SCDT: FC-NNC-structured Complex Decision Technique for Gene Analysis Using Fuzzy Cluster based Nearest Neighbor Classifier

Sudha V.[1], Girijamma H. A.[2]
[1]Department of IS&E, RNS Institute of Technology, India
[2] Department of Computer Science & Engineering, RNS Institute of Technology, India

Article Info

Article history:

Received Feb 12, 2018
Revised Jun 17, 2018
Accepted Jun 20, 2018

Keyword:

Fuzzy classification
Gene analysis
Gene selection
Machine learning
Micro array data

ABSTRACT

In many diseases classification an accurate gene analysis is needed, for which selection of most informative genes is very important and it require a technique of decision in complex context of ambiguity. The traditional methods include for selecting most significant gene includes some of the statistical analysis namely 2-Sample-T-test (2STT), Entropy, Signal to Noise Ratio (SNR). This paper evaluates gene selection and classification on the basis of accurate gene selection using structured complex decision technique (SCDT) and classifies it using fuzzy cluster based nearest neighborclassifier (FC-NNC). The effectiveness of the proposed SCDT and FC-NNC is evaluated for leave one out cross validation metric(LOOCV) along with sensitivity, specificity, precision and F1-score with four different classifiers namely 1) Radial Basis Function (RBF), 2) Multi-layer perception(MLP), 3) Feed Forward(FF) and 4) Support vector machine(SVM) for three different datasets of DLBCL, Leukemia and Prostate tumor. The proposed SCDT &FC-NNC exhibits superior result for being considered more accurate decision mechanism.

Corresponding Author:

Sudha V.,
Department of Information Science & Engineering,
RNS Institute of Technology, Bengaluru, India.
Email: sudhavinayakam@gmail.com

1. INTRODUCTION

The accuracy of diagnosis is the basis for the perfect treatment process to be adopted especially in the case of fatal disease like cancers, leukemia and prostrate tumor etc. Along with the histopathology, medical radiology and imaging techniques, the micro-array data analysis could be proven quite helpful as well as rightful if efficient techniques of analysis are evolved [1]. The accuracy of disease classification or early diagnosis depends upon, how accurately the gene of significance is selected.

The DNA-microarray data analysis is challenging in both aspects of statistically and computationally as it possesses non-linear noises along with high dimensionality of low sample data [2]. Many efforts towards disease diagnosis particularly cancer, tumor etc, classification have been seen in literature [3]-[10]. The section 2 describes the insights of related work. Various machine learning approaches are used for the classification which includes radial basis function (RBF), artificial neural network (ANN), support vector machine (SVM) etc. by forming the problem as binary classification. The problem of dimension reduction for searching most significant gene is being formulated as many problem spaces which includes 1) Mixed integer programming (MIP), 2) Bio-inspired optimization (BIO), 3) Mining association rules (MAR), and last but not the least 4) Ensemble technique (ET) [8].

The clinically comprehensive method requires handling high dimensional data with veracity and noises to handle ambiguity during the right gene candidate selection. This paper proposes a mechanism of structured complex decision technique (SCDT) for fuzzy clustering neighborhood cluster (FC-NC). Section 3 describes complete system model for SCDT & FC-NC, Section 4 describes about three different microarray datasets. Section 5 illustrates results and analysis followed by conclusion in Section 6.

1.1. Background

The accurate clustering of the data is a challenging and open research problem for classifications specially using supervise learning. An extensive survey is conducted to understand the effectiveness of clustering techniques particularly for medical data like micro array gene dataset [11]. For the purpose of tumor diagnosis, the approach of profiling the gene accuracy is comparatively of higher reliability with more accuracy than that of the method adopted by the medical imaging technique of morphological analysis of tumor. Traditionally adopted supervised learning approaches falls into pitfall of accuracy due to fewer samples of cancer types exist into the training dataset of gene expression as well the overheads due to higher data-dimensionality due to large gene expression.

In the work of Lipowang et al aims to select few numbers of genes to classify the cancer from the microarray data to meet the goal of balancing trade-off among the accuracy as well as minimization of the computational complexity or overheads[3]. They have used “feature importance ranking scheme” for the accurate or significant gene selection and formulated the classification problem as typical cluster of binary classification problem. The machine learning approaches used in their work are mix use of fuzzy neural network (FNN) and SVM. The dimension reductions obtained were getting same accuracy only by selecting 28 genes as compared to 16,063 genes of traditional method of that time. The typical dataset explored for the observations includes 1) Lymphoma Data, 2) SRBCT Data, 3) Liver Cancer Data, and 4) GCM data. They recommended considering the cooperation aspects between the genes to minimize the gene subset for more accurate prediction.

Further, the work which has refer to this includes the work by Chien-Pang et al who have introduced a method hybridized using genetic algorithm and dynamically setting up the parameter for significant gene selection and then further uses SVM for verification purposes to predict gene selectionefficiency [12]. The dimension reduction and feature selection is the core problem to be handled as gene expression microarray (GEMA) consist of hundred to sometime thousands of the features in a very small sample size. These high numbers of features in a small sample of GEMA makes it of very high dimension data. The conventional methods adopted for feature selection which is also called as gene selection in case of the GEMA analysis for the classicization of the dieses includes 1) Gain & Relief, 2) Chi Squares, 3) Fisher Score, and 4) Lasso etc.

The geneselection method is classified into three categories 1) supervised, 2) unsupervised and 3) semi-supervised on the basis of corresponding data types of 1) fully labeled, 2) unlabeled and 3) partially labeled respectively for classification or prediction of classes as described by [13]. Further, the feature available into the GEMA samples are categorized into two critical selections namely redundancy and relevancy. The Figure 1, shows the typical classification based on the combination of these two critical information’s.

