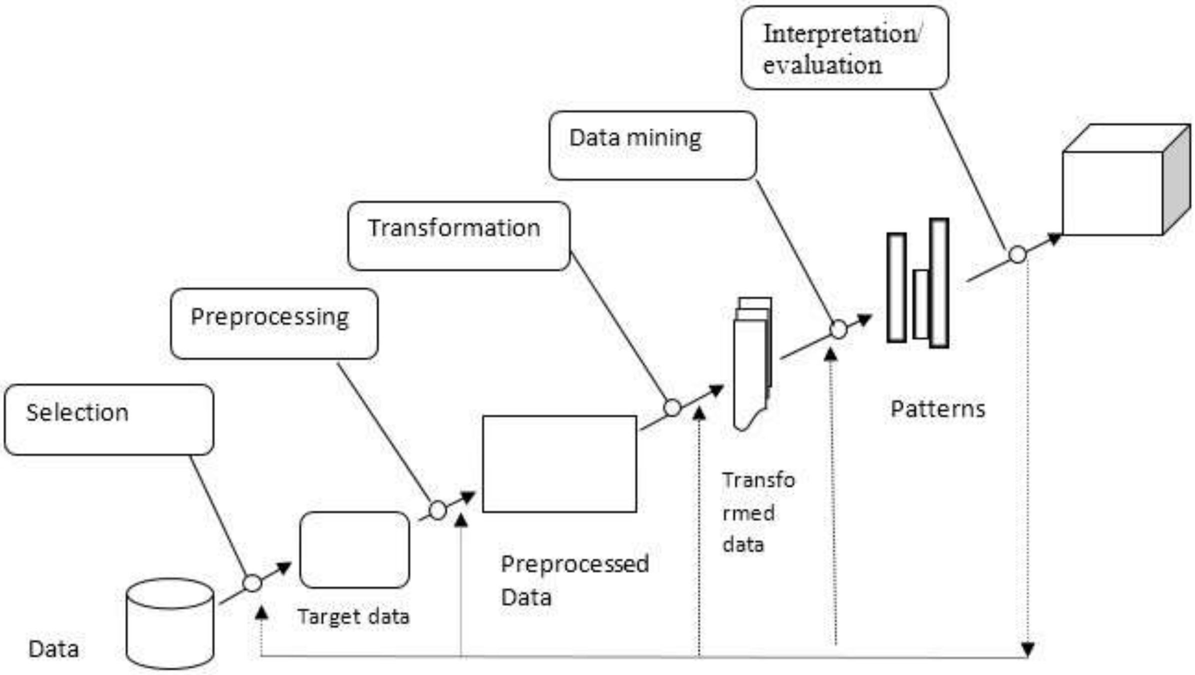
KEYWORDS: K-means, Hierarchical clustering, DBSCAN, PAM, Model based, PCA, i-relief, gene data, Breast cancer, clustering validation.

INTRODUCTION

DNA microarray datasets used to analyze the cell functions of genes. These datasets are composed of a small no of thousand of samples and gene expression level in the sample. This will lead to poor generalization in the classification or clustering process [1].This is already known that only a small no. of genes is enough to diagnose a disease. Data mining is the main step for KDD (knowledge discovery process) in which some intelligent methods are applied so that patterns can be extracted. BC (Breast cancer) is the most common type of cancer among the women with high mortality rate. The most effective way to reduce the breast cancer mortality rate is to detect it before. To predict the results of the disease is the most essential and challenging task where data mining applications are developed. For medical researchers, data mining techniques has become an important tool to identify patterns and to predict the outcome of a disease using historical data [2].

