Research Journal of Applied Sciences, Engineering and Technology15(5): 190-196, 2018
DOI:10.19026/rjaset.15.5851
ISSN:2040-7459; e-ISSN: 2040-7467

Submitted:February 5, 2018 Accepted:March 2, 2018 Published:May 15, 2018

Research Article

Self-Organizing Maps and Principal Component Analysis to Improve Classification Accuracy

Hicham Omara, Mohamed Lazaar and Youness Tabii
Lirosa Laboratory, Faculty of Science, Abdelmalek Essaadi University, Tetouan, Morocco

Abstract: The aim of this study is to perform the Kohonen Self-Organizing Map (SOM) using Principal Component Analysis (PCA). SOM is an algorithm commonly used to visualize and classify datasets, due to its ability to project large data into a smaller dimension. However, their performance decreases when the size of the problem becomes too big. Therefore, reducing the size of the data by removing irrelevant or redundant variables and selecting only the most significant ones according to certain criteria has become a requirement before any classification, this reduction should give the best performance according to a certain objective function. Many researchers have tried to solve this problem. This study presents a new approach to improve SOM based on PCA. The experimental analysis of real data from the UCI machine learning repository shows an improvement of the proposed SOM compared to a traditional approach. More than 2% of the improvement in the accuracy of the classification is observed.

Keywords: Classification, feature extraction, feature selection, principal component analysis, self-organizing maps

INTRODUCTION

In recent years, the data is exponentially expanded, so their characteristics, consequently, reducing the size of the data by removing irrelevant or redundant variables and selecting only the most significant according to some criterion has become a requirement before any classification, this reducing should give the best performance according to some objective function (Devaraj *et al*., 2002; Dudoit *et al*., 2002; Narayanan *et al*., 2004). In general, the performance of a classifier decreases when the dimensionality of the problem becomes too large.

Several approaches are used in classification, to name a few, Hopfield network, K-means, Support Vector machine; most of them are inspired by biological neural networks. Among these, Kohonen Self-Organizing Maps (SOM) are popularly and widely used for the classification. SOM is one type of the neural networks commonly used for visualizing and classifying of multidimensional data. It is applied in various areas: medicine, financial, ecological, engineering, law enforcement and other fields (Ettaouil*et al*., 2013, 2012; Kohonen, 1998; Pavel and Olga, 2011). However, certain topological constraints of the SOM are fixed before the training phase; the dimension of neurons has a great effect on the classification performance that we had to discuss in this study. The interesting question is which features should be used. Given a set of d features; how do we select an optimal subset of l features such that? Consequently, the execution time for classification the data decreases and the accuracy increases (Arauzo-Azofra *et al*.,2011).

One approach to solve this problem is to use feature selection that consists of choosing a subset of input variables and deleting redundant or irrelevant entities from the original dataset. It is divided into three categories; filters, wrappers and embedded or hybrid selectors (Blum and Langley, 1997; Ding and Peng, 2005). The filters extract features from the data without any learning involved by ranking all features and chosen top ones (Guyon and Elisseeff, 2003; Ruiz *et al*., 2012). There were several and widely used filters in literature, such as Information Gain (IG) (Wang *et al*., 2006), Minimum Redundancy Maximum Relevance (mRMR) (Ding and Peng, 2005), Relief F (Kira and Rendell, 1992). The wrappers use classifying algorithm to evaluate which features are useful; it means that the features were selected taking the classification algorithm into account (Gheyas and Smith, 2010; Kohavi and John, 1997). The third field of feature selection approaches is embedded methods. It takes advantage of the two models by using their different evaluation criteria in different search stages (Guyon and Elisseeff, 2003; Maldonado *et al*., 2011; Mundra and Rajapakse, 2010).

Corresponding Author: Hicham Omara, Lirosa Laboratory, Faculty of Science, Abdelmalek Essaadi University, Tetouan, Morocco, Tel.: 00212660730436

The second approach used which called feature extraction that replaces the set of n features by a set of m features; each one is a combination of the original feature. A well-known dimensionality reduction technique is Principal Component Analysis (Abdi and Williams, 2010). PCA tries to find a linear subspace of lower dimensionality, such that the largest variance of the original data is kept. However, note that the largest variance of the data does not necessarily represent the most discriminative information (Jolliffe, 1972).

This research opts for the classification of real-world data from the UCI Machine Learning Repository using SOM and PCA. Accuracy rate is used to evaluate this algorithm. The aim of our study is to reduce the number of features and demonstrate the importance of feature selection to improve classification. The experimental analysis shows the speed up of the proposed SOM training process in comparison to a classical approach.