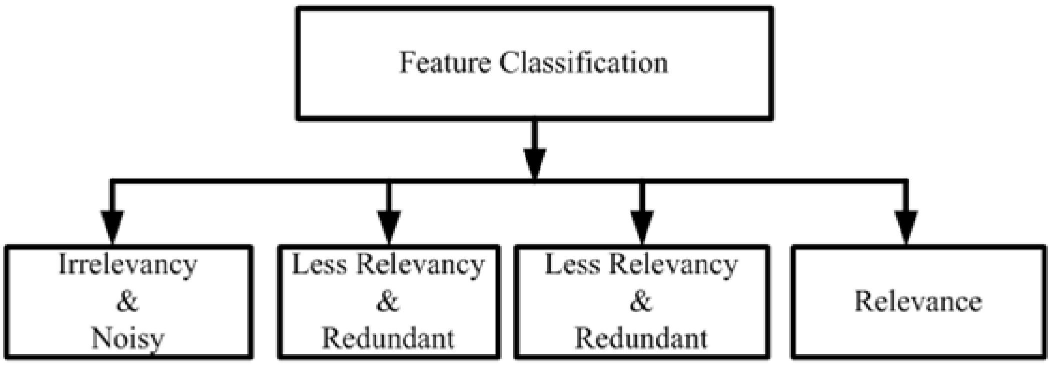


Figure 1. Gene featureselection or feature classicization basis

Recently the focus of research is very active as when a keyword of ‘gene selection’ given into IEEE Xplore a digital library then approximately 51 journals was found only from 2016 till 3rd February 2018. Tang et al in their method of feature selection from GEMA have introduced an improvised mutual information correlation (MIC) to handle the distortion due to noise in gene and challenges of multivariate distribution estimation by adopting relevance boosting and enhancement of the feature enhancement [14]. Table 1 list the trends of the methods used for the gene selection.

Table 1. Trend of the Method Adoption for Gene Selection

Sl. No and references	Gene Selection / Classification Method	Dieses Classification & Dataset used
[15] Zhang et al. (2016)	● minimum redundancy feature selection method (mRMR) ● Multiple Kernel Machine (MKL) learning method	● Glioblastomamultiforme ● Cancer Genome Atlas(TCGA) database
[16] Azzawi et al. (2016)	● Two gene selection methods ● Gene expression programming (GEP)-based model	● Lung cancer ● Real microarray lung cancer datasets
[17] Mallik et al. (2017)	● maximal-relevance and minimal-redundancy	● Epigenetic Biomarker discovery ● Multi-Omics Prostate Carcinoma (PC) dataset
[18] Huerta et al. (2016)	● Genetic Algorithm ● Tabu Search ● Support Vector Machine	● Tumor classification ● Diffuse Large B-cell Lymphoma
[19] Saha et al. (2016)	● Fuzzy C-means	● Hypothetical condition of Yeast ● Yeast Sporulation, Yeast Cell Cycle, Arabidopsis, Human Fibroblast Scrum, Rat CNS
[20] Montiel (2016)	● Simulated annealing ● Support vector machine	● Leukemia database ● Colon Cancer database
[21] Nguyen (2016)	● Type-2 Fuzzy logic	● diffuse large B-cell lymphoma, leukemia cancer, and prostate
[22] Jin and Win (2016)	● Swarm intelligence	● Tumor classification ● Gene Microarray dataset
[23]Ray et al. (2016)	● Self-Organizing Map	● Gene Microarray dataset
[24] Wang et al. (2016)	● Matrix factorization	● Gene Microarray dataset
[25] Han et al. (2017)	● Particle Swarm Optimization	● SRBCT Data
[26] Li and Wang (2017)	● K-means algorithm	● ALL, GCM, LYM, NC160, MLL, HBC
[27] Feng et al. (2017)	● Principle Component Analysis	● PDDA-GE Dataset
[28] Omar et al. (2018)	● Feature selection principle	● Gene expression dataset
[29] Harikiran et al. (2015)	● segmentation of microarray images	● Gene Microarray dataset
[30] Hore et al. (2016)	● Image segmentation	● Alpert dataset

There are various studies being carried out in existing system towards analyzing microarray data using different forms of clustering approach. Existing mechanism of clustering are immensely iterative in its approach which evidently calls for computational complexity. Such complexity issues have never being addressed by any researchers till day. One of the effective mechanisms to resist such complexity problem is to design and develop a novel technique with very limited set of iteration unlike conventional machine learning approaches. As microarray data consists of higher number of information, there is a need of a system that can read all the explicit features of the database in order to perform an effective classification. Adoption of fuzzy-based inference system is one such approach where accuracy in classification and complexity can be balanced. But existing approaches towards fuzzy logic also doesn’t seem to offer much convincing outcomes towards classification posing as one impediment towards existing research works. The next section outlines the system model of proposed solution.

2. SYSTEM MODEL: SCDT & FC-NNC

The proposed system models SCDT & FC-NNC consist of $DS_i \in$ {DLBCL(DS_1), Leukemia(DS_2), Prostate Tumor (DS_3)}, where i=1,2,3. The individual dataset characteristics are shown in the Table 1a, 1b, and 1c of each $DS_{1,}DS_2$ and DS_3. The snapshot visualization of each dataset is shown in Table 2

Table 1(a). Description of DLBCL(DS_1) Dataset

Dataset name	Total Gene	Total Sample	DLBCL	FL
DLBCL(DS_1)	5470	77	58	19

Table 1(b). Description of Leukemia(DS2)

Dataset name	Total Gene	Total Sample	ALL	AML
Leukemia(DS_2)	5328	72	47	25

Table1 (c). Description of Prostate Tumor (DS_3)

Dataset name	Total Gene	Total Sample	ALL	AML
Prostate Tumor (DS_3)	10510	102	52	50

Table 2. Snapshot of each dataset DS_1,DS_2 and DS_3

	1	2	3	4
1	59	17480	3	384
2	267	12086	52	-325
3	66	8611	-7	491
4	-37	24197	25	-694
5	109	15109	38	-108
6	71	9059	-23	-220
7	31	29480	31	-5868
8	148	8305	-21	-96
9	84	10321	2	-4933
10	53	10599	-11	-266
11	72	15842	-32	-5193

DLBCL(DS_1)

	1	2	3	4
1	88	15091	7	311
2	283	11038	37	134
3	309	16692	183	378
4	12	15763	45	268
5	168	18128	-28	118
6	71	34207	65	154
7	55	30801	43	80
8	-2	25147	338	269
9	268	15272	29	188
10	219	21801	-36	-39
11	82	18167	-8	115

Leukemia(DS_2)

	1	2	3	4
1	6.1000	-0.1000	11.9000	14.4000
2	1	0	2	4
3	22	2	51	52
4	14	6	15	21
5	13	4	39	25
6	20	1	23	29
7	16	8	47	33
8	13	0	29	15
9	32	8	96	37
10	18	14	58	32

Prostate Tumour(DS_3)

2.1. Gene Selection Method: Conventional and Proposed SCDT

Three conventional methods for the gene sections includes 1) Two sample T test(2STT), 2) Entropy test(ET) and 3) Signal-to-Noise Ratio(SNR) is evaluated with random sample size section(S_s) for gene ranking and visualizing top-k gene, where k=3. Along with proposed Structured Complex Decision Technique (SCDT). The section 3.1.1 describes 2STT.

