International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 25 No. 1 Jun. 2016, pp. 308-319

http://www.ijisr.issr-journals.org/

# MULTI VARIANT GENE SELECTION APPROACH BASED HIGH DIMENSIONAL SUB SPACE CLUSTERING OF BREAST CANCER DATA SET FOR EFFICIENT CLASSIFICATION USING FUZZY RULE SETS AND MULTI GENE IMPACT MATRIX

***N. MAGENDIRAN[1] and S. SELVARAJAN[2]***

[1]Associate Professor / CSE, Paavai Engineering College, Namakkal, Tamilnadu, India

[2]Principal, Muthayammal College of Engineering, Rasipuram, Tamilnadu, India



**ABSTRACT:** The breast cancer is the most threatening factor of women's lifestyle and the reason of the disease has many factors, but still the gene factor has more influence in the generation of breast cancer where the early diagnosis and prevention is essential. There are many approaches has been discussed in the literature, but the identification and selection of a set of genes which influence the disease is still complicated one. We propose a multi variant approach for gene selection which is performed by performing high dimensional subspace clustering. With the given data set, the method generates a set of rules and unlike generic fuzzy rules the method splits the range values into the number of parts and based on that the rules are generated. Also, according to the different range values, the method generates a multi gene impact matrix where the frequency of range values of each rule is stored. The data set is clustered according to the generated rules and from the generated rules the gene selection is performed. For the gene selection, we compute the multi gene frequency measure which represents how depth the gene has an impact on the classification of disease. The proposed method produces efficient classification of genes in the influence of breast cancer and produces efficient results.

**KEYWORDS:** Gene Selection, High Dimensional Clustering, Multi Gene Impact Matrix, Fuzzy Rule Sets.

## 1 INTRODUCTION

The growth of data sets in, their dimension increases the challenges in clusters them, where the higher dimensional space requires more sophisticated approaches to cluster the data sets. In any high dimensional space, identifying the subspace is the most important task which has to be performed in an efficient manner. For a breast cancer data set there are a number of genes influencing or taking part in the appearance of the cancer in the women. To identify them or to cluster such data set the genes are the most important factor which participates in the clustering approach.

Gene selection in high dimensional breast cancer data set clustering is the most important task and how the gene selection is performed is the big question here. Not all the genes have a great impact, but all the genes has some impact in the cause of breast cancer. To find a strategic approach to selecting the gene selection there must be some efficient approach to be there. Unfortunately the existing approach misses the case of gene selection in a modern sophisticated approach and has no efficient solution to perform the task of gene selection.

The breast cancer can be classified into many cases and to identify the exact subspace we must come up with more efficient measures and gene selection approaches. For example, a subset of genes may be the cause of a specific type of cancer, but they may not have any impact in the presence of another type of cancer. So the gene selection is the most important task which could be used to predict the future appearance of breast cancer. So for the prediction of the breast cancer the gene selection approach can be used which helps early detection and cure of cancer in many ways.

The general fuzzy rule sets are nothing but the range of values for each rule. It contains a number of rules for each case of breast cancer and has a range of values for each gene participate in the appearance of cancer cells. Unlike generic one, we generate a modern fuzzy rule set which has many numbers of rules and for each rule, each gene value is split into a number of subdivisions and we generate that many numbers of rules to fill the rule sets.

From the generated fuzzy rule sets we compute a multi gene impact matrix, which represent the impact of genes for the occurrence of cancer. From the rule set with the data set, of each subdivision and rules available, the impact matrix is generated by the values of the data set. We compute the number of possible occurrences of cancer appearance based on the genes selected or the values of genes. According to the appearance a single value for the pattern of the genes and values is generated which decides the impact of the gene and their values for the occurrence of cancer in the human body.

## 2 RELATED WORKS

There are many approaches has been discussed for the gene selection of breast cancer and we discuss a few of them here in this section for better understanding of breast cancer gene selection.

Stable Gene Selection from Microarray Data via Sample Weighting [1], proposed a framework of sample weighting to improve the stability of feature selection methods under sample variations. Their experiments show that the sample weighting algorithm improves the stability of gene selection. It evaluated with SVM-REF, Relief classifiers.

A Novel Approach for Single Gene Selection Using Clustering and Dimensionality Reduction [2], proposed Hybrid Fuzzy C Means-Quick Reduct algorithm for single gene selection. Average Correlation Value (ACV) is calculated for the high class discriminated genes. The algorithm is evaluated using WEKA classifier with leukemia cancer data set.

Relevant and Significant Supervised Gene Clusters for Microarray Cancer Classification [3] , have developed mutual information based supervised gene clustering (MSG) algorithm to form the reduced gene clusters for cancer classification. The approach has been evaluated using different micro array cancer data sets with different classifiers like Naïve bayes, K-nearest rule and SVM.

Gene-Expression-Based Cancer Subtypes Prediction through Feature Selection and Transductive SVM [4] proposed combined gene selection and transductive support vector machine (TSVM). Their method identified the potential genes and uses the TSVM to improve the prediction accuracy compared to standard inductive SVM. Experimental results confirm the effectiveness compared to the ISVM and low-density separation method in the area of semi supervised cancer classification as well as gene-marker identification.

