One-Class Classification with Extreme Learning Machine

Qian Leng, 1 Honggang Qi, 1 Jun Miao, 2 Wentao Zhu, 2 and Guiping Su 1

1 School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100190, China
2 Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China

Correspondence should be addressed to Jun Miao; jmiao@ict.ac.cn

Received 13 August 2014; Revised 8 November 2014; Accepted 10 November 2014

1. Introduction

One-class classification [1, 2] has received much interest during recent years, which has also been known as novelty or outlier detection. Different from normal classification, data samples from only one class, called the target class, are well characterized, while there are no or few samples from the other class (also called the outlier class). To reveal the necessity of one-class classification, we take the case of online shopping service as an example. In order to recommend goods users want, it is convenient to track the users’ history shopping lists (positive training samples), while collection of negative training samples is challenging because it is hard to say which one users dislike. Other applications include machine fault detection [3], disease detection [4], and credit scoring [5]. The goal is to “teach” the classifier through observing target samples so that it can be applied to select unknown samples similar to the target class and reject samples which deviate significantly from the target class.

Various types of one-class classifier have been designed and applied in different fields; see [6] for a comprehensive review. Early attempt to obtain a one-class classifier is by estimating the probability density functions based on training data. Parzen density estimation [7, 8] superposes kernel functions on individual training samples to estimate the probability density function. Naive Parzen density estimation, similar to Naive Bayes approach used for classification, fits a Parzen density estimation on each individual feature and multiplies the results for final density estimation. A test sample is rejected if its estimated probability is below a threshold. However, estimating the true density distribution usually requires a large number of training samples.

A simpler task is to find the domain of the data distribution. Schölkopf et al. [9] constructed a hyperplane which is maximally distant from the origin to separate the regions that contain no data. An alternative approach is to find a hypersphere [10] instead of a hyperplane to include the most target data with the minimum radius. Both approaches are cast out in the form of quadratic programming, while some approaches [11–13] are of linear programming. One-class LP classifier [11] minimizes the volume of the prism, which is cut by a hyperplane that bounds the data from above with some mild constrains on dissimilarity representations. Lanckriet et al. [13] propose the one-class minimax probability machine that minimizes the worst case probability of misclassification of test data, using only the mean and covariance matrix.
of the target distribution. When kernel methods are used, the aforementioned domain-based classifiers [2] can obtain more flexible descriptions. Recently, a minimum spanning tree based one-class classifier [14] was proposed. It considered graph edges as additional set of virtual target objects. By constructing a minimum spanning tree, recognition of a new test sample is determined by the shortest distance to the closest edge of that tree.

Autoencoder neural network is one of the reconstruction methods [1] to build a one-class classifier. The simplest architecture of such model is based on the single-hidden layer feed-forward neural networks (SLFNs). Usually, the hidden layer contains a smaller number of nodes than the number of input nodes which works like an information bottleneck. The classifier reproduces the input patterns at the output layer through minimizing the reconstruction error. However, standard backpropagation (BP) algorithm is used to train the networks, which is quite time-consuming. Extreme learning machine [15, 16] is originally developed to address the slow learning speed problem of gradient based learning algorithms for its iterative tuning of the networks’ parameters. It randomly selects all parameters of the hidden neurons and analytically determines the output weights. It is stated [17, 18] in theory that ELM tends to provide the best generalization performance at extreme learning speed since it is a simple tuning-free algorithm.

In this paper, the proposed one-class classifier based on ELM is constructed for situations where only the target class is well described. The proposed one-classifier utilizes the unified ELM learning theory [17], which leads to extreme learning speed and superior generalization performance. Moreover, the classifier further lessens the human intervention since it is not limited to specific target labels. Both random feature mappings and kernels can be adopted for such classifier which makes it more flexible to unique target descriptions. Constructing the proposed classifier for three different specific-designed artificial datasets demonstrates the classifier’s ability to describe universal target class distributions. When real-world datasets are evaluated, the proposed one-classifier is competitive over a variety of one-class models and learns hundreds of times faster than autoencoder neural network for one-class classification.