2.1.1. Two sample T-test (2STT)

In this process, two independent samples of data is taken and in order to know the whether the average difference among these two samples are significant or not, the 2-Sample-T-test (2STT) is done. In the context of gene selection, the 2STT is performed on each gene and the expression levels are separated on the basis of class variable. If the value of '**abs(t)'** is found more that indicates that the gene is more important.nIf the two-data size of n_1 and n_2 with their sample mean as μ_1 and μ_2 as well σ_1 and σ_2 be their sample standard deviation, then the value of t is computed by Equation 1.

$$T = \frac{(\mu_1 - \mu_2)}{\sqrt{\left(\frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}\right)}} \tag{1}$$

2.1.2 Entropy Test (ET)

The cases where the assumption is that classes are normally distributed relative entropy(RE) or Kullback-Liebler distance or divergence test is conducted using Equation 2. The gene having highest value of entropy is selected for the input of classification module.

$$\frac{1}{2}\left[\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} - 2\right) + \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)(\mu_1 - \mu_2)^2\right] \tag{2}$$

2.1.3. Signal to Noise Ratio (SNR)

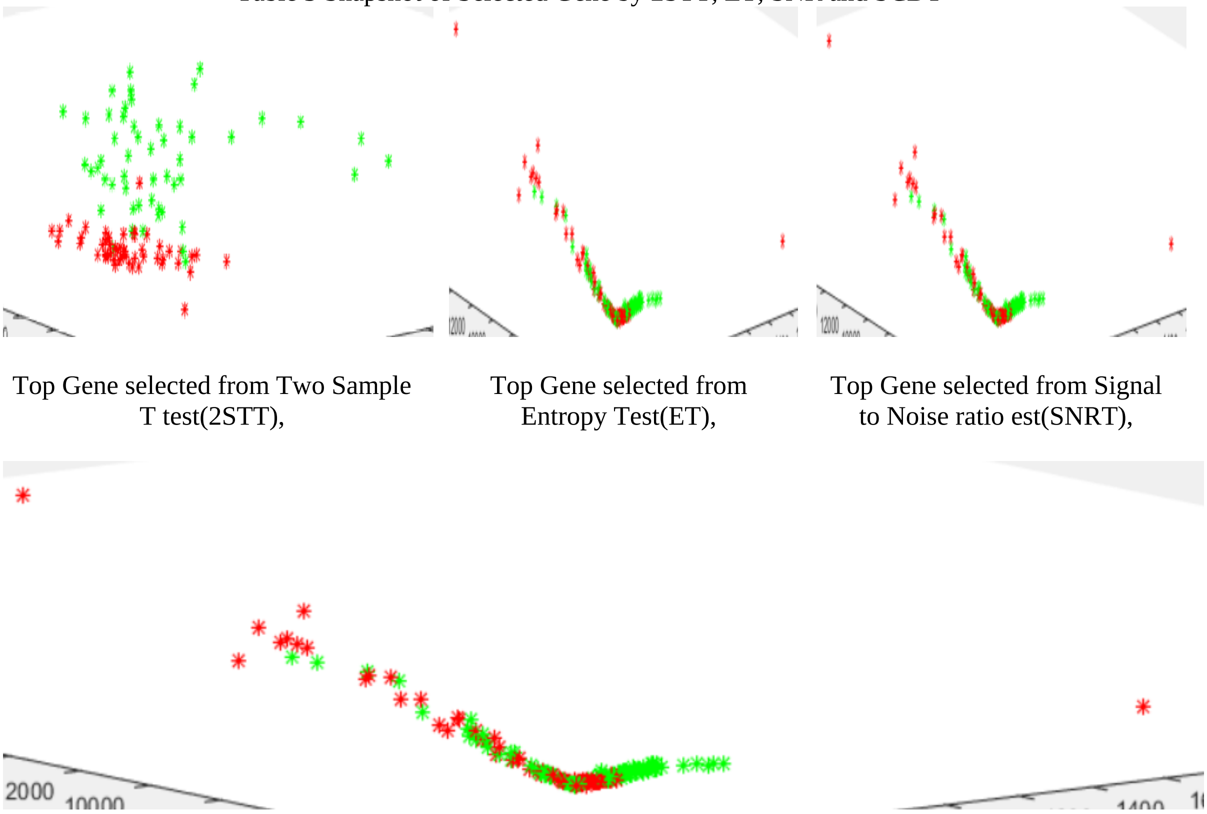
SNR defines the relative class separation metric by means of signal quality and noise

2.2. Gene Raking Algorithms

The proposed SCDT gene selection algorithm takes input from the three-different algorithm namely 2STT, ET and SNR for greatest ranking selection of gene for the classification purpose. On the execution of above algorithm, the snapshot of each individual method is taken and shown in the table 3

```
SCDT: Gene Ranking Algorithm: GR-Algorithms
Create Empty vector for 2STT, ET, SNR
for each Gi
[v1, v2]←f(All, FL)
[mu1m mu2]←$f_{mean}$(v1, v2)
[sd1, sd2]← $f_{std}$(v1,v2)
[n1, n2]←$f_{len}$(v1,v2)
2STT ←formula
ET← formula
SRN← formula
Process for Proposed SCDT
Normalization of 2STT, ET, SRN
[2STT, ET, SRN]← [2STT /fmax(2STT), ET /fmax(ET), SRN /fmax(SRN)]
SCDT←$f_{avg}$(2STT, ET, SRN)
```

Table 3 Snapshot of Selected Gene by 2STT, ET, SNR and SCDT



Top Gene selected from Two Sample T test(2STT),

Top Gene selected from Entropy Test(ET),

Top Gene selected from Signal to Noise ratio est(SNRT),

Top Gene selected from proposed SCDT

2.3. Classification based on the Selected Gene

The system is evaluated for five different classifications in which four namely 1) Radial Basis function, 2) MLP, 3) Feed forward, 4) SVM and finally FC-NNC is used

2.3.1. Radial Basis Function (RBF)

The radial basis network (RBN) approximates the function by adding additional layer to the hidden layer of RBN unless it reaches or achieves specified (or targeted) mean square objectives.

