Journal of Engineering and Applied Sciences 14 (23): 8828-8834, 2019
ISSN: 1816-949X

Recursive Feature Elimination and Gravitational Search Algorithm for Classification of Medical Data

[1]P. Penchala Prasad, [1]F. Sagayaraj Francis and [2]S. Zahoor-Ul-Huq
[1]Department of CSE, G. Pulla Reddy Engineering College, 510002 Kurnool, Andhra Pradesh, India
[2]Department of CSE, Pondicherry Engineering College, Puducherry, India
p.prasad71@gmail.com, 9030551881

Abstract: Medical data classification is the challenging task due to noisy data or missing data are present in the dataset. The feature selection techniques play the important part in the classification process. The more relevant features help to provide the efficient classification of medical data which is essential for the disease detection. In this research, Recursive Feature Elimination with the Gravitational Search Algorithm (RFE-GSA) is proposed for efficient classification of the data. The Recursive Feature Elimination (RFE) method helps to remove the irrelevant features from the medical data and rank them in order of importance that helps to reduce the computation cost of the proposed method. The ranked features from the RFE are given as input to the GSA which select the feature for the classification. The GSA is fast convergence and that helps to find the relevant features in the data. The features selected from the RFE-GSA is provided as input to the Radial Basis Function (RBF) for the classification. The performance of the RFE-GSA method is high compared to the other existing method. The proposed RFE-GSA method has the accuracy of the 98.24% in the breast cancer dataset in UCI dataset and the state-of-art method has achieved the accuracy of 96.87%.

Key words: Feature selection, gravitational search algorithm, medical data classification, recursive feature elimination and UCI dataset, Radial Basis Function (RBF), UCI

INTRODUCTION

Healthcare data classification methods helps the clinicians to assist in the diagnosis and treatment in medical diseases (Nguyen *et al*., 2015). The various data mining techniques are applied in the field of the medical and health field to effectively classify the medical data. Method with high precious and reliable helps the doctor to select the necessary treatment with the accurate prediction of disease (Yang *et al*., 2018). The main concerns with the data streaming processing method is related to the length of the stream (usually high), non-stationary distribution that causes the concept drift and the data are collected at high speed (Junior and Nicoletti, 2019). The challenges in the medical data classification is to handle small datasets and limited annotated samples, especially when using supervised method that requires to label the data and larger training examples (Frid-Adar *et al*., 2018). There are some incomplete datasets are present that contains missing values. Missing values in the datasets are common and this affects the performance of the classification. For instance, the popular benchmark dataset for the medical data is UCI benchmark dataset, 45% of the datasets are suffer from the missing values (Tran *et al*., 2018a, b). Many classification algorithm considered that there are no large disproportions between the objects of the different classes. However, in the practical terms, there are large number of disproportion are present, the object in the one class is varied from the objects from another class. Such issues are said to be imbalance data classification (Koziarski *et al*., 2019; Liu and Zio, 2019). The large repository of image and the quality of the image is the major challenges for the image retrieval and for the data classification in medical research (Zhang *et al*., 2017a, b). The deep learning techniques are applied to improve the classification accuracy and this also increases the computation time of the method (Renuka and Annadhason, 2019). Increasing in the technology of the medical increase the different modalities are generated in massive numbers and this requires effective method for classification (Zhang *et al*., 2017a, b). In this research, the RFE-GSA method is proposed to increase the performance of the medical data classification. The RFE-GSA select the important features from the medical dataset and the RBF is applied to classify the data. The experimental results shows that the proposed RBF-GSA method has the higher performance compared to the other existing method.

Corresponding Author: P. Penchala Prasad, Department of CSE, G. Pulla Reddy Engineering College, 510002 Kurnool, Andhra Pradesh, India, p.prasad71@gmail.com, 9030551881

Literature review: The classification of the medical data is the important task and machine learning techniques are applied to achieve the efficient classification. Many methods has been proposed has been applied to the medical data to improve the classification accuracy. The latest researches in the classification of medical data are surveyed in this study.

Shen *et al.* (2016) applied the Support Vector Machines (SVM) for the medical data classification with the fruit fly optimization is proposed for the tuning parameter. The FOA-SVM method is evaluated in the UCI medical dataset in term of sensitivity, specificity, accuracy and AUC. The inner parameter optimization process dynamically adjusted the SVM parameters using the FOA technique. The GSA can be applied to feature selection techniques for effective performance of the classification.