Previous methods of identifying patters in data are bayes theorem and regression analysis but now as data set has grown in size and complexity there is a very high dimensional data in microarray so data analysis has been done by automated data preprocessing, cluster analysis ,support vector machine(SVM). Different subtype of cancer respond differently to treatment and therefore it is essential to classify the cancer so that accurate or better treatment can be given to patient and it also reduces the medical cost associated with the unnecessary treatment.
Various approaches for diagnosis (to detect the particular type of cancer) and prognosis (to predict how it will behave in future) exist that are based on the microarray gene expression data but the problem with using this expression data is its higher dimensionality; few samples for too many genes but this problem can be overcome by using data preprocessing techniques such as feature extraction. There are some issues in microarray data, the curse of dimensionality [3] and the no. of irreleavant features present which can be overcome by feature extraction and dimensionality reduction technique. Accurate prediction of disease is a key to examine patients for prognosis and treatment. Prognostic factors are integrated by determining outcome through patients age, survival time and survival rate. This integration makes a prognostic system which is used to predict outcome of a new patient. Earlier approach was used in prognostic system is TNM system[8] which include three factors : metastatis ,tumor extent and lymph node involvement but the outcome prediction of this system is limited as it includes only survival rate so a new computer based prognostic approach has been proposed which can include multiple prognostic factors. In this approach we partition patients or cluster them from a large cancer dataset into groups such that patients includes in the same group have same survival rate than patients are in different group but before clustering it is

important to extract relevant features to remove noise and outliers. Feature selection is performed to mainly selection of relevant and beneficial features, it may have other motivations including data reduction which will limit storage requirement and which will increase the speed of the algorithm.



FEATURE SELECTION

The large number of features tends to poor generalization and high execution time. The main idea behind gene selection is to eliminate the genes which do not have any significance. Irrelevant features can be removed by feature selection .Several methods for feature selection exist such as PSO (Particle swarm optimization), I-relief, PCA (Principle Component Analysis). There are also some wrapper based and filter based methods which refined the features.

I-Relief

I-RELIEF[9] is one of the approach of feature selection which is used to select a small number of feature subset so that the performance of the learning algorithm has been optimized. It is a new algorithm of feature selection that succeeds in dealing with the problems of filter based methods which are not able to remove redundant features and when some of the features evaluated separately get low ranking scores but when integrated with other features then may provide censorious information. I-RELIEF is used to identify a hybrid signature by combining the both genetic marker and clinical markers. This approach adopted the feature weighting strategy which assigns each feature a real valued number not a binary one in order to show the relevance of the learning problem. It uses the performance of a non linear classifier. Below is a brief review of I-RELIEF.

Let D = $\{(\boldsymbol{xn}, \boldsymbol{yn})\}$ for n = 1 to N is a training dataset where x_n is a nth sample of training data and y_n is a class label, it may be either metastasis or no metastasis. We define two different functions nearest miss (NM(x_n)) and nearest hit (NH(x_n)) for x_n and a margin for x_n as p_n = d(x_n - NM(x_n)) - d(x_n - NH(x_n)) such that average margin may b maximized.

$$max \sum_{n=0}^{n} p\boldsymbol{n}(\boldsymbol{w}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1)$$

$$=\max \sum_{n=1}^{N} \quad \sum_{i=1}^{l} w_{i} \left(\left| x^{(i)}\boldsymbol{n} - NM^{(i)}(x\boldsymbol{n}) \right| - \left| x^{(i)} - NH^{(i)}(x\boldsymbol{n}) \right| \right) \ldots\ldots\ldots\ \ldots.. 2)$$

$$s.t \ ||w||_2^2 = 1, w \geq 0, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3)$$

PCA

PCA [2] is a technique of feature extraction and dimensionality reduction. It was invented by Pearson(1901) and

Hoteling (1933). It was first applied in ecology by Goodall (1954) under the name "factor analysis" Principal component analysis (PCA) uses orthogonal transformation to convert a set of correlated variables into linearly uncorrelated variables. PCA provides picture in lower dimensional because in high dimensional dataset visualization is difficult. It takes a data matrix of n objects by p variables which may be correlated and summarizes it by uncorrelated axes or principal components that are linear combination of the original variables. Patterns are to be find to reduce the dimensions of dataset with minimal loss of information. The transformation is performed in such a way that the object with maximum variance becomes the first principal component, next highest variance become the second principal component and so on. Each component is orthogonal to the preceding component. This is done by taken only first few principal components so that dimension is reduced.