$\boldsymbol{f}_{\mathrm{rbn}}$(Input vector, Target class value) : →RBN

2.3.2. Multilayer Perception (MLP)

It is basically a type of neural network that uses feed forward-based learning mechanism with presence of three distinct layers of nodes. It is also known for its adoption of supervised learning approach that is termed as Back propagation algorithm. MLP is known for its capability to understand the distinction of linear and non-linear data. MLP is also known for its utilization of sigmoid function that is empirically represented by

$y(v_i)=\tanh(v_i)$ and $y(v_i)=(1+e^{-vi})^{-1}$

The preliminary component is basically a hyperbolic tangent with a range of [-1 1] while the second component is basically representation of logistic function with a range of [0 1].

2.3.3. Feed Forward

A feed forward learning process is one of the frequently used training algorithms which govern the orientation of the information restricted to a single direction. The operation of feed forward approach is carried out both in single and multiple-layer perceptions where both of them are associated with pros and cons. The pros factor of single and multiple-layered perception is its simplicity and capability to solve complex problems respectively. On the other side, the cons factor of single and multiple-layered perception is its consumption of higher computational time and includes increasing iterations respectively.

2.3.4. Support Vector Machine (SVM)

It is also a kind of machine learning concept that uses supervised learning approaches with a target of applying them for performing regression or performing classification operation. SVM is capable of performing both linear and non-linear classification quite effectively irrespective of its input type of higher degree of dimensionality. In order to apply this algorithm, it is required for labeling all the data. Implementation of the regression, identification of outliers, and classification is carried out using hyper plane in support vector machine. This scheme is also capable of controlling the computational load that allows simpler processing of dot product using a variable using kerne function $k(x, y)$.

2.3.5 Fuzzy Clustering Neighborhood Cluster (FC-NNC)

The prime intention of this is to embedded the better degree of freedom in both the inference (Mamdani and Takagi Sugeno) models in Fuzzy logic in order to ensure enhance capability to address uncertainties. This version of fuzzy logic has more capability as compared to existing one as it offers more practicality in the inference process. In conventional fuzzy logic based implementation, the crisp inputs are given to fuzzifier which is further forwarded as fuzzy sets to the inference block that is controlled by a set of fuzzy rules. The fuzzy outcomes are then forwarded to the defuzzifier in order to obtain crisp outputs. The proposed FC-NNC performs the similar step till inference block but after that it is significantly amended. The fuzzy inputs in FC-NNC are subjected to a special form of output processing. In this case, a type reducer obtains the input of fuzzy output sets, processes it and then forwards it to defuzzifier block. There are two outputs obtained in FC-NNC process i.e. one of crisp output and another is type-reduced set.

3. MICROARRAY DATA SET

Basically, microarray can be said to be a collection of different number of spots of DNA, where these information is utilized for computing the degree of expression associated with gene. Usually, the process of representation of gene expression data is carried out using expression matrix, where the information retaining columns represent single experimental data while all the rows exhibits complete collection of experimental data. Basically, it is an archive of various forms of data in microarray that consists of essentially the information of gene expression. The prime purpose of this database is to perform an effective management of the data wth an aid of the data index assisting in generating a query. Different forms of micro-array data are ArrayTrack, ImmGen database, ArrayExpress, GeneNetwork, MUSC, UPSC-BASE, Stanford Microarray database. All these databases offer voluminous information of gene expression that is used for public utilization. Consisting of more than 60, 000 samples of data there are more than millions of profiles in gene expression.

4. RESULTS AND ANALYSIS

This section discusses about the results being obtained from the proposed system. The complete analysis of the outcome is carried out with an aid of F1-Score; Leave one out cross validation metric, Sensitivity, Specificity, and Precision. The section also elaborates about these methods individually and illustrates the proposed analysis of results

4.1. Analysis of F1-Score

In binary classification, the F1-score generally measures the accuracy of the test which is computed on the basis of precession and recall values. The value of F1-score is computed by Equation 3, where P= precision, S= Sensitivity. The best value of F1-score is considered as 1 and the worst one as 0.

$$\text{F1-Score}= 2 \times \left[\frac{(P\times S)}{(P+S)}\right] \qquad (3)$$

The outcome shown in Figure 2 highlights that F1-Score of proposed SCDT is much better than existing approaches with respect to SNR, entropy, and t-test. Whereas, proposed FC-NCC offer better performance in comparison to existing machine learning techniques i.e. RBF, MLP, Feed-forward, and support vector machine, etc. A closer look into the performance will only show that proposed SCDT offers better F1-score in comparison to FC-NCC. The value of F1-Score for Proposed SCDT is found at 1 for least number of gene from the gene range of 5 -30. Even at the incremental values of the gene, the F1-score reduces but again becomes consistent at 30 gene at highest level of score 1. That shows that the proposes SCDT achieves best value of F1-score. At the same time at the lower gene the 2STT, SNR and then Entropy exhibit the better performance in reducing order. Whereas on the higher selection of gene SNR gets better results as compared to the 2STT and entropy (both exhibit same value). Figure 3 shows performance of F1-score vs changing value of numbers of gene.

