Turkish Journal of Electrical Engineering & Computer Sciences

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

Research Article

Turk J Elec Eng & Comp Sci
(2017) 25: 3409 – 3420

doi:10.3906/elk-1606-122

A fast feature selection approach based on extreme learning machine and coefficient of variation

Ömer Faruk ERTUĞRUL[1,*], **Mehmet Emin TAĞLUK**[2]

[1]Department of Electrical and Electronics Engineering, Batman University, Batman, Turkey
[2]Department of Electrical and Electronics Engineering, İnönü University, Malatya, Turkey

Received: 08.06.2016 • **Accepted/Published Online:** 02.02.2017 • **Final Version:** 30.07.2017

Abstract: Feature selection is the method of reducing the size of data without degrading their accuracy. In this study, we propose a novel feature selection approach, based on extreme learning machines (ELMs) and the coefficient of variation (CV). In the proposed approach, the most relevant features are identified by ranking each feature with the coefficient obtained through ELM divided by CV. The achieved accuracies and computational costs, obtained with the use of features selected via the proposed approach in 9 classification and 26 regression benchmark data sets, were compared to those obtained with all features, as well as those obtained with the features selected by a wrapper and a filtering method. The achieved accuracy values obtained with the proposed approach were generally higher than when using all features. Furthermore, high feature reduction ratios were obtained with the proposed approach, including the achieved feature reduction ratios in epilepsy, liver, EMG, shuttle, and abalone. Stock data sets were 90.48%, 90%, 70.59%, 66.67%, 75%, and 77.78%, respectively. This approach is an extremely fast process that is independent of the employed machine-learning methods.

Key words: Feature selection, extreme learning machine, coefficient of variation

1. Introduction

Today's modern systems are equipped with processors involving a high level of computation, storage, and communication capacity, besides the systems' low cost and weight. These characteristics facilitate the modeling of various efficient systems that are very similar to the real world. However, in real life, because of their incomplete or insufficient information, all system parameters may not be well-defined to perfectly model a realistic system. Machine learning (ML) algorithms, in this sense, are developed to model successful systems with deficient data. Employing only relevant features, instead of the whole data set (i.e. the irrelevant, redundant, or noisy features in the data set are eliminated), is generally a solution for increasing the accuracy and speed of applied ML algorithms. Thus, the computational cost of the ML stage may be decreased, and, in some cases, the accuracy of ML may be increased [1–3]. Additionally, after the feature selection process, the system can be better optimized owing to the use of less equipment and computational work such as electrode reduction [4]. It was also used to reduce instrumental costs and processes involving huge data sets such as gene classification [5,6]. The ambition for feature selection is to identify the best descriptive feature or feature subset to improve computational and storage costs, while retaining almost the same or higher classification accuracy [1–3,7–10]. Therefore, the selection of the most definitive features can be defined as an important step in the ML process.

*Correspondence: omerfarukertugrul@gmail.com

Feature selection methods are classified into filtering, wrapping, and embedding methods, according to the employed feature ranking scheme. Filtering-based methods attempt to measure the qualities of each feature from the given data set by utilizing a discriminating criterion, and all features in the data set are ranked based on the obtained score. One thing to bear in mind is that the filtering feature selection process is fast enough, scalable, and independent of ML methods [2,3]. On the other hand, Guyon and Elisseeff suggested that a variable, which is useless or has a minor effect when employed alone, may provide a significant improvement when used together with other variables [1]. Hence, feature ranks obtained by filtering methods may not always provide the best feature subset.

Wrapping-based methods put into practice each particular subset according to their effectiveness on a given predictor to determine the best feature subset. Despite wrapping-based methods providing a selection of the best relevant features, they lead to a high computational cost, depending on the rate of trials (that is: 2^N trial for N feature) [2,3]. To reduce the computational cost, some future selection schemes, such as forward and backward selection, have been proposed [2]. Such search methods may make the system faster than traditional wrapper methods, because they do not try certain subsets. Nonetheless, these search strategies may cause an overfitting problem and/or add irrelevant features, which can make prediction worse or more complex [11]. In embedded methods, features in a data set are selected automatically with a learning method. Therefore, the produced results are suitable for this particular ML method. Embedded methods, such as decision tree and weighted Naïve Bayes, require lower computational cost compared to wrapper methods. However, the accuracy of these algorithms depends on ML methods [2,3,7].

From the literature, it can be understood that developing a fast, consistent, and independent feature selection algorithm with higher accuracy is still desirable. It has been thought that the extreme learning machine (ELM), which has an extremely fast training stage with a high generalization capacity, is a good candidate for a feature selection process [12–14]. Several studies investigating the applicability of ELM to feature selection have been published [15–18]. One of these studies, which employs the wrapper process, computes the impact of a feature by the difference between achieved accuracies with and without that feature in the data set [15]. Additionally, a relatively new embedded method, which offers an additional feature selection layer to standard ELM, has been proposed [16]. In the literature, ELM has been generally employed as an ML method, and the relevant features were selected based on the wrapper method [17] or the filter method [18]. Although extensive research has been carried out on feature selection by means of ELM to obtain the most relevant subset of features with a high speed, accuracy, and compression ratio, no single study exists that entirely overcomes the disadvantages of filtering, wrapping, and embedded methods, mentioned above.