Identifying Gene Pathways Associated with Cancer Characteristics viaSparse Statistical Methods [5] proposed statistical method for uncovering gene pathways that characterize cancer heterogeneity. They define a set of activities of pathways from microarray gene expression data based on the Sparse Probabilistic Principal Component Analysis (SPPCA). It creates a novel gene-gene associations relating to the cancer phenotypes. This method analysis breast cancer gene expression data.

An optimal gene selection based on search mechanism has been adapted for diagnosis of cancer in Optimal Search-Based Gene Subset Selection for Gene Array Cancer Classification [6]. The method introduces tabu search (TS) to gene selection from high-dimensional gene array data. TS were shown to be promising tool for gene subset selection.

A Family wise error rate based gene selection approach has been discussed in Gene Selection for Sample Classifications in Microarray Experiments" DNA and cell Biology [7], which uses two or multiple samples for the classification of genes. The method reduces the false positive ratio which has been evaluated using colon cancer data set.

Gene Expression Profiles for Predicting Metastasis in Breast Cancer: A Cross-Study Comparison of Classification Methods [8], performs a comparative analysis of different methods. The method uses voting mechanism for the prediction of metastasis. The voting based method produces more efficiency with the breast cancer patients.

Comparison of feature selection methods for cross-laboratory microarray analysis [9], investigate four feature selection methods t-Test, Significance Analysis of Microarrays (SAM), Rank Products (RP) and Random Forest (RF) across breast cancer and lung cancer microarray data which consists of three cross lab data sets each. Their results show that SAM has the best classification performance. RF also gets high classification accuracy, but it is not as stable as SAM. The Test performance is the worst among the four methods.

Data Mining Techniques for the Identification of Genes with Expression Levels Related to Breast Cancer Prognosis [10], applied data mining techniques to study the gene expression values of breast cancer patients with known clinical outcomes. Created the classification models for clinical practice to support therapy prescription. With nine algorithms of feature

selection they extracted a group of subsamples of data, which was analyzed with different classification algorithms for comparison purpose. They used five learning algorithms implemented in YaLE or WEKA. Classifying a patient as "good prognosis" when she is in a state that will develop metastasis (i.e., A FN error) is much more serious than classifying a patient as" poor prognosis" when she is not in a state that will develop metastasis (i.e., a FP error). The algorithm classified ill patients more accurately (lower FN and higher TP) at the expense of the classification of healthy patients (higher FP and lower TN).

All the above discussed methods have the problem of gene selection efficiency and could not perform any prediction about the breast cancer identification and the reason for them.

## 3 PRELIMINARIES IN HIGH DIMENSIONAL CLUSTERING

Let the data set Ds has N number of data points from {Dp1,Dp2,…., Dpn} where each data point has Q dimensions from { D1,D2,…,Dq} and the number of attribute types from A1 to Am. Clustering such high dimensional data set can be performed based on the similarity measure HSim.

The high dimensional similarity measure Hsim is computed by computing the similarity or closure value of data points at each dimension as follows.

Hsim = {S1(Dpx,Dpy), S2(Dpx,Dpy),S3(Dpx,Dpy),…, Sq(Dpx,Dpy)} ---- (1).

From the equation (1), the variables S1,S2, S3 represent the similarity measure on particular dimension and to identify the cumulative similarity measure, we can perform the averaging method or standard deviation or any other mathematical approach.

Finally the cumulative similarity measure can be computed using the equation (2).

$$CSim = \frac{\sum_{i=1}^{Q} HSim(i)}{Q} \quad \text{-- (2).}$$

The equation (2) shows how the similarity between two data points can be computed, but when we cluster data points with high dimensions, in order to assign a data point to a group of data point Cs, the similarity measure is performed with all the data points.

For example there exist C number of clusters with each cluster has variable number of data points Dp, then to identify the group of data point to which the testing sample Ts belongs to, we must compute the similarity measure k as follows:

$$K = \frac{\sum_{i=1}^{Size(C)} CSim}{Size(C)} \quad \text{---------- (3).}$$

Here K is the similarity measure which represents closeness value towards the cluster Ci, similarly the closeness of the data point towards each cluster Cx can be used to identify the final class of the data point.

## 4 GENE SELECTION FROM HIGH DIMENSIONAL SPACE

The breast cancer data set is categorized into many classes, as of the data points are high dimensional they are clustered based on any similarity measure. Even though the clustering is performed based on all the dimensional similarity, there is only little number of dimensions and their values which decides the closure of the data point.

If each dimension is about a gene and the data point has numerous numbers of genes, among them only a small set of gene values which decides the category of the data point. When we classify the data sample towards number of classes, in order to come into a class, the gene values at a data point has to be close to the gene values of the data points available in the cluster. By identifying such small set of genes, which has more influence in getting assigned to the class, the causes and the specific small set of genes can be monitored in disease prediction.