Wood *et al.* (2019) presented the private key fully homomorphic encryption technique to develop the encryption method for private classification using Naive Bayes. This technique helps the user to classify their data without the direct access of the model. This prototype has been applied to the breast cancer classification with privacy preserving. The optimization technique can be applied to increase the accuracy of the classification with data security.

Tran *et al.* (2018a, b) proposed an improvement on the ensemble method by combining the genetic feature selection technique and the imputation. High quality training data are created using the imputation method. The number of missing patterns are reduced by the feature selection that increases the classification speed and greatly increase the new instance fraction by the ensemble. The proposed method results and analysis has the better classification accuracy for most cases. The GSA techniques can be applied to enhance the performance of imputation in the missing data.

Baccour (2018) combines the Multi-Criteria Decision Making (MCDM) methods Techniques for Order Preference by Similarity to Ideal Solution (TOPSIS) and VIKOR (a Serbian name). These techniques are modified for the classification based on the three sets namely attribute, class and objects. Hence, the new classifier is proposed named as ATOVIC. The ATOVIC criteria are replaced by the features and alternatives are replaced by objects. ATOVIC is applied on the UCI benchmark dataset to predict the heart disease. The experimental results shows that the proposed method has the higher accuracy and true positive rates. The effectiveness of the method in the big data is need to improve and computation time is need to be reduced.

Sajjad *et al.* (2019) established a method using convolution neural network for the brain tumor classification. The brain tumor is classified from the Magnetic Resonance Image (MRI) using the deep learning technique. Then, the data augmentation technique is applied to missing data problem for the MRI. The augmented data is used for the pre-train CNN Model that is fine tuned for the brain tumor classification. The computation time of the developed method is high and need to be reduce for the effective performance.

MATERIALS AND METHODS

The medical data classification is the challenging task due large number of data are present in the datasets. Machine learning method are applied to effective data classification for disease identification. This study aims to select the features using RFE-GSA method for the RBF to classify the medical data. The risk factor of the diseases is identified using the RFE-GSA method. The features selected by the RFE-GSA is applied to the RBF to classify the data. The benchmark of UCI respiratory data is used to evaluate the proposed RFE-GSA performance in medical data classification. The proposed RFE-GSA method description is given in this study. The block diagram of the proposed method is shown in Fig. 1.

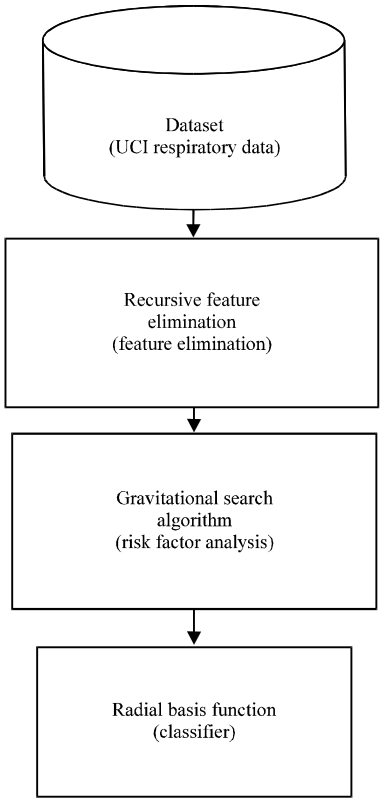

Fig. 1: The block diagram of the proposed method

Inputs:

Training set T
Set of p features $F = \{f_1, ..., f_p\}$
Ranking method $M(T, F)$

Outputs:

Final ranking R

Code:

Repeat for i in $\{1 : p\}$
Rank set F using $M(T, F)$
$f^* \leftarrow$ last ranked feature in F
$R(p - i + 1) \leftarrow f^*$
$F \leftarrow F - f^*$

Fig. 2: The pseudo code of recursive feature elimination

Recursive feature elimination: The RFE features elimination technique (Granitto *et al*., 2006) is the recursive process that order the features based on their importances. The pseudo code of the RFE is shown in (Fig. 2). The feature importance are measured at each iteration and the less relevant one is removed. The group of features are eliminated each time to speed up the process. The recursive is required to measure the relative feature importance that can substantially changes when process the different feature subset in stepwise elimination process. The features are eliminated in the inverse order to finalize the feature rank. The feature selection process consists of taking the first n feature from ranking.