WORKING OF PCA

1. The first step of PCA is to obtain covariance matrix. For this step variance of each dimension and covariance between dimensions is needed. The diagonal terms should be variances in covariance matrix and other terms will be covariance. Covariance and variance can be calculated by:

$$\mathrm{Cov}_{\mathrm{xy}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{(n-1)} = \frac{\sum(xy)-n\overline{xy})}{(n-1)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 4)$$

$$\mathrm{Var}_{\mathrm{x}} = \frac{\sum_{i=1}^{n}(x_{\mathrm{i}}-x)^2}{n} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 5)$$

2. Second step is to get eigen values by solving a function determinant

$(A - \lambda I) = 0$ ……………………………………………………………………6)

The calculated eigen values will be the sum of variance. Sum of eigen values will always equal to the sum of variance.

3. After getting eigen values there is a need to calculate the eigen vector by solving a matrix X in such that :

$[A - \lambda I] * [X] = [0]$ ……………………………………………………………......7)

4. Once eigenvectors are found the next step is to sort them in decreasing order i.e. from highest to lowest, the eigenvector with highest eigenvalue will be the first principal component. Now it can be decided that which components can be ignored. Components which has lower eigen values has lesser significance can be ignored. However it will lose some information but it doesnot lose much information. This will give final dataset with lower dimension.

5. The next step is to obtain coordinates of data point in the direction of eigen vector. We obtain this by multiplying centered data matrix to the eigen vector matrix. Variance of the projection on the line of principal components is to be obtained which is equal to the eigen values of the principal components. First eigen vector is able to explain about 99% of the variance.