The coefficient of variation (CV), which shows the ratio of standard deviation of data to its mean value, measures the variability of the data (or feature), independent of the unit of measurement used for the data [19]. The CV alone has been used in previous studies for feature selection [20] and feature extraction [21,22]. A small value of CV means that the data has small variations or scattering between samples in the data [20]. Therefore, the best feature was defined in accordance with the minimum value of CV [20].

In view of this information, an efficient approach based on ELM and CV values was proposed for feature selection. CV values that were calculated from data sets and coefficients, which were determined from trained ELM, were used as feature ranks. The results achieved from the given examples showed that the proposed approach is simple to employ and possesses very high speed, high accuracy, and a high feature reduction ratio.

2. Materials and methods

In order to evaluate and validate the proposed approach, 9 classification benchmark data sets and 26 different regression data sets were employed. Details of each of these data sets are given in Table 1.

Table 1. Employed data sets.

Name	Performed task	#Features	#Observations	#Classes	Details
Butterfly image	Classification	27	190	19	[23,24]
EMG	Classification	17	80	2	[25,26]
Liver	Classification	7	345	2	[26–28]
Epilepsy	Classification	42	200	2	[29]
Pima Indian diabetes	Classification	8	762	2	[26,30]
Hepatitis	Classification	19	127	2	[26]
Image segmentation	Classification	19	2315	7	[26]
Satellite image	Classification	36	6435	7	[26]
Statlog (shuttle)	Classification	9	58,000	7	[26]
Forest fire	Regression	12	517	-	[26]
CASP 5-9	Regression	9	45,731	-	[26]
Abalone	Regression	8	4177	-	[26]
Delta ailerons	Regression	5	7129	-	[12]
Auto-Price	Regression	15	159	-	[26]
Bank-8FM	Regression	8	6481	-	[12]
Boston housing	Regression	13	506	-	[26]
Breast cancer	Regression	32	194	-	[26]
California housing	Regression	8	20,640	-	[12]
Census-8L	Regression	8	22,784	-	[12]
Census-8H	Regression	8	22,784	-	[12]
Census-16L	Regression	16	22,784	-	[12]
Census-16H	Regression	16	22,784	-	[12]
CPU-small	Regression	12	8192	-	[26]
CPU	Regression	21	8192	-	[26]
Diabetes child	Regression	2	43	-	[12]
Delta elevators	Regression	6	9517	-	[12]
Elevators	Regression	18	16,599	-	[12]
Kinematics	Regression	8	8192	-	[12]
Machine-CPU	Regression	6	209	-	[26]
Puma-8NH	Regression	8	6677	-	[12]
Puma-32H	Regression	32	4938	-	[12]
Pyrimidines	Regression	28	74	-	[12]
Servo	Regression	4	167	-	[26]
Stocks	Regression	9	950	-	[12]
Triazines	Regression	60	186	-	[12]

2.1. Extreme learning machine

In ELM, weights and biases in the hidden layer are assigned arbitrarily, and weights in the output layer are calculated analytically through generalized Moore–Penrose pseudoinverse matrix [31,32]. Therefore, ELM has an extremely fast training stage and high generation capacity [12–14], and, unlike gradient-based learning algorithms, there is no risk of falling into any local minima in ELM [33,34]. In general, the output of a neural

network that has a single hidden layer can be expressed as

$$y_k = \sum_{j=1}^{m} \beta_{j,k} g\left(\sum_{i=1}^{n} w_{i,j} x_i + b_j\right) \tag{1}$$

where x_i denotes the input of the system and y_k denotes the k'th output of the system. n, m , and k indicate the number of neurons in the input, hidden, and output layers, respectively. $w_{i,j}$ and $\beta_{j,k}$ denote the weights assigned to input and output neurons, respectively. b_j denotes the bias values that are applied to the neurons in the hidden layer. $g(.)$ is the activation function implemented for every single neuron [35]. Eq. (1) can be rewritten as

$$\mathbf{H}\beta = \mathbf{y} \tag{2}$$

Here $\mathbf{H}$ is the hidden layer output matrix [12]. The weights of output neurons, $\beta_{1\cdots m,1\cdots k}$, can be computed with the generalized Moore–Penrose pseudoinverse matrix as

$$\hat{\beta}{=}\mathbf{H}^{+}\mathbf{y}, \tag{3}$$

where $\mathbf{H}^{+}$ denotes the generalized Moore–Penrose pseudoinverse matrix of $\mathbf{H}$, as given in [33], and $\boldsymbol{y}$ is the desired output of the system.