PROPOSED MODEL

The SOM-PCA proposed is divided into two main steps. In the first, the network was trained by the classical SOM. The neurons resulted from the training phase, were used as input for PCA; to transform them to a new set of vectors with the low dimension. So, the dataset will be reduced to a smaller number of dimensions with low information loss. Figure 1 shows a flowchart of this model.

Self-organizing maps: The SOM often consists of a regular grid of map units. Each unit is represented by a vector $W_j = (w_{j1}, w_{j2}, ..., w_{jd})$, where d is input vector dimension. The units are connected to adjacent ones by neighbourhood relation.

The SOM is trained iteratively. At each training step, a sample vector S_i is randomly chosen from the input data set, a metric distance is computed for all weight vectors n to find the reference vector W_{bmu} that satisfies a minimum distance or maximum similarity criterion following the Eq. (1). The neuron with the most similar weight vector to the input pattern is called the Best Matching Unit (BMU):

$$W_b(t) = \arg\min_{1 \le i \le n} \|S(t) - W_i(t)\| \quad (1)$$

where, n is the neurons number in the map in instant t. The weights of the BMU b and its neighbours are then adjusted towards the input pattern, following Eq. (2):

$$W_i(t+1) = W_i(t) + \beta_{b,i}(t)\|S - W_i\| \quad (2)$$

One of the main parameters influencing the training process is the neighbourhood function between

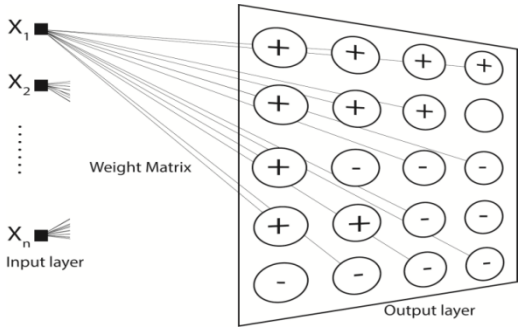

Fig. 1: SOM Representation (rectangular topology). The hidden layer is composed by neurons (+ represents benign tumour and-malignant one)

the winner neuron b and neighbour neuron i. This function is positive and symmetric defines a distance-weighted model for adjusting neuron vectors. It can be calculated using the Eq. (3):

$$\beta_{b,i}(t) = \exp\left(\frac{\|r_b - r_i\|}{2\sigma_i^2(t)}\right) \quad (3)$$

where, $\|r_b - r_i\| \cong \|W_b - W_i\|$ and are positions of the BMU neuron i on the Kohonen map. The function decreases monotonically with time. This function can introduce zones of influence around each winner neuron, the weightings of each neuron are changed, but the degree of change decreases with the distance on the map between the positions of the neuron to neuron winner and to make updated.

Feature selection using PCA: Principal Component Analysis (PCA) was a powerful statistical tool for reducing the dimensionality of multivariate data sets in many areas such as image analysis, data compression, time series prediction and analysis of biological data by finding a new set of variables (Abdi and Williams, 2010). The new set of variables, called Principal Components (PCs), is characterized by his dimension that is smaller than the original counterpart and is ordered by the fraction of the total information each retains. These PCs have been chosen so that the first principal component must have the greatest possible variance; the second component is computed under the constraint of being orthogonal to the first component and having the greatest possible inertia and so on.

In our study, we consider the use of PCA in extracting relevant features from the neurons vectors w_j; were j is the j^{th} weight vector from the n neurons resulted after the SOM training process; and that have a d features (dimension). Therefore, we have an array matrix W with the size of $n \times d$:

$$W = [w_1, w_2, ..., w_n]$$

These vectors are now subjected to principal components analysis. To transform them into a new set of the vector with derived dimensions ($l < d$), but in this case, their information content is ranked and stored in the first dimensions. So, the dataset will be reduced to a smaller number of dimensions with low information loss. The transformation is based on the matrix computation:

$$R = T.W \tag{4}$$

Under the constraints that $R^tR = T^tW^tWT$ is a diagonal matrix and that $T^tT = I$ is an identity matrix. Matrix R has the same dimension as W and related by a linear transformation T. R will have the properties that most of their information content is stored in the first dimensions and T should be chosen so that R represents the largest variance for the input data.

