ISSN print: 1224-5984
ISSN online: 2248-3942

Rom Biotechnol Lett. 2019; 24(3): 554-562
doi: 10.25083/rbl/24.3/554.562

Received for publication, July, 28, 2017
Accepted, April, 21, 2018

Original paper

Identifying Autism Biomarkers from Microarray Data using Intuitionistic Fuzzy Normalization and Feature Selection

R. PREMA[1], K. PREMALATHA[2]

[1]Assistant Professor/MCA, New Horizon College of Engineering, Bangalore, India
[2]Professor/CSE, Bannari Amman Institute of Technology, Sathyamangalam, India

Abstract

Analysis of gene expression data is essential in autism in order to retrieve the required information. Gene expression data generally contain a large number of genes but a small number of samples. The complicated relations among the different genes make analysis more difficult and removing irrelevant genes improve the quality of results. In this work, two novel techniques were used to identify the significant genes related to Autism. Initially, the dataset was normalized using Fuzzy Set (FS) and Intuitionistic Fuzzy Set (IFS) logics. The Minimum Redundancy Maximum Relevance with Mean based Ranking (mRMR-μR) was proposed for feature selection. To identify the performance of the proposed work, the experimental results were compared with Min-Max and Z-Score normalization methods and state-of-the art feature selection methods. The classifiers, Support Vector Machine (SVM) and k-Nearest-Neighbor (kNN) were used to identify the accuracy of selected features. The experimental results showed that the proposed feature selection method gave cent percent classification accuracy for the top-fifty selected genes using intuitionistic fuzzy normalized data.

Keywords Autism; Gene Expression Data; Feature Selection; Classification; Minimum Redundancy Maximum Relevance; Intuitionistic Fuzzy Normalization.

To cite this article: PREMA R, PREMALATHA K. Identifying Autism Biomarkers from Microarray Data using Intuitionistic Fuzzy Normalization and Feature Selection. *Rom Biotechnol Lett*. 2019; 24(3): 554-562. DOI: 10.25083/rbl/24.3/554.562

✉ *Corresponding author: R. PREMA, Assistant Professor/MCA, New Horizon College of Engineering, Bangalore, India. E-mail: premabit@gmail.com
K. PREMALATHA, Professor/CSE, Bannari Amman Institute of Technology, Sathyamangalam, India. E-mail: kpl_barath@yahoo.co.in

Introduction

A microarray dataset is a repository contains microarray gene expression data. The raw microarray data are images that are transformed into gene expression data matrices where rows represent genes, columns represent various samples such as tissues or experimental conditions and numbers in each cell that characterize the expression level of the particular gene in the particular sample[1].

In the data mining process, one of the most essential stages of pre-processing is normalization. Normalization is a method used to standardize the range of independent features of data. In many applications, the available features are continuous values, where each feature is measured in a different scale and has a different range of possible values. Autism dataset contains continuous gene expression values. So, it should be normalized before gene selection. In this paper, the gene expression values are normalized based on fuzzy logic using fuzzy sets that defines the intervals of a continuous random variable. In this paper, two different fuzzy techniques (FS and IFS) were used to normalize the data.

Data dimensionality reduction is one of the important machine learning tasks while facing data with enormity on size, missing values and noise [2, 3]. Gene expression dataset contains thousands of gene expression values, many of which may be irrelevant or redundant for classification [4]. Leaving out relevant attributes or keeping irrelevant attributes may affect the performance of the classification algorithm. Therefore statistical methods are required to identify a reduced search space are commonly used for classification [5].

This paper presented a filter approach mRMR-µR to identify the biomarker genes associated with Autism disorder. It identified k most informative genes out of the M original genes which were closely associated with the disease, where $k < M$. The experimental results were compared with tTest [6], Feature Correlation (FC) with class [7], Signal to Noise Ratio (SNR) [8], F-statistic and minimum Redundancy Maximum Relevance (mRMR) [9]. The rest of the paper is organized as follows. FS and IFS normalization process is explained in Section 2. Section 3 discusses the F-statistic and mRMR algorithm. Section 4 explains the proposed method mRMR-µR. Experimental results and susceptibility genes related to Autism are presented in Section 5.