2.2. Proposed feature selection approach

Referring to Eq. (1), it has been reported that any infinitely differentiable function can be chosen as a transfer function ($g(.)$) in the hidden layer [12–14]. In this study, in the framework of linear mathematics, let us say a linear separable function, which complies with $(a\,x + b) = a g(x) + g(b)$, can be chosen for $g(.)$. In this case, Eq. (1) may be rewritten as

$$y_k = \sum_{j=1}^{m} \beta_{j,k} \left(\sum_{i=1}^{n} w_{i,j} g(x_i) + g(b_j)\right) \tag{4}$$

where $w_{i,j}$ and b_j are randomly assigned constants. The term $\beta_{j,k}$ is the weight calculated to go between the j'th neuron and the k'th output of the system. This equation may be unwrapped as

$$\begin{aligned} y_k \;=\; & \beta_{1,k}(w_{1,1}g(x_1) + \cdots + w_{i,1}g(x_i) + \cdots + w_{n,1}g(x_n) + g(b_1)) + \cdots \\ & +\beta_{j,k}(w_{1,j}g(x_1) + \cdots + w_{i,j}g(x_i) + \cdots + w_{n,j}g(x_n) + g(b_j)) + \cdots \\ & +\beta_{m,k}(w_{1,m}g(x_1) + \cdots + w_{i,m}g(x_i) + \cdots + w_{n,m}g(x_n) + g(b_m)) \end{aligned} \tag{5}$$

For simplicity, since b_1, $b_2, \ldots b_j b_m$ are constants, the constant terms can be replaced by a single constant value, such that $\psi_k = \beta_{1,k}g(b_1) + \cdots + \beta_{j,k}g(b_j) + \cdots + \beta_{m,k}g(b_m)$. Therefore, Eq. (5) can be reformed such that

$$\begin{aligned} y_k \;=\; & (\beta_{1,k}w_{1,1} + \ldots + \beta_{j,k}w_{1,j} + \cdots + \beta_{m,k}w_{1,m})g(x_1) + \ldots \\ & +(\beta_{1,k}w_{i,1} + \cdots + \beta_{j,k}w_{i,j} + \cdots + \beta_{m,k}w_{i,m})g(x_i) + \cdots \\ & +(\beta_{1,k}w_{n,1} + \cdots + \beta_{j,k}w_{n,j} + \cdots + \beta_{m,k}w_{n,m})g(x_n) + \psi_k \end{aligned} \tag{6}$$

Then Eq. (7) may be put in compact form as

$$y_k = \alpha_{1,k} g(x_1) + \cdots + \alpha_{i,k} g(x_i) + \cdots + \alpha_{n,k} g(x_n) + \psi_k \tag{7}$$

Or, equivalently, $y_k = \sum_{i=1}^{n} \alpha_{i,k} g(x_i) + \psi_k$. Here the coefficients denoted by $\alpha_{i,k} = \beta_{1,k} w_{i,1} + \cdots + \beta_{j,k} w_{i,j} + \cdots + \beta_{m,k} w_{i,m}$ are employed in ranking the j'th feature of the desired k'th output of the system. The entire process can also be expressed in matrix form as

$$y_k = \begin{bmatrix} \beta_{1,k} \\ \beta_{2,k} \\ \beta_{3,k} \\ \vdots \\ \beta_{m,k} \end{bmatrix}^T \begin{bmatrix} w_{1,1} & w_{2,1} & \cdots & w_{n,1} \\ w_{1,2} & w_{2,2} & \cdots & w_{n,2} \\ w_{1,3} & w_{2,3} & \cdots & w_{n,3} \\ \ddots & \ddots & \vdots & \ddots \\ w_{1,m} & w_{2,m} & \cdots & w_{n,m} \end{bmatrix} g\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \right) + \psi_k \tag{8}$$

The set of coefficients defined in this way for the particular kth output of the system can be expressed in matrix form as

$$\left[\alpha_{1,k},\ \alpha_{2,k}, \ldots \alpha_{i,k}, \ldots \alpha_{n,k}\right] = \begin{bmatrix} \beta_{1,k} \\ \beta_{2,k} \\ \beta_{3,k} \\ \vdots \\ \beta_{m,k} \end{bmatrix}^T \begin{bmatrix} w_{1,1} & w_{2,1} & \cdots & w_{n,1} \\ w_{1,2} & w_{2,2} & \cdots & w_{n,2} \\ w_{1,3} & w_{2,3} & \cdots & w_{n,3} \\ \ddots & \ddots & \vdots & \ddots \\ w_{1,m} & w_{2,m} & \cdots & w_{n,m} \end{bmatrix} \tag{9}$$

Finally, the system equation becomes

$$y_k = \left[\alpha_{1,k},\ \alpha_{2,k}, \ldots \alpha_{i,k}, \ldots \alpha_{n,k}\right] g\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \right) + \psi_k \tag{10}$$