There are several ways of obtaining the solution of this problem. In this study, we try to construct T using covariance method. Before calculating the covariance matrix S we need to centering data in matrix W_c as follow:

$$W_c = W - h\bar{w}^t$$

where, h is an $n \times 1$ column vector of ones $h_i = 1$ for $i = 1, 2, \cdots, n$; and $\bar{w}_j$ is a vector of dimensions $p \times 1$ that contains the empirical mean along each column j = 1,..., p of W and defined as:

$$\bar{w}_j = \frac{1}{n}\sum_{i=1}^{n} w_{ij}$$

The covariance matrix S is now, defined by outer product of W_c with itself:

$$S_{n \times n} = \frac{1}{n-1} W_c^t \cdot W_c \tag{5}$$

The eigenvalues of S for the given data should be calculated. Those m eigenvectors corresponding to the m largest eigenvalues of S define a linear transformation from the n-dimensional space to an m-dimensional space in which the features are uncorrelated. An eigenvalue and eigenvector of a matrix $S_{n \times n}$ are a scalar λ and a nonzero vector e so that:

$$Se = \lambda e$$

Let $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_n\}$ provided that $\lambda_1 \geq ... \geq \lambda_n \geq 0$ be the set of eigenvalues of S and $V = \{e_1, e_2, ..., e_n\}$ with $\det(V) \neq 0$ their corresponding eigenvectors, called the principal axes. Then:

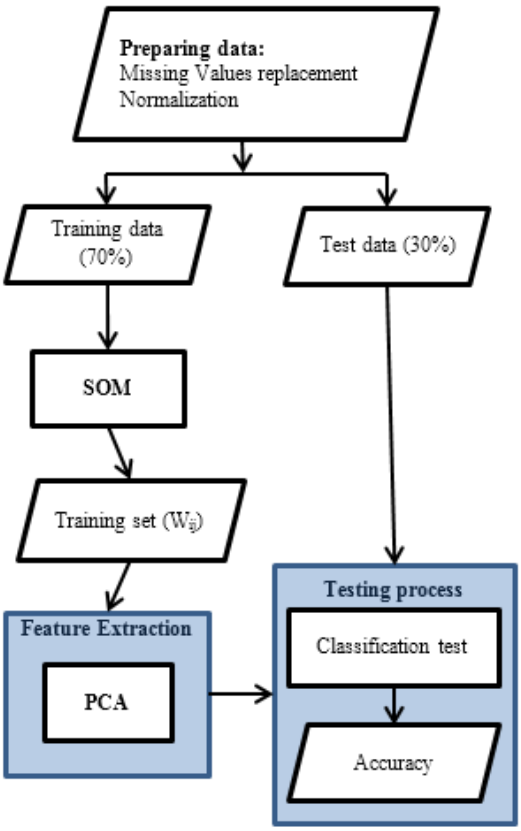

Fig. 2: Flowchart of the different issues discussed in this study

$$S = V \Lambda V^{-1}$$

The problem in using PCA as the dimensional reduction is to define the number of principal components needed to get a good representation of the data. Different methods exist for predicting this value (Abdi and Williams, 2010; Jolliffe, 1972; King and Jackson, 1999) including Kaiser's stopping rule (Kaiser, 1960) that retains and interprets any component where its eigenvalue greater than 1.00. Scree test (Cattell, 1966) which trace the eigenvalues in descending order of their magnitude in relation to their number of factors and determines where they stabilize (D'agostino and Russell, 2005). Percentage of variance explained (Jolliffe, 1972; Shaharudin and Ahmad, 2017); this technique retains components that account for at least of the total variance. Cumulative Percentage of Variance extracted retains components where certain percentages of the cumulative have been suggested;

In this study, the Cumulative Percentage of Variance explained method was used following the equation:

$$\rho = \sum_{i=1}^{n} \lambda_i \Big/ \sum_{i=1}^{m} \lambda_i \tag{6}$$

where,

$\sum_{i=1}^{n} \lambda_i$ =The full variance of all data set

$\sum_{i=1}^{m} \lambda_i$ = The variance of subset of size m

The choice of the subset of m characteristics represents a good estimate of the n-dimension space if the ratio ρ is sufficiently large or greater than a threshold, usually at least 70%. This method is inexpensive in calculation when it is applied directly to the total data; however, if PCA is applied on the neurons, it reduces enormously the computations (Fig. 2).

DATASETS DESCRIPTION

The performance of the proposed SOM-PCA method has experimented on the variety of real classification problems. The specification of these problems is listed in Table 1. All datasets are available from the UCI Machine Learning Repository. Table 1 summarizes the number of features, instances and classes for each dataset used in this study.

