JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A METHOD FOR FINDING AND ELIMINATING OUTLIERS IN DATA MINING CLASSIFICATION TASKS

Karuna C. Khobragade[1], Pankaj B. Dhumane[2]

[1]Assistant Professor, Computer Science Department, Dr. Khatri Mahavidyalaya, Tukum, Chandrapur, Maharashtra, India 442401

[2]Assistant Professor, Computer Science Department, Sardar Patel Mahavidyalaya, Chandrapur, Maharashtra, India 442401

Abstract: Data mining is technique used to extract interesting data patterns from datasets. A vital piece of data mining is outlier detection, that detects objects exhibiting behaviour different from the rest. One common issue in data mining field is identifying outliers from a set of patterns. An outlier can significantly affect the result. Thus, in order to avoid skewing, outliers must be eliminated or resolved prior to the analysis. Data mining seeks to uncover patterns within data that deviate from expected behaviour in order to detect outliers. In this paper, we have suggested a system that uses the widely-used statistical technique known as IQR to find out outliers and use the Winsorizing method to handle them. Under this architecture, the pre-processed dataset is used to create RBFN that is prepared by teaching a learning-based optimisation model. As a result of this study, our proposed framework may serve as a useful substitute tool for data mining classification tasks where outliers preprocessing is required.

Keywords: Outlier Detection, Data Patterns, Winsorizing, RBNN, PCA and IQR

Introduction: Data mining's classification work is handled as supervised learning in machine learning (ML). To build a machine learning classification method, a precise dataset for training is required, however, there is a chance of error when allocating raw data to the right class. One of the most important elements of the data set in the classification process that represents an unusual case that could have a significant influence on the classifier's accuracy [1]. Outliers are frequently found in real-world data [2,5]. With so many variables being collected, outliers can be detected in any statistical data. Any value that deviates significantly from the rest of the data is called an outlier. Errors in data processing, measurement, sampling, data gathering, and volatility in our data can all lead to outliers. Outliers are emerging among a range of sources and conceal themselves in a range of proportion during the creation, gathering, processing as well as an analysis of data. While outlier provide significant challenges for statistical analysis, they can also be useful in gathering important data related to our research work. The process of extracting specific information or data vector from a larger data set is known as "outlier detection and elimination," and it is essential in almost all quantitative fields, including ML. In any measurable field, including machine learning, quality of information is equally as crucial as the categorisation model's quality. So it is necessary to recognise and deal with them.

There are two main categories of outlier detection techniques:

1) Non-parametric and
2) Parametric.

Non-parametric methods include proximity-based modelling and linear regression while the parametric approaches contain Statistical and probabilistic modelling. In recent years, numerous methods for identifying outliers have been proposed. In order to find outliers for smart farming, inter-quartile range technique has been used[6].

In the present work, IQR statistical technique has been utilised for identify outliers in the data set and address them through the Winsorizing procedure. RBFN may evaluate any without a break multidimensional function for classification problems. To enhance the appearance of RBFNs during training, the dispersion of networks and centre of gravity should be made visible. Rapid input selection is used in this work to eliminate unnecessary neural inputs before creating an effective RBFNN model. In order to create the optimal model, the parameters of the RBFN are adjusted empirically through Teaching-Learning-Based-Optimization (TLBO).

RBFN classifier:

The feed-forward network, referred to as the RBFN, is the network most frequently used for function approximation and offers some advantages over a multilayer neural network.
1. First layer: feeds the input neurons,
2. Second layer: creates a radial activation function using hidden units,
3.Third layer: induces the output applying weighted sum hidden units. The RBF network uses supervised learning to build new and potentially useful models.as shown in fig.1