The coefficients $\alpha_{1,k}, \ldots \alpha_{n,k}$, obtained in this way, are the coefficients ranking the input elements of the network. For a multi-input multioutput (MIMO) system, the global equation then becomes $\mathbf{y} = \mathbf{A} g\left([\mathbf{X}]^T\right) + \psi$, where $\boldsymbol{A}$ denotes the coefficients' matrix, $[\mathbf{X}]$ denotes the vector of elements fed to the input of the system, and ψ is a scalar obtained by β, obtained through ELM multiplied by the transpose of the bias vector employed for the hidden layer of the neural network. Then we minimize the risk of these coefficients ($\alpha_{1,k}, \ldots, \alpha_{n,k}$) by adjusting them by a meaningful quantity such as unitized risk factor so-called variation coefficient [36]. The variation coefficient or CV, which is sometimes called relative standard deviation, is defined as a standardized measure of dispersion of a probability distribution in probability theory. The reason for adjusting the coefficients with CV ($CV = {}^{\sigma}/_{\mu}$) can be explained as follows:

Let the CV calculated over data set x_i fed to the ith input of the system be denoted by $CV(x_i)$. By dividing $\alpha_{i,k}$ coefficients by $CV(x_i)$, it is revealed that if the probabilistic risk with input x_i in identifying decision boundary is high, i.e. $CV(x_i)$ is high, then the rank of x_i must be lower as compared to x_i with low $CV(x_i)$. In other words, in the example of feature data, any x_i that has a high $CV(x_i)$ will have a minimum effect on the decision boundary and vice versa. From this information, in a feature selection process, feature ranking (FR) for y_k can be designed as

$$FR_{x_i} = \left[\left| \frac{\alpha_{1,k}}{CV\,(x_1)} \right|,\ \left| \frac{\alpha_{2,k}}{CV\,(x_2)} \right|,\ \cdots \left| \frac{\alpha_{i,k}}{CV\,(x_i)} \right|,\ \cdots \left| \frac{\alpha_{n,k}}{CV\,(x_n)} \right| \right] \tag{11}$$

In this work, we consider a multi-input but single output (MISO) system, such that $k = 1$. The number of elements in x_i was equal to the number of observations in the data set, taken into account for the training of ELM. For example, in this study, the data sizes of the butterfly, EMG, epilepsy, diabetes, and liver data sets were 190, 80, 86, 768, and 583, respectively.

2.3. Employed procedure

To comparatively evaluate the proposed approach, FM and WM were performed on real world data sets of different specifications [23–30], and were analyzed in terms of both accuracy and computational cost. The motivation behind the use of WM was the success of the best accuracy [17]. The use of FM was in being fast and providing the relevance of each individual feature one-by-one [15]. In the wrapper method, the most relevant feature subsets were constructed by determining the most relevant feature, then the two most relevant features, then the three most relevant features, and so on. In these methods, the ranking was achieved in accordance with the success rate, whereas in the proposed approach, the ranking was achieved.

Firstly, data sets were classified/estimated with kNN, taking into account all the features. Then filter, wrapper, and the proposed approach were conducted on the data sets for feature selection, and the obtained ranked feature sets were classified/estimated by kNN. In this study, k nearest neighbor (kNN) (with 10-fold cross-validation) was employed as the classifier/estimator in the system, due to its simplicity and being a nonparametric algorithm [37]. Another reason for choosing kNN was to certify that the obtained results did not depend on the classification system. The optimum value of k was determined by the value (in the range of 1 to 10) that yields the highest accuracy.

3. Results and discussion

The most relevant features that were selected from the butterfly data set with the three methods (the proposed approach, FM, and WM) are provided in Table 2. In this table, μ, σ, and H show the mean, standard deviation, and entropy of the images that were filtered with TEM filters. A TEM filter is formed by multiplying a TEM vector by the transpose of a TEM vector, e.g., $S_7^T E_7$. Each TEM vector is characterized in accordance with a mission to extract the targeted configuration in the image, e.g., the average energy level (L), edge (E), and spot (S) [23].

As seen from this table, among the extracted features, *spots of images* were dominantly involved in selecting the most relevant features in the proposed approach and WM. Additionally, in WM, as mentioned earlier, the accuracies of all possible feature subsets were calculated. In this case, the relation order was equal to the number of features in the subset that had the highest accuracy. In FM, features were established by

Table 2. Relevance order.

Relevance order	Proposed approach	FM	WM
1	H($S_7^T E_7$*Image)	σ($E_7^T L_7$*Image)	σ($E_7^T L_7$*Image)
2	σ($S_7^T S_7$*Image)	μ($S_7^T L_7$*Image)	σ($E_7^T S_7$*Image), μ($S_7^T E_7$*Image)
3 4-4	μ($S_7^T S_7$*Image)	μ($E_7^T E_7$*Image)	σ($E_7^T S_7$*Image), μ($S_7^T E_7$*Image), μ($S_7^T S_7$*Image)
4	H($S_7^T S_7$*Image)	μ($E_7^T L_7$*Image)	σ($S_7^T E_7$*Image), μ($E_7^T S_7$*Image), μ($S_7^T L_7$*Image), μ($S_7^T S_7$*Image)
5	σ($S_7^T E_7$*Image)	μ($L_7^T L_7$*Image)	σ($E_7^T S_7$*Image), μ($S_7^T E_7$*Image), μ($S_7^T S_7$*Image), σ($L_7^T S_7$*Image), μ($S_7^T S_7$*Image)

utilizing the edge and level of 1-D TEM vectors. Then this data set was classified into 19 species by way of kNN using these ranked features. The accuracies of the FM, WM, and proposed approaches were obtained in association with the number of features employed and comparatively plotted in Figure 1, where the dashed line shows the obtained accuracy level when all the features were used. As seen in Figure 1, the performances of the proposed, FM, and WM approaches are quite similar. Furthermore, the accuracies obtained with the proposed, FM, and WM approaches in the other employed data sets were calculated and plotted in Figure 2 as a function of the most relevant feature numbers. From this figure, it is obvious that the accuracies obtained with the selected features are larger than the accuracies obtained with the use of all features. This means that the same or higher accuracies can be achieved by using a lower number of features.