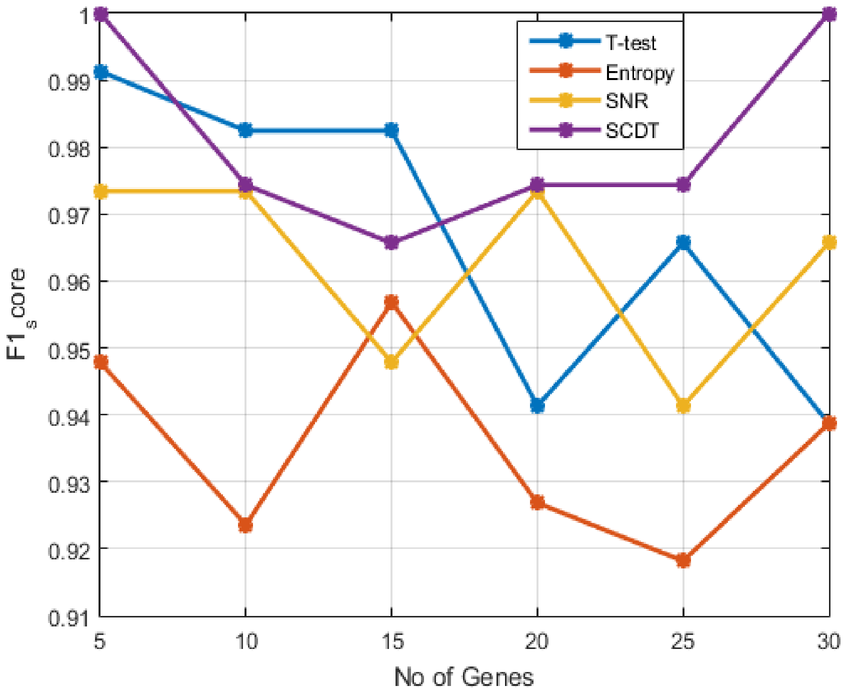


Figure 2. Performance of F1-score vs changing value of numbers of gene (a) 2STT, (b) Entropy test, (c) SNRT, (d) Proposed SCDT

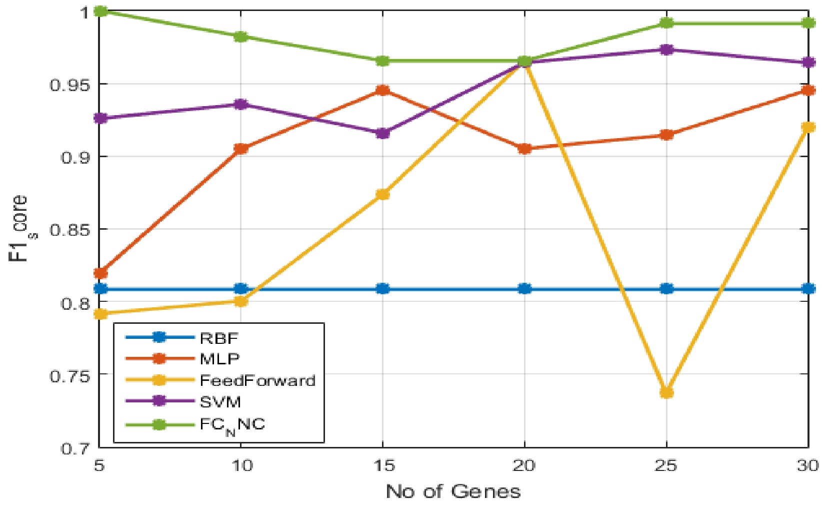


Figure 3 Performance of F1-score vs changing value of numbers of gene (a) RBF, (b) MLP, (c) Feed forward, (4) SVM, (5) Proposed FC-NCC

4.2. Analysis of Leave One Out Cross Validation Metric (LOOCV)

Usually, in gene expression dataset of microarray the number of samples are very small, therefore to provide exhaustive training leave one out cross validation method(LOOCV) is used. In LOOCV the entire dataset is divided into ‘K’ random and distinct subset. The K-1 is used for training and kth sample is used for the testing purpose. The accuracy of LOOCV is computed by Equation 4, where A is the count of correctly classified samples.

$$LOOCV = \frac{A}{K} \tag{4}$$

A comparison in the trends of SCDT and FC-NCC shows that SCDT offers increasing value of LOOCV in comparison to FC-NCC over increasing number of genes. This is another clear indicating that proposed SCDT could offer better degree of information while attempting to perform classification of the disease or any other form of abnormality in microarray data. Till now, the trend of SCDT is found to offer similar form of consistency for both F1-score and LOOCV. Performance of LOOCV vs changing value of Numbers of Gene as shown in Figure 4. Figure 5 shows performance of LOOCV vs changing value of numbers of gene

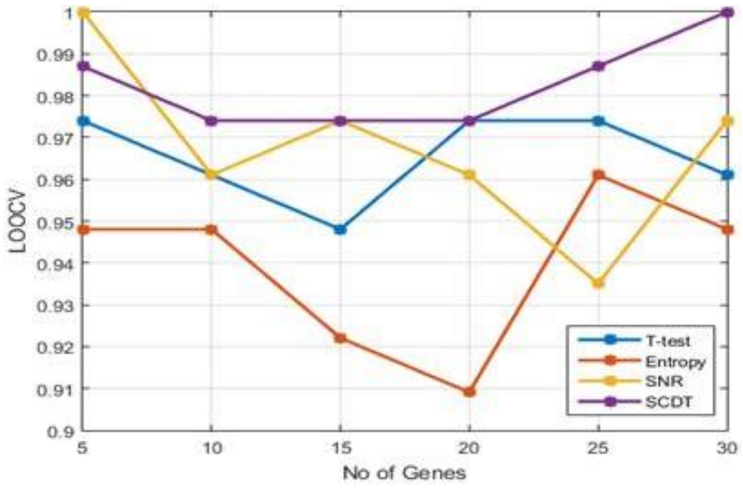


Figure 4. Performance of LOOCV vs changing value of numbers of gene (a) 2STT, (b) Entropy Test, (c) SNRT, (4) Proposed SCDT

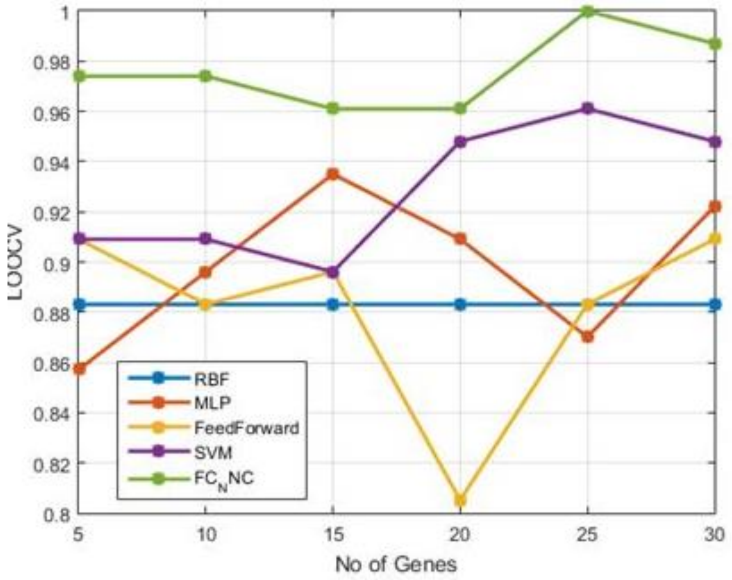


Figure 5. Performance of LOOCV vs changing value of numbers of gene (a) RBF, (b) MLP, (c) Feed forward, (4) SVM, (5) Proposed FC-NCC

4.3. Analysis of Precession

Precision is one of the elementary parameter used in pattern recognition as well as in classification problems. The computation of the precision is carried out as follows Equation 5.

$$P = \frac{RI - EI}{EI} \tag{5}$$

The above expression shows that precision P is calculated by dividing the difference of relevant information RI and extracted information EI with extracted information EI. This expression is always interpreted with respect to probability. The outcomes obtained are as shown in Figure 6 and Figure 7.