6. From the choosen components feature vector matrix are to be formed by taking the eigenvectors that are choosen.

7. Final step is to derive a new data set by simply multiplying vector transposed to the original dataset transposed.

$FinalData = RowFeatureVector\ X\ RowDataAdjust$

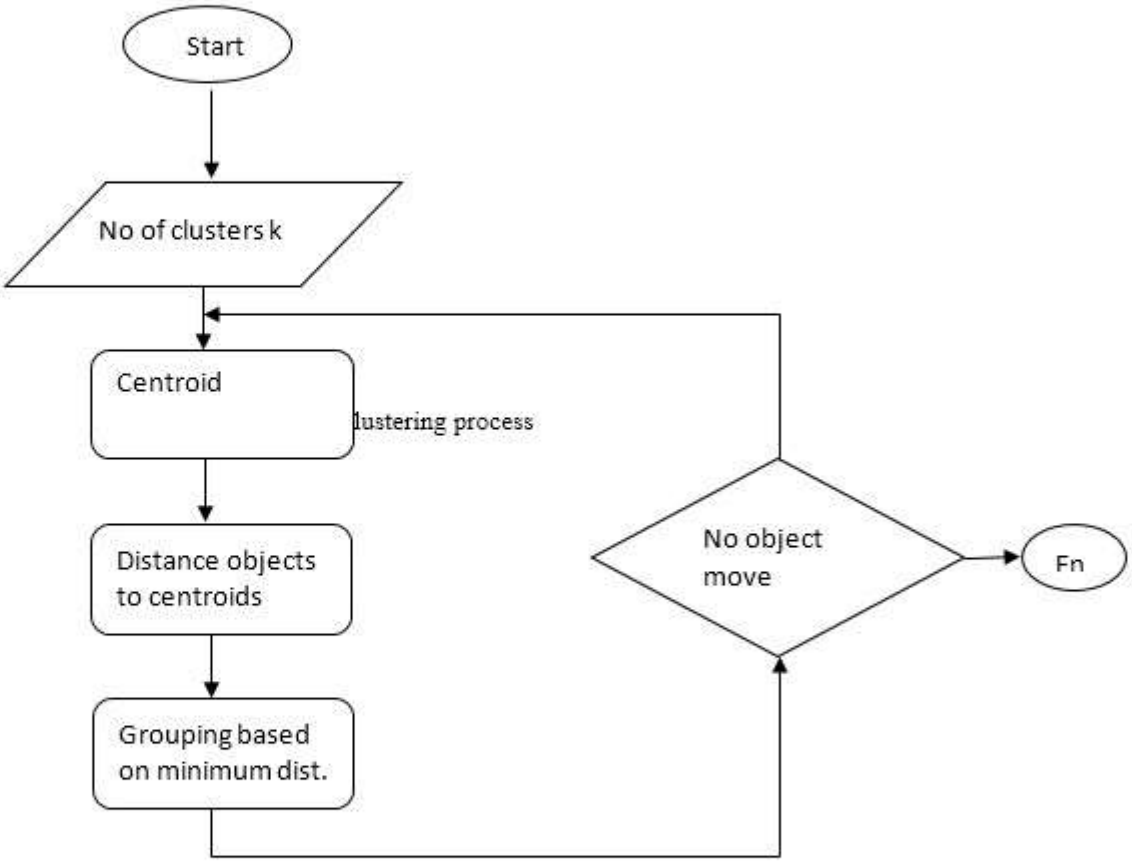
CLUSTERING

Clustering is a process to find a group of similar type of objects. It is unsupervised learning process because predefined classes are not known in clustering [4]. Before applying clustering methods datasets have been classified during classification process by applying different classification techniques. Most commonly used clustering methods are k-means, hierarchical, portioning approaches.

K-means clustering: Initially, it will take the no of parts which is equal to required no of clusters in conclusive result [5]. Ultimately required no of cluster will be selected so that points are mutually far from each other and then each component is examined and assign to that cluster that is based on minimum distance. Every time when a cluster is added, the centroid position is calculated and it repeats until all the components are grouped into final clusters required. The basic idea of k-means is to use cluster means to represent cluster. The data elements are to be assigned to the closest cluster. This algo usually comes together in a small no of cycles. However it also has bad results:

First, the no. of groups in a gene expression dataset is not known before. To identify the best no of groups one has to run the algo again and again with distinct values of k and results of clustering are compared. For a large gene expression dataset which contain thousands of genes, this process may not be practical.

Second, gene expression data usually contain a huge amount of noise; However, the k-means set of instructions forces each gene into a group, which may cause the algorithm to be sensitive to noise.



Algorithm:
Let $X = \{x1, x2, x3, \ldots \ldots \ldots \ldots .. x_n$ be the set of data points
1. The first step is to select the first partition of k cluster.
2. Calculate the distance between each of the data point and center of cluster.
3. Assign the data point to the cluster center whose distance from the cluster center is the minimum value of all of the cluster center.
4. Again calculate the new cluster center by the following formula:
$v_i = \left(\frac{1}{c_i}\right) \sum_{j=1}^{c} x_i$ ……………………………………………………………8)
5. The distance between each data point and new cluster center obtained is recalculated.
6. It stops when no data point is reassigned and repeat from step 3.
This clustering will provides better results when data sets are well separated from each other.