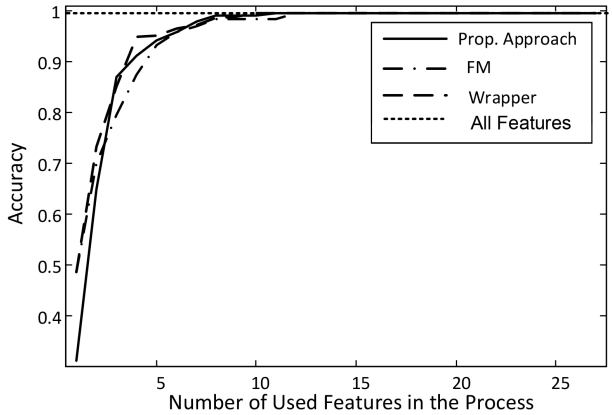

Figure 1. Accuracy results of the butterfly data set.

In the proposed approach, the feature ranks were determined by utilizing the output weights of the trained ELM, multiplied by the inverse of the coefficient of variation calculated over features. The features selected in this way are assumed as eventually influenced by all the features in the data set. Hence, besides being simple and easily applicable, the proposed approach overcomes some of the disadvantages of FM and WM, such as computational cost and the effect of removed features. This can be understood as the relevant order or accuracy with respect to the number of employed features obtained with the proposed approach, given in Figures 1 and 2. The total number of features, the selected number of the most relevant features, and their accuracies are summarized in Table 3. It is obvious from Table 3 that the classification accuracies achieved by using the selected features in diabetes, liver, EMG, shuttle, segmentation, and hepatitis data sets via the