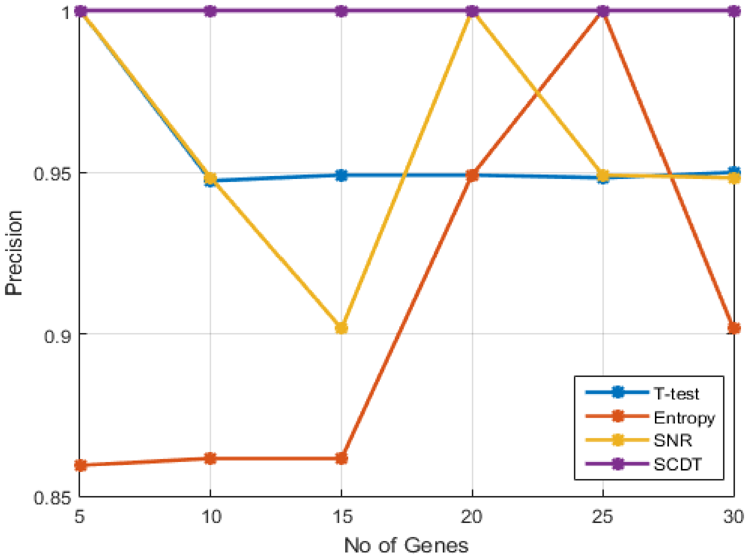


Figure 6. Performance of precision vs changing value of numbers of gene (a) 2STT, (b) Entropy test, (c) SNRT, (4) Proposed SCDT

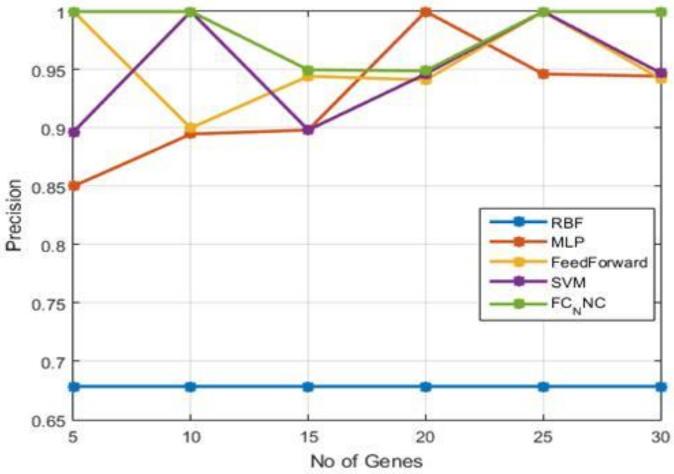


Figure 7. Performance of precision vs changing value of numbers of gene (a) RBF, (b) MLP, (c) FeedForward, (4) SVM, (5) Proposed FC-NCC

A closer look into the pattern of the curve shown in Figure 6 and Figure 7 shows that both SCDT and FC-NCC offers similar linear pattern of precision with increasing number of genes. This outcome shows that irrespective of any number of genes, the proposed system using any form of fuzzy logic (not the conventional singleton one) will always yield similar consistency in its outcome, which is quite predictable in itself. The predictability in precision performance offers value added performance when attempting to perform classification of any form of clinical abnormalities in microarray data.

4.4. Analysis of Sensitivity

Sensitivity is another frequently used performance parameter for assessing classification performance. It is used for calculating amount of positive outcome considered to be accurately identified. The calculation of sensitivity is carried out in following manner:

$$Sen = \frac{X}{X+Y} \tag{6}$$

In Equation 6, Sensitivity is computed by considering X which is true positive identification of some clinical abnormality and Y which is false negative identification. The graphical outcome of sensitivity is as follows Figure 8 and Figure 9.

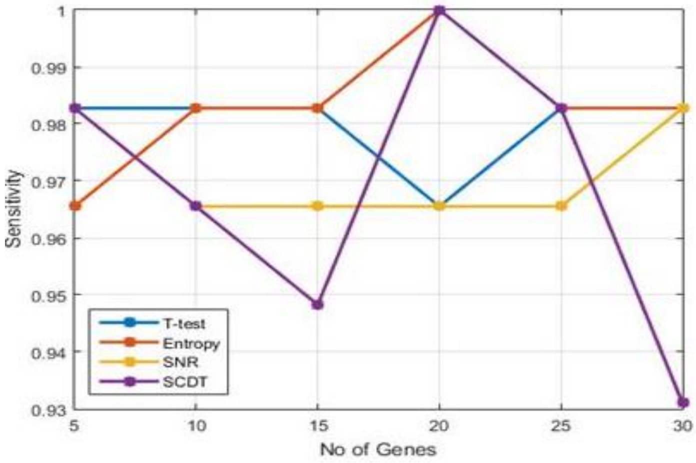


Figure 8. Performance of Sensitivity vs changing value of Numbers of Gene (a) 2STT, (b) Entropy Test, (c) SNRT, (4) Proposed SCDT

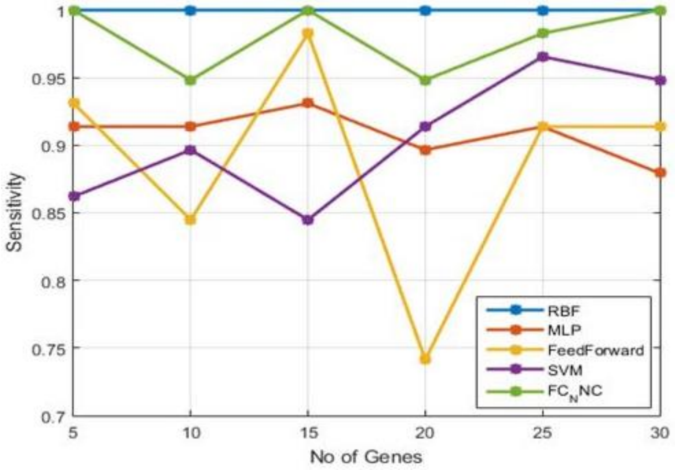


Figure 9. Performance of Sensitivity vs changing value of Numbers of Gene a.) RBF, b) MLP, c) Feed Forward, 4) SVM and 5) Proposed FC-NCC

The study outcome shows higher value of sensitivity for both the proposed system of SCDT and FC-NCC with increase of number of genes. This proves that success rate of identification process of the proposed system is good in comparison to any form of existing approaches.

