Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2018) 26: 1141 – 1152
© TÜBİTAK
doi:10.3906/elk-1708-105

Effect of intuitionistic fuzzy normalization in microarray gene selection

Prema RAMASAMY[1,*], **Premalatha KANDHASAMY**[2]
[1]Anna University, Chennai, India
[2]Department of Computer Science & Engineering, Faculty of Information & Communication Technology,
Bannari Amman Institute of Technology, Erode, India

Received: 09.08.2017 • **Accepted/Published Online:** 03.04.2018 • **Final Version:** 30.05.2018

Abstract: Analysis of gene expression data is essential in microarray gene expression 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 improves the quality of results. This paper presents two fuzzy preprocessing techniques, using a fuzzy set (FS) and intuitionistic fuzzy set (IFS), to normalize datasets. In the feature selection part, four statistical methods were used. Using three publicly available gene expression datasets, the fuzzy normalization techniques were compared with two standard normalization techniques (min-max and Z-score) as well as raw gene expression. The classifiers of support vector machine, k-nearest-neighbor, and random forest were used to identify the accuracy of selected features. The experimental results show that the genes selected using FS- and IFS-normalized datasets give high classification accuracy; in addition, IFS outperforms FS normalization.

Key words: Gene expression data, feature selection, classification, intuitionistic fuzzy normalization

1. Introduction

1.1. Gene expression data

A microarray dataset is a repository containing 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 characterize the expression level of the particular gene in the particular sample [1].

1.2. Data normalization and feature selection

One of the most essential stages of preprocessing 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 on a different scale and has a different range of possible values. Microarray datasets contain continuous gene expression values. Therefore, an effective normalization technique should be applied to preprocess the expression data [2]. In this paper, two fuzzy normalization techniques were tested. These results were compared with two popular normalization techniques, min-max and Z-score, abbreviated as MM and ZS, respectively.

After the preprocessing of the data, a feature selection approach is used to select the most significant

[*] Correspondence: premabit@gmail.com

features. There are many feature selection approaches to assist in classification of samples [3–7]. They are classified into four categories, namely as filter approach, wrapper approach, embedded approach, and hybrid approach. A filter approach applies a statistical measure to assign a score to each feature without using a learning algorithm [8]. A wrapper approach uses learning techniques to evaluate the accuracy produced by the use of the selected features in the classification [9]. An embedded approach combines the feature selection step and classifier construction. A hybrid approach is a combination of both filter and wrapper-based methods [10].

A gene expression dataset contains thousands of gene expression values, many of which may be redundant or irrelevant for classification [11]. Leaving out relevant attributes or keeping irrelevant attributes may affect the performance of the classification algorithm. Therefore, statistical methods are required to find the most important genes before classification. In this paper, four different filter selection methods were used for gene selection.

2. FS and IFS normalization

Fuzzy sets were introduced by Lotfi Zadeh in 1965 as an extension of the classical notion of sets [12]. Fuzzy set theory can be used in a wide range of domains in which information is incomplete or imprecise, such as bioinformatics.

A fuzzy set A of a nonempty set X is defined as a set of ordered pairs, $\langle x, \mu_A(x)\rangle$, where $x \in X$ and $\mu_A\left(x\right)$ is the membership function of the fuzzy set A. A membership function is a curve that defines how each point in the input space (X) is mapped to a membership value between 0 and 1. A fuzzy set is a collection of objects with graded membership, i.e. having degrees of membership [12].

In this study, datasets were transformed by exploitation of a fuzzy membership function rather than by using their absolute expression values. A fuzzy membership function that is used to represent vague, linguistic terms is the Gaussian function, which is given in Eq. (1):

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

where m and k are the center 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 nonempty set, all gene values with respect to all the samples were fuzzified with three fuzzy qualifiers, low, medium, and high. The maximum and minimum values of each gene were used to define these three fuzzy sets. The center and width of each fuzzy set was calculated. Then the Gaussian membership function (Eq. (1)) was applied to all the genes in each fuzzy set. The raw gene values were replaced with these FS-normalized values. Thus, each gene value was 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.

Fuzzification determines the degree of membership. The term “intuitionistic fuzzification” refers to formulating the membership and nonmembership functions of IFS. In practice, due to the insufficiency of the information available, the evaluation of membership and nonmembership values is not always possible. Therefore, an indeterministic part remains, known as hesitation [13].

Let A be an IFS of nonempty set X defined as $\{< x, \mu_A\left(x\right), \gamma_A(x) > | \, x \epsilon X\}$ where $\mu_A(x) : X \to [0,1]$ and$\gamma_A(x) : X \to [0,1]$ such that $0 \leq \mu_A(x) + \gamma_A(x) \leq 1$ and $\mu_A(x)$ and $\gamma_A(x)$ denote the degree of membership and nonmembership, respectively.

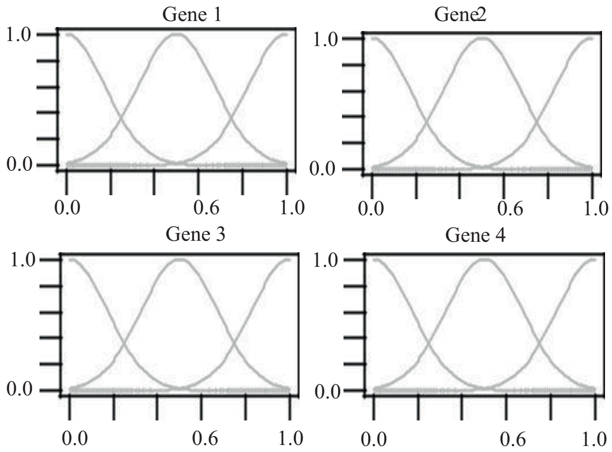


Figure 1. Membership functions of four random genes.