## 5 MULTI VARIANT GENE SELECTION APPROACH

The proposed multi variant gene selection using fuzzy sets and impact matrix approach has the following stages, namely preprocessing, rule generation, Multi Gene Impact Matrix Generation, Multi Variant Gene Selection. We discuss each of them in detail in this section.

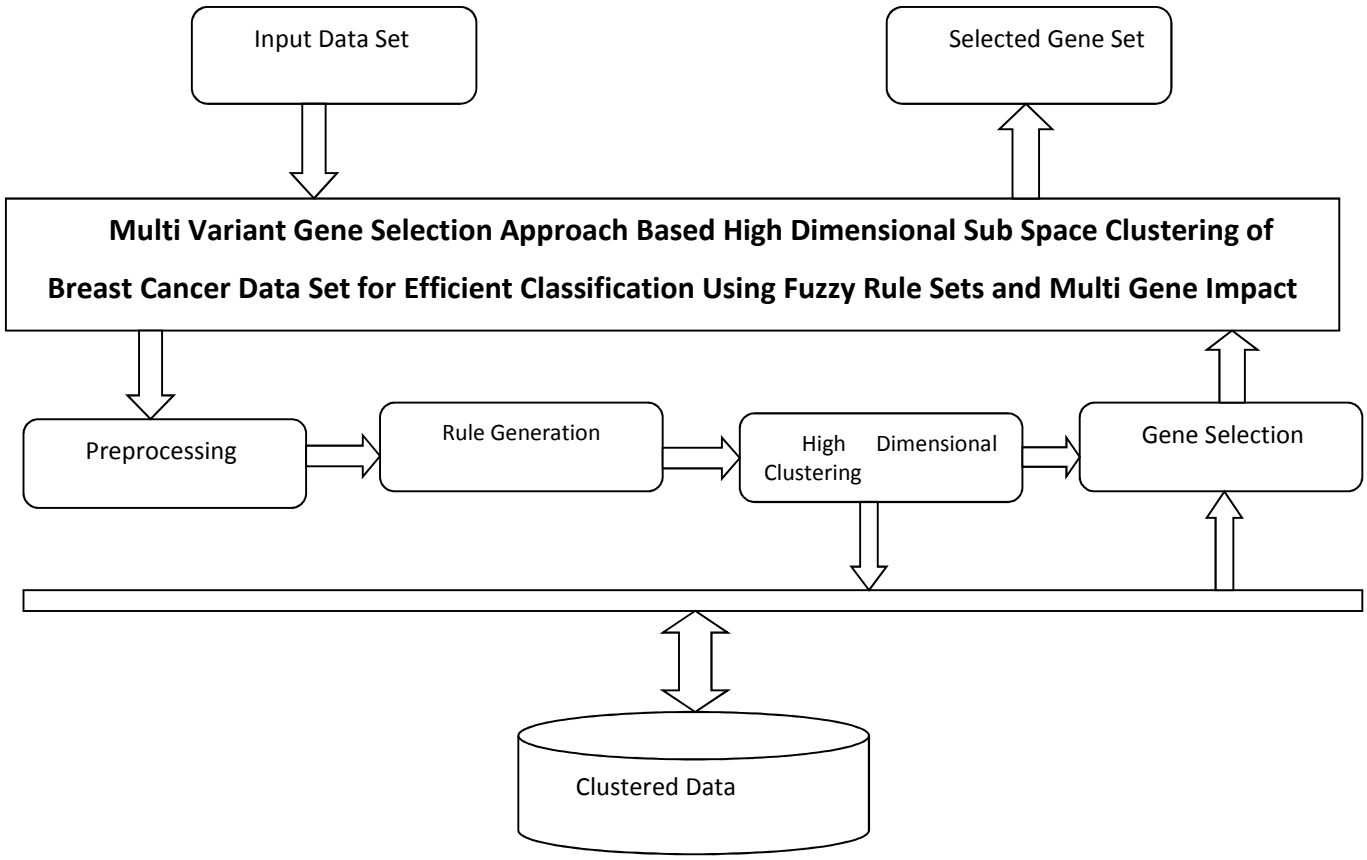



***Figure 1: Proposed System Architecture***

The Figure 1 shows the architecture of the proposed multi variant gene selection approach and its functional components which will be discussed in detail in this section.

### 5.1 PREPROCESSING

The given data set Ds, has N number of genes with M number of instances. At this stage, we identify the number of genes present in the data set and for each of the gene from each of the data points in the data set, the presence of value is identified. If there is any data point gene with missing values are identified and the data point will be removed from the data set. The noise removed data set is given for the rule set generation which will be used for further processes.

Procedure:

Input: Data Set Ds.

Output: Noise Removed Data Set NDS.

Identify Number of genes available in the data set Ds.

Gene set GS = $\sum_{i=1}^{size(Ds)} \sum Gene(Ds(i)) \nexists Gs$ --- (4)

The equation (4) identifies the distinct genes available in the data set Ds.