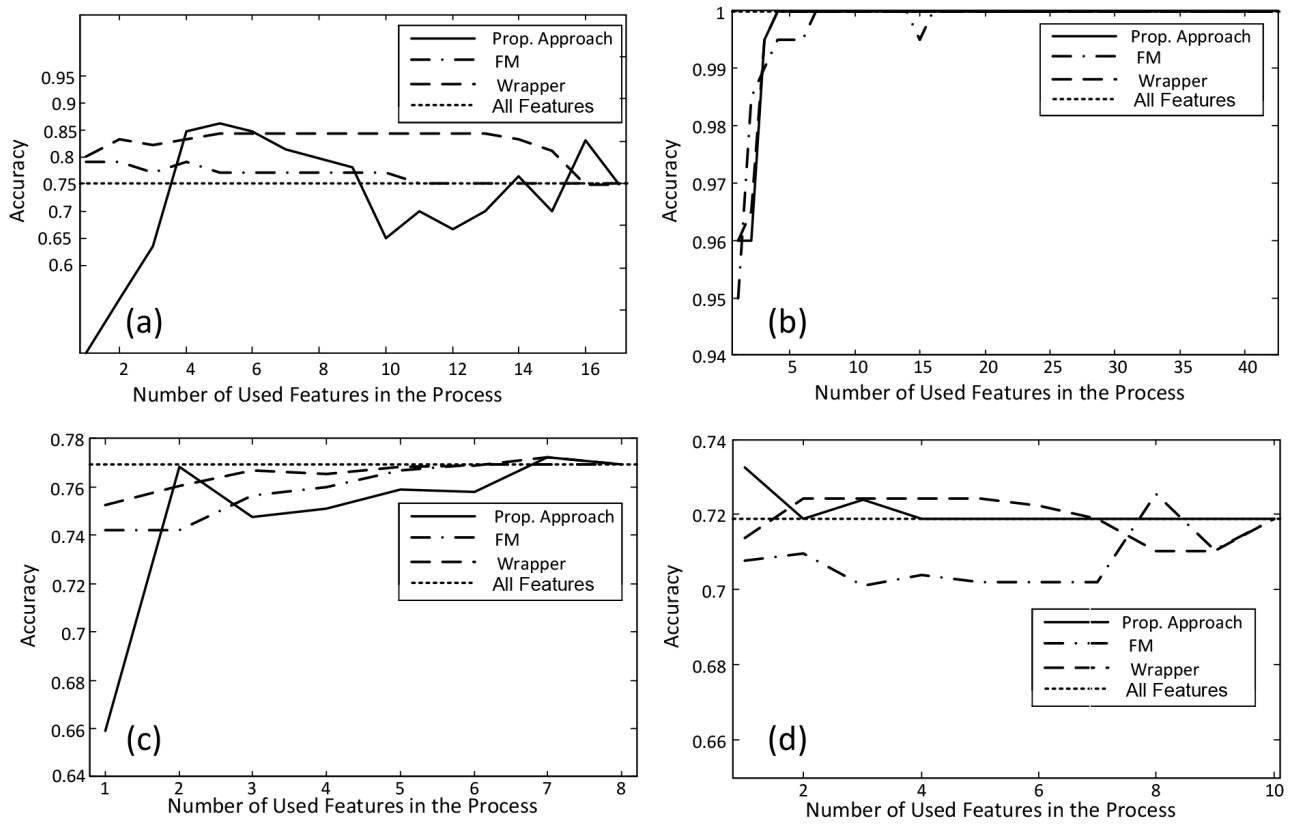

Figure 2. Accuracy results of the (a) EMG data set, (b) epilepsy data set, (c) diabetes data set, and (d) liver data set.

proposed approach are higher than those obtained by using all the features. The accuracies obtained in the epilepsy and butterfly data sets were 100% and 99.47%, respectively, when employing the selected features and all the features. The results obtained with regression data sets are given in Table 4.

Table 3. General performance of the proposed approach in classification data sets.

	Number of features	Number of features selected	Feature reduction ratio (%)	Selection time (s)	Obtained accuracy	
					All features	Selected features
Epilepsy	42	4	90.48	0.0624	100	100
Butterfly	27	11	59.26	0.0624	99.47	99.47
Diabetes	8	7	12.50	0.0156	76.95	77.21
Liver	10	1	90.00	0.0780	71.87	73.24
EMG	17	5	70.59	0.0156	75	87.5
Shuttle	9	3	66.67	0.1406	99.45	99.85
Segmentation	36	32	11.11	0.2031	88.5	89.1
Hepatitis	19	2	89.47	0.1406	66.25	73.75

As seen in Table 4, when all the features were used, the RMSEs obtained for 21 data sets are quite low. However, when the selected features were used, the RMSEs obtained for 23 data sets were generally at the same level, or even less. Differences in feature reduction ratios in both classification and regression data sets can be explained with the complexity of the employed data sets and the significance of the features. The feature

Table 4. General performance of the proposed approach in regression data sets.

Data set	Number of features	Number of features selected	Feature reduction ratio (%)	Selection time (s)	Obtained RMSE	
					All features	Selected features
Forest fire	12	5	58.33	0.016	0.0044	0.0044
CASP 5-9	9	3	66.67	0.016	0.0018	0.0017
Abalone	8	2	75.00	0.031	0.0087	0.0077
Delta ailerons	5	3	40.00	0.078	0.0088	0.0080
Autoprice	15	9	40.00	0.031	0.0059	0.0055
Bank-8FM	8	3	62.50	0.078	0.0024	0.0022
Boston housing	13	1	92.31	0.078	0.0302	0.0186
Breast cancer	32	13	59.38	0.016	0.1021	0.0992
California housing	8	3	62.50	0.031	0.0500	0.0198
Census-8L	8	6	25.00	0.031	0.0136	0.0082
Census-8H	8	4	50.00	0.078	0.0119	0.0091
Census-16L	16	4	75.00	0.031	0.0118	0.0081
Census-16H	16	10	37.50	0.016	0.0131	0.0087
CPU small	12	2	83.33	0.031	0.0053	0.0042
CPU	21	6	71.43	0.031	0.0055	0.0053
Diabetes child	2	2	0.00	0.016	0.0090	0.0090
Delta elevators	6	6	0.00	0.078	0.0204	0.0204
Elevators	18	1	94.44	0.078	0.0096	0.0047
Kinematics	8	6	25.00	0.016	0.0105	0.0096
Machine CPU	6	4	33.33	0.016	0.0048	0.0048
Puma-8NH	8	2	75.00	0.078	0.1133	0.0907
Puma-32H	32	19	40.63	0.094	0.1194	0.1188
Pyrimidines	27	25	7.41	0.078	0.0070	0.0043
Servo	4	4	0.00	0.031	0.0156	0.0156
Stocks	9	2	77.78	0.016	0.0485	0.0114
Triazines	60	37	38.33	0.078	0.0008	0.0006

reduction ratio does not only show the decrease in computational cost of the employed ML, but also shows lower memory requirement and the pertinent grade of the features.

In order to make clear that the ranking method is independent from ELM, since the latter was employed in the formulation of the proposed approach, kNN was used in the classification/estimations stage. To demonstrate this idea and to show that the proposed approach can go with other linear and nonlinear machine learning algorithms as a preprocess, the features selected by the proposed approach (for the liver data set) were classified/estimated by the nearest mean (NM), Naïve Bayes (NB), iteration-based artificial neural network (ANN), support vector machine (SVM), and instance-based kNN classifiers. The classification accuracy levels achieved with the use of ranked features, obtained through the proposed approach, and all other features are given in Table 5.