For each intuitionistic fuzzy set in X, there exists an indeterministic part (or hesitation margin) $\pi_A(x)$. Then the degree of nonmembership can be given as:

$$\gamma_A(x) = 1 - \mu_A(x) - \pi_A(x). \tag{2}$$

Let $D = [d_{ij}]_{M\times N}$ be the IFS matrix, where $d_{ij} = \{\mu_{ij}, \gamma_{ij}, \pi_{ij}\}$. The following matrix shows the representation of IFS gene expression data:

$$D = \left| \begin{array}{llll} \mu_1(g_1), \gamma_1(g_1), \pi_1(g_1) & \mu_1(g_2), \gamma_1(g_2), \pi_1(g_2) & \cdots & \mu_1(g_N), \gamma_1(g_N), \pi_1(g_N) \\ \mu_2(g_1), \gamma_2(g_1), \pi_2(g_1) & \mu_2(g_2), \gamma_2(g_2), \pi_2(g_2) & \cdots & \mu_2(g_N), \gamma_2(g_N), \pi_2(g_N) \\ \vdots & \vdots & \cdots & \vdots \\ \mu_M(g_1), \gamma_M(g_1), \pi_M(g_1) & \mu_M(g_2), \gamma_M(g_2), \pi_M(g_2) & \cdots & \mu_M(g_N), \gamma_M(g_N), \pi_M(g_N) \end{array} \right| .$$

Using these triplets, the IFS-normalized value can be described by using the following implications:

$d_{ij} = \mu_{ij}$, if $\mu_{ij} \geq \gamma_{ij}$,

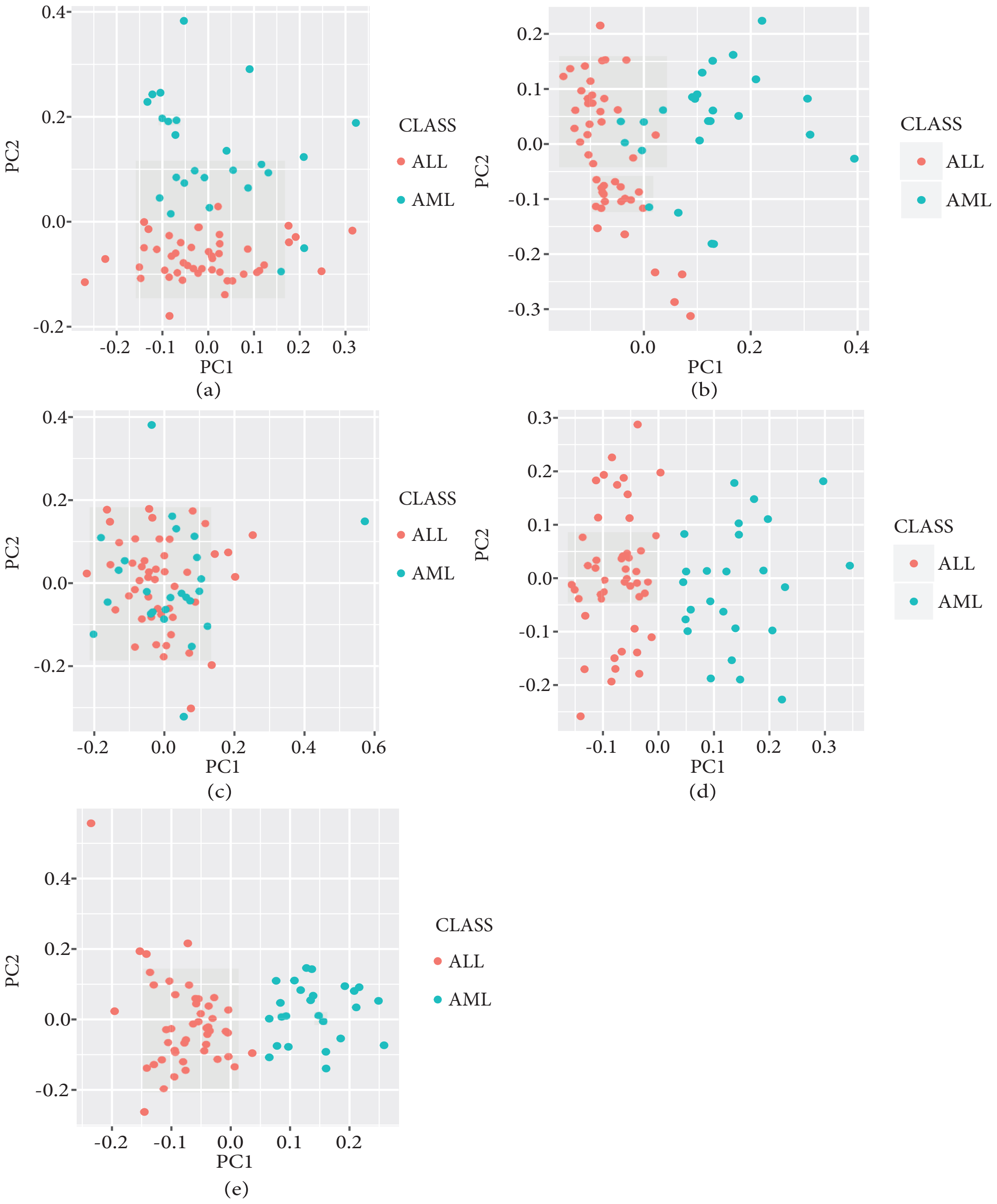
$d_{ij} = -\gamma_{ij}$, if $\mu_{ij} < \gamma_{ij}$.

