International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 3 (2018) pp. 1716-1721

Improved Bacterial Foraging Optimization based Twin Support Vector Machine (IBFO-TSVM) Classifier for Risk Level Classification of Coronary Artery Heart Disease in Diabetic Patients

Rajkumar R[1], Anandakumar K [2] and Bharathi A[3]

[1]*Department of Computer Applications, Sri Krishna Arts and Science College, Coimbatore, India.*
[2]*Department of Computer Applications, Bannari Amman Institute of Technology, Tamil Nadu, India.*
[3]*Department of Information Technology, Bannari Amman Institute of Technology, Tamil Nadu, India.*

Abstract

Data mining is an ever demanding research arena for the computer science researchers. Data mining in health sector is gaining lot of research scope particularly in decision support system, applied soft computing and expert systems and applications. This research work aims in performing an effective feature selection method using improved bacterial foraging optimization for risk level classification of coronary artery heart disease in diabetic patients using twin support vector machine. Conventional machine learning algorithms are most widely used for classification task particularly in data mining. It is a fact that machine learning algorithms without the help of feature selection strategy consumes more time for performing the classification task both in training phase and testing phase. Hence the proposed IBFO is employed in order to select the features that help the classifier to perform better in terms of elapsed time. Dataset has been obtained from UCI machine learning repository. Implementations are carried out using MATLAB tool. The results are compared with the existing techniques. From the results it is evident that the proposed IBFO-TSVM classifier performs better in terms of classification accuracy and also elapsed time.

Keywords: data mining, machine learning, bacterial foraging, feature selection, classification, support vector machine, accuracy

INTRODUCTION

Data mining is the progression of hauling out concealed knowledge from existing data which is capable enough to divulge the patterns and relationships among large amount of data in a single or several datasets. Data mining is employed in many real world applications which include crime detection, risk evaluation and market analysis. Various industries like banking, insurance, and marketing are using data mining for cut down costs, and augment profits. Cardiovascular diseases are among the most common reasons of death all over the world. One major type of these diseases is coronary artery heart disease (CAHD). Twenty five percent of people, who have CAHD, die suddenly without any previous symptoms. CAHD is one of the most imperative types of diseases affecting the heart, and possibly lead to severe heart attacks in patients. Being aware of disease symptoms, can aid in time treatment, and reduce the severity of disease's side effects. The motivation of the research work starts from these preliminaries. The problem statement is quite obvious. Patients who have diabetes are more prone to CAHD. There are several machine learning algorithms and data mining techniques are employed in this research arena. Many of the algorithm deals only with proposing a classifier. Only very few literatures are found that focuses on feature selection task before performing the classification task. Hence this research work concentrates on employing the efficient feature selection strategy by employing an improved bacterial foraging optimization technique. After that twin support vector machine classifier is used to perform the classification. The main objective of this research work is to reduce the elapsed time of the overall classification task during the testing phase. It is evident that without making use of feature selection strategy, the classifier will consume more time to perform the classification task. Hence there is a wide scope for employing an appropriate feature selection mechanism. This paper is organized as follows. This section gives a quick view of the significance, motivation, problem statement. Section 2 discusses on the related works. Section 3 portrays the proposed work. Section 4 narrates the dataset taken with results and discussions. Section 5 provides the concluding remarks of the paper.

RELATED WORKS

Several computer aided diagnosis methodologies have been proposed in the literature for the diagnosis of CAHD. More specifically, the use of approaches like artificial neural networks [1], [2] and [3]), Naïve Bayes [4], support vector machines [5], decision trees [6] have been previously reported. Even though these approaches produce good classification accuracy, the interpretation of results is hard. They are popularly known as ''Black Box'' method since they focus only on the classification accuracy. Although rule based classifier systems, reported in Tsipouras et al. [7] and Adeli and Neshat [8] produces interpretable rules, they lack the robustness in the missing data. Different approaches have been discussed in Grzymala-Busse and Hu [9] and Su et al. [10]. A decision tree is a classifier that can be expressed as a recursive partition of the instance space [11], [12].

Cao et al, 2017 coalesced proposed an unified kernel framework and proposed a novel ℓ 2 , 1 norm balanced multiple kernel feature selection that is proximal based optimization algorithm for efficiently learning the model. Furthermore the authors deployed a multiple kernel oversampling (MKOS) in order to generate synthetic instances

in the optimal kernel space induced by ℓ 2, 1 MKFS in order to reimburse for the class imbalanced distribution. The authors claimed that their experimental results on multiple UCI data and two real medical application and demonstrated jointly operating nonlinear feature selection and oversampling with ℓ 2 , 1 norm multi-kernel learning framework which can lead to a promising classification performance.