Perform Noise Removal

For each gene Gi from Gs

For each data point $Ds_i$ from Ds

If $Ds_i(Gi)==0$ then //the gene value is 0 then remove the data point

$Ds = \sum Ds - Ds_i$ --- (5)

The equation (5) removes the noisy data point from the data set Ds.

Else

End

End

NDS = ∑NDS+ $Ds_i$ //Add the correct data point to the data set.

End

The above algorithm performs the preprocessing of input data set by identifying the unique genes, and for each data point, the presence of all the gene values are verified, if there is any missing cases or has no value then they are removed from the data set.

### 5.2 RULE GENERATION

The rule generation is performed based the values of genes present in the data set. Rule generation is performed using the preprocessed data, for each gene value we identify the maximum and minimum values and from them number of range values are generated by splitting them. Based on generated multiple range values, we generate a number of rules using which the multi gene impact matrix is generated and clustering is performed.

Procedure:

Input: Input Data Set Ds.

Output: Rule Set Rs.

Identify set of genes from data set Ds.

Gene set GS = $\sum_{i=1}^{size(Ds)} \sum Gene(Ds(i)) \nexists Gs$ --- (6)

The equation (6) identifies the set of distinct genes from the data set.

For each of the gene identified, the minimum and maximum values are identified as follows.

For each gene gi from Gs

Compute minimum and maximum values of gene Gi.

Gmin = $\text{Min}(\sum_{i=1}^{size(Ds)} Ds(i)(Gi)$ ---- (7)

The equation (7) computes the minimum value of each gene Gi from the data set Ds.

Gmax = $\text{Max}(\sum_{i=1}^{size(Ds)} Ds(i)(Gi)$ ---- (8)

The equation (8) computes the maximum value of the gene Gi from the data set Ds.

End

For each gene Gi generate set of range values

RV=$\sum_{i=1}^{size(Gs)} \forall(Range\ Values)$ --------- (9).

The equation (9), generates set of range values between minimum and maximum values of each gene Gi present in the gene set Gs.

End

For each Gene Gi and range values

Generate rules Rs.

Rs =$\sum_{i=1}^{size(Gi)} \sum Rv(u)$ ---------- (10).

The Equation (10), generates number of rules for each gene Gi from gene set Gs and the number of rule for each gene Gi is depend on the number of range values generated by the equation (9).

End

The above algorithm first computes the minimum and maximum values for each of the gene present in the data point. With the computed minimum and maximum values, we generate set of range values for each gene attribute. Based on computed range values, number of rules is generated to perform clustering. The total number of rules generated by the rule generation approach could be computed as follows:

Total number of rules $X = M \times 2^N$ ---- (11).

The equation (11) computes the total number of rules computed using the gene set Gs, where M represent the number of range values, N represent the number of combination.

### 5.3 MULTI VARIANT IMPACT MATRIX

The multi variant impact matrix is generated using the rule set generated and based on the rule generated with the data set available we compute the multi variant impact matrix. For each data point of the data set, we perform matching of a rule with the other gene values. Based on the matching rule, we compute the multi variant support values and with the total number of data points, we compute the impact factor. Similarly for each of the rules identified we generate the impact matrix and the generated impact matrix is used to cluster the data points of the data set.

Procedure:

Input: Rule Set Rs.

Output: Multi variant impact matrix MVIM.

For each rule Ri from Rs

Compute number of data points match with the range values of genes.

Gene count Gsup.

$Gsup = \sum_{i=1}^{size(Ds)} Ds(i).Gi.value > Ri.\min \&\& \; Ds(i).Gi.value < Ri.max$ --- (12).

The equation (12) computes the gene support value which represent the total number of data points has the value falls within the range of the gene value of the rule.

Compute impact value $Gimp = \frac{Gsup}{size(Ds)}$ ------- (13).

The equation (13) computes the impact value of each gene using computed gene support value and the total number of data set.

End

$MVIM(i) = Ri + Gimp$ ------ (14).

The equation (14) adds the impact value to the static variable and adds to the multi variant impact matrix, which contains the impact value of all the genes.

The above algorithm computes the impact matrix, where for each data point, the impact matrix is computed using the rule set available.

### 5.4 HIGH DIMENSIONAL CLUSTERING

The clustering is performed based on the multi variant impact matrix generated and with the input data set. For each data point, from the rule set available, we find the matching rule and from the impact matrix the values of genes are identified. For each level, we find the matching cluster or sub space to identify the cluster to which the data point belongs. With the data points of each cluster or sub space the input data point is computed with the multi variant similarity measure and the deviation with the multi variant impact factor is performed. If the deviation is more than the process will look for the next cluster and finally a more closure data cluster is identified and the data point is assigned to the selected class label.

Procedure:

Input: Rule Set Rs, Data set Ds, MVIM.

Output: Cluster Cs.

For each data point Di

Identify the matching rule Ri.

For each level of the gene value

Identify the cluster Cs.

Compute multi variant similarity MVS.