The raw values were replaced with these IFS-normalized values. These FS- and IFS-normalized datasets were given as the input for feature selection.

3. Results

The normalization techniques were examined with three different gene expression datasets: Leukemia, Colon, and DLBCL. The Leukemia dataset contains 72 samples, with 47 acute lymphoblastic leukemia (ALL) samples and 25 acute myeloid leukemia (AML) samples [14]. The Colon dataset contains 62 samples in two classes. Among them, 40 samples are tumor and 22 normal [15]. There are 77 samples in the DLBCL dataset, among which 58 samples belong to diffuse large B-cell lymphoma (DLBCL) and 19 to follicular lymphoma (FL) [16]. The statistical analysis was performed with R packages.

The raw gene expression datasets were preprocessed with each of the four methods: MM, ZS, FS, and IFS. The top 15% of genes were selected with maximum variance in each method and principal component analysis (PCA) transformation was applied to them. A scatter plot of the coordinates corresponding to the first two principal components (PC1 and PC2) of each sample was visualized. A good preprocessing method is expected to show a clear clustering of samples of the same class and separation between samples of different classes. Figures 2, 3, and 4 show the PCA scatter plots of the Leukemia, Colon, and DLBCL datasets, respectively, using all normalization methods. From these figures, it can be observed that samples from different classes were clearly separated using IFS normalization. It showed the best clustering of samples among all preprocessing techniques. MM did not perform well on any of the datasets. ZS normalization showed good performance only on the Leukemia dataset, and both FS and IFS normalization performed well on all the datasets.



(a)

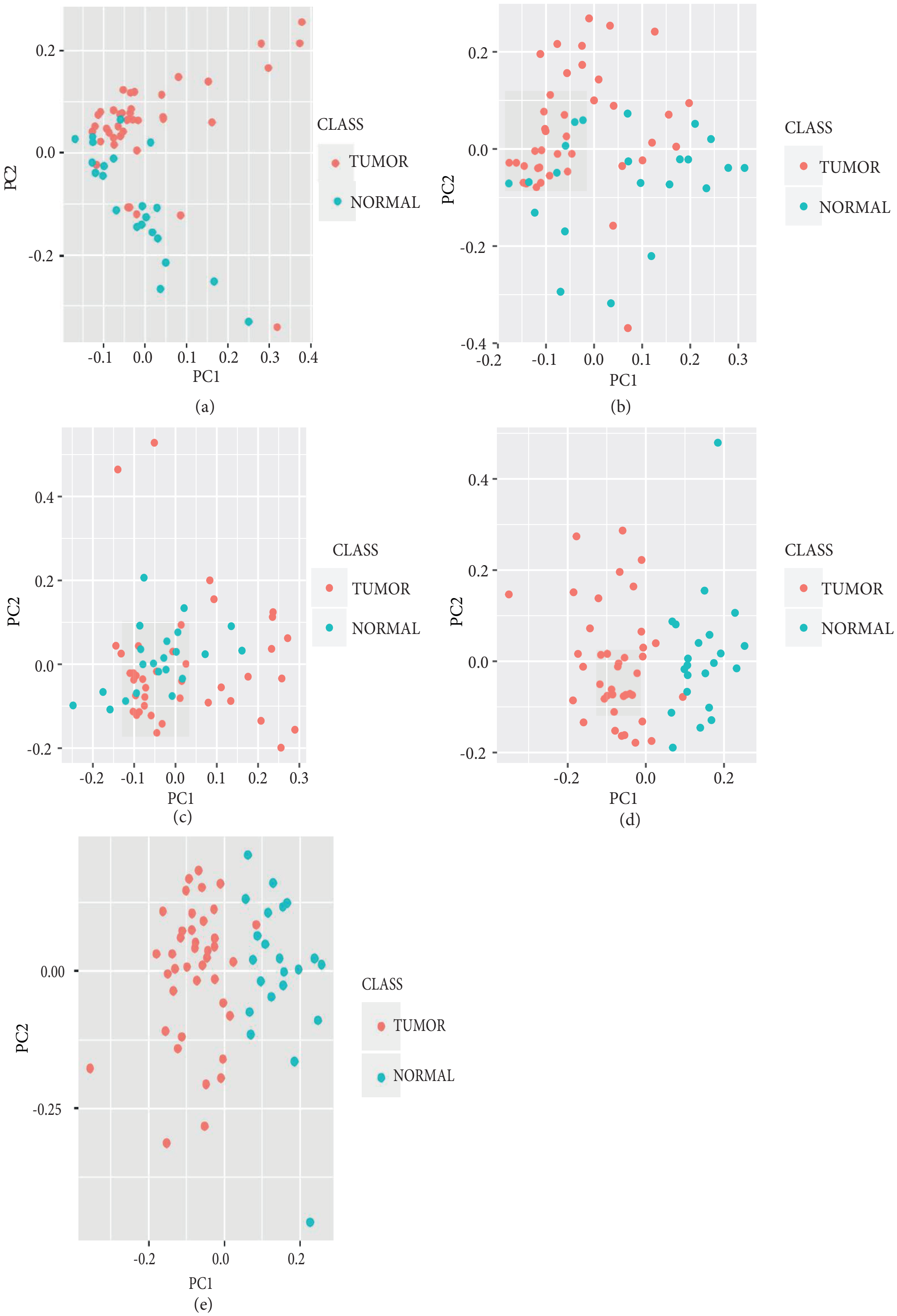
(b)

(c)

(d)

(e)

Figure 2. PCA scatter plots for Leukemia data: (a) NN; (b) MM; (c) ZS; (d) FS; (e) IFS.