Feature selection with gravitational search algorithm: The GSA is developed based on the law of gravity and the gravity is implemented for each individual in the swarm algorithm. The gravity is act on every particles and it differ from the other physical force. The gravity has the significant impact on the particle and its motion in the universe. The particle motions are determined by the gravity force and one particle may have high mass attract to another particle with low mass that minimize the distance and change the direction. The gravity force act between two particles is depend on the distance and mass. The particles positions and velocity are varying due to the act of gravitational force. Based on this process, GSA is initially proposed to optimization in the continuous function (Wang *et al*., 2019; Zhao *et al*., 2018).

Each particle in GSA are considered as individual. The gravity forces on the individuals are depend on the expressed masses in terms of fitness function, gravitational constant and their distance. The gravitational effects on the individuals are act on each other based on the mass and distance that changes the direction and motion of the particles. The higher mass in the system shows the higher fitness function of the individuals. The position and mass are considering for every particle to calculate the quality and composition of a solution, respectively. Each element mass is increase by continuously changes the position based on the gravitational forces. The feature from the RFE is applied to the GSA in the feature importance order as x(t):

$$F_{ij}^{d}(t) = \frac{G(t)\left(M_i(t)\times M_j(t)\right)}{R_{ij}(t)+\in\left(x_j^d(t)-x_j^d(t)\right)} \quad (1)$$

where, G(t) is gravitational constant with iteration t, the two individual masses are $M_i(t)$ and $M_j(t)$. The Euclidean distance between the particles are denoted as $R_{ij}(t)$ which is defined as $R_{ij}(t) = \bullet\bullet X_i(t, X_j(t)\bullet\bullet_2$ and • is a little constant. The constant in gravitationalis defined in Eq. 2:

$$G(t) = G_0 \times e^{-\alpha\frac{t}{T}} \quad (2)$$

where, the initial value G_0 is and is constant. The current iteration are denoted as t and T, respectively. The mass $M_i(t)$ of individual X_i is termed as in Eq. 3:

$$m_i(t) = \frac{f_i(t)-w(t)}{b(t)-w(t)} \quad (3)$$

$$M_i(t) = \left(m_i(t)\right)/\left(\sum_{1=1}^{n} m_1(t)\right) \quad (4)$$

where, $f_i(t)$ is the manifests fitness value of the particle X_i. w(t) and b(t) denotes the best and worst fitness values in the iteration. For particle X_i, the gravitational force $F^d(t)$ from other particles in the dth dimension is shown in Eq. 5:

$$F_i^d(t) = \sum_{j\in k_b, j\neq i} rand_j F_{ij}^d(t) \quad (5)$$

where, K_b is the best value of K in the population and K value is limited from the initial n to 2. The $rand_j$ is the uniform random variables that is present between the value of [0, 1] for each particle X_j. The acceleration $a_i^t(t)$ of particle X_i in the dimension of d is applied for gravitational force as given in Eq. 6:

$$a_i^d(t) = \frac{F_i^d(t)}{M_i(t)} \quad (6)$$

The velocity $v^d_i(t+1)$ of each particles X_i is updated for position change in the next iteration t+1 is shown in Eq. 7 and 8:

$$v_i^d(t+1) = rand_i v_i^d(t) + a_i^d(t) \quad (7)$$

$$x_i^d(t+1) = x_i^d(t) + v_i^d(t+1) \quad (8)$$

Radial basis function networks: The overlapping and locally-tuned receptive field is the general structure that are present in the cerebral cortex region. Moody and Darken (Jang and Sun, 1993; Liu *et al*., 2019) developed a network based on the biological receptive field for mapping function. The RBFN schematic diagram is shown in Fig. 1, the ith receptive field unit is given in Eq. 9:

$$w_i = R_i(\vec{x})\, R_i\left(\|\vec{x}-\vec{C_1}\|/\sigma_i\right),\ i = 1, 2, ..., H \quad (9)$$

where, $\vec{x}$ is the input vector from the N-dimension and $\vec{C_i}$ has the same dimension. The receptive filed unit is denoted as H and $R_i(.)$ is the receptive field response with a single maximum at the origin. The Gaussian function is considered as R(.):

$$R_i(\vec{x}) = \exp\left|-\frac{\|\vec{x}-\vec{C_1}\|^2}{\sigma_i^2}\right| \quad (10)$$