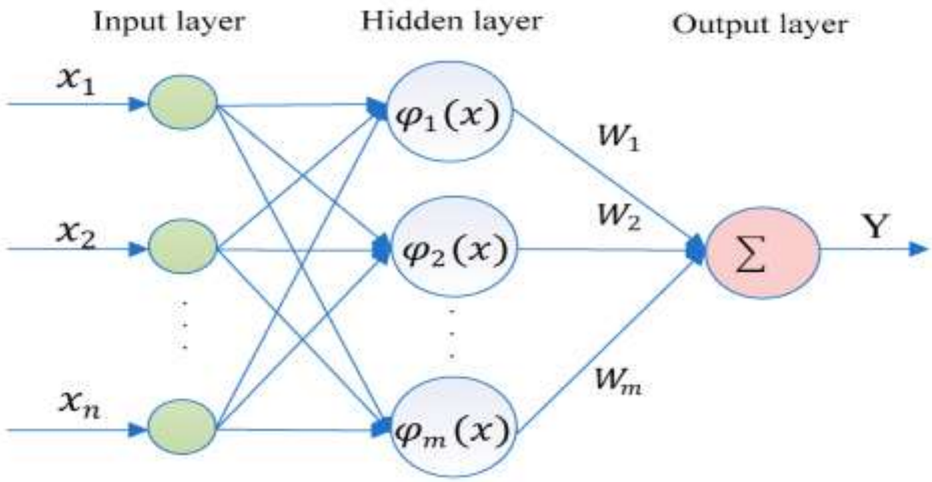

Fig.1 Radial basis neural network

Using a Gaussian kernel, the second layer carries out nonlinear transformation, as stated by the following expression:

$$\phi^i(x) = \exp\left(\frac{||x-\mu^2||}{2\sigma^2 i}\right)^2 \quad (1)$$

where, μi is center, σi is spread and ϕi an output for the ith neuron. Through the interconnection of intermediate and last layers a Weighted connections wi are formed. Summing unit, known as an output layer, provides reaction of network to external environment.

Final Processing unit, also referred to as the summing unit, gives network's reaction to external stimuli.

Outlier detection: To find outliers, many techniques are employed [11, 12]. A technique that is (similar) comparable to the Histogram, Boxplot, Scatter Plot, and IQR approaches is visualisation. In handling outliers, there are two approaches:

i) **Managing Error Outliers:** All errors in this approach are expected to be resolved or discarded. To prevent major data loss from deletion, the original record should be preserved whenever it includes important information. If data contain any error, then the good technique is to just delete those entries.

ii) **Managing Non-Error Outliers:** Non-error outliers can handled using one of three techniques: keep, delete and recode.

a) **Keep:** We should be mindful that outliers have the potential to skew the outcomes of our actual work when we keep them.

b) **Delete:** Simplest solution is to remove any observations that are unexpected or abnormal.
c) **Recode:** Use the Winsorizing method to handle outliers and prevent significant data loss.

Principal Component Analysis (PCA): PCA serves to minimise dimensionality of large-scale data or to compress many attributes into fewer ones[13]. Suppose a data matrix X with nXq, q column representing a particular attribute and a row denoting a distinct test repetition.

The transformation is specified mathematically by a set of x-dimensional coefficients or weights $C_{(k)} = (C_1, \ldots, C_q)_{(k)}$ having size of who change the vector row $r_{(i)}$ to $s_{(i)} = (s_1, \ldots, s_l)_{(i)}$ such that $s_{m(i)} = r_{(i)} * C_{(m)}$, where $i = 1, \ldots, n$ and $m = 1, \ldots, l$.

a) Component First:

For maximize, $c_{(1)}$ has to be satisfied,

$$C_{(l)} = \arg \max_{\|c\|=1} \left\{ \sum_i (s_1)^2_{(i)} \right\} = \arg \max_{\|c\|=1} \left\{ \sum_i \left(rx_{(i)} * c \right)^2 \right\} \quad (2)$$

In matrix form we write it as

$$C_{(l)} = \arg \max_{\|c\|=1} \{\|RC\|^2\} = \arg \max_{\|c\|=1} \{C^T R^T RC\} \quad (3)$$

$(C)_1$ also satisfies the criteria since it is a unit vector.

$$C_{(l)} = \arg\max \left\{\frac{C^T R^T RC}{C^T C}\right\} \quad (4)$$

A score for a data vector's first main component can be found in the converted coordinates as follows: $s1_{(i)} = r_{(i)} * C_{(l)}$

b) Other Components :

By dividing the original m-1 principal components by R, the mth component is calculated.