(a)

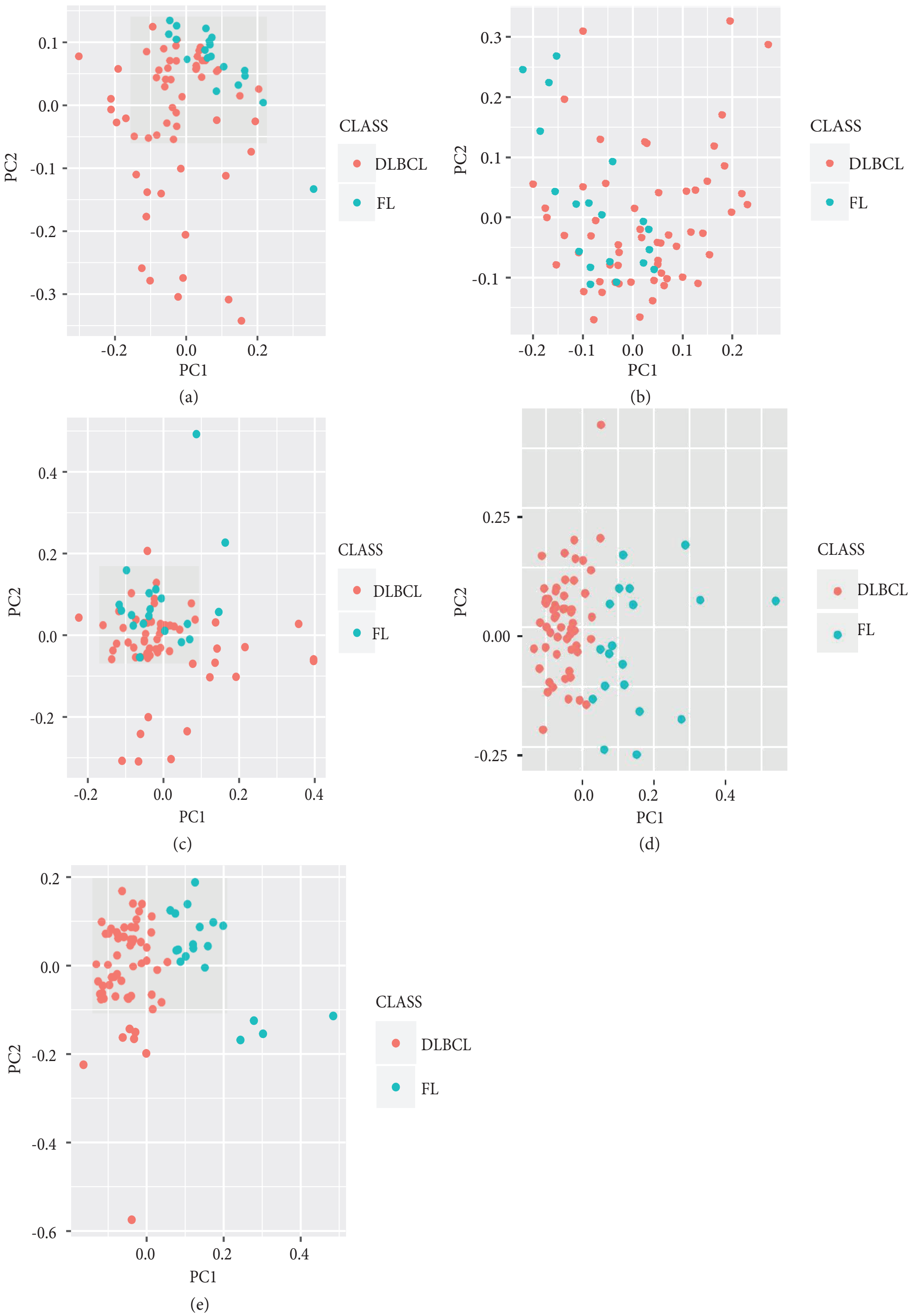
(b)

(c)

(d)

(e)

Figure 3. PCA scatter plots for Colon data: (a) NN; (b) MM; (c) ZS; (d) FS; (e) IFS.



(a)

(b)

(c)

(d)

(e)

Figure 4. PCA scatter plots for DLBCL data: (a) NN; (b) MM; (c) ZS; (d) FS; (e) IFS.

To demonstrate the performance of the fuzzy normalization techniques, the top 25 and 50 genes were selected using the T-statistic [17], SNR [18], F-statistic, and mRMR [19] feature selection methods from the Leukemia, Colon, and DLBCL datasets. To determine the accuracy of these features, the well-known classifiers of support vector machine (SVM), k-nearest neighbor (kNN), and random forest (RF) were 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 instances considered for determination of similarity with classes was three for kNN. In RF, the number of trees used was 500. A large number of trees are used because RF does not overfit when the number of trees is increased. Figures 5a–5f, 6a–6f, and 7a–7f show the classification accuracy of the top 25 and 50 for the Leukemia, Colon, and DLBCL datasets, respectively.