The data given in Table 5 suggest that the proposed approach is independent of the ML method and can be employed as a preprocess before any ML method. In general, as mentioned earlier, the ML method can obtain higher accuracies with less computational cost and storage capacity when using selected features of the proposed approach. With this approach, the requirement for storage capacity and computation time is decreased due to using less features. For example, the mean processing times of the proposed method, when NB

Table 5. Accuracies achieved by the employed ML methods in the liver data set.

ML method	Number of features	Number of used features	Feature reduction (%)	All features		Selected features	
				Obtained accuracies	Process time (s)	Obtained accuracies	Process time (s)
NM	10	1	90	58.59	0.0881	54.37	0.0827
NB	10	5	50	67.84	0.0944	70.92	0.0835
kNN	10	1	90	71.87	0.1182	73.24	0.0872
ANN	10	10	0	72.38	1.2758	72.38	1.2758
SVM	10	1	90	71.36	0.3664	71.36	0.2939

was used for classification, were 0.0466 s, 0.0835 s, and 0.0944 s when using 1, 5, and 10 features, respectively. Some minor accuracy differences may have been caused by cross-validation errors. Additionally, the required time lengths of the employed feature selection processes, which show the computational costs of the feature selection process, are summarized in Table 6. The simulations were conducted with a personal computer with an Intel Core i7-2600 CPU, 3.4 GHz, and 4 GB RAM.

Table 6. Processing time for feature selection (s).

Data set	Proposed approach	FM	WM
Butterfly	0.0624	0.6396	28,969.87
EMG	0.0156	7.888	16,630.46
Epilepsy	0.0624	11.84	61,689.12
Diabetes	0.0156	6.0684	187.56
Liver	0.0780	7.0512	522.135

The feature selection process with the proposed approach takes less time compared to other employed feature selection approaches, as shown in Table 6. It is obvious that the proposed approach is almost ten times faster than FM and much faster than WM. This may be because FM requires selecting and applying the optimum discriminating criteria, and WM requires 2^N trials in determining the most relevant feature, whereas the proposed approach requires only one trial in which the solution of the ELM matrix is calculated. Additionally, a positive correlation was found between the required times for feature selection in the proposed approach, FM, and WM with the number of features. This means that a data set with a high number of features requires more time to select relevant features. Therefore, the increase in processing time in the epileptic data set (compared to the other data sets), as seen in Table 6, can be due to the increase in the number of features employed in the process. The required time for feature selection by WM took nearly 17 hours for this data set, which shows that WM is a time-consuming process.

4. Conclusion

In this study, a new feature selection algorithm was proposed and demonstrated for a variety of data sets. The results obtained by the proposed approach and the FM and WM methods were analyzed according to their accuracies and processing speed. From the obtained results it can be concluded that the proposed approach is superior to the existing methods, because it possesses a significant feature reduction ratio and faster processing time. Although under some conditions the accuracies can be similar to those of the epilepsy and butterfly data sets, under other conditions, the classification accuracy achieved by using selected features of the proposed approach is better than that achieved by all the features. For example, for the diabetes, liver, and EMG data sets, the accuracy obtained with the selected features was better than that obtained with all the features.

Besides generality, the proposed approach was proven to be adequately selective, extremely fast, and possessing a high classification accuracy. Since the presented approach uses ELM, it takes into account the effect of all features and reflects onto the mechanism of ranking. Moreover, via *CV* values, the method selects the most stable features to better standardize the ranking scheme. Such important characteristics result in the superiority of the method in feature selection. Most importantly, the proposed approach is simple to design and easily applicable to various problems.

References

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

[2] Ladha L, Deepa T. Feature selection methods and algorithms. Int J Comput Sci Eng 2011; 3: 1787-1797.

[3] Dash M, Liu H. Feature selection for classification. Intell Data Anal 1997; 1: 131-156.

[4] Atyabi A, Luerssen M, Fitzgibbon S, Powers DM. Evolutionary feature selection and electrode reduction for EEG classification. In: IEEE 2012 Evolutionary Computation Congress; 10–15 June 2012; Brisbane, Australia. New York, NY, USA: IEEE. pp. 1-8.

[5] Sanchez-Monedero J, Cruz-Ramirez M, Fernández-Navarro F, Fernández JC, Gutiérrez PA, Hervás-Martínez C. On the suitability of Extreme Learning Machine for gene classification using feature selection. In: IEEE 2010 Intelligent Systems Design and Applications Conference; 29 November–1 December 2010; Cairo, Egypt. New York, NY, USA: IEEE. pp. 507-512.

[6] Melita NT, Popescu I, Holban S. A genetic algorithm approach to DNA microarrays analysis of pancreatic cancer. Adv Electr Comput En 2008; 8: 43-48.

[7] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007; 23: 2507-2517.

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

[9] Fodor IK. A Survey of Dimension Reduction Techniques.. Washington, DC, USA: US Department of Energy, 2002.

[10] Dubois D, Fargier H, Bonnefon JF. On the qualitative comparison of decisions having positive and negative features. J Artif Intell Res 2008; 32: 385-417.

[11] Hawkins DM. The problem of overfitting. J Chem Inform Comput Sci 2004; 44: 1-12.