Al-Rajab et al, 2017 inspected the accuracy and time complexity of high performance genetic data feature selection and classification algorithms for cancer diagnosis. The authors proposed a three-phase approach. In the phase 1 and 2 the feature selection algorithms are examined and classification algorithms employed separately. The phase 3 of their research evaluated the performance examined the performance. The authors claims that the combination of PSO and SVM surpassed other algorithms in accuracy and performance, and was faster in terms of time analysis.

Deniz et al, 2017 investigated the success of a multiobjective genetic algorithm (GA) combined with state-of-the-art machine learning (ML) techniques for the feature subset selection (FSS) in binary classification problem (BCP). Their multiobjective evolutionary algorithm includes two phases, FSS using a GA and applying ML techniques for the BCP. The authors preferred GA for the first phase of their algorithm for intelligently detecting the most appropriate feature subset. In their second phase of the algorithm, the fitness of the selected subset is decided by using state-of-the-art ML techniques; Logistic Regression, Support Vector Machines, Extreme Learning Machine, K-means, and Affinity Propagation. The authors evaluated the performance of the multiobjective evolutionary algorithm with comprehensive experiments and compared with state-of-the-art algorithms, Greedy Search, Particle Swarm Optimization, Tabu Search, and Scatter Search. The authors proposed algorithm has been observed to be robust and it performed better than the existing methods on most of the datasets.

Zhang et al, 2015 proposed a novel feature selection method based on Class-Separability (CS) strategy and Data Envelopment Analysis (DEA). In order to confine the relationship between features and the class, class labels are separated into individual variables and relevance and redundancy are explicitly handled on each class label. Super-efficiency DEA is employed to evaluate and rank features via their conditional dependence scores on all class labels, and the feature with maximum super-efficiency score is then added in the conditioning set for conditional dependence estimation in the next iteration, in such a way as to iteratively select features and get the final selected features. The authors claims that theirproposed feature selection method performs better than that of chosen methods.

Cheng et al, 2016 proposed a feature selection method that simultaneously embedding the low-rank constraint, sparse representation, global and local structure learning into a unified framework. Initially the authors utilized the conventional regression function to form a novel regression framework by introducing a low-rank constraint and a relaxation term. And then the authors employed an l21-norm regularization term to filter out the redundant and irrelative features. In addition, a hypergraph based regularization term has been utilized to construct a Laplacian matrix that will be used in enhancing the inherent association of data. Above and beyond, the authors proposed a novel optimization algorithm to solve the objective function. To end with, the authors fed the reduced data got by the proposed feature selection method into Support Vector Machines (SVM) in term of classification accuracy. Their experimental results showed that their proposed method achieved the best classification performance, compared with the state-of-the-art feature selection methods on real multi-view dataset.

PROPOSED WORK

Technical Background of Improved Bacterial Foraging Optimization

The lifecycle of E-Coli bacteria is classified as chemotaxis, reproduction, elimination and dispersal. By and large, E. coli runs by its 8–10 rotating 'whip-like flagella', and alters the track by the flagella stirring towards tumbling (clockwise) or run forward (anti-clockwise). All the creatures proliferate themselves and the population then could increase a lot, at the same time as some of them also would be dead for numerous reasons like being old, or get out from the viable environment. Consequently, the population size remains stable, and the eminence of the population augments over and over again. The operation process of E. coli is primarily based on the flagella that probably will travel the bacteria in forward tracks or random tracks. In point of fact, bacterium travels with the swinging between those two modes until it obtains the 'nutrients'. The nutrient in veracity application is the object to search for. The process of chemotaxis plays a significant role and is mathematically represented as in Equation 1:

$$\theta_i(j+1,k,l)=\theta_i(j,k,l)+C(i)\frac{\Delta(i)}{\sqrt{\Delta^T(i)\Delta(i)}} \quad \ldots (1)$$

Where $\theta_i(j,k,l)$ is the position of ith bacterium at jth chemotactic kth reproductive and lth elimination and dispersal step. C(i) indicates the step size of ith bacterium, which is a constant. These preliminaries are common for BFO . $\Delta(i)$ represents the track angle of ith bacterium whose elements belong to [−1, 1]. As far as improved BFO is concerned the proposed work makes use of the below mathematical model found in Equation 2.

$$C_j(i)=C_{\min}+\frac{iter_{\max}-iter}{iter_{\max}}(C_{\max}-C_{\min}) \quad \ldots (2)$$

Where $iter_{max}$ is the maximum iteration, and iter stands for the current iteration, j is the current number of iteration. When $C_{max} = C_{min}$, the chemotaxis step becomes a constant.