Wisconsin breast cancer: The dataset was collected by Dr. William H. Wolberg (1989-1991) at the University of Wisconsin-Madison Hospitals. It contains 699 instances whose 458 (65.5%) instances of them areBenign and 241 (34.5%) instances are Malignant, characterized by nine features, which are used to predict benign or malignant disease. This data contains 16 instances with single missing value.

Heart-Statlog: The dataset is based on data from the Cleveland Clinic Foundation and it contains 270 instances belonging to two classes: the presence or absence of heart disease. It is described by 13 features.

Cardiotocography Data Set: The dataset consists of measurements of Fatal Heart Rate (FHR) and Uterine Contraction (UC) features on cardiotocograms classified by expert obstetricians. 2126 fatal cardiotocograms (CTGs) were automatically processed and the respective diagnostic features measured. The CTGs were also classified by three expert obstetricians and a consensus classification label assigned to each of them. Classification was both with respect to a morphologic pattern (A, B, C. ...) and to a fatal state (N, S, P). Therefore the dataset can be used either for 10-class or 3-class experiments available in UCI Machine Learning Repository.

RESULTS AND DISCUSSION

In order to show the efficiency of the proposed method, SOM-PCA has experimented on the variety of real benchmark classification problems downloaded from the UCI Machine Learning Repository (a short description of each data set is shown in Table 1) and it is evaluated in terms of accuracy and it is compared to classical SOM. In our topology, the hidden layer consists of 25 neurons (rectangular topology 5×5). The output layer was determined by one neuron that can be 0 or 1. The general architecture of the proposed network is shown in Fig. 1. A summary of the parameters used is described in Table 2. Firstly, All datasets were prepared for the classification, the missing values were replaced by median value (Acuña and Rodriguez, 2004), the data were normalized using min-max normalization (Sola and Sevilla, 1997; Jain and Bhandare, 2011), the datasets were divided into two, 70% is employed for training process and 30% for testing process and all the weights have initialized to random numbers. Then the training process will be done.

When the training process is complete for the training data, the last weights of the network have been saved to be ready for the feature extraction procedure using the PCA algorithm and then apply the test dataset.

To evaluate SOM-PCA, we used the classification accuracy as follow:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where,

TP (True Positives) = The correctly classified as positive cases

TN (True Negative) = Correctly classified as negative cases

FP (False Positives) = Incorrectly classified as negative cases

Table 1: Short description of datasets

Dataset	N°. of attributes	N°. of instances	Class distribution
WBC	11	699	B: Benign M: Malignant
Heart-Statlog	13	270	0: Presence of heart disease 1: Absence of heart disease
CTG	23	2126	N = normal S = suspect P = pathologic

Table 2: Parameters for SOM-PCA

Parameters	Value
Number of input neurons	5×5
Learning rate	0.9
Radius	20
Distance Metric	Euclidean
Normalization attributes	True
Initialization	Random sample
Number of iterations	1000

Table 3: Numerical result obtained by calculating accuracy for SOM and proposed method

Datasets	WBC	Heart-Statlog	CTG
SOM	96.56	79.01	86.20
Proposed method	97.14	81,48	89.02

FN (False Negative) = Incorrectly classified as positive cases

Table 3 the best results obtained for the accuracy of classifier using $75\% \leq \rho \leq 95\%$ for feature reduction. These results are gotten from Fig. 3 to 5 on a percentage basis. In these figures, the horizontal axis represents the number of PCs and the vertical axis represents accuracy of classification (the gray curve) and Cumulative Percentage of Variance explained (black curve) on percentage basis.

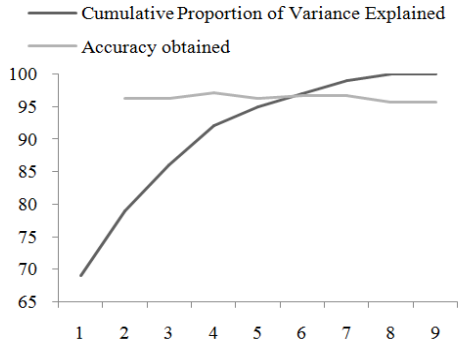

Fig. 3: The cumulative proportion of variance explained and accuracy classification obtained for WBC dataset