[12] Huang GB, Zhu QY, Siew CK. Extreme learning machine: theory and applications. Neurocomputing 2006; 70: 489-501.

[13] Huang GB, Zhu QY, Siew CK. Extreme learning machine: a new learning scheme of feedforward neural networks. In: IEEE 2004 Neural Networks Conference; 25–29 July 2004; Budapest, Hungary. New York, NY, USA: IEEE. pp. 985-990.

[14] Huang GB, Wang DH, Lan Y. Extreme learning machines: a survey. Int J Mach Learn Cybern 2011; 2: 107-122.

[15] Zhai MY, Yu RH, Zhang SF, Zhai JH. Feature selection based on extreme learning machine. In: IEEE 2012 Machine Learning and Cybernetics Conference; 15–17 July 2012; Xian, China. New York, NY, USA: IEEE. pp. 157-162.

[16] Benoît F, Van Heeswijk M, Miche Y, Verleysen M, Lendasse A. Feature selection for nonlinear models with extreme learning machines. Neurocomputing 2013; 102: 111-124.

[17] Termenon M, Graña M, Barrós-Loscertales A, Ávila C. Extreme learning machines for feature selection and classification of cocaine dependent patients on structural MRI data. Neural Process Lett 2013; 38: 375-387.

[18] Salcedo-Sanz S, Pastor-Sánchez A, Prieto L, Blanco-Aguilera A, García-Herrera R. Feature selection in wind speed prediction systems based on a hybrid coral reefs optimization—extreme learning machine approach. Energ Convers Manag 2014; 87: 10-18.

[19] Abdi H. Coefficient of variation. In: Salkind NJ, editor. Encyclopedia of Research Design. Thousand Oaks, CA, USA: SAGE Publications, 2010. pp. 169-171.

[20] Kumar D, Unikrishnan P. Class specific feature selection for identity validation using dynamic signatures. J Biom Biostat 2013; 4: 1000160-1–1000160-5.

[21] Nakariyakul S, Casasent D. Hyperspectral feature selection and fusion for detection of chicken skin tumors. Proc SPIE 2004; 5271: 128-139.

[22] Habib M, Rokonuzzaman M. Distinguishing feature selection for fabric defect classification using neural network. J Multimed 2011; 6: 416-424.

[23] Ertuğrul ÖF, Kaya Y, Kaycı L, Tekin R. A vision system for classifying butterfly species by using Law's texture energy measures. Int J Comput Vis Mach Learn Data Min 2015; 1: 20-28.

[24] Kaya Y, Kayci L, Tekin R, Ertuğrul ÖF. Evaluation of texture features for automatic detecting butterfly species using extreme learning machine. J Exp Theor Artif Intell 2014; 26: 267-281.

[25] Ertuğrul OF, Tağluk ME, Kaya Y, Tekin R. EMG signal classification by extreme learning machine. In: IEEE 2013 Signal Processing and Communications Applications Conference; 24–26 April 2013; Haspolat, Turkey. New York, NY, USA: IEEE. pp. 1-4.

[26] Bache K, Lichman M. UCI Machine Learning Repository. Irvine, CA, USA: University of California, 2013.

[27] Ramana BV, Babu P, Surendra M, Venkateswarlu NB. A critical study of selected classification algorithms for liver disease diagnosis. Int J Database Manag Syst 2011; 3: 101-114.

[28] Ramana BV, Babu P, Surendra M, Venkateswarlu NB. A critical comparative study of liver patients from USA and India: an exploratory analysis. Int J Comput Sci 2012; 9: 506-516.

[29] Andrzejak RG, Lehnertz K, Mormann F, Rieke C, David P, Elger CE. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state. Phys Rev E 2001; 64: 061907-1–061907-8.

[30] Tağluk ME, Ertuğrul ÖF. A joint generalized exemplar method for classification of massive datasets. Appl Soft Comput 2015; 36: 487-498.

[31] Huang GB. An insight into extreme learning machines: random neurons, random features and kernels. Cognit Comput 2014; 6: 376-390.

[32] Wang R, Kwong S, Wang X. A study on random weights between input and hidden layers in extreme learning machine. Soft Comput 2012; 16: 1465-1475.

[33] Huang GB, Ding X, Zhou H. Optimization method based extreme learning machine for classification. Neurocomputing 2010; 74: 155-163.

[34] Ertuğrul ÖF. Forecasting electricity load by a novel recurrent extreme learning machines approach. Int J Electr Power Energ Syst 2016; 78: 429-435.

[35] Rong HJ, Ong YS, Tan AH, Zhu Z. A fast pruned-extreme learning machine for classification problem. Neurocomputing 2008; 72: 359-366.

[36] Milani AS, Eskicioglu C, Robles K, Bujun K, Hosseini-Nasab H. Multiple criteria decision making with life cycle assessment for material selection of composites. Express Polymer Lett 2011; 5: 1062-1074.

[37] Guo G, Wang H, Bell D, Bi Y, Greer K. Using kNN model for automatic text categorization. Soft Comput 2006; 10: 423-430.