Mvs = $\sum_{i=1}^{size(Csi)} \forall(dp(csi)) \sim Di$ ---- (15)

The equation (15) computes the multi variant similarity value for each data point and the matching rule Ri.

Compute impact deviation Idev = Ri(Imp)-Cs(Imp) ---- (16).

The equation (16) computes the impact deviation between the rule and the cluster.

If Idev< Th then

Assign di with the class Cs.

End

End.

End

The above discussed high dimensional clustering algorithm performs the grouping of similar gene data points in to a class. For each data point available, for each of the gene present in the data point, a multi variant similarity is computed and based on that an impact deviation is computed. Finally if the impact deviation is less than the threshold, then the data point is assigned to the class.

### 5.5 GENE SELECTION APPROACH

With the available information, the other gene values with the data points available in the other cluster are identified. The gene values which are not more similar to the other cluster data point are identified. For each gene present in the data point, we identify the impact of the gene in the impact matrix and using the values of genes the impact of the gene in the other cluster is identified. For the number of data points from the each cluster, the data points closure to the available gene value is identified and the number of data points with the same gene value of other cluster is identified. Using these details, the impact of the gene value is computed.

Procedure:

Input: MVIM, Data point Dp, Cluster Cs.

Output: Gene selection Gs.

For each cluster Cs

For each gene value Gv of Dp

For each cluster cs

Compute number of data points present within the range.

Nodp = $\sum_{i=1}^{size(Cs)} Gs == Cs(i)(gv)$

Compute number of data points present within range in another cluster.

Nodpo = $\sum_{i=1}^{size(OCs)} Gs == OCs(i)(gv)$

If Nodpo>Nodp

Else

Add to list Gs.

End

End

End

The above discussed procedure shows how the gene selection is performed. In this procedure, for each cluster available, for each gene value, we identify the number of data points falls within the range values and identify the number of data points falls in the range of other cluster. Using both the values, that if the gene has more number of data samples in the current category than the other category then the gene is selected.

## 6 RESULTS AND DISCUSSION

The proposed multi variant gene selection approach has been implemented and tested with Matlab environment and the proposed method has produced efficient results in clustering and the method has produced efficient results in time complexity and accuracy of clustering. Also the method has produced efficient results in gene selection and reduces the time complexity also.

***Table 1: Details of the data set being used***

| Dataset | Number of Data Points (N) | Attributes (d) | Attribute Values (AA) | Classes (K) |
|---|---|---|---|---|
| UCI | 286 | 9 | Multi variant | 2 |
| Wisconsin | 569 | 32 | Multi Variant | 2 |
| BCSC | 765 | 13 | Multi variant | 2 |
| Vanteer | 24481 | 19 | 1 | 2 |

The Table 1, shows the details of the data set being used to evaluate the performance of multi variant gene selection approach. The data sets listed in Table1 has number of data points and each has different number of classes. The data point has more number of dimensions and each dimension has different attribute type. The method has been evaluated with each of the data set where each has various dimensions and the number of samples with data types is also different.

Each data set has been used to evaluate the performance of the proposed method on clustering accuracy. From the data set 80 percent of the data set has been used as training set and remaining 20 percent of data points are used to perform evaluation. The method produces efficient results with each of the data sets and the gene selection approach has produced less time complexity with higher detection accuracy.