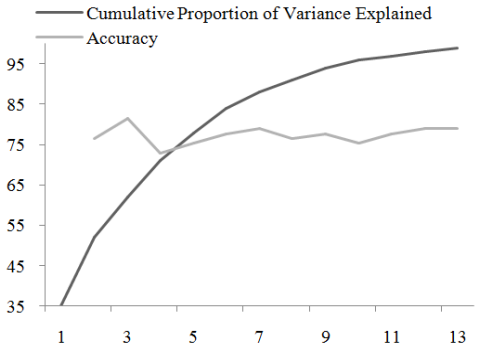

Fig. 4: The cumulative proportion of variance explained and accuracy classification obtained for Heart-Statlog dataset

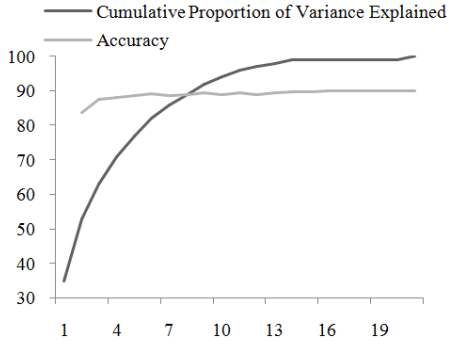

Fig. 5: The cumulative proportion of variance explained and accuracy classification obtained for Cardiotocography dataset

These figures demonstrate that by using proposed method, the accuracy is almost unchanged and even increased; it is clear that there is a slight improvement in the classification rate; the maximum value is obtained when the cumulative is between 75% and 95% and after it begins to decrease. In other words, when the number of contributed variables increases the classification rate decreases, therefore, we can only keep variables whose cumulative is less than 95% and the remained features have no effect on the classification rate.In the rest of this part, the results in detail for each dataset.

Breast cancer dataset: Figure 3 shows the cumulative sum of explained variance over different feature selection for the breast cancer dataset (black curve) and the accuracy obtained (grey curve). The grey curve shows that most of the variance (79% of the variance) can be explained by the two first principal components. The third, fourth and fifth principal component still bears some information (16%) while the remaining principal components can carefully be dropped without losing too much information. Together, the first five principal components contain 95% of the information. Now, take a look at the grey curve; we can see that the value of accuracy is around 97% when using 5 features. On classifying the dataset employing original features, it is noted that the classification accuracy of 95.85% is obtained. On applying the proposed method, the accuracy is increased to 97.14%. The highest accuracy is reported for this dataset when the proposed SOM-PCA approach is employed with 5 components.

Heart-Statlog: Figure 4 the cumulative of variance explained over different feature selection for the Heart-Statlog shows that most of the variance can be explained by the eight first principal components. The first eight principal components contain 91% of the information. In opposite, the best accuracy 81.48% is obtained with first three components. Compared to 79% of accuracy obtained by classifying the dataset employing original features, With SOM-PCA, the accuracy is increased slightly to 81.48%. The highest accuracy is reported for this dataset when the proposed SOM-PCA approach is employed with three components.

Cardiotocography dataset: From the Fig. 5 the five first components accounts for 77% of the variance. The remaining components contribute with gradually decreasing variance and we assume this smaller variation is mostly unimportant. The value of accuracy is around 89% when using 5 features and it kept its value almost fixed along the rest components. The accuracy obtained using all original features are 79.93%. So, applying the proposed method, the accuracy is increased significantly to 97.14%. The

highest accuracy is reported for this dataset when the proposed SOM-PCA approach is employed with 5 components.

CONCLUSION

This study presents a result of direct classification of variety of datasets using self-organizing maps algorithm. A novel approach based on the Self Organizing Maps and principal component analysis to address the problem of classification. The main innovation is to reduce the dimension of the neurons detected after the SOM training; the new dataset will represented the map with high accuracy. From the numerical results, the improved method gives better accuracy and low time for training, by reducing the dimension of the map and so decreasing the memory size to store the map. The presented method considers the datasets with low dimension and can be extended to treat the data with high dimension. Up to 2% of improvement is obtained using SOM-PCA compared to classical SOM; it can be concluded that this method can be a solution to some problems where very few numbers of training samples exist and feature reduction is needed to apply unsupervised classifiers.

REFERENCES

Abdi, H. and L.J. Williams, 2010. Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat., 2(04): 433-459.

Acuña, E. and C. Rodriguez, 2004. The Treatment of Missing Values and its Effect on Classifier Accuracy. In: Banks D., F.R. McMorris, P. Arabie and W. Gaul (Eds.), Classification, Clustering and Data Mining Applications. Springer, Berlin, Heidelberg, pp: 639-647.