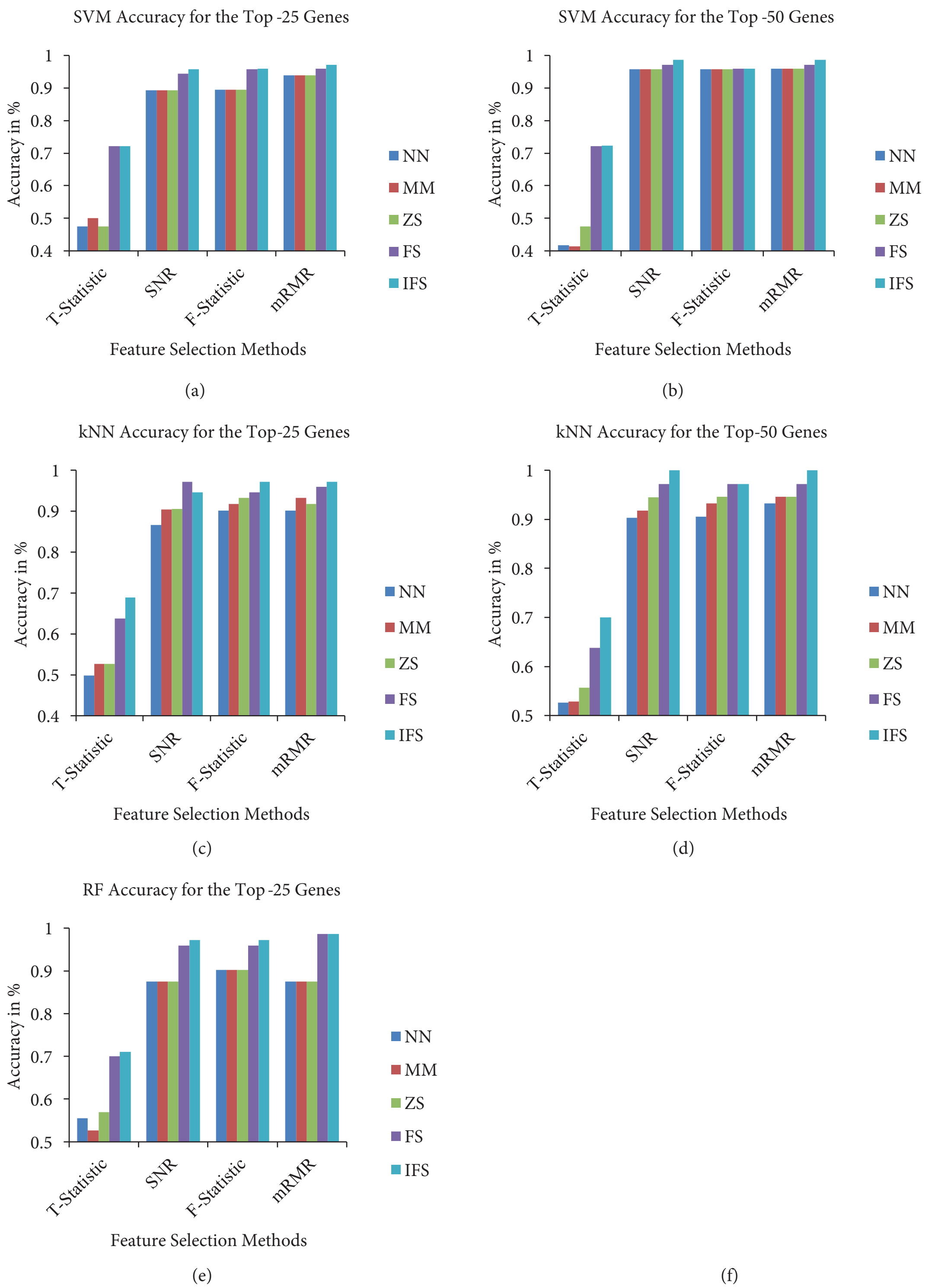
From Figures 5a and 5b, it can be observed that SVM was not more affected by MM or ZS normalization. Using the top 25 genes, MM and ZS normalization produced the same accuracy of 89.29%, 89.46%, and 93.81% as NN for SNR, F-statistic, and mRMR respectively. However, both FN and IFS outperformed MM and ZS for all the feature selection methods. mRMR with IF normalization had the best accuracy of 97.14% for the SVM classifier. The kNN classifier had different accuracies for different normalization techniques, as shown in Figures 5c and 5d. This classifier depends on distance calculations and is affected by normalization since after normalization all the dimensions have the same weight and no one dominates the others. It shows statistical improvement after using a normalization technique. It can be inferred that ZS normalization performed better than MM normalization for the kNN classifier. Both fuzzy normalization methods outperformed the other two techniques. With the top 25 genes, mRMR and F-statistic with IF normalization had the best accuracy of 97.14% for the kNN classifier. In Figures 5e and 5f, MM and ZS normalization show no difference with the RF classifier. However, the accuracy performance was improved after applying normalization. The RF classifier had the highest accuracy of 98.57% in the Leukemia dataset using mRMR with IF normalization.

For the Colon dataset, SVM was affected by all the normalization methods, as shown in Figures 6a and 6b, and the IFS normalization method outperformed all other methods. It can also be observed that MM reported better accuracy than ZS for both SVM and kNN. Figure 6e shows that RF was the least affected with the top 25 genes. With the top 25 genes, kNN and RF had the maximum accuracy of 93.81% and 93.57%, respectively, using mRMR with IFN. SVM achieved the highest accuracy of 96.91% with SNR and mRMR in the Colon dataset.