The w_i is the radial basis function that is based on the ith hidden units with the input vector present near to the center. The RBFN output can be obtained in the two ways. The weighted sum of the function value is the simple method is related to each receptive field:

$$f(\vec{x}) = \sum_{i=1}^{H} f_i w_i = \sum_{i=1}^{H} f_i R_i(\vec{x}) \quad (11)$$

where, f_i is the function value and i-th receptive field. The lateral connections between the receptive field with the weighted average strength to generate network as in Eq. 12:

$$f(\vec{x}) = \frac{\sum_{i=1}^{H} f_i w_i}{\sum_{i=1}^{H} w_i} = \frac{\sum_{i=1}^{H} f_i R_i(\vec{x})}{\sum_{i=1}^{H} R_i(\vec{x})} \quad (12)$$

The GSA learning algorithm is proposed to find the feature and reduce the square errors in the model output. The features are given as the parameters $(\vec{C_i}, \sigma_i \text{ and } f_i)$ in the RBFN.

Experimental design

Dataset description: The UCI respiratory medical dataset are used to evaluate the performance of the RFE-GSA method. The datasets include the Pima Indians diabetes, Parkinson, Wisconsin breast cancer and thyroid cancer are used. The description about the different datasets with number of instance and number of classes are shown in Table 1.

Table 1: The dataset description

Dataset	No. of instance	No. of features	No. of classes	Missing values
Wisconsin breast cancer (Wisconsin)	699	9	2	Yes
Pima Indians diabetes (Pima)	768	8	2	No
Parkinson	195	22	2	No
Thyroid	215	5	3	No

Performance measure: The RFE-GSA method is evaluated using the four parameters such as accuracy, sensitivity, specificity and Area Under Curve (AUC). These four parameters are measured from the outcome of the RFE-GSA method. The formula for measuring the performance such as accuracy, sensitivity, specificity and AUC:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \times 100 \quad (13)$$

$$\text{Sensitivity} = \frac{TP}{TP+FP} \times 100 \quad (14)$$

$$\text{Specificity} = \frac{TN}{FP+TN} \times 100 \quad (15)$$

RESULTS AND DISCUSSION

The classification of medical data involves in the difficult process due to the presence of noisy and missing data in dataset. Many research has been conducted to improve the efficiency of the medical data classification. In this research, the RFE-GSA method is proposed to increase the efficiency of the medical data classification. The RFE-GSA method select the important features from the data and applied to the RBF method classify the data. This study, discuss about the performance of the proposed RFE-GSA method in the four UCI dataset such as breast cancer dataset (Wisconsin) diabetes datasets (Pima) Parkinson disease dataset and thyroid dataset.

Breast cancer diagnosis problem: The breast cancer dataset is used to evaluate the effectiveness of the proposed RFE-GSA and analysis its effectiveness, as given in Table 2. The effectiveness of the different techniques in the breast cancer classification is shown in Fig. 3. The RFE-GSA method performance is analyzed with state-of-art method to investigate the performance. This shows that the RFE-GSA has the more performance compare to the other existing methods. The accuracy of the proposed RFE-GSA method is achieved as 98.24% and the accuracy of the state-of-art method is 96.9%.

Table 2: Performance of several methods in breast cancer dataset

Metrics	PSO-SVM (Shen *et al*., 2016)	Grid-SVM (Shen *et al*., 2016)	GA-SVM (Shen *et al*., 2016)	BFO-SVM (Shen *et al*., 2016)	FOA-SVM (Shen *et al*., 2016)	RFE-GSA
Accuracy	96.27	96.24	95.64	95.57	96.90	98.24
AUC	96.26	96.02	95.29	96.09	96.87	98.71
Sensitivity	96.24	96.62	96.27	94.57	96.86	100
Specificity	96.59	95.45	94.32	97.61	96.89	97.43

Table 3: The RF-GSA method in diabetes dataset

Metrics	PSO-SVM (Shen *et al*., 2016)	Grid-SVM (Shen *et al*., 2016)	GA-SVM (Shen *et al*., 2016)	BFO-SVM (Shen *et al*., 2016)	FOA-SVM (Shen *et al*., 2016)	RFE-GSA
Accuracy	76.50	76.48	76.26	76.47	77.46	88.13
AUC	71.46	71.19	71.14	71.21	72.34	93.06
Sensitivity	54.18	53.59	54.12	53.82	55.07	87.56
Specificity	88.74	88.80	88.16	88.61	89.62	98.61