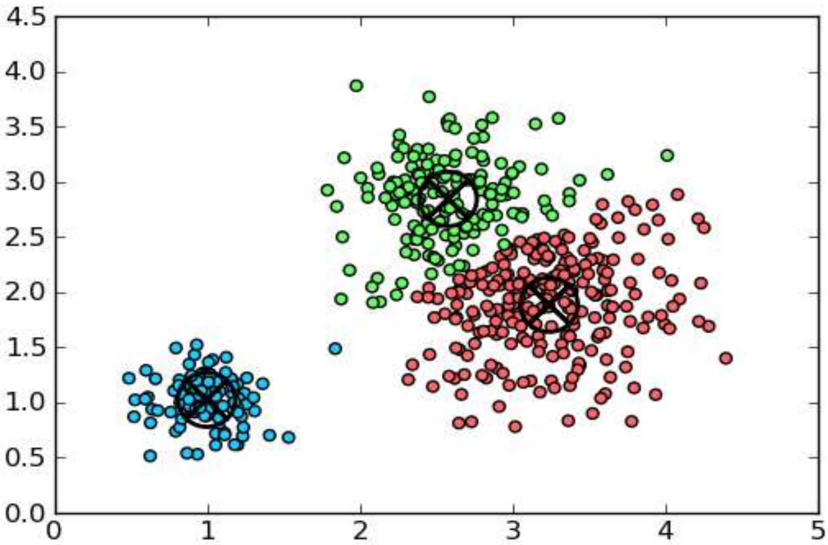


Fig: k-means results for and no. of clusters=3

Hierarchical clustering: This approach builds a hierarchy of cluster. It groups together genes and provides a way to represent the dataset graphically which allow users a careful examination of whole dataset and get a first impression of the distribution of data. It can be agglomerative (bottom up approach) and divisive (top down approach). This approach proceeds by either combining smaller clusters into larger one and by splitting cluster into smaller clusters. It starts by assigning each item to its own cluster. The most similar pair of clusters has been find and merged into a single cluster so that number of clusters can be reduced and then the distance is calculated between the new and each of the old cluster. This process repeats until all the elements are clustered into a single one. This method doesnot require predefined number of clusters but it requires a termination condition [6]. For agglomerative approaches, different measures of group related closeness such as single link, complete link and minimum-variance, get different merge success plans. For divisive approaches, the extremely important problem is to decide how to split groups at each step. Some are based on heuristic (experience-based thinking) methods such as the pre-decided toughening algorithms while many others are based on graph methods (related to ideas about how things work or why they happen). As this clustering algorithm is of two types: agglomerative and divisive which are just reverse of each other. Agglomerative hierarchical clustering- This algorithm is based on the closest distance measure of the distance of all the pair wise between the data points, it works by grouping the data one by one[4]. The distance between the data points will be recalculated again and which of distance is to be chosen based on some methods like single linkage, complete linkage, average linkage, centroid distance.