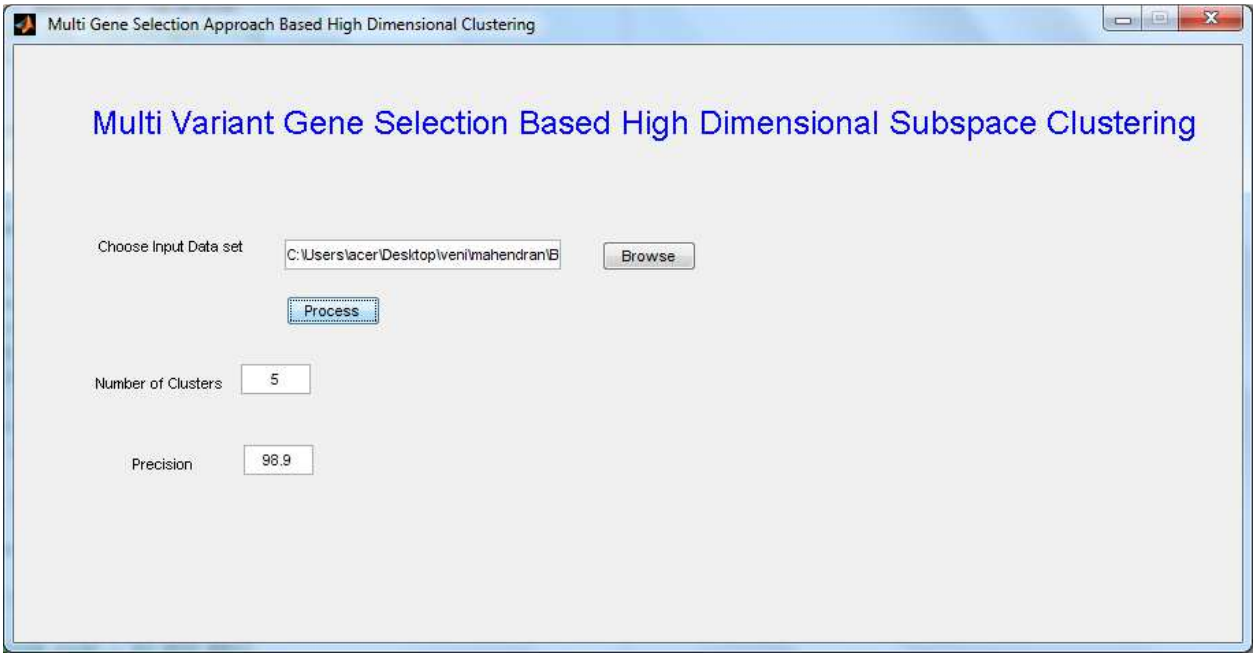



***Figure 1: Snapshot of the result produced***

The Figure 1 show the snapshot of results produced by the proposed method and it show the proposed method has produced efficient results in clustering.

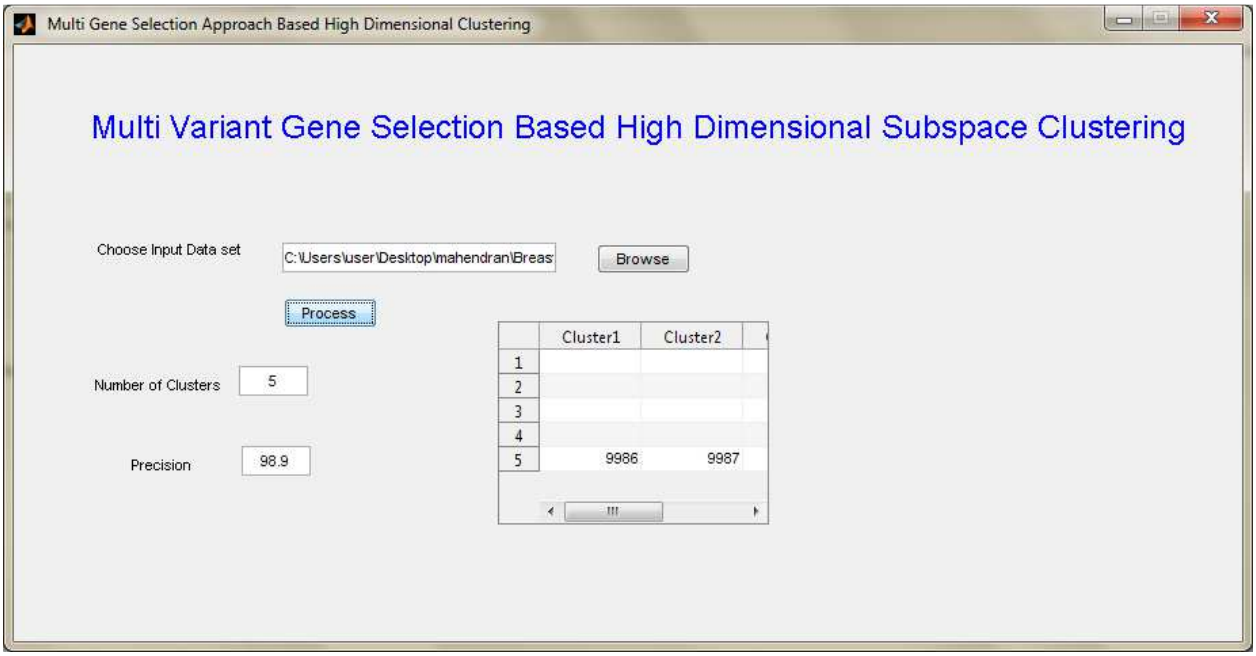
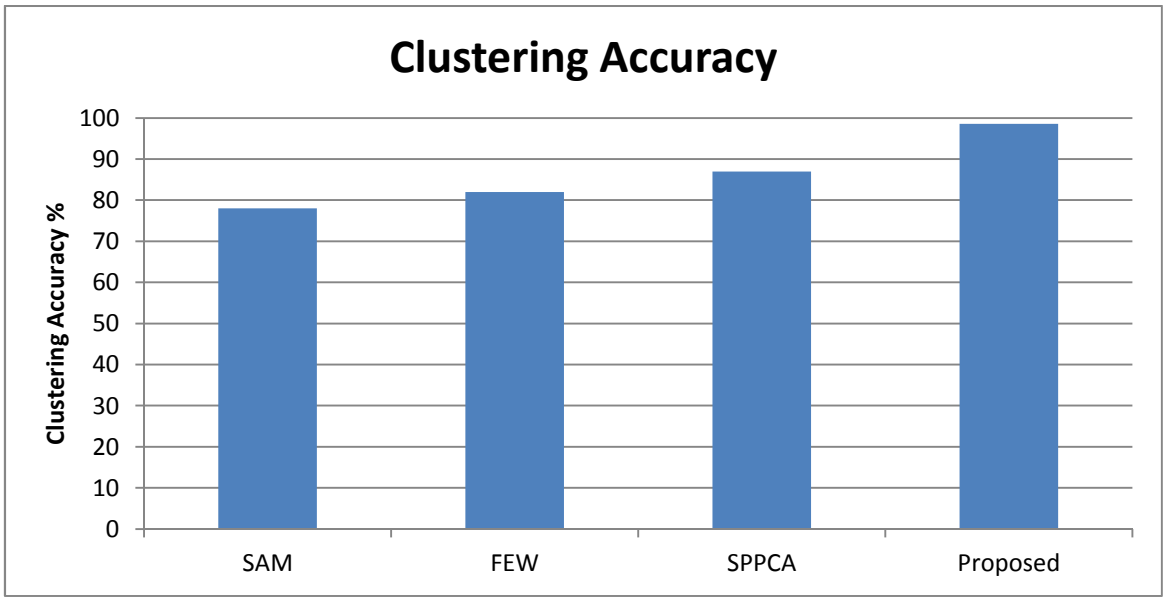