Table 4: Parkinson's dataset

Metrics	PSO-SVM (Shen *et al*., 2016)	Grid-SVM (Shen *et al*., 2016)	GA-SVM (Shen *et al*., 2016)	BFO-SVM (Shen *et al*., 2016)	FOA-SVM (Shen *et al*., 2016)	RFE-GSA
Accuracy	96.27	96.24	95.64	95.57	96.90	93.58
AUC	96.26	96.02	95.29	96.09	96.87	100
Sensitivity	96.24	96.62	96.27	94.57	96.86	98.36
Specificity	96.59	95.45	94.32	97.61	96.89	98.36

Table 5: Thyroid disease

Metrics	PSO-SVM (Shen *et al*., 2016)	Grid-SVM (Shen *et al*., 2016)	GA-SVM (Shen *et al*., 2016)	BFO-SVM (Shen *et al*., 2016)	FOA-SVM (Shen *et al*., 2016)	RFE-GSA
Accuracy	95.26	94.99	95.94	94.4	96.38	98.21

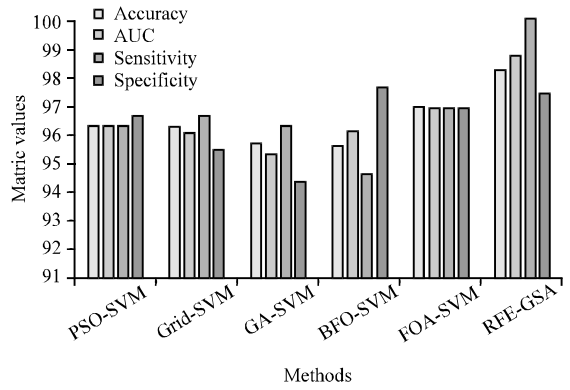

Fig. 3: The various methods in medical data classification

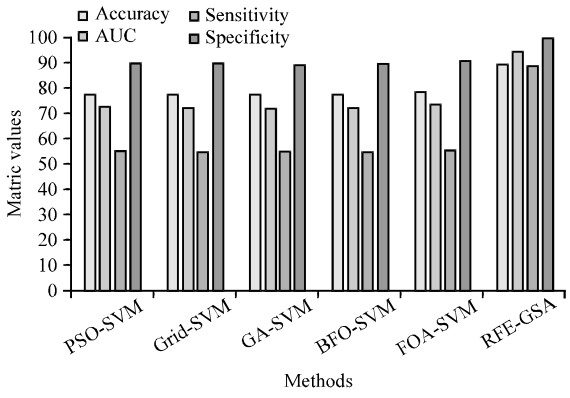

Fig. 4: The various techniques in diabetes classification

Diabetes disease diagnosis problem: The diabetes dataset is used to evaluate the effectiveness of the RFE-GSA method in the medical data classification. Table 3 displays that the RFE-GSA method has the higher performance in the medical data classification compared to existing method. The comparison of the different techniques in the diabetes classification is shown in Fig. 4. The RFE-GSA method has the AUC of 93.06 while the existing method has the AUC of 72.34. This shows that the RFE-GSA has the higher efficiency in the medical data classification.

Parkinson's disease diagnosis problem: The proposed RFE-GSA is evaluated with the Parkinson's dataset and measure the accuracy, AUC, sensitivity and specificity. The proposed RFE-GSA is compared with other existing method in the classification in Parkinson's dataset as shown in Table 4. This shows that the RFE-GSA technique has the higher performance than other existing method. The proposed RFE-GSA accuracy is achieved as 93.58% while state-of-art method is achieved as 96.9%. This shows that the proposed RFE-GSA method achieves higher performance compared to existing methods.

Thyroid disease diagnosis problem: The proposed method RFE-GSA is tested on the thyroid dataset to analysis the performance in classification (Table 5). The

proposed RFE-GSA method has the higher accuracy of 98.21% while existing method has 96.38% accuracy. This shows that RFE-GSA has the higher efficiency in the medical data classification than the other existing methods. The analysis of the RFE-GSA in medical data classification of four datasets shows that the RFE-GSA is high efficient compared to other previous methods. This shows that the RFE-GSA can be applicable to the classification for effective performance.