4.5. Specificity

The proposed system uses specificity as the final performance parameter in order to assess the classification performance. This performance factor is used for accurately rejecting the false positive cases that could offer anomaly in the outcome. The calculation of specificity is carried out in following manner:

$$Spe = \frac{A}{A+B} \tag{7}$$

In the above expression, the variable A represent number of true negatives while B represents number of false negative. The graphical outcome of specificity is as follows in Figure 10 and Figure 11.

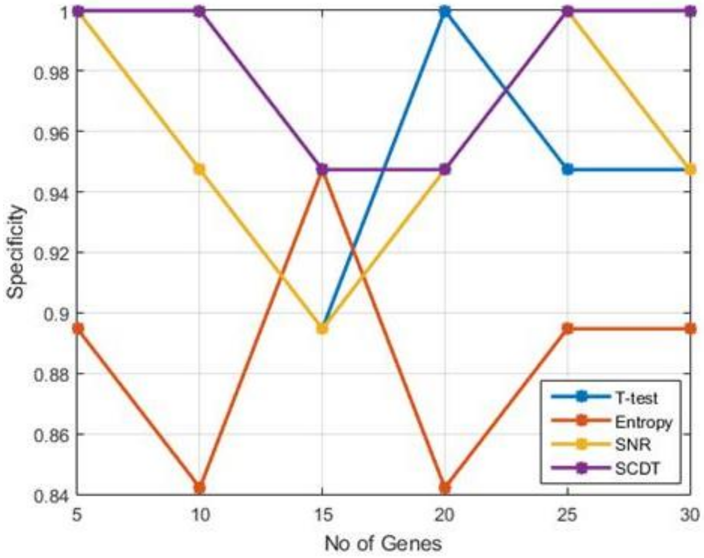


Figure 10. Performance of specificity vs changing value of numbers of gene a.) 2STT, b) Entropy Test, c) SNRT and 4) Proposed SCDT

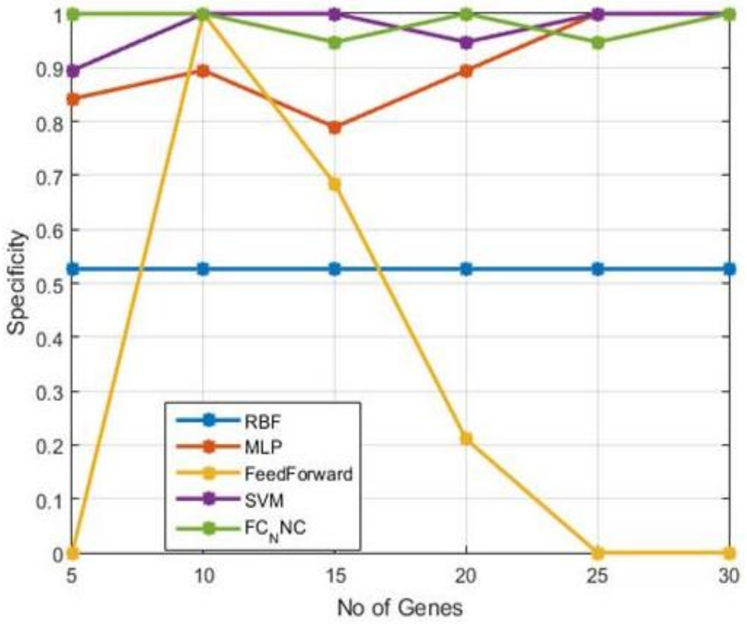


Figure 11. Performance of sensitivity vs changing value of numbers of gene (a) RBF, (b) MLP, (c) Feed forward, (4) SVM, (5) Proposed FC-NCC

Higher value of specificity is a direct interpretation that proposed system doesn’t have many chances to offer false negatives and hence the better form of accuracy within the proposed system is always retained. In this case, FC-NCC offers better performance in contrast to SCDT with respect to productiveness in is linearity in its outcome in comparison to other systems also.

5. CONCLUSION

This paper discusses about a simplistic modeling of solving classification problem considering the case study of microarray data. The approach is essentially meant for classifying cancer-based gene expression using modified version of fuzzy logic. The contribution of the proposed system is that it overcomes the dependencies of larger rule set unlike conventional fuzzy logic for assisting in better performance of clustering. Different performance parameters associated with accuracy has been considered for assessing the classification performance of the proposed system where the outcome shows that proposed system offers better classification accuracy in contrast to other existing system.

REFERENCES

[1] A. Mohammadi, M.H. Saraee, M. Salehi, “Identification of disease-causing genes using microarray data mining and Gene Ontology”, *BMC Medical Genomics*, vol.4, pp.4-12, Jan 2011

[2] J. Tang and S. Zhou, "A New Approach for Feature Selection from Microarray Data Based on Mutual Information," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 6, pp. 1004-1015, Nov 2016

[3] L. Wang, F. Chu and W. Xie, "Accurate Cancer Classification Using Expressions of Very Few Genes," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 1, pp. 40-53, Jan.-March 2007.