Fuzzy Pre-processing

1.1. Fuzzy based Normalization

A fuzzy set A of a non empty set X is defined as $\langle x, \mu_A(x)\rangle : x \in X, \mu_A(x)$, is the membership function of the fuzzy set A. Fuzzy set is a collection of objects with graded membership i.e. having degrees of membership [10, 11]. In this paper, Autism dataset was transformed by exploitation of a fuzzy membership function rather than by using their absolute expression values. A membership function is a curve that defines how each point in the input space is mapped to a membership value between 0 and 1. A fuzzy membership function that is used to represent vague, linguistic terms is the Gaussian which is given in Equation (1).

$$\mu_A(x) = exp\left(-\frac{(x-m)^2}{2(k)^2}\right) \quad (1)$$

where m and k are centre and width of the fuzzy set A respectively. Here, all the sample values for each gene were considered as a set. To find the membership function of this non empty set, each gene values with respect to all the samples were fuzzified into three fuzzy qualifiers, low, medium and high. By applying the Gaussian membership function (Equation (1)), each gene values were normalized to a scale of 0 to 1, where 1 is the highest expression level and 0 is the lowest. Figure 1 shows the membership values of four random genes.

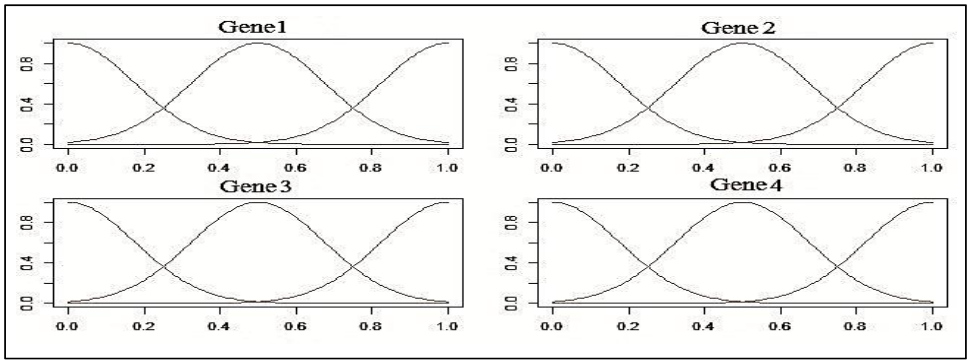

Figure 1. Membership Functions of Four Random Genes

1.2. Intuitionistic Fuzzy based Normalization

Fuzzification determines the degree of membership. In practice, due to insufficiency of the information available, the evaluation of membership and non-membership values is not always possible. Therefore an indeterministic part of which hesitation survives remains [10, 12]. The term intuitionistic fuzzification functions refer to formulating membership and non-membership functions of an IFS.

The Gaussian membership function applied by using the formulas given below.

$$\mu_A(x) = exp\left(-\frac{(x-m)^2}{2(k)^2}\right) \quad (2)$$

$$\gamma_A(x) = 1 - \left(exp\left(-\frac{(x-m)^2}{2(k)^2}\right)\right) \quad (3)$$

$$\pi_A(x) = \ 1 - \mu_A(x) - \ \gamma_A(x) \quad (4)$$

where $\mu_A(x)\&\gamma_A(x)$ represents membership and non-membership functions respectively and m and k are centre and width of the fuzzy set A respectively.

Let $D = [d_{ij}]_{m\times n}$, where $d_{ij} = \{\mu_{ij}, \gamma_{ij}, \pi_{ij}\}$ which denote membership, non-membership and hesitancy degree of the IFS matrix. The following matrix shows an example of IFS representation.

$$D = \begin{Bmatrix} \left(\mu_{i1}(x_1), \left(\gamma_{i1}(x_1)\right), \pi_{i1}(x_1)\right) & \left(\mu_{i1}(x_2), \gamma_{i1}(x_2), \pi_{i1}(x_2)\right) \cdots & \left(\mu_{i1}(x_n), \gamma_{i1}(x_n), \pi_{i1}(x_n)\right) \\ \left(\mu_{i2}(x_1), \gamma_{i2}(x_1), \pi_{i2}(x_1)\right) & \left(\mu_{i2}(x_2), \gamma_{i2}(x_2), \pi_{i2}(x_2)\right) & \left(\mu_{i2}(x_n), \gamma_{i2}(x_n), \pi_{i2}(x_n)\right) \\ \vdots & \vdots \quad \cdots & \vdots \\ \vdots & \vdots & \vdots \\ \left(\mu_{im}(x_1), \gamma_{im}(x_1), \pi_{im}(x_1)\right) & \left(\mu_{im}(x_2), \gamma_{im}(x_2), \pi_{im}(x_2)\right) \ldots & \left(\mu_{im}(x_n), \gamma_{im}(x_n), \pi_{im}(x_n)\right) \end{Bmatrix}$$