CONCLUSION

Medical data classification is the challenging task due to the missing data in the medical dataset. Although, some research was carried out to improve the classification of the medical data, still there is need to efficiency in classification technique. In this research, the RFE-GSA technique is proposed to increase the efficiency of the classification. The RFE method helps to decrease the irrelevant features from the datasets. The GSA is used to select the important features from the data. The RBF method uses the features selected by the RFE-GSA for medical data classification. The RFE-GSA technique is evaluated with the four UCI respiratory medical dataset and compared with previous methods. The proposed RFE-GSA method is compared with existing method to analyze the performance. The experimental outcome shows that the RFE-GSA has the higher performance than the other existing methods. The RFE-GSA achieve 98.21% accuracy in thyroid dataset while existing method achieved 96.38% in the same dataset.

RECOMMENDATION

The future research of this method will be involves in developing the model to overfitting while selecting the feature that improves the effectiveness.

REFERENCES

Baccour, L., 2018. Amended fused TOPSIS-VIKOR for classification (ATOVIC) applied to some UCI data sets. Expert Syst. Appl., 99: 115-125.

Frid-Adar, M., I. Diamant, E. Klang, M. Amitai and J. Goldberger *et al.*, 2018. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing, 321: 321-331.

Granitto, P.M., C. Furlanello, F. Biasioli and F. Gasperi, 2006. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemom. Intell. Lab. Syst., 83: 83-90.

Jang, J.S.R. and C.T. Sun, 1993. Functional equivalence between radial basis function networks and fuzzy inference systems. IEEE. Trans. Neural Netw., 4: 156-159.

Junior, J.R.B. and M.D.C. Nicoletti, 2019. An iterative boosting-based ensemble for streaming data classification. Inf. Fusion, 45: 66-78.

Koziarski, M., B. Krawczyk and M. Wozniak, 2019. Radial-based oversampling for noisy imbalanced data classification. Neurocomputing, 343: 19-33.

Liu, J. and E. Zio, 2019. Integration of feature vector selection and support vector machine for classification of imbalanced data. Appl. Soft Comput., 75: 702-711.

Liu, Y., X. Wang and L. Wang, 2019. Interval uncertainty analysis for static response of structures using radial basis functions. Appl. Math. Modell., 69: 425-440.

Nguyen, T., A. Khosravi, D. Creighton and S. Nahavandi, 2015. Medical data classification using interval type-2 fuzzy logic system and wavelets. Appl. Soft Comput., 30: 812-822.

Renuka, S. and A. Annadhason, 2019. Mil based lung CT-image classification using CNN. Health Technol., 1: 1-9.

Sajjad, M., S. Khan, K. Muhammad, W. Wu and A. Ullah *et al.*, 2019. Multi-grade brain tumor classification using deep CNN with extensive data augmentation. J. Comput. Sci., 30: 174-182.

Shen, L., H. Chen, Z. Yu, W. Kang and B. Zhang *et al.*, 2016. Evolving support vector machines using fruit fly optimization for medical data classification. Knowl. Based Syst., 96: 61-75.

Tran, C.T., M. Zhang, P. Andreae, B. Xue and L.T. Bui, 2018a. An effective and efficient approach to classification with incomplete data. Knowl. Based Syst., 154: 1-16.

Tran, C.T., M. Zhang, P. Andreae, B. Xue and L.T. Bui, 2018b. Improving performance of classification on incomplete data using feature selection and clustering. Appl. Soft Comput., 73: 848-861.

Wang, Y., Y. Yu, S. Gao, H. Pan and G. Yang, 2019. A hierarchical gravitational search algorithm with an effective gravitational constant. Swarm Evol. Comput., 46: 118-139.

Wood, A., V. Shpilrain, K. Najarian and D. Kahrobaei, 2019. Private naive bayes classification of personal biomedical data: Application in cancer data analysis. Comput. Boil. Med., 105: 144-150.

Yang, S., J.Z. Guo and J.W. Jin, 2018. An improved Id3 algorithm for medical data classification. Comput. Electr. Eng., 65: 474-487.

Zhang, J., Y. Xia, Y. Xie, M. Fulham and D.D. Feng, 2017b. Classification of medical images in the biomedical literature by jointly using deep and handcrafted visual features. IEEE. J. Biomed. Health Inf., 22: 1521-1530.

Zhang, R., J. Shen, F. Wei, X. Li and A.K. Sangaiah, 2017a. Medical image classification based on multi-scale non-negative sparse coding. Artif. Intell. Med., 83: 44-51.

Zhao, F., F. Xue, Y. Zhang, W. Ma and C. Zhang *et al*., 2018. A hybrid algorithm based on self-adaptive gravitational search algorithm and differential evolution. Expert Syst. Appl., 113: 515-530.