The rest of the paper is organized as follows. Section 2 briefly reviews extreme learning machine. In Section 3, we first describe the hypersphere perceptron as a one-class classifier and then introduce our proposed ELM based one-class classifier. Section 4 describes the experiments conducted on both artificial and real-world datasets. Finally, Section 5 presents the conclusion of the work.

2. Brief Review of ELM

ELM aims to reach not only the smallest training error but also the smallest norm of output weights [16] between the hidden layer and the output layer. According to Bartlett’s theory [19], the smaller norm of weights is, the better generalization performance of networks tends to have. Thus, better generalization performance can be expected for ELM networks. In [17], equality constraints are used in ELM, which provides a unified solution for regression, binary, and multiclass classifications.

2.1. Equality-Optimization-Constraints-Based ELM. Given N training data \((x_i, t_i)^N_{i=1}\), where \(x_i = [x_{i1}, \ldots, x_{in}]^T \in \mathbb{R}^n\) is the individual feature vector with dimension \(n\) and \(t_i \in \mathbb{R}^m\) is the desired target output, in the one-class classification case, single output node \((m = 1)\) is enough. The ELM output function can be formulated as

\[
f(x) = h(x)^T \beta = \sum_{j=1}^{L} \beta_j G(w_j, b_j, x),
\]

where \(\beta = [\beta_1, \ldots, \beta_L]^T\) is the vector of the output weights between the hidden layer and the output layer, \(w_j = [w_{j1}, \ldots, w_{jn}]^T\) is the input weights connecting input nodes with the \(j\)th hidden node, \(b_j\) is the bias of the \(j\)th hidden node, \(h(x) = [G(w_j, b_j, x), \ldots, G(w_L, b_L, x)]^T\) is the input feature vector of the hidden layer with respect to input \(x\), and \(G(w, b, x)\) is the activation function (e.g., sigmoid function \(G(w, b, x) = 1/(1+ \exp(-(w^T x + b)))\)) satisfying ELM universal approximation capability theorems [20, 21]. In fact, \(h(x)\) is a known nonlinear feature mapping which maps the training data \(x\) from the \(n\)-dimensional input space to the \(L\)-dimensional ELM feature space [17]. The goal of ELM is to minimize the norm of output weights as well as the training errors, which is equivalent to

\[
\text{min} \quad L_{\text{ELM}} = \frac{1}{2} \|\beta\|^2 + C \sum_{j=1}^{N} \|\xi_j\|^2 \quad \text{s.t.} \quad h(x_i)^T \beta = t_i - \xi_i, \quad i = 1, \ldots, N,
\]

where \(\xi_i\) is the slack variable of the training sample \(x_i\) and \(C\) controls the tradeoff between the output weights and the errors. Based on the Karush-Kuhn-Tucker (KKT) theorem [22], the corresponding Lagrange function of the primal ELM optimization (2) is

\[
L_{\text{D-ELM}} = \frac{1}{2} \|\beta\|^2 + C \sum_{j=1}^{N} \|\xi_j\|^2 - \sum_{j=1}^{N} \alpha_j \left( h(x_i)^T \beta - t_i + \xi_i \right);
\]

the following optimality conditions of (3) should be satisfied:

\[
\frac{\partial L_{\text{D-ELM}}}{\partial \beta} = 0 \implies \beta = \sum_{j=1}^{N} \alpha_j, \quad h(x_i) = H^T \alpha, \quad (4a)
\]

\[
\frac{\partial L_{\text{D-ELM}}}{\partial \xi_i} = 0 \implies \alpha_i = C \xi_i, \quad i = 1, \ldots, N, \quad (4b)
\]

\[
\frac{\partial L_{\text{D-ELM}}}{\partial \alpha_i} = 0 \implies h(x_i)^T \beta - t_i + \xi_i = 0, \quad i = 1, \ldots, N, \quad (4c)
\]