$$\hat{R}_m = R - \sum_{t=1}^{m-1} R\, C_{(t)} C_{(t)}^T \quad (5)$$

$$C_{(m)} = \arg \max_{\|c\|=1} \left\{ \left\| \hat{R}_m C \right\|^2 \right\} = \arg\max \left\{\frac{C^T \hat{R}_m{}^T \hat{R}_m C}{C^T C}\right\} \quad (6)$$

The rest of the eigen vectors of $R^T R$, as well as their associated eigen values, which provide maximum values, are thus produced. Thus, the weight matrices or vectors are eigen vectors of $R^T R$.

The expression $s_{m(i)} = r_{(i)} C_{(m)}$ represents the mth primary component of $r(i)$.The analogous vector of the original variables is $C_{(m)}$ in translated Spatial references, or $C(m)$serves as an eigen vector for $R^T R$ as well as original variables' corresponding vector and $r_{(i)} * C_{(m)} C_{(m)}$.

Thus, the entire principal or primary component breakdown of X given by:

$$S = R * C \quad (7)$$

Teaching–learning based optimization: R.V. Rao et al. (2011) presented TLBO, the population-driven meta-heuristic method of optimization that mimics a classroom setting to maximize a given objective function. In a classroom, an educator works hard to educate each learner in the class [14].

There are two stages to TLBO's operation: a) Teacher phase b) Student phase

a) Teacher Phase: Using their skills, the instructor attempts to raise an average of the class in that subject which they teach. Assume that there are "S" subjects, "L" learners, and Np,q, the mean result of students in particular subjects "p" at given iteration q. The most desirable comprehensive outcome is $Avg_{total1} = pbest_{1.1}$. The sum of all the subjects learnt by the entire student body could be regarded as an influence of most successful student, k-best. But typically a educator is regarded as extremely educated somebody who guides students to achieve better outcomes, the algorithm takes into account the top learner as the instructor. The following formula calculates the discrepancy among teacher's corresponding outcome for every subject and current average performance for each topic:

$$\text{Difference1\_Avg1}_{spl}=ri(Avg_{s,pbest1}- T_F N_{sl}) \quad (8)$$

Where $Avg_{s,pbest1}$ indicates best learner. The average value can be determined as T_F which is nothing but Teaching factor and ri tends to stochastic value from 0 to 1. Then T_F value calculation is given by the equation (9) given below

$$T_F= \text{round}\ [rand(0,1)\{2-1\} +1] \quad (9)$$

Following a number of tests, it is shown that the approach works best when T_F falls between {1,2}. To make the process simpler, the T_F must be between 1 to 2, according to rounded up Standards given by Eq. (9), so an algorithm works a lot better whenever T_F value is any of 1 or 2.

b) Student Phase: To expand their knowledge, students randomly connect to another students. The student picks up new information if another student knows more than they do. The processes of learning that took place throughout this time are given below.
Consider two students, X and Y, at random so that $T'_{total1=X,l} \neq T'_{total1=Y,l}$
(where $T'_{total1=X,l}$ and $T'_{total1=Y,l}$ are updated values of function $T_{totaal1=X,l}$ $T_{totaal1=Y,l}$ of X and Y, respectively). Minimisation problems are solved using Equations (10) and (11) respectively shown below.

$$T''_{m,X,l}=T'_{m,X,l} + ri(T'_{m,X,l} - T'_{m,Y,l}),\ If\ T^{l}_{total1-X,l} < T^{l}_{total1-Y,l} \quad (10)$$

$$T''_{m,X,l}=T'_{m,X,l} + ri(T'_{m,Y,l} - T'_{m,X,l}),\ If\ T^{l}_{total1-Y,l} < T^{l}_{total1-X,l} \quad (11)$$

If it provides a greater feature value, $P''_{m,x,l}$ is acceptable.
Similarly, Maximisation problems are solved using Equations (12) and (13) as shown below.