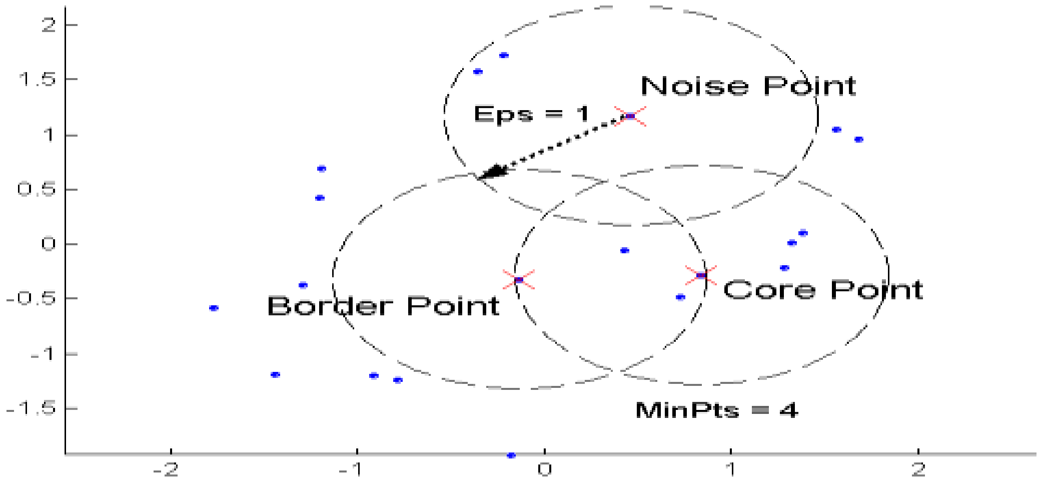
ALGORITHM

1. Start with the disjoint clustering which has level L(0)=0 and sequence no. m=0.
2. Calculate the least distance of pairs of cluster in the current cluster acc. To d[(a),(b)] = min d[(p),(q)] , where min is cluster of overall pairs
3. Sequence no. is incremented and two clusters a & b are merged into a cluster and formed a new cluster.
4. Now set the level of this clustering to L(m) = d[(a) , (b)].
5. Distance matrix is updated by deleting rows and columns which corresponds to clusters (a) & (b) and rows and columns are added which corresponds to newly formed cluster. The distance between the new cluster and the old cluster is defined in such a way that d[k , (a,b)] = min (d[(k) , (a)] , d[(k) , (b)]).
6. When all the data points are in a cluster then stop otherwise repeat from step 2.

Model-based clustering: Model based clustering [7] provides a statistical structure to model the group related structure of gene expression data. This group related analysis is based on the idea that the watched/followed data comes from a population consisting of more than two subpopulations. Each subpopulation model based clustering is separately modeled and overall population with a limited mixed model is modeled as a weighted sum of a mixture or a subset using a finite mixture model. This algorithm find good close guesses of model guidelines that best fits the data. It can be partitioned clustering or hierarchical clustering depends upon the structure they guess about dataset these algorithms are near to density based in which they grow particular groups so that the prejudiced model is improved and sometimes they start with the fixed no. of groups and do not use the same idea of density. Partitioning clustering datasets have been directly decomposed into a set of disjoint clusters or assign different objects into groups of cluster

PAM: It is a partitioning method which operates on dissimilarity matrix. It comes from selecting a predetermined number or medoid and each data object is assigned to the nearest medoid. It replaces mediods with other non mediods until all objects are classified as mediod. PAM is more powerful and expensive algorithm than k means because it is adapted that cluster center the centroid it doesnot centroid mean. So it is extension to k-means meant to handle things efficiently that are not part of a main group. The implementation of PAM is clustering large application.

DBSCAN: DBSCAN is a density based spatial clustering algorithm. It discover clusters of arbitarary shape and is defined as a maximal set of density connected points. The points that are alone in a region are marked as outliers or noise. It continues to grow as along as the specified cluster exceeds the number of parameters in the number of neighborhood objects. Density based places a high concentration area that are separated from each other by a region of low density where density is the no of points specified within a radius Eps. In DBSCAN algorithm a point is a core point if it has more points than specified no of points which is Minpts within Eps and core points are at the interior of a cluster. A border point which has points fewer than the Minpts within Eps but it always lies in the neighborhood of core point. A point is a noise point which is neither a core point nor a border point.



CHALLENGES IN GENES CLUSTERING:

Cluster analysis is the first step of data mining knowledge discovery [8]. The purpose of the clustering of gene expression data is to clarify the nature of the data structure and gain some early insight into data distribution. So a good clustering algorithm should depend on prior knowledge that is generally not available prior to cluster analysis. By microarray experiments, in many cases due to the complicated procedures of gene expression data it contains a huge amount of noise. For gene expression data clustering algorithm should be able to extract the useful information from a high level of noise. Gene expression data is often highly connected and the cluster is able to close to each other or the one embedded with other. Therefore gene based clustering algorithm should be effective.