Minimum Redundancy Maximum Relevance

The mRMR is a feature selection approach that tends to select features with a high correlation with the class (output) and a low correlation between themselves. For continuous features, the F-statistic can be used to calculate correlation with the class (relevance) and the Pearson correlation coefficient can be used to calculate correlation between features (redundancy). Thereafter, features are selected one by one by applying a greedy search to maximize the objective function, which is a function of relevance and redundancy. Two commonly used types of the objective function are MID (Mutual Information Difference criterion) and MIQ (Mutual Information Quotient criterion) representing the difference or the quotient of relevance and redundancy, respectively.

The F-statistic value of a gene variable g_i in K classes denoted by h has the following form,

$$F(g_i, h) = [\sum_k n_k(\bar{g}_k - \bar{g})^2/(K-1)]/\sigma^2 \quad (5)$$

$$\sigma^2 = [\sum_k (n_k - 1)\sigma_k^2]/(N-K) \quad (6)$$

where $\bar{g}$ is the mean value of g_i in all samples, $\bar{g}_k$ is the mean value of g_i within the k^{th} class, σ^2 is the pooled variance, n_k is the number of samples in k^{th} class, σ_k is the variance of the k^{th} class and N denotes the number of samples.

The Pearson correlation coefficient value between two variables is given in Equation (7).

$$P(X,Y) = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}} \quad (7)$$

where N is the number of pairs of values, $\sum XY$ is the sum of the products of paired values, $\sum X$ is the sum of X values, $\sum Y$ is the sum of Y values, $\sum X^2$ is the sum of squared X values and $\sum Y^2$ is the sum of squared Y values.

mRMR-μR for Feature Selection

For a given dataset d of dimension m, with a set of M genes $G = \{g_1, g_2, \ldots, g_M\}$, the problem is to select an optimal subset of relevant genes k where (i) k is a proper subset of G and (ii) for k, a classifier gives the best possible classification accuracy. In other words, to identify a subset of genes where for any pair $\{g_i, g_j\} \in k$, the gene-gene correlation is minimum and gene-class mutual information is maximum.

The proposed method used F-statistic (Equation (5)) to identify the relevance between the genes and the class variables. It used Pearson correlation coefficient (Equation (7)) as the score of redundancy between the genes. The method considered both feature–feature correlation and feature–class mutual information to determine an optimal set of genes.

Initially, the feature–class mutual information was computed and the feature with highest mutual information was selected. The remaining k - 1 features were built in an incremental way, which maximized $max_{i\in k}\left[F(i,h) - \frac{1}{|k|}\sum_{j\in k}|P(i,j)|\right]$. To identify the significant genes, the average of gene-class mutual information

and the average of gene-gene correlation were calculated. The genes which have greater value than average mutual information and lesser value than average correlation were considered in the above mentioned maximization condition. Then the proposed method selected the gene that has the maximum difference. Algorithm 1 and 2 showed the proposed mRMR-µR feature selection method.

Algorithm 1 mRMR-µR Feature Selection

Input: Gene expression dataset *d* with *m* dimension,
Output: Top-*k* features dataset F'.