$$T''_{m,X,l}=T'_{m,X,l} + ri(T'_{m,X,l} - T'_{m,Y,l}),\ If\ T^{l}_{total1-Y,l} < T^{l}_{total1-X,l} \quad (12)$$

$$T''_{m,X,l}=T'_{m,X,l} + ri(T'_{m,Y,l} - T'_{m,X,l}),\ If\ T^{l}_{total1-X,l} < T^{l}_{total1-Y,l} \quad (13)$$

Here we are concentrating only on creating centers and spreads, even if TLBO is used for the creation of widths, centers, weights etc. which simultaneously join resultant nodes and kernel. This ensures that TLBO individually represented effectively. The individual's length and the search space grow too big if all of these parameters are kept, which leads to an extremely slow rate of convergence. For stochastic search we merely encode the widths & centers into an individual because the performance of RBFNs is mainly dictated by widths & centers of kernel.
if Kmax be the maximum number of kernel then individual structure can be representd as shown in following Figure 2.

Center				Width				Bias
C1	C2		C_{Kmax}	W1	W2		W_{Kmax}	B

Fig. 2

As per the Fig.2, it shows that each one individual is consist of three parts and they are centers, widths and bias. The particular person is $2K_{max}+1$.
The equation of fitness which guides the search is defined by Eq. (14) shown below

$$f(x) = \frac{1}{I}\sum_{i=1}^{I}\left(OT_n - \phi(\vec{x}_i)\right)^2 \quad (14)$$

Here I indicates total number of individual, an actual output is denoted by OTn, and the expected output is denoted by $\Phi(\vec{x}_i)$. Once the centres and widths are determined, the second learning phase's weight calculation method reduces to computing an intuitive linear solution. In this work, a setoff optimal weight is determined using the pseudo inverse technique.

Since the training data is used to determine the centres, widths, and bias, so these can be written as follows:

$$D=(WT) * \Phi$$

$$\text{Where } WT=(\Phi^T \Phi)^{-1} \Phi^T Y \qquad (15)$$

Detection of Outlier using IQR: By using this technique, one can determine the variation in the data by calculating the Inter Quartile Range. Any result outside the limit (25th percentile - 1.5x interquartile range) to the (75th percentile + 1.5x interquartile range) is considered an outlier, with the definition of IQR being the 75th percentile - 25th percentile [9,15].

Winsorizing is a technique that minimise outlier's impact in given data set, after the outlier has been detected. When each variable's distribution is winsorized, the score of the kth percentile alone is used to replace any number of a variable that is above or below that percentile. This process applies to the distribution of all the variables. In our project, all of the outliers found by the IQR technique were recoded using winsorizing. All the values or numbers which are less than (25^{th} percentile-1.5*IQR) is replaced by (25^{th} percentile-1.5*IQR) while the values which are more than (75^{th} percentile + 1.5*IQR) is substituted by(75^{th} percentile + 1.5*IQR).

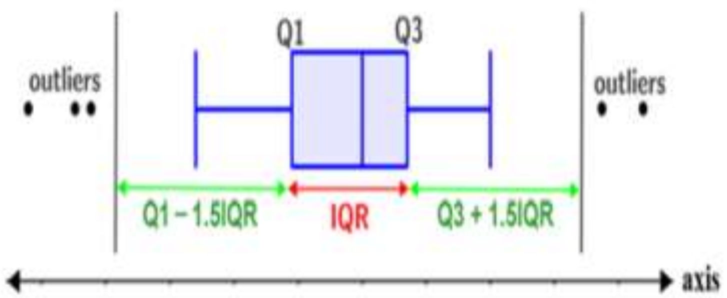

Fig: 3 IQR Method for Outlier Detection

PCA-based Dimensionality Reduction: By projecting each statistic to the most efficient initial several main elements, the PCA reduces dimensionality and produces lower-dimensional data containing as much variation as is practical[13].