Like the Leukemia dataset, DLBCL was not affected by normalization for both SVM and RF. For SVM, MM and ZS normalization produced the same accuracy of 71.79%, 83.04%, 81.96%, and 83.21% as NN for T-statistic, SNR, F-statistic, and mRMR, respectively. It was also found that ZS performed better than MM for kNN classification (Figures 7a–7d). Classifiers kNN and RF had the maximum accuracy of 93.57% using F-statistic and mRMR. For the DLBCL dataset, SVM showed the highest accuracy of 97.51% with F-statistic using IF normalized data.

As a general conclusion, from Figures 5, 6, and 7, it can be inferred that the mRMR feature selection method gave higher accuracy than the other methods. In most cases, SVM performed better than the other classification methods due to its suitability for high-dimensional data. Both FS and IFS produced higher accuracy compared to the other two normalization techniques, and IFS gave significant improvement over the FS normalization method. Thus, the fuzzy normalization methods improved the quality of the feature selection methods. It can also be observed that the datasets with FS and IFS normalization showed the best performance for all feature selection methods.

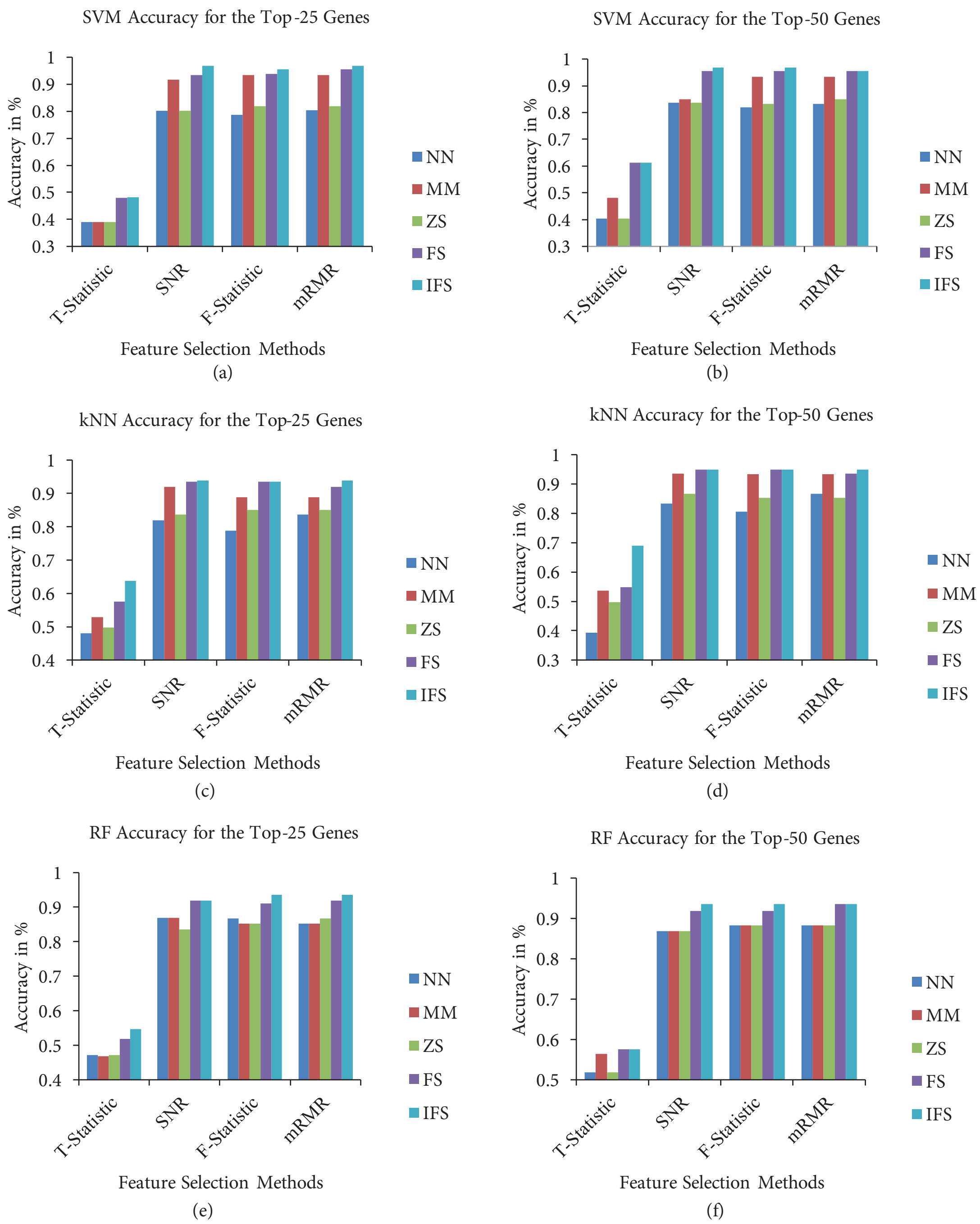
A heat-map is a two-dimensional representation of data in which values are represented by colors. Heat-



(a) (b) (c) (d) (e) (f)

Figure 5. Classification for Leukemia dataset: a, b) SVM; c, d) kNN; e, f) RF.