1. *id_left=d*
2. *relevance_list*=F-statistic (*d*);
3. *[R,id]=max(relevance_list)*
4. F'*=d(:id)*
5. *id_left=id_left−*F';
6. *for m=2:k do*
7. $relevance\_list = F - statistic(id\_left)$;
8. $redundancy_{list} = \frac{1}{|F'|}\sum_{i \in F'} PearsonCorrelation(i, id_{left})$
9. *[R,id]=**µ-Ranking***(*relevance_list, redundancy_list*)
10. $F' = F' \cup d(:id)$
11. *id_left=id_left-*F'
12. *endfor // (Top-k)*

Algorithm 2 µ-Ranking

Input: id_left, redundancy_list, relevance_list,
Output: Selected Feature id

1. *mean1=avg(relevance_list)*
2. *mean2=avg(redundancy_list)*
3. *max_rel_list = min_red_list =* ϕ
4. *for i=1:len(id_left) do*
5. *if relevance_list(i)>mean1*
6. *max_rel_list=max_rel_list∪d(:i)*
7. *end for*
8. *for j=1:id_left do*
9. *if redundancy_list(j)<mean2*
10. *min_red_list=min_red_list∪d(:j)*
11. *end for*
12. *diff_list =* ϕ
13. *subList= max_rel_list∩ min_red_list*
14. *for k=1:len(subList)do*
15. *diff_list=diff_list∪ (relevance(sublist(k))-redundancy(sublist(k)))*
16. *end for*
17. *return id(max(diff_list))*

Experimental Results

This paper analyzed the Autism dataset from experiment GSE25507 available at Gene Expression Omnibus (GEO) [14]. This dataset consists of 54613 genes with 69 samples of Autistic patients and 77 healthy children from the general population. Statistical analysis wasperformed with R packages [14]. The proposed algorithm selected a subset of relevant features from the Autism dataset. To evaluate the performance of mRMR-µR, the well known classifiers SVM and kNNwere employed. The selected genes were utilized for training the classifiers. The performance was evaluated using 10-fold cross validation. The radial basis kernel function was used for SVM classifier. The number of instance considered for determination of similarity with classes as three for kNN. The performance of the proposed method was compared with feature selection methods, tTest, SNR, FC, F-statistic and mRMR.

To demonstrate the performance of the proposed method, the top 5, 10, 15, 20, 30, 40 and 50 genes were selected as features for Autism classification. Each feature selection methods produced different set of features using the dataset without normalization. The same set of features was produced by the dataset using Min-Max and Z-Score normalization methods. But the dataset with FS and IFS normalization techniques gave different set of features for all above mentioned feature selection methods. To know

the accuracy of these features, SVM and kNN classifiers were used. From the figures 2B and 2C, it was observed that the Min-Max and Z-Score normalization makes no difference with SVM classifier. It gave same accuracy with the case that is no normalization. But the figures 2D and 2E showed that FS and IFS normalized dataset gave significant improvement in accuracy for all feature selection methods. The kNN classifier which depends on distance calculations was affected from the normalization since after the normalization all the dimensions had the same weight and no one dominate the others. From the figure 3, it was observed that kNN gives different accuracy for different normalization methods. And it gave a statistical improvement using normalization techniques. Also both fuzzy normalization methods outperformed the other two techniques.

As a general conclusion, from the figures 2 and 3, it was inferred that the mRMR-µR feature selection method gave higher accuracy than other methods. Both FS and IFS produced the highest accuracy compared to the other two normalization techniques. And IFS can give significant improvement than FS normalization method. Thus, fuzzy normalization methods improve the quality of the feature selection methods. Also it was observed that the dataset