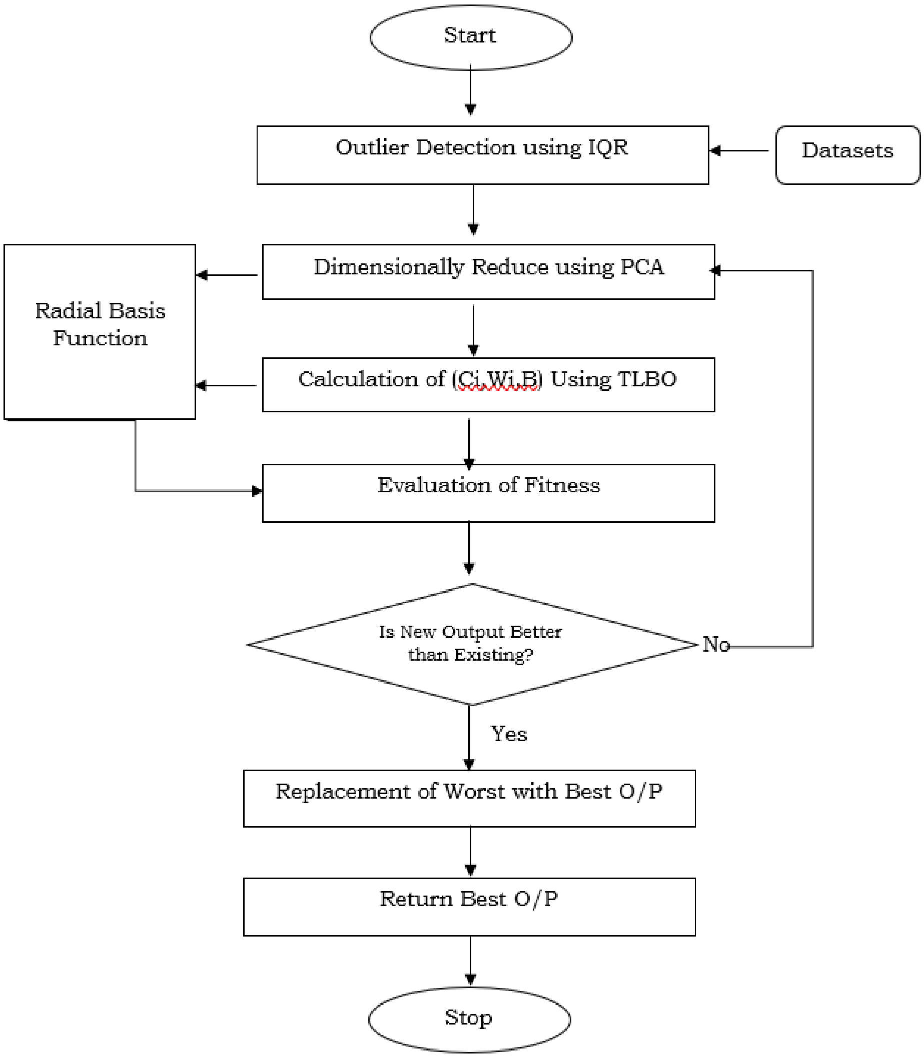

Fig: 4 Flowchart of Proposed Process

Literature Review:

We will examine a few of the works that are closely related to this one in this part. The performance of the TLBO approach on a scheduling problem was examined by Baykasoğlu et al. [16]. It was observed that TLBO is more effective than another heuristic method. The TLBO algorithm can be utilised to maximise the coordination of Directional Over Current Relay(DOCR) relays inside loop structure of power, as demonstrated by Singh et al. [17]. By altering the target function, they reduce the amount of time required for the primary as well as emergency relays to function. The outcomes were compared to the Plug setting and time dial setting ideal values determined by the modified differential evolution algorithm. Machine learning techniques have been applied by Rao et al. [18] for the parameter optimisation procedure. In their study, they have utilised two distinct multi-objective issues. Their results demonstrate that the new strategy improved the population size, generation number, and computation time. TLBO has been suggested by Kiziloz et al. [19] as an FSS (Feature Subset Selection) process. Using this method, no parameter needs to be adjusted all over the optimisation process. The efficiency of the majority of conventional meta-heuristic algorithms may suffer from the additional work needed to adjust their parameters. A novel method for concurrent selection of project having an aim for optimising overall expected benefit of portfolio was proposed by Kumar et al. [20]. In comparison to the current algorithms, three algorithms were developed: TLBO, TS (Turbo Search), and hybrid TLBO-TS[20]. An effectiveness of the algorithms is evaluated on several data sets. To determine the ideal weights for FLANN model, Naik et al. used TLBO refined using gradient descent conditioning [21]. To solve challenges in mechanical design, Rao et al. introduced TLBO[22]. When compared to other optimisation strategies, the TLBO strategy performs better based on the optimal solution, integration rate, and computational effort. The results suggest that for the mechanical structure or design optimisation problems discussed, TLBO outperforms alternative optimisation techniques in terms of effectiveness and efficiency. Rao and colleagues formulated a huge nonlinear optimisation problem to find global solutions[23]. The suggested approach is predicated on the observation that teachers have an impact on students' performance in the classroom. A range of benchmark issues with different features are used to assess the method's efficiency, and the results is contrasted by those derived from different public-based techniques. Rao et al. have previously Contributed to both constrained as well as unconstrained tasks [24]. The method they used focusses on the effect that a teacher has on the work that students produce in the classroom. The method is tested on 35 restricted and 25 unrestricted benchmark functions with a variety of features. Efficiency derived from proposed algorithm is shown more accurate than TLBO approach when compared to other optimisation strategies already in use. Rao et al. introduced the TLBO approach along with its non-dominated and dominated multi-objective variants. A multi-objective constrained problem and two sample benchmark method, one set with constraints and one without are used to illustrate the algorithm[25]. By merging TLBO and RBF networks, Dash et al. [26] created a new method for creating a model for checking missing values and unrelated features. In order to address the no-Gaussian problem, Guo et al. suggested a novel method for managing a nonlinear data set that makes use of PCA and RBFNN. This proposed PCA - RBFNN method considered as robust due to its useful addition Compared to the linear PCA method [27]. Rousseeuw et al. have presented a number of robust methods and outlier detection tools [28]. The interquartile range method was used by Vinutha et al. in to identify outliers. This method divides the input range into continuous scale quartiles, which are then examined to identify outliers. After the outliers have been gathered, an remove / eliminate with value filter is used[29]. To measure the weight of multiple indexes in multivariate data and determine the influence of different attributes on forecast results, Yang et al. initially created one novel index weight evaluation method in conjunction with information entropy [30].

Result and Conclusion: A technique for classifying data has been proposed in this study. Initially, the task of preprocessing such as detecting and handling an outlier is take place. The statistical technique of IQR is used for identifying or detecting outliers present in data. Winsorizing method then applied for normalise an outlier data. Secondly, PCA is subsequently applied to reduce dimensionality. Finally Developing an RBFN-based classifier that has been trained using TLBO.

References:

[1] S. Haykin, Neural Networks and Learning Machines, third ed., Pearson Education, Inc., Upper Saddle River, New Jersey, 2008.

[2] V. Hodge, J. Austin, A survey of outlier detection methodologies, Artif. Intell. Rev. 22 (2) (2004) 85–126.

[3] C.C. Aggarwal, P.S. Yu, Outlier detection for high dimensional data, ACM SIGMOD Rec. 30 (2) (2001) 37–46.

[4] V. Hodge, J. Austin, A survey of outlier detection methodologies, Artif. Intell. Rev. 22 (2) (2004) 85–126.

[5] R.R. Wilcox, Outlier Detection. Encyclopedia of Statistics in Behavioral Science, John Wiley & Sons, Ltd, 2005.

[6] S.K.S. Durai, M.D. Shamili, Smart farming using machine learning and deep learning techniques, Decis. Anal. J. 3 (2022) 100041

[7] Y. Kuvvetli, M. Deveci, T. Paksoy, H. Garg, A predictive analytics model for COVID-19 pandemic using artificial neural networks, Decis. Anal. J. 1 (2021) 100007.

[8] P. Chhajer, M. Shah, A. Kshirsagar, The applications of artificial neural net works, support vector machines, and long–short term memory for stock market prediction, Decis. Anal. J. 2 (2022) 100015.

[9] M. Seyedan, F. Mafakheri, C. Wang, Cluster-based demand forecasting using Bayesian model averaging: An ensemble learning approach, Decis. Anal. J. 3 (2022) 100033.