CLUSTER VALIDATION TECHNIQUES:

Evaluation of clustering results is one of the important issues in cluster analysis [4]. Three approaches to investigate the group related validity: External criteria is user specific which is based upon that results of clustering algorithm is calculated on the basis of pre-specified structure. Internal criteria that figure out the results of clustering algorithms in terms of amount that includes the vectors of the dataset themselves and relative criteria whose basic idea is the process of figuring out the clustering structure in comparison with other schemes of clustering and results the same algorithm but they have different values. Internal and external criteria are based on statistical methods. The disadvantages of these methods are its calculation complex difficulty whereas in relative criteria one or more clustering algorithm are executed many times with different input guidelines. The purpose of the relative criteria is to select the best clustering algorithm from different results. The basis of comparison is a measure of effectiveness.

Two criteria proposed for clustering evaluation and selection of an optimal clustering scheme are: compactness and separation and three common approaches to measure the distance between different clusters are single linkage which is used to measure the distance between nearest members of clusters, complete linkage which calculates the distance between the most distant members and comparison of centroids which is used to measure the distance between the centers of clusters.

Some Validity indices are introduced which is used for measuring "goodness" of a clustering result comparing to other ones which were created by other clustering algorithms, or by the same algorithms but using different parameter values[9]

1. RMSSTD: The RMSSTD (root mean square standard deviation) index which is the variance of the cluster defined in below eq ,it is used to measure the homogeneity of clusters. The purpose of the clustering process is to identify homogeneous groups, lower the RMSSTD value, better is the clustering.
2. RS: R Squared index is used to estimate dissimilarity of clusters. It calculates the uniformity of degree between groups[7]. Value of RS lies between 0 to 1 where 0 indicates that there is no difference among clusters and 1 indicates the lot of difference between clusters.

$$RS = \frac{SS_t - SS_w}{SS_t}\text{, where } \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 9)$$

$$SS_t = \sum_{j=1}^{d} \sum_{k=1}^{n_j} (x_k - \bar{x}_j)^2\,,\quad SS_w = \sum_{\substack{i=1\ldots\ldots nc\\ j=1\ldots\ldots d}}^{d} \sum_{k=1}^{n_j} (x_k - \bar{x}_j)^2 \;\ldots\ldots 10)$$

Where, SSb is the sum of sq between groups , SSw is the sum of sq within groups and SSt is the total sum of square of whole dataset.
3. SD validity index: This index is basically the average scattering and total separation of clusters and is calculated by the variance of the dataset and clusters so it can be used to calculate the homogeneity and compactness of cluster. Variance is defined in below eq.
Variance of dataset: variance of cluster:

$$\sigma_x^p = \frac{1}{n} \sum_{k=1}^{n} (x - \bar{x}^p)^2 \quad \ldots\ldots 11) \qquad \sigma_{v_i}^p = \frac{1}{||C_i||} \sum_{k=1}^{n} (x_k^p - \bar{v}_i{}^p)^2 \ldots 12)$$

$$\sigma(x) = \begin{bmatrix} \sigma_x^1 \\ . \\ \sigma_x^d \end{bmatrix} \qquad \sigma(v_i) = \begin{bmatrix} \sigma_{v_k}^1 \\ . \\ \sigma_{v_i}^d \end{bmatrix}$$

And the average scattering of cluster is defined as:

$$Scatt = \frac{1}{n_c} \sum_{i=1}^{n} \frac{||\sigma(v_i)||}{||\sigma(x)||} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 13)$$

4. SPR index: SPR of the new group is the difference between the combined SSw(sum of sq within group) of the new group and the sum of the combine SSw(sum of sq within group) of the groups joined to get the new group(loss of evenness) , divided by the combined SSt(total sum of square for whole datset) for the dataset. It can be used to measure the loss of homogeneity after merging the two clusters of a single algorithm step. If the index value is 0, the new cluster, obtained by merging the two completely homogeneous clusters. If the value is higher, the new cluster is obtained by merging clusters of two different kinds.