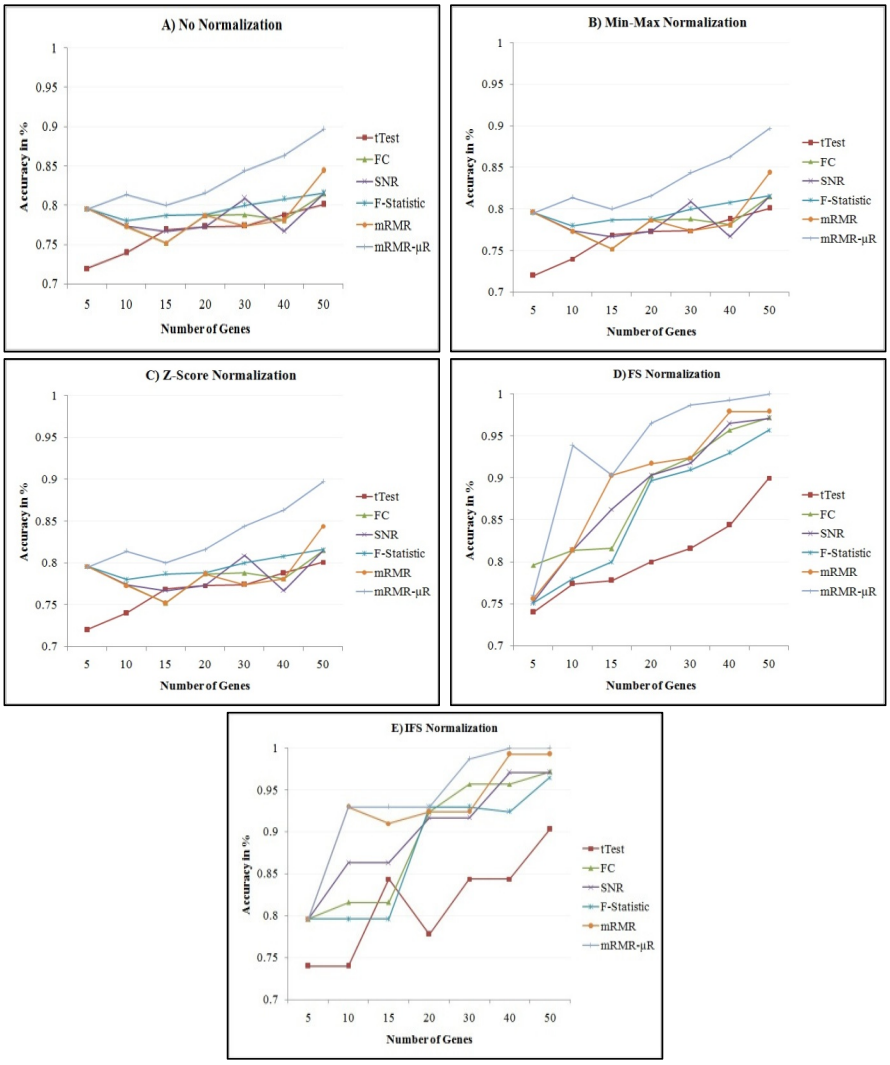

Figure 2. SVM Classification

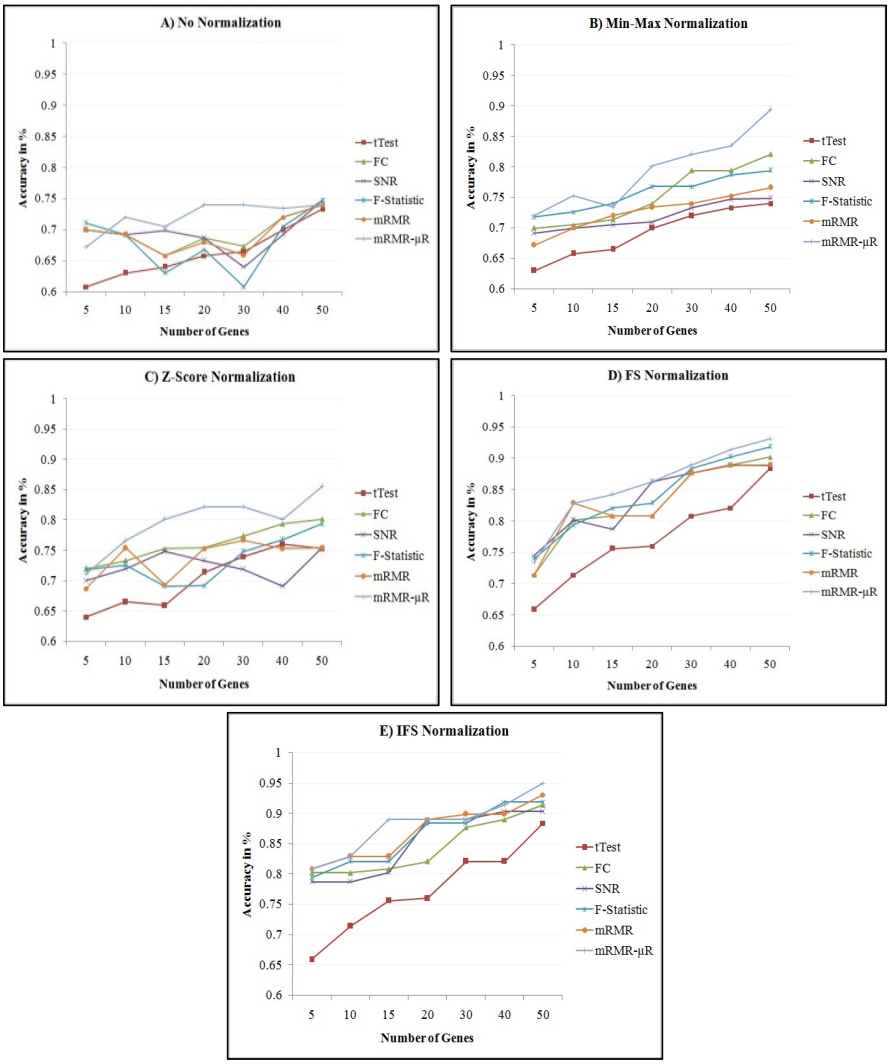

Figure 3. kNN Classification

with FS and IFS normalization outperformed for all feature selection methods.

Receiver Operating Characteristic (ROC) curve displays the relationship between the proportion of true positive (Sensitivity) and false positive (1-Specificity) binary classifications. The curve closer to left-hand border and then the top border of the ROC space gives more accuracy. The top-50 genes obtained from each feature selection method are used to plot ROC. Figures 4 and 5 shows the ROC curve plotted for top-50 genes obtained by SVM and kNN.

Area Under Curve (AUC) is a combined measure of sensitivity and specificity. It measures the overall performance of a classifier and is interpreted as the average value of sensitivity for all possible values of specificity. It takes any value between 0 and 1. If the maximum AUC is 1 then it is perfectly accurate and if it is 0 then it incorrectly classifies all samples. Table 1 shows the AUC for the ROC obtained from the figures 4 and 5.