[10] M. Bansal, A. Goyal, A. Choudhary, A comparative analysis of K-nearest neigh bour, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning, Decis. Anal. J. (2022) 100071.

[11] H. Wang, M.J. Bah, M. Hammad, Progress in outlier detection techniques: A survey, IEEE Access 7 (2019) 107964-108000.

[12] H.P. Kriegel, P. Kröger, A. Zimek, Outlier detection techniques, Tutor. KDD 10 (2010) 1–76.

[13] https://en.wikipedia.org/wiki/Principal_component_analysis.

[14] Teaching Learning based Optimization (TLBO) - GeeksforGeeks

[15] D.L. Whaley III, The Interquartile Range: Theory and Estimation (Doctoral dissertation), East Tennessee State University, 2005.

[16] A. Baykasoğlu, A. Hamzadayi, S.Y. Köse, Testing the performance of teaching learning based optimization (TLBO) algorithm on combinatorial problems: Flow shop and job shop scheduling cases, Inform. Sci. 276 (2014) 204–218.

[17] M. Singh, B.K. Panigrahi, A.R. Abhyankar, Optimal coordination of directional over-current relays using teaching learning-based optimization (TLBO) algorithm, Int. J. Electr. Power Energy Syst. 50 (2013) 33–41.

[18] R.V. Rao, V.D. Kalyankar, Parameters Optimization of Advanced Machining Processes using TLBO Algorithm 20, EPPM, Singapore, 2011, pp. 21–31.

[19] H.E. Kiziloz, A. Deniz, T. Dokeroglu, A. Cosar, Novel multiobjective TLBO algorithms for the feature subset selection problem, Neurocomputing 306 (2018) 94–107.

[20] M. Kumar, M.L. Mittal, G. Soni, D. Joshi, A hybrid TLBO-TS algorithm for integrated selection and scheduling of projects, Comput. Ind. Eng. 119 (2018) 121–130.

[21] B. Naik, J. Nayak, H.S. Behera, A TLBO based gradient descent learning functional link higher order ANN: An efficient model for learning from non-linear data, J. King Saud Univ.-Comput. Inf. Sci. 30 (1) (2018) 120–139.

[22] R.V. Rao, V.J. Savsani, D.P. Vakharia, Teaching–learning-based optimization: a novel method for constrained mechanical design optimization problems, Comput. Aided Des. 43 (3) (2011) 303–315.

[23] R.V. Rao, V.J. Savsani, D.P. Vakharia, Teaching–learning-based optimization: an optimization method for continuous non-linear large-scale problems, Inform. Sci. 183 (1) (2012) 1–15.

[24] R.V. Rao, V.J. Savsani, J. Balic, Teaching–learning-based optimization algorithm for unconstrained and constrained real-parameter optimization problems, Eng. Optim. 44 (12) (2012) 1447–1462.

[25] R.V. Rao, Teaching-learning-based optimization algorithm, in: Teaching Learning Based Optimization Algorithm, Springer, Cham, 2016, pp. 9–39.

[26] C.S.K. Dash, A.K. Behera, S. Dehuri, S.B. Cho, Building a novel classifier based on teaching learning based optimization and radial basis function neural networks for non-imputed database with irrelevant features, Appl. Comput. Inform. 18 (1/2) (2020) 151–162.

[27] Y. Guo, K. Li, Z. Yang, J. Deng, D.M. Laverty, A novel radial basis function neural network principal component analysis scheme for PMU-based wide-area power system monitoring, Electr. Power Syst. Res. 127 (2015) 197–205.

[28] P.J. Rousseeuw, M. Hubert, Robust statistics for outlier detection, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 1 (1) (2011) 73–79.

[29] H.P. Vinutha, B. Poornima, B.M. Sagar, Detection of outliers using interquartile range technique from intrusion dataset, in: Information and Decision Sciences, Springer, Singapore, 2018, pp. 511–518.

[30] Y. Yang, C. Fan, L. Chen, H. Xiong, IPMOD: An efficient outlier detection model for high-dimensional medical data streams, Expert Syst. Appl. 191 (2022) 116212.