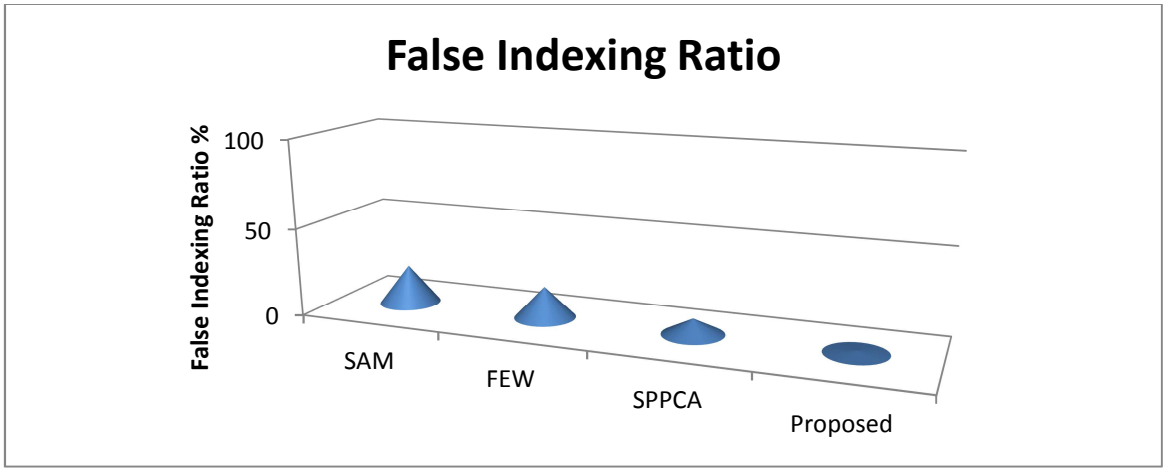

***Figure 2: Snapshot of the result produced***

The Figure 2, shows the snapshot of the result produced and the number of genes participated in each cluster is identified and displayed. The above figure show the result of gene selection produced by the proposed approach and it shows that the selected number of genes for the data points of cluster1 and cluster 2.



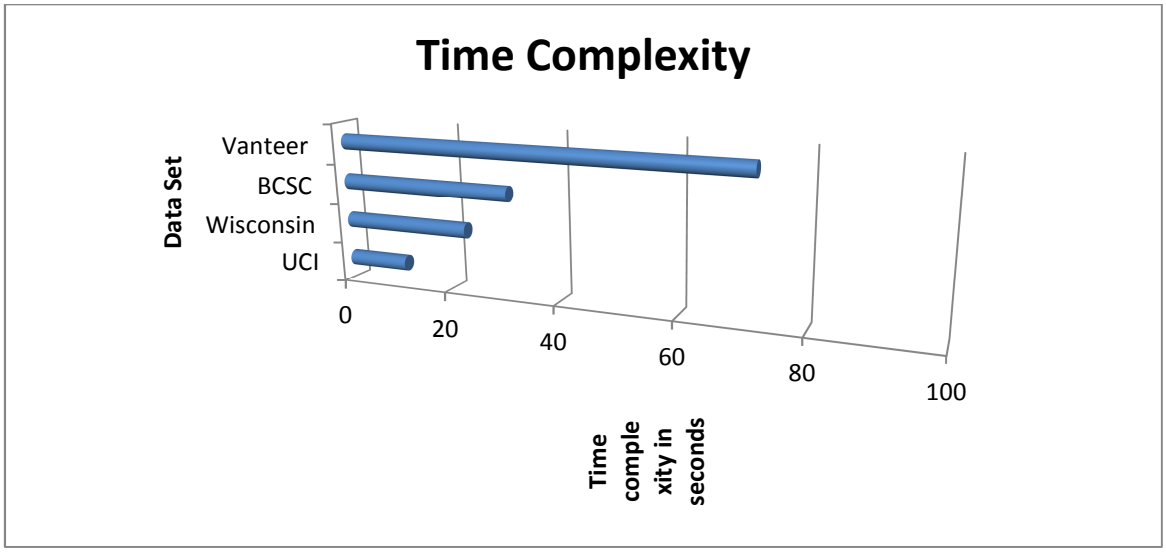

***Graph1: Comparison of clustering accuracy***

The Graph 1, shows the comparison of clustering accuracy produced by different methods. It shows clearly that the proposed method has produced efficient clustering than other approaches.