Reproduction procedure in IBFO is a means to update the population. The fitness individuals with better seeking capacity are taken proliferate the new offspring. In point of fact, the recitals of the bacteria are assessed with the help of 'health'. This recital metric is a quantification of nutrients the

bacteria have obtained during its lifetime. For i[th] bacterium at its j[th] chemotactick[th] reproductive and l[th] elimination and dispersal step, the health of it is computed using the formula found in Equation 3:

$$J^{i}_{health} = \sum_{j=1}^{N_c+1} J(i, j, k, l)$$

$$\theta_i + S_r(j,k,l) = \theta_i(j,k,l), i = 1,...,S_r \quad \ldots (3)$$

Superior values of J^{i}_{health} denotes the lower health of that bacterium to minimize the goal. Depending on the values of J^{i}_{health}, the general performance of the bacteria will be ranked once sorting operation is carried out. As mentioned in the Equation 3, the first half of the bacterium with the smaller values of J^{i}_{health} probably will save and train for reproduction. The rest half with the higher J^{i}_{health} will get eliminated (die) from the population. S_r denotes the half of the population size. As far as conventional BFO is concerned, the elimination and dispersal is determined by a given constant P_{ed}. During random generated probability $P<P_{ed}$ is satisfied, the bacterium will get placed by a random location within the optimization domain, if not it stays where it is. The above said process is represented in the Equation 4 as:

$$\begin{aligned} &If\ P < P_{ed}, then \\ &\theta_i(j,k,l) = rand\ X\ (\max - \min) + \min \\ &else \\ &\theta_i(j,k,l) = \theta_i(j,k,l) \end{aligned} \quad \ldots (4)$$

Where max and min are the upper and lower boundary of the optimization problem, and rand∈[0, 1].

Feature Selection using IBFO

The finest set of features will certainly improve the performance of classification. The cost allocated to each features in a precise iteration and variable represents the momentous performance of it in the optimization process over and above the appearance of it in population. In point of fact that, two indexes: 'Archive' (Arc) and 'Cost' (CT) are predefined to save the incidence of features and implication of features, respectively. The matrix Arcis used to save usage of features in population. While the vector CT is made use of imitate the performance of features. Perceptibly, the unobserved features need to be allocated with a higher probability in favoring the prospective investigation while the features already being selected would have a lower probability, which will be realized with the help of Arc. Furthermore, as far as the classification performance is concerned, the features pertaining to elevated classification accuracy would be assigned with a higher CT. The chemotaxis process in bacterial based algorithms includes two independent modes: 'Running' and 'Tumbling'. Two different cost vectors CT and Arc are defined to allot the features with illustrious significance indexes to portray the selection process both in 'Running' and 'Tumbling'. Such tactics have been incorporated over the bacteria based algorithms for feature selection. It is presumed that the total number of features is H. The performance scores used to record the performance of features in each particle are defined as $CT = \{CT_{f1}, ...CT_{fH}\}$. At first, all the costs representing the performance are equal to 0, $CT_0 = \{0, ...0\}$. At the time of performing optimization, the performance scores of features get updated ad infinitum according to the certain statute. The evidence of the higher performance of feature is obtained in the classification performance after the feature is added to the candidate subset. Moreover, during the feature is added in the population by replacing one of feature, then the performance of classification gets increased. At that point of time, it may be foreseens that the newly added feature probably obtains better performance than the older one, and the cost of the newly added feature needs to be assigned with higher CT than the beforehand replaced one.