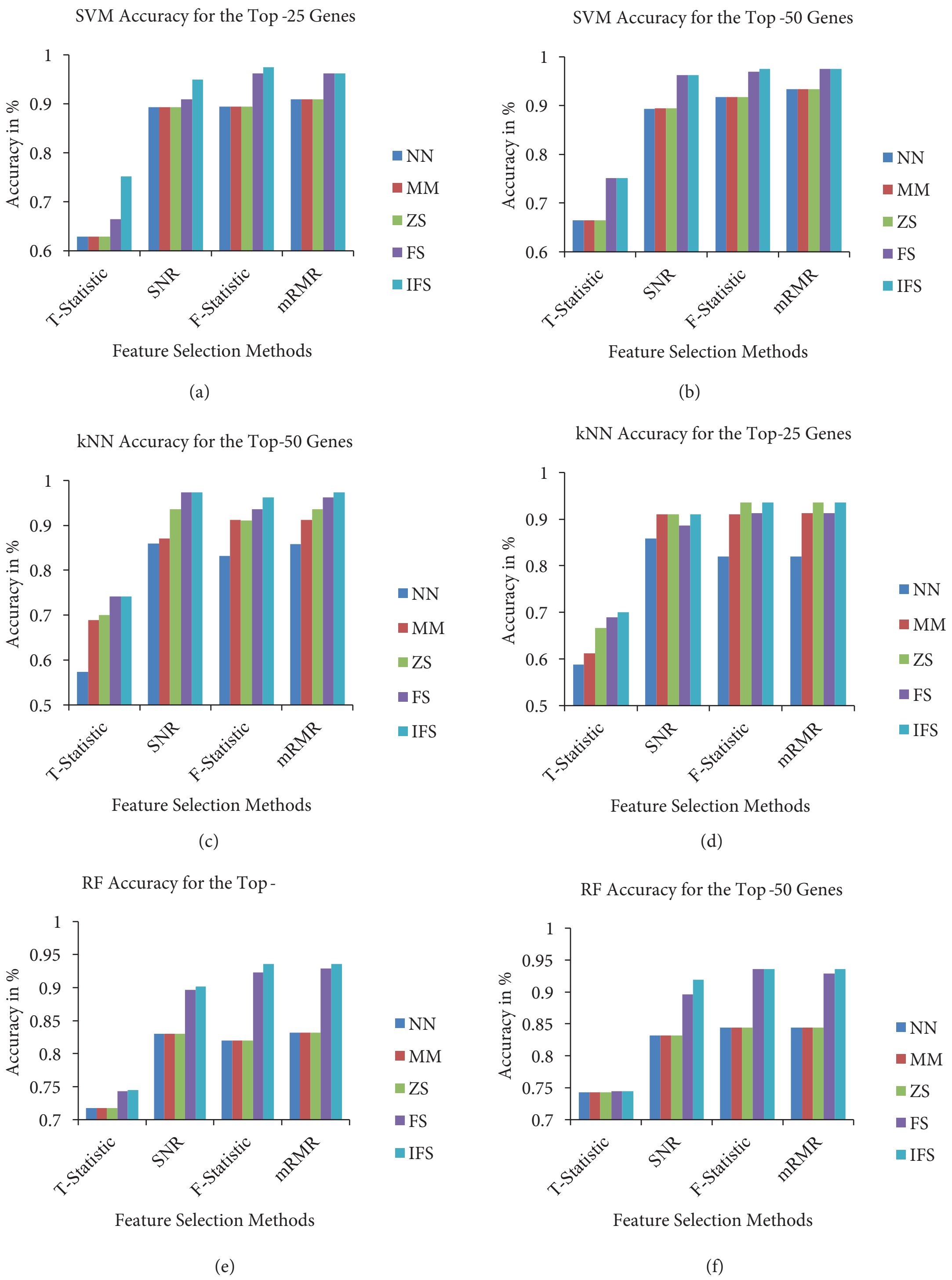
maps originate from 2D displays of the values in a data matrix. Larger values are represented by small dark squares (pixels) and smaller values by lighter squares. Each row shows the expression levels of one selected



(a) (b) (c) (d) (e) (f)

Figure 6. Classification for Colon dataset: a, b) SVM; c, d) kNN; e, f) RF.

feature, and each column is a sample. Figure 8 shows the heat-maps depicting the predictive performance of the top 50 ranked features selected by mRMR using IFS normalization for the datasets Leukemia, Colon, and DLBCL. From Figure 8a, it can be observed that there is a visible border between the 47 observations of the ALL group and the remaining 25, representing the AML samples. Figure 8b depicts a cut between two classes (tumors, healthy). The good performance of the selected features for the DLBCL dataset is also shown in Figure 8c.