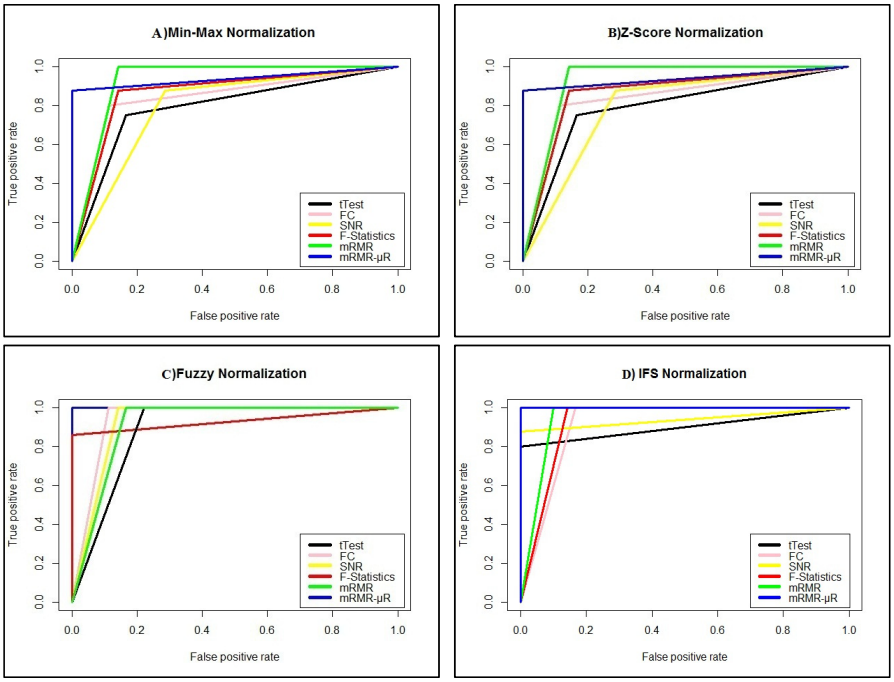

Figure 4. ROC Curve for SVM Classification

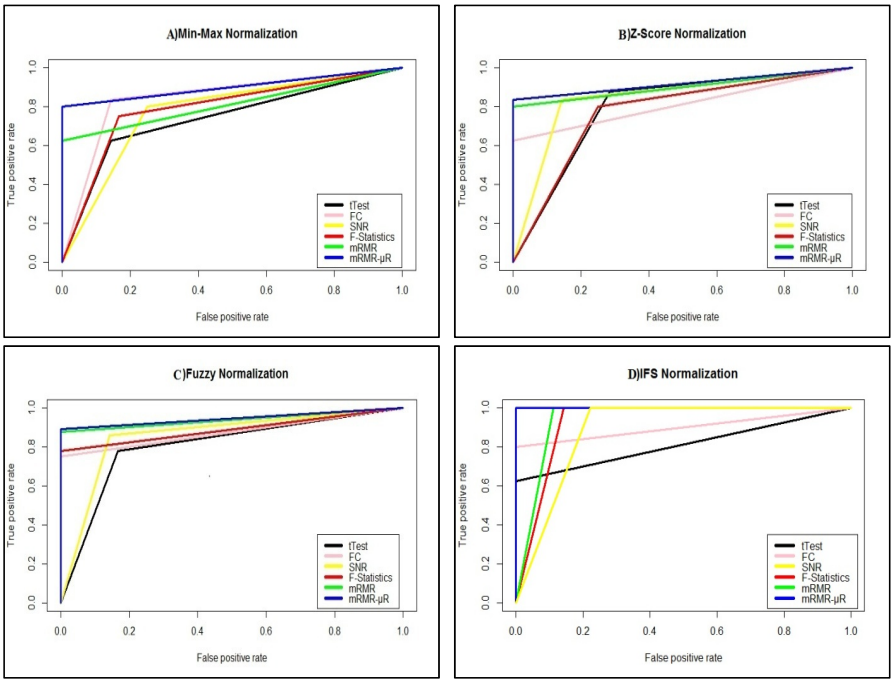

Figure 5. ROC Curve for kNN Classification

Table 1. Area Under ROC Curve

Feature Selection Methods	Normalization	Classifier	
		SVM	**kNN**
tTest	Min-Max	0.792	0.74
	Z-Score	0.792	0.753
	FS	0.889	0.805
	IFS	0.9	0.813
FC	Min-Max	0.844	0.821
	Z-Score	0.844	0.802
	FS	0.944	0.875
	IFS	0.917	0.9
SNR	Min-Max	0.795	0.749
	Z-Score	0.795	0.754
	FS	0.929	0.857
	IFS	0.937	0.889
F-Statistic	Min-Max	0.866	0.794
	Z-Score	0.866	0.794
	FS	0.926	0.889
	IFS	0.929	0.929
mRMR	Min-Max	0.929	0.766
	Z-Score	0.929	0.754
	FS	0.917	0.938
	IFS	0.95	0.944
mRMR-µR	Min-Max	0.938	0.894
	Z-Score	0.938	0.856
	FS	1	0.944
	IFS	1	1

From the results, it was observed that the proposed feature selection method outperforms the state-of-the art feature selection methods with minimum number of genes using both FS and IFS normalized dataset. The top-50 genes selected by mRMR-µR were analyzed for the significance of genes related to Autism. Table 2 shows the ten candidate genes related to Autism obtained from mRMR-µR gene selection method using IFS normalization.

Table 2. Candidate Genes Related to Autism

Probe Id	Gene Name	Gene Related to
209298_s_at	*ITSN1*	Alzheimer Disease
225540_at	*MAP2*	Alzheimer Disease
222419_x_at	*UBE2H*	Autistic Disorder
202794_at	*INPP1*	Autistic Disorder
210383_at	*SCN1A*	Autistic Disorder
238333_s_at	*PAOX*	Alzheimer Disease
238835_at	*AVPR1A*	Bipolar Disorder
207210_at	*GABRA3*	Bipolar Disorder
214157_at	*GNAS*	Autistic Disorder
204261_s_at	*PSEN2*	Bipolar Disorder

A heatmap is a two-dimensional representation of data in which values are represented by colors. Heatmaps originated in 2D displays of the values in a data matrix. Larger values were represented by small dark squares (pixels) and smaller values by lighter squares. Each row has the expression levels of one selected feature, and each column is a sample. Figure 6 shows heatmap depicting the predictive performance of top-50 ranked features selected by proposed method. There is a visible border between the 69 observations of the Autism group and the remaining 77 representing the control.