5. CD index: It is used to calculate the distance between the two groups that are combined. Distance is accordingly representative selected for the hierarchical clustering performed by measuring each time .
Using these indices no of groups/clusters can be determined that is present in a dataset

RESULTS AND DISCUSSION

Clustering algorithms	Advantages	Disadvantages
k-means	This is efficient when dealing with large scale datasets. It is simple and fast and typically will converge with no of iterations is less.	No prior information of no of genes in a gene expression dataset so users have to run the algorithm again and again at every time different value of k to identify ideal no of clusters.

		For a thousand of genes this process may not be practical. It contains large amount of noise .
Hierarchical	It is easy to implement and gives best results in some cases. In case of hierarchical clustering apriori information is not required	In hierarchical, agglomerative approach is not robust. Sometimes a small disruption of the dataset is greatly changed the structure of hierarchy tree diagram. This approach has high calculation complexity. Sometimes it is difficult to find the correct no of clusters in a dendogram.
Model based	This algorithm yields the roughly calculated probability that a data object will belong to a cluster k.	It depends on the idea that dataset fits a particular distribution.
PAM	It can handle outliers efficiently. Rather than group centers, it selects to represent each group by its medoid.	These algorithms have very high complexity. For a small no of objects, the no of partitions is enormous.
DB SCAN	It can handle noise/outliers and clusters of arbitrary shape. It requires only one scan of the input dataset.	DBSCAN does not work to good in case clusters of varying densities/with high dimensional data

CONCLUSION & FUTURE SCOPE

In this survey paper, we analysis different type of clustering techniques and feature selection in breast cancer gene dataset. After our analysis with all above mention algorithm we found the k-means algorithm having best accuracy in the case of gene data as many author already demonstrate. During observation we also analysis that this algorithm have less space complexity as compare to rest four algorithm of clustering mention in this papers.
PCA is best feature extraction technique as per our survey because it reduced the dimension of the data and extract maximum features for machine learning. Clustering technique provide best output with PCA only. Other algorithm having some very good advantage e.g DBSCAN (low time and space complexity), hierarchical clustering (Reducing space complexity).

FUTURE SCOPE

In future research can use hybrid approach to solve real life problem because as our observation single algorithm of any field is not sufficient to handle real life problem. As per our observation hybridation definitely enhance the efficiency of clustering technique as two algorithms solve each other problem.

REFERENCES

[1] Tang C., Zhang L., Zhang A. and Ramanathan M. Interrelated two-way clustering: An unsupervised approach for gene expression data analysis. In *Proceeding of BIBE2001: 2nd IEEE International Symposium on Bioinformatics and Bioengineering*, pages 41–48, Bethesda, Maryland, November 4-5 2001.

[2] Tang, Chun and Zhang, Aidong. An iterative strategy for pattern discovery in high-dimensional data sets. In *Proceeding of 11th International Conference on Information and Knowledge Management (CIKM 02)*, McLean, VA, November 4-9 2002.

[3] Tavazoie, S., Hughes, D., Campbell, M.J., Cho, R.J. and Church, G.M. Systematic determination of genetic network architecture. *Nature Genet*, pages 281–285, 1999.

[4] Tefferi, A., Bolander, E., Ansell, M., Wieben, D. and Spelsberg C. Primer on Medical Genomics Part III: Microarray Experiments and Data Analysis. *Mayo Clin Proc.*, 77:927–940, 2002.

[5] Thomas J.G., Olson J.M., Tapscott S.J. and Zhao L.P. An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles. *Genome Research*, 11(7):1227–1236, 2001.

[6] Troyanskaya, O.,Cantor M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman R. Missing value estimation methods for dna microarrays. *Bioinformatics*, In press.

[7] Abba,M. et al. (2005) Gene expression signature of estrogen receptor a status in breast cancer. BMC Genomics, 6, 74–81.

[8] Brenton,J. et al. (2005) Molecular classification and molecular forecasting of breast cancer: ready for clinical application? J. Clin. Oncol., 23, 7350–7360.

[9] Dalton,W. et al. (2006) Cancer biomarkers–an invitation to the table. Science, 312, 1165–1168.