***Graph2: Comparison of false indexing ratio.***

The graph 2, shows the false indexing ratio produced by different methods and it shows clearly that the proposed method has produced less false ratio than other methods.




***Graph 3: Comparison of time complexity***

The graph 3, shows the time complexity produced by the proposed method of clustering different data sets. It shows clearly that the proposed approach has produced less time complexity in all the data sets where each of them varies with the dimensions and number of samples.

***Table 2: Comparative analysis results on vanteer data set***

| Methods | Accuracy | False Indexing | Time Complexity |
|---|---|---|---|
| SAM | 81 | 19 | 93 |
| FEW | 84.5 | 15.5 | 81 |
| SPPCA | 89.3 | 10.7 | 63 |
| Proposed | 97.8 | 2.2 | 24 |

The Table 2 shows the comparative results produced by the different methods on clustering accuracy, false indexing, and time complexity while using the vanteer data set. It shows clearly that the proposed method has produced efficient results than other methods.

## 7 CONCLUSION

We proposed a multi variant gene selection approach based on multi variant impact matrix to develop the clustering accuracy. We preprocess the data points and generate the rule sets according to number of range values available with the data point. Using the data points and the rule sets we generate the multi factor impact matrix which is used to identify the class of data points. Similarly, based on the multi variant impact matrix and rule sets, a set of gene is selected based on computed similarity measure values. The proposed method increases the accuracy of clustering and reduces the time complexity highly. Also the proposed method reduces the false indexing ratio produced by other methods also.

## REFERENCES

[1] Lei Yu, Yue Han, and Michael E. Berens, ”Stable Gene Selection from Microarray Data via Sample Weighting”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 1, pp. 262-272, 2012.

[2] E.N.Sathishkumar, K.Thangavel, T.Chandrasekha, “A Novel Approach for Single Gene Selection Using Clustering and Dimensionality Reduction”, International Journal of Scientific & Engineering Research, Volume 4, Issue 5, pp. 1540-1545, 2013.

[3] PradiptaMaji and Chandra Das, “Relevant and Significant Supervised Gene Clusters for Microarray Cancer Classification”, IEEE Transactions on Nanobioscience, vol. 11, no. 2, pp. 161-168, 2012.

[4] UjjwalMaulik, AnirbanMukhopadhyay, DebasisChakraborty, "Gene-Expression-Based Cancer Subtypes Prediction Through Feature Selection and Transductive SVM", IEEE Transactions on Biomedical Engineering, vol. 60, no. 4, pp. 1111-1117, 2013.

[5] Shuichi Kawano et al., "Identifying Gene Pathways Associated with Cancer Characteristics viaSparse Statistical Methods" IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 4, pp. 966-972, 2012.

[6] Jiexun Li, Hua Su, Hsinchun Chen, and Bernard W. Futscher, " Optimal Search-Based Gene Subset Selection for Gene Array Cancer Classification" IEEE Transactions on Information Technology in Biomedicine, vol. 11, no. 4, pp 398-405, 2007.

[7] CHEN-AN TSAI et al. "Gene Selection for Sample Classifications in Microarray Experiments" DNA and cell Biology, Volume 23, Number 10, pp 607-614, 2004.

[8] Mark Burton, MadsThomassen, Qihua Tan and Torben A. Kruse, "Gene Expression Profiles for Predicting Metastasis in Breast Cancer: A Cross-Study Comparison of Classification Methods" The Scientific World Journal Volume 2012, Article ID 380495, 11 pages, 2012.

[9] Hsi-Che Liu et al., "Comparison of feature selection methods for cross-laboratory microarray analysis" IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2013.

[10] Gabriele Giarratana et al., "Data Mining Techniques for the Identification of Genes with Expression Levels Related to Breast Cancer Prognosis" Ninth IEEE International Conference on Bioinformatics and Bioengineering, pp 295-300, 2009.

[11] Marina Bessarabova et al., "Bimodal gene expression patterns in breast cancer" BMC Genomics, Supplementary 1, 2010.

[12] Vitoantonio Bevilacqua et al., "Comparison of data-merging methods with SVM attribute selection and classification in breast cancer gene expression" BMC Bioinformatics 2012.

[13] Jonathan Tyrer, Stephen W. Du_y and Jack Cuzick, "A breast cancer prediction model incorporating familia land personal risk factors" Statistics In Medicine Statist. Med., pp 1111–1130, 2004.

[14] Jorgen Aaroe et al., "Gene expression profiling of peripheral blood cells for early detection of breast cancer" Breast Cancer Research, 2010.