[4] M. Dashtban, M. Balafa, “Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts”, Genomics, vol.9(2), pp.91-107, Mar 2017

[5] B.A. Garro, K. Rodríguez, R.A. Vázquez, “Classification of DNA microarrays using artificial neural networks and ABC algorithm”, Applied Soft Computing, vol.38(5), pp. 48-60, Jan 2016

[6] H. Salem, G. Attiya, N. E-Fishawy, “Classification of human cancer diseases by gene expression profiles”, Applied Soft Computing, vol. 50(12), pp. 4-34, Jan 2017

[7] S. Y. Hsieh and Y. C. Chou, "A Faster cDNA Microarray Gene Expression Data Classifier for Diagnosing Diseases," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 1, pp. 43-54, Jan-Feb 2016

[8] T. Nguyen and S. Nahavandi, "Modified AHP for Gene Selection and Cancer Classification Using Type-2 Fuzzy Logic," in *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 2, pp. 273-287, April 2016

[9] M.K. Khormuji, M. Bazrafkan, “A novel sparse coding algorithm for classification of tumors based on gene expression data”, Medical & biological engineering & computing, vol. 54(6), pp.869-76, Jun 2016

[10] S.A Ludwig, S. Picek, D. Jakobovic, “Classification of Cancer Data: Analyzing Gene Expression Data Using a Fuzzy Decision Tree Algorithm”, InOperations Research Applications in Health Care Management, Springer, Cham, vol. 262, pp. 327-347, Jan 2018

[11] V. Sudha and H. A. Girijamma, “Appraising Research Direction & Effectiveness of Existing Clustering Algorithm for Medical Data”, *International Journal of Advanced Computer Science and Applications,* vol. 8(3), pp. 343-351, Mar 2017

[12] Lee CP, Leu Y. A novel hybrid feature selection method for microarray data analysis. Applied Soft Computing. 2011 Jan 1;11(1):208-13.

[13] J. C. Ang, A. Mirzal, H. Haron and H. N. A. Hamed, "Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 5, pp. 971-989, September 1 2016.

[14] J. Tang and S. Zhou, "A New Approach for Feature Selection from Microarray Data Based on Mutual Information," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no. 6, pp. 1004-1015, November 1 2016

[15] Y. Zhang, A. Li, C. Peng and M. Wang, "Improve GlioblastomaMultiforme Prognosis Prediction by Using Feature Selection and Multiple Kernel Learning," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no. 5, pp. 825-835, September 1 2016.

[16] H. Azzawi, J. Hou, Y. Xiang and R. Alanni, "Lung cancer prediction from microarray data by gene expression programming," in IET Systems Biology, vol. 10, no. 5, pp. 168-178, 10 2016.

[17] S. Mallik, T. Bhadra and U. Maulik, "Identifying Epigenetic Biomarkers using Maximal Relevance and Minimal Redundancy Based Feature Selection for Multi-Omics Data," in IEEE Transactions on NanoBioscience, vol. 16, no. 1, pp. 3-10, Jan. 2017.

[18] E. Bonilla-Huerta, A. Hernández-Montiel, R. Morales-Caporal and M. Arjona-López, "Hybrid Framework Using Multiple-Filters and an Embedded Approach for an Efficient Selection and Classification of Microarray Data," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no. 1, pp. 12-26, Jan.-Feb. 1 2016.

[19] S. Saha, A. K. Alok and A. Ekbal, "Use of Semisupervised Clustering and Feature-Selection Techniques for Identification of Co-expressed Genes," in IEEE Journal of Biomedical and Health Informatics, vol. 20, no. 4, pp. 1171-1177, July 2016.

[20] L. A. Hernandez Montiel, "Hybrid Algorithm Applied on Gene Selection and Classification from Different Diseases," in IEEE Latin America Transactions, vol. 14, no. 2, pp. 930-935, Feb. 2016.

[21] T. Nguyen and S. Nahavandi, "Modified AHP for Gene Selection and Cancer Classification Using Type-2 Fuzzy Logic," in IEEE Transactions on Fuzzy Systems, vol. 24, no. 2, pp. 273-287, April 2016. doi: 10.1109/TFUZZ.2015.2453153

[22] C. Jin and S. W. Jin, "Gene selection approach based on improved swarm intelligent optimisation algorithm for tumour classification," in IET Systems Biology, vol. 10, no. 3, pp. 107-115, 6 2016.

[23] S. S. Ray, A. Ganivada and S. K. Pal, "A Granular Self-Organizing Map for Clustering and Gene Selection in Microarray Data," in IEEE Transactions on Neural Networks and Learning Systems, vol. 27, no. 9, pp. 1890-1906, Sept. 2016.

[24] D. Wang, J. X. Liu, Y. L. Gao, C. H. Zheng and Y. Xu, "Characteristic Gene Selection Based on Robust Graph Regularized Non-Negative Matrix Factorization," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no. 6, pp. 1059-1067, November 1 2016.

[25] F. Han et al., "A Gene Selection Method for Microarray Data Based on Binary PSO Encoding Gene-to-Class Sensitivity Information," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 14, no. 1, pp. 85-96, Jan.-Feb. 1 2017.

[26] J. Li and F. Wang, "Towards Unsupervised Gene Selection: A Matrix Factorization Framework," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 14, no. 3, pp. 514-521, May-June 1 2017.

[27] C. M. Feng, Y. L. Gao, J. X. Liu, C. H. Zheng and J. Yu, "PCA Based on Graph Laplacian Regularization and P-Norm for Gene Selection and Clustering," in IEEE Transactions on NanoBioscience, vol. 16, no. 4, pp. 257-265, June 2017.

[28] Omara, Hicham, Mohamed Lazaar, and YounessTabii. "Effect of Feature Selection on Gene Expression Datasets Classification Accuracy." International Journal of Electrical and Computer Engineering (IJECE) 8.5 (2018).

[29] Harikiran, J., P. V. Lakshmi, and R. Kiran Kumar. "Multiple feature fuzzy c-means clustering algorithm for segmentation of microarray images." International Journal of Electrical and Computer Engineering (IJECE) 5.5 (2015).

[30] Hore, Sirshendu, et al. "An integrated interactive technique for image segmentation using stack based seeded region growing and thresholding." International Journal of Electrical and Computer Engineering (IJECE) 6.6 (2016): 2773.

BIOGRAPHIES OF AUTHORS

Sudha V., currently working as Assistant Professor in the Department of Information Science & Engineering, RNS Institute of Technology, Bengaluru and having teaching experience of 11 year. Her research interests are in the field of Data mining, Data analytics and machine learning algorithms.

Girijamma H. A., currently working as Professor in the Department of Computer Science & Engineering, RNS Institute of Technology, Bengaluru and having teaching experience of 23 year. Her research interests are in the field of automata, fuzzy logic and machine learning algorithms.