Based on the conditions said above, the cost mentioning the performance of feature in each individual are obtained. The fitness is usually made use for evaluating the performance which ranges from 0 to 1 i.e. error rate. Consequently, the smaller value of error rate means the better performance of the feature combination. When the accumulation of the i[th] feature f_i in the m[th] bacterium, the current error rate $F_{(f_i,m)}$, then the cost of i[th] feature in the m[th] bacterium will be increased as given in the Equation 5. Otherwise, if the adding feature f_i results in performance decreasing, the costs of i[th] feature in that bacterium will be decreased as given in the Equation 6.

$$If\ F_{(F_i,m)} < Fit_{(f_i,m)}\ then$$

$$CT(f_i,m) = CT(f_i,m) + \frac{|Fit(f_s,m) - F(f_s,m)|}{Fit(f_s.m)} \quad \ldots (5)$$

Else

$$CT(f_i,m) = CT(f_i,m) - Fit(f_s,m) * |Fit(f_s,m) - F(f_s,m)| \quad \ldots (6)$$

In view of the fact that the performance is evaluated by classification error rate that ranges from 0 to 1, the increasing point of the costs be liable to be superior than decreasing point as the fitness is lesser than 1, that needs to be decreased due to the overwhelming removing of features from the candidates. The quality of features needs to be evaluated by the number times that the feature has been selected in good subsets. The parameter Q_{fi} is made use for pointing to the distribution of the features within the total population, and the CT will be updated using Equation 7 and Equation 8.

$$Q_{fi} = \left(\frac{G_i}{G_i + B_i}\right) / \max_j \left(\frac{G_j}{G_j + B_j}\right) \quad \ldots (7)$$

$$CT(f_i,m) = CT(f_i,m) + Q_{fi} \quad \ldots (8).$$

where m = 1,…, NoP, and NoP is the number of population. G_i is the number of times that the i[th] feature has been employed in the subsets whose classification error is smaller

than the average fitness that represents better performance, and B_i is the number of times that the i^{th} feature has been employed in the subsets whose performance is lower than the average level.

Algorithm for Cost Mechanism based on Performance

For each bacterium

If classification error rate of is smaller than previous $F_{(f_i,m)}$ after adding the k^{th} feature and removing of the s^{th} feature, **then**

The cost of k^{th} feature in m^{th} bacterium is updated using Equation 5

The cost of s^{th} feature in m^{th} bacterium is updated using Equation 6

//update the CT in the m^{th} bacterium by comparing the fitness function

For all bacteria

Evaluate the classification error rate (fitness) of all bacteria, and calculate the quality of features according to Equation (7)

Update the performance ranking matrix W using Eq. (8).

// update the matrix W according to feature occurrence in better variables

Twin Support Vector Machine Classifier

TSVM is used in this research work for risk level classification of CAHD problem that relaxes the requirement that the hyperplanes are parallel in conventional SVM,and aims to seek a pair of nonparallel proximal hyperplanes using the equation (9)

$$f_1(x): w_1'x+b_1=0 \text{ and } f_2(x): w_2'x+b_2=0 \ldots (9)$$

such that each is closer to its own class and is as far as possible from the other, where w_1, w_2 and $b_{1,} b_2$ are the common vectors and bias terms in the equation of the above mentioned two hyperplanes, respectively.

As like conventional SVM it is presumed that the matrix $X_1 \in R^{m\times n}$ as the labeled data belonging to "þ1" class, and $X_2 \in R^{m_2\times n}$ as the labeled data belonging to "+1" class, where $m_1+m_2=m$. Hence in order to acquire the above two proximal hyperplanes present in the equation (9), the optimization problems for TSVM will be coined as

$$\min \frac{1}{2} \| X_1 w_1 + e_1 b_1 \| 2 + c_1 e'_2 \xi \quad \ldots (10)$$
$$s.t \quad -(x_2 w_1 + e_2 b_1) + \xi \geq e_2, \qquad \xi \geq 0,$$

and

$$\min \frac{1}{2} \| X_2 w_2 + e_2 b_2 \| 2 + c_2 e'_1 \eta \quad \ldots (11)$$
$$s.t \quad -(x_1 w_2 + e_1 b_2) + \eta \geq e_1, \qquad \eta \geq 0,$$

where c1and c2 are the penalty parameters, and ξ,η are the slack vectors. It can be observed that the first term in the objective function present in the equation (10) is used to craft "+1" labeled occurrences proximate to the hyperplane $w'_1 x + b_1 = 0$, while the second term and constraints force "- 1" labeled instances bounded in the hyperplane $w'_1 x + b_1 = -1$. In order to obtain the solutions of problems (10) and (11), it is derived the CAHD classification as dual problems as

$$\max_{\alpha} e'_2 \alpha - \frac{1}{2} \alpha' G (H'H)^{-1} G' \alpha \quad (12)$$
$$s.t \ 0 \leq \alpha \leq c_1 e_2.$$