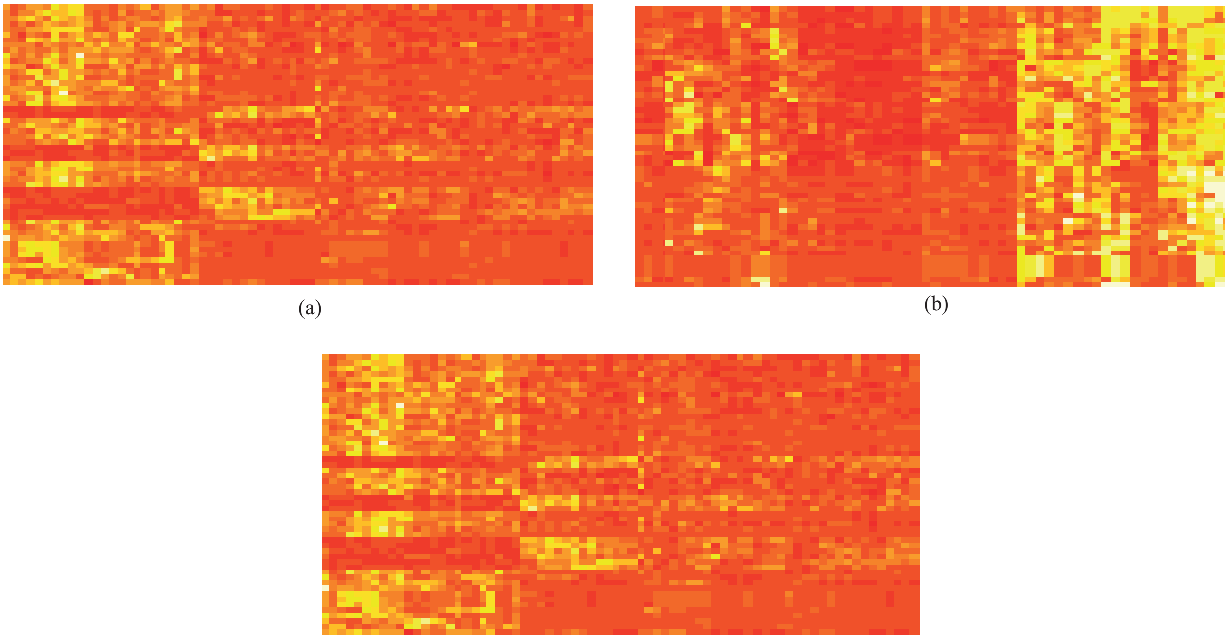


(a) (b) (c) (d) (e) (f)

Figure 7. Classification for DLBCL dataset: a, b) SVM; c, d) kNN; e, f) RF.

4. Conclusion

This paper provides information on the performance of different preprocessing techniques for microarray datasets. Two novel fuzzy normalization techniques were used to normalize three datasets and compared with



(a)

(b)

(c)

Figure 8. Heat-maps: a) Leukemia dataset; b) Colon dataset; c) DLBCL dataset.

two other popular preprocessing methods with respect to two important criteria. First, PCA transformation was applied on the datasets and visualized with scatter plots. The plots showed that samples from different classes were clearly separated using IFS normalization. Secondly, the state-of-the-art feature selection methods T-statistics, SNR, F-statistic, and mRMR were used to select the top 50 genes for all the normalized datasets as well as the raw datasets. To analyze the performance of the selected genes, SVM, kNN, and RF classifiers were used. The experimental results demonstrate that the classification accuracy was improved with the genes selected with IFS-normalized datasets. The heat-maps were also visualized. The results illustrate that FS and IFS normalization techniques can be used to increase the quality of gene selection. In addition, it was demonstrated that IFS can yield significant improvement compared to the FS normalization method.

References

[1] Rinaldis ED, Lahm A. DNA Microarrays: Current Applications. Norfolk, UK: Horizon Bioscience, 2007.

[2] Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY et al. The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 2006; 24: 1151-1161.

[3] Bhattacharyya DK, Kalita JK. Network Anomaly Detection: A Machine Learning Perspective. 1st ed. Boca Raton, FL, USA: CRC Press, 2013.

[4] Hoque N, Bhattacharyya DK, Kalita JK. MIFS-ND: A mutual information-based feature selection method. Expert Syst Appl 2014; 41: 6371-6385.

[5] Min N, Hu Q, Zhu W. Feature selection with test cost constraint. Int J Approx Reason 2014; 55: 167-179.

[6] Tabakhi S, Moradi P, Akhlaghian F. An unsupervised feature selection algorithm based on ant colony optimization. Eng Appl Artif Intel 2014; 32: 112-123.

[7] Jenson R, Shen Q. New approaches to fuzzy-rough feature selection. IEEE T Fuzzy Syst 2009; 17: 824-838.

[8] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003; 3: 1157-1182.

[9] Blum AL, Langley P. Selection of relevant features and examples in machine learning. Artif Intell 1997; 97: 245-271.

[10] Hsu HH, Hsieh CW, Lu MD. Hybrid feature selection by combining filters and wrappers. Expert Syst Appl 2011; 38: 8144-8150.

[11] Horng JT, Wu LC, Liu BJ, Kuo JL, Kuo WH, Zhang JJ. An expert system to classify microarray gene expression data using gene selection by decision tree. Expert Syst Appl 2009; 36: 9072-9081.

[12] Zadeh LA. Fuzzy sets. Inform Control 1965; 8: 338-353.

[13] Atanassov KT. Intuitionistic Fuzzy Sets: Theory and Application. Heidelberg, Germany: Physica Verlag, 1999.

[14] Golub TR,Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999; 286: 531-537.

[15] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. P Natl Acad Sci USA 1999; 96: 6745-6750.

[16] Yang K, Cai Z, Li J and Lin G. A stable gene selection in microarray data analysis. BMC Bioinformatics 2006; 7: 228-243.

[17] Chandra B, Manish G. An efficient statistical feature selection approach for classification of gene expression data. J Biomed Inform 2011; 44: 529-535.

[18] Sahu B, Mishra D. Feature selection for cancer classification signal-to-noise ratio approach. Int J Sci Eng Res 2011; 2: 1-7.

[19] Ding C, Peng HC. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol 2005; 3: 185-205.