Arauzo-Azofra, A., J.L. Aznarte and J.M. Benítez, 2011. Empirical study of feature selection methods based on individual feature evaluation for classification problems. Expert Syst. Appl., 38(7): 8170-8177.

Blum, A.L. and P. Langley, 1997. Selection of relevant features and examples in machine learning. Artif. Intell., 97(1-2): 245-271.

Cattell, R.B., 1966. The Scree test for the number of factors. Multivar. Behav. Res., 1(02): 245-276.

D'agostino, R.B. and H.K. Russell, 2005. Scree Test. In: Encyclopedia of Biostatistics. John Wiley and Sons, Ltd.

Devaraj, D., B. Yegnanarayana and K. Ramar, 2002. Radial basis function networks for fast contingency ranking. Int. J. Elec. Power, 24(05): 387-393.

Ding, C. and H. Peng, 2005. Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biol., 3: 185-205.

Dudoit, S., J. Fridlyand and T.P. Speed, 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc., 97: 77-87.

Ettaouil, M., M. Lazaar and G. Youssef, 2012. Vector quantization by improved kohonen algorithm. J. Comput., 4: 111-117.

Ettaouil, M., M. Lazaar and Y. Ghanou, 2013. Architecture optimization model for the multilayer perceptron and clustering. J. Theor. Appl. Inf. Technol., 47: 64-72.

Gheyas, I.A. and L.S. Smith, 2010. Feature subset selection in large dimensionality domains. Pattern Recog., 43(01): 5-13.

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

Jain, Y.K. and S.K. Bhandare, 2011. Min max normalization based data perturbation method for privacy protection. Int. J. Comput. Commun. Technol., 2: 45-50.

Jolliffe, I.T., 1972. Discarding variables in a principal component analysis. I: Artificial data. J. R. Stat. Soc. C-Appl., 21(02): 160-173.

Kaiser, H.F., 1960. The application of electronic computers to factor analysis. Educ. Psychol. Meas., 20(1): 141-151.

King, J.R. and D.A. Jackson, 1999. Variable selection in large environmental data sets using principal components analysis. Environmetrics, 10(01): 67-77.

Kira, K. and L.A. Rendell, 1992. A practical approach to feature selection. Proceedings of the 9th International Workshop on Machine Learning(ML92). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp: 249-256.

Kohavi, R. and G.H. John, 1997. Wrappers for feature subset selection. Artif. Intell., 97(1-2): 273-324.

Kohonen, T., 1998. The self-organizing map. Neurocomputing, 21: 1-6.

Maldonado, S., R. Weber and J. Basak, 2011. Simultaneous feature selection and classification using kernel-penalized support vector machines. Inform. Sciences, 181(1): 115-128.

Mundra, P.A. and J.C. Rajapakse, 2010. SVM-RFE With MRMR filter for gene selection. IEEE T. NanoBiosci., 9(01): 31-37.

Narayanan, A., E.C. Keedwell, J. Gamalielsson and S. Tatineni, 2004. Single-layer artificial neural networks for gene expression analysis. Neurocomputing, 61: 217-240.

Pavel, S. and K. Olga, 2011. Visual analysis of self-organizing maps. Nonlinear Anal-Model., 16(4): 488-504.

Ruiz, R., J.C. Riquelme, J.S. Aguilar-Ruiz and M. Garcia-Torres, 2012. Fast feature selection aimed at high-dimensional data via hybrid-sequential-ranked searches. Expert Syst. Appl., 39(12): 11094-11102.

Shaharudin, S.M. and N. Ahmad, 2017. Choice of Cumulative Percentage in Principal Component Analysis for Regionalization of Peninsular Malaysia Based on the Rainfall Amount. In: Mohamed Ali, M., H. Wahid, N. MohdSubha, S. Sahlan, M. Md. Yunus and A. Wahap (Eds.), Modeling, Design and Simulation of Systems. AsiaSim, 2017. Communications in Computer and Information Science, Springer, Singapore,752: 216-224. https://link.springer.com/chapter/10.1007/978-981-10-6502-6_19

Sola, J. and J. Sevilla, 1997. Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE T. Nucl. Sci., 44(3): 1464-1468.

Wang, Z., V. Palade and Y. Xu, 2006. Neuro-fuzzy ensemble approach for microarray cancer gene expression data analysis. Proceeding of the 2006 International Symposium on Evolving Fuzzy Systems, pp: 241-246.