and

$$\max_{\beta} e'_1 \beta - \frac{1}{2} \beta' H (G'G)^{-1} H' \beta \quad (13)$$
$$s.t \ 0 \leq \beta \leq c_1 e_1.$$

where H=[A e_1] ,G=[B e_2] and j=[Xe]. By observing (12) and (13), it can be observed that TSVM provides a solution to the pair of smaller sized rather than a large one which is present in the conventional support vector machine. Once the solutions α and β of problems (12) and (13) are obtained, the nonparallel proximal hyperplanes (9) can be coined by

$$\begin{bmatrix} w_1 \\ b_1 \end{bmatrix} = -(HH)^{-1} G' \alpha \text{ and } \begin{bmatrix} w_2 \\ b_2 \end{bmatrix} = -(GG)^{-1} H' \beta \quad (14)$$

A new unseen instance x is assigned to label "+1" or "- 1", depending on which of the proximal hyperplanes (9) it lies closer to.

RESULTS

Dataset

The UCI (University of California, Irvine) – ML (Machine Learning) - PIMA dataset [30] contains 768 data samples and 8 numerical features per sample. All patients are females at least 21 years old of Pima Indian heritage. The variables are allocated into two classes. The first class is labelled as "negative to diabetes and CAHD" that involves 500 samples and the remaining 268 samples is labelled as "positive to diabetes and CAHD".

Performance Metrics

- Accuracy is the performance metric that projects the classification correctness of the classifier.
- Elapsed time is the performance metric that gives the total time taken for execution (testing).

Results

True positive, true negative, false positive, false negative, accuracy, elapsed time are given in the Table 1.

Table 1. Results

Algorithm	**True Positive**	**True Negative**	**False Positive**	**False Negative**	**Accuracy (%)**	**Elapsed Time (milliseconds)**
Neuro Fuzzy Classifier [18]	263	496	6	3	98.82813	223320
IBFO-SVM	263	497	7	3	98.69792	38475
IBFO-TSVM	262	496	***6***	***2***	***98.95833***	***21894***

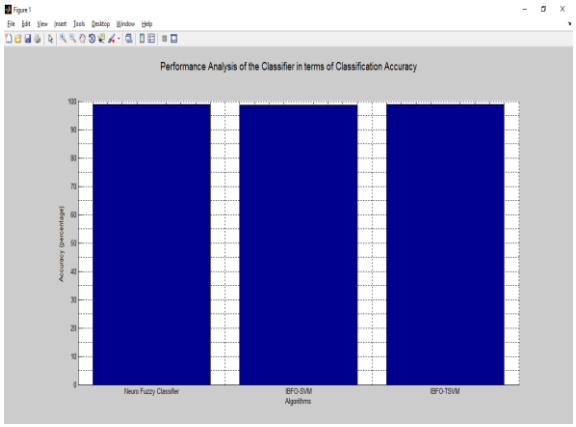

Figure 1. Classification Accuracy Analysis

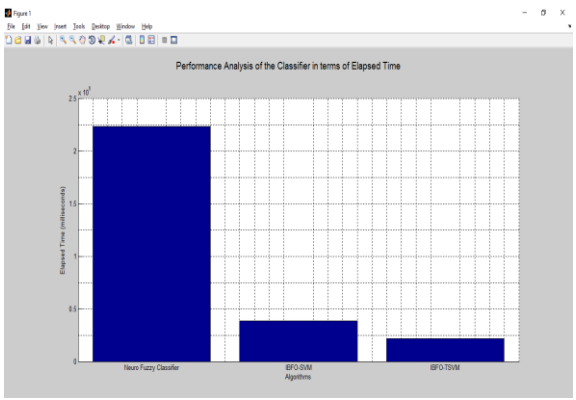

Figure 2. Elapsed Time Analysis

It is noteworthy that out of 268 CAHD patients, the proposed IBFO-TSVM correctly classified 263 patients. The accuracy of the classifier is also slightly increased. From the results it is evident that the proposed IBFO-TSVM consumes less time to perform the overall classification task during the testing phase. The proposed classifier takes very less time i.e., one tenth of the time when compared with the Neuro Fuzzy Classifier which is our previous work. The simulation results are presented in the Fig.2.