where \(H = [h(x_1), \ldots, h(x_N)]^T\) is the hidden layer output matrix and \(\alpha = [\alpha_1, \ldots, \alpha_N]^T\) is the vector of Lagrange variables. Substituting (4a) and (4b) into (4c) we have

\[
\left( \frac{1}{C} + HH^T \right) \alpha = T.
\]
Here I is the identity matrix and $T = [t_1, \ldots, t_N]^T$. Substituting (5) into (4a), we get

$$\beta = H^T \left( \frac{1}{C} + HH^T \right)^{-1} T. \quad (6)$$

The ELM output function (1) can be further derived as

$$f(x) = h(x)^T \beta = h(x)^T H^T \left( \frac{1}{C} + HH^T \right)^{-1} T. \quad (7)$$

If the hidden nodes’ feature mapping $h(x)$ is unknown to users, kernel methods that satisfy Mercer’s condition can be adopted: $K(x_i, x_j) = h(x_i) \cdot h(x_j)$. The ELM kernel output function can be written as

$$f(x) = h(x)^T \beta = h(x)^T H^T \left( \frac{1}{C} + HH^T \right)^{-1} T \quad (8a)$$

and the kernel matrix for ELM is

$$\Omega_{ELM} = HH^T : \Omega_{ELM},j = h(x_j) \cdot h(x_j) = K(x, x_j). \quad (8b)$$

2.2. Advances of ELM. Extreme learning machine has gained much more popularity since its advent. It has been able to avoid the problem of time complexity which classic learning techniques are confronted with while providing better generalization performance with less human intervention. Because of such attractive features, researchers have extended the basic ELM to several different directions and many variants of ELM have been developed. For instance, online sequential ELM (OS-ELM) [23, 24] can learn the sequential coming data (one by one or chunk by chunk) with a small effort to update the output weights. The training data are discarded after being learned by the network and the output weights need not be retrained, which is especially efficient for time-series problems. Other typical works include fully complex ELM [25, 26], incremental ELM (I-ELM) [20, 21], sparse ELM [27], ELM with elastic output [28, 29], and ELM ensembles [30–32]. See [33] for further details on the many ELM variants.

When uncertainty is present in the dataset, integration of fuzzy logic system and extreme learning machine tends to enhance the generalization capability of ELM. In [34], a neurofuzzy Takagi-Sugeno-Kang (TSK) fuzzy inference system is constructed utilizing extreme learning machine. The number of inference rules is previously determined by the $k$-means method. One ELM is used to obtain the membership of each fuzzy rule and multiple ELM are used to obtain the consequent part. Rong et al. [35] show that type-1 fuzzy inference system (type-1 FLS) is equivalent to a generalized SLFN. Hence, the hidden nodes work as the antecedent part and the output weights as the consequent part. Then, extreme learning machine is directly applied to the type-1 FLS and the corresponding online sequential fuzzy ELM has also been developed. Deng et al. [36] further extend the idea to type-2 fuzzy inference system (type-2 FLS) because of type-2 FLS’s superiority in modeling high level uncertainty. With the most widely used interval type-2 FLS, the parameters of the antecedents are randomly initialized according to the ELM mechanism. The Moore-Penrose generalized inverse is used to initialize the parameters of the consequents and the parameters are finally refined by Karnik-Mendel algorithm [37]. Many applications have also been investigated in the literature. For example, the hybrid model of ELM with interval type-2 FLS has been applied for permeability prediction [38].

3. The Proposed One-Class Classifier

3.1. Support Vector Data Description. For a better understanding of one-class classifiers, support vector data description (SVDD) [10] is discussed here for one-class classification process. SVDD defines a spherically shaped boundary around the complete target set and is intuitively appealing since it regards the target class as a self-closed system. Let $X = \{x_i, i = 1, \ldots, N\}$ be the training set, and $x_i \in R^n$ is drawn from the target distribution