Conclusion

This paper provided the information on performance of different pre-processing techniques and different feature selection algorithms to identify biomarker genes of Autism. Two novel fuzzy normalization techniques were used to normalize the data. The novel mRMR-µR algorithm was proposed to identify the subset of significant genes from Autism dataset. To analyze the effect of FS and IFS normalization, they were compared with Min-Max and Z-Score normalization techniques. Based on mean score ranking, mRMR-µR identified subset of genes which were

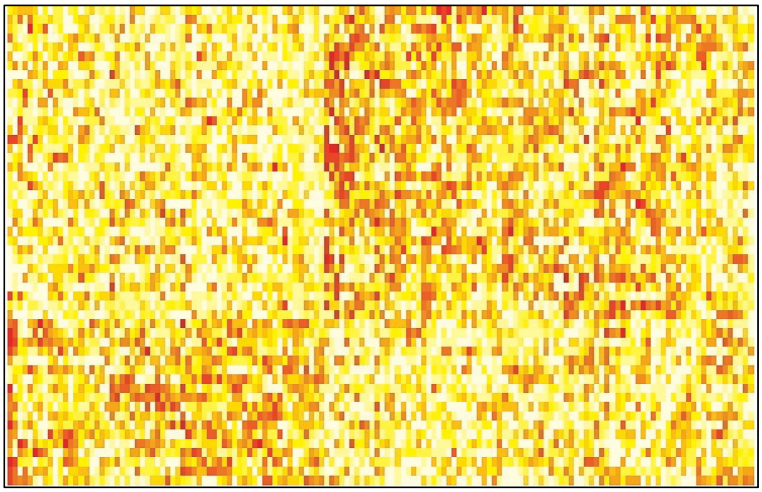

Figure 6. Heatmap for the Top-50 Genes

strongly relevant to the class and also the genes were non redundant. The ten significant genes related to Autism were identified by top-50 genes from mRMR-μR using IFS normalized data. To analyze the performance of the proposed work, kNN and SVM classifiers were used for Autism dataset. The experimental results were compared with the state-of-the art feature selection methods tTest, SNR, FC, F-statistic and mRMR. The experimental results demonstrated that the classification accuracy was significantly improved withmRMR-μR feature selection.

References

1. E.D. RINALDIS, DNA Microarrays: Current Applications, Norfolk, 2007.
2. J.M. PENA, J.A. LOZANO, P. LARRANAGA AND I. INZA, Dimensionality reduction in unsupervised learning of conditional Gaussian networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23 No. 6, pp. 590-603, (2001).
3. I. GUYON AND A. ELISSEEFF, An introduction to variable and feature selection, The Journal of Machine Learning Research, Vol. 3, pp. 1157-1182, (2003).
4. J.T. HORNG, L.C. WU, B.J. LIU, J.L. KUO, W.H. KUO AND J.J. ZHANG, An expert system to classify microarray gene expression data using gene selection by decision tree, Expert Systems with Applications, Vol. 36,- No. 5, pp. 9072-9081, (2009).
5. A. AZOFRA, J.L. AZNARTE AND J.M. BENÍTEZ, Empirical study of feature selection methods based on individual feature evaluation for classification problems, Expert Systems with Applications, Vol. 38 No. 7, pp. 8170-8177, (2011).
6. B. CHANDRA AND G. MANISH, An efficient statistical feature selection approach for classification of gene expression data, Journal of Biomedical Informatics, Vol. 44 pp. 529-535, (2011).
7. S. OSOWSKI, Methods and tools in data mining, BTC, Warsaw, (2013).
8. B. SAHU AND D. MISHRA, Performance of Feed Forward Neural Network for a Novel Feature Selection Approach, International Journal of Computer Science and Information Technologies, Vol. 2 No. 4, pp. 1414-1419, (2011).
9. C. DING AND H.C. PENG, Minimum Redundancy Feature Selection from Microarray Gene Expression Data, Proceedings of the Second IEEE Computational Systems Bioinformatics Conference, pp. 523-528, (2003).
10. R. PREMA, AND K. PREMALATHA, Effect of intuitionistic fuzzy normalization in microarray gene selection, Turkish Journal of Electrical Engineering and Computer Sciences, Vol. 26, No. 3, 2018.
11. LA. ZADEH, Fuzzy sets, Information and Control, Vol. 8, pp. 338-353, (1965).
12. K.T. ATANASSOV, Intuitionistic fuzzy sets:theory and application, Springer, Vol. 35, (1999).
13. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25507
14. http://www.r-project.org/