CONCLUSION

This research work mainly focuses on proposing an effective feature selection. It is noteworthy that efficient feature selection strategy will certainly contribute for the effectiveness of the classifier in terms of elapsed time. In this paper an improved bacterial foraging optimization strategy is used for performing the feature selection task. Twin support machine is chosen for performing the task of classification. Dataset is obtained from the UCI machine learning repository and the results projects that the proposed IBFO-TSVM obtains better classification accuracy and reduced elapsed time.

REFERENCES

[1] Patil, S. B., & Kumaraswamy, Y. S.: Intelligent and Effective Heart Attack Prediction System using Data Mining and Artificial Neural Network. European Journal of Scientific Research. 31, 642–656 (2009)

[2] Resul, D., Ibrahim, T., & Abdulkadir, S.: Effective Diagnosis of Heart Disease through Neural Networks Ensembles. Expert Systems with Applications. 36, 7675–7680 (2009)

[3] Ture, M., Kurt, I., & Kurum, A. T.: Comparing Performances of Logistic Regression Classification and Regression Tree and Neural Networks for Predicting Coronary Artery Disease. Expert Systems with Applications. 34, 366–374 (2008)

[4] M.C., Shin, D., Shin, D.: A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms. In: Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, pp. 183-187 (2009)

[5] Andreeva, P.: Data Modeling and Specific Rule Generation via Data Mining Techniques. In: Proc. International Conference on Computer Systems and Technologies. pp. 17–23 (2006)

[6] Palaniappan, S., Awang, R.: Intelligent Heart Disease Prediction System using Data Mining Techniques. In: Proc. of IEEE/ACS International Conference on Computer Systems and Applications, pp. 108–115 (2008)

[7] Tsipouras, M. G., Exarchos, T. P., Fotiadis, D. I., Kotsia, A. P., Vakalis, K. V., Naka, K.K., Michalis, L. K.: Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling.

IEEE Transactions on Information Technology in Biomedicine. 12, 447-458 (2008)

[8] Adeli, A., Neshat, M.: A Fuzzy Expert System for Heart Disease Diagnosis. In: Proc. International Multiconference of Engineering and Computer Scientists. pp. 134–139 (2010)

[9] Grzymala-Busse, J.W., Hu, M.: A comparison of several approaches to missing attribute values in data mining. In: Second International Conference on Rough Sets and Current Trends in Computing. pp. 378–385 (2012)

[10] Su, X., Khoshgoftaar, T.M., Greiner, R.: Using Imputation Techniques to Help Learn Accurate Classifiers. IEEE International Conference on Tools with Artificial Intelligence. pp. 437-444 (2008)

[11] Weihong, W., Li, Q., Han, S., Lin, H.: A Preliminary Study on Constructing Decision Tree with Gene Expression Programming. In: Proc. First International Conference on Innovative Computing Information and Control. pp. 222–225 (2006)

[12] Rokach, L., Maimon, O.: Data Mining with Decision Trees Theory and Applications. World Scientific Publishing (2008)

[13] P. Cao, X. Liu, J. Zhang, D. Zhao, M. Huang, O. Zaiane, "ℓ2,1 norm regularized multi-kernel based joint nonlinear feature selection and over-sampling for imbalanced data classification," Neurocomputing, vol. 234, pp. 38-57, 2017.

[14] M. Al-Rajab, J. Lu, Q. Xu, "Examining applying high performance genetic data feature selection and classification algorithms for colon cancer diagnosis," Computer Methods and Programs in Biomedicine, vol. 146, pp.11-24, 2017.

[15] A. Deniz, H. E. Kiziloz, T. Dokeroglu, A. Cosar, "Robust multiobjective evolutionary feature subset selection algorithm for binary classification using machine learning techniques," Neurocomputing, vol. 241, pp. 128-146, 2017.

[16] Y. Zhang, C. Yang, A. Yang, C. Xiong, X. Zhou, Z. Zhang, "Feature selection for classification with class-separability strategy and data envelopment analysis," Neurocomputing, vol. 166, pp. 172-184, 2015.

[17] X. Cheng, Y. Zhu, J. K. Song, G. Wen, W. He, "A novel low-rank hypergraph feature selection for multi-view classification," Neurocomputing, vol. 253, pp. 115-121, 2016.

[18] R. Rajkumar, K. Ananadakumar, A. Bharathi, "Risk Level Classification of Coronary Artery Heart Disease in Diabetic Patients using Neuro Fuzzy Classifier," International Journal of Computational Intelligence Research, vol. 13, pp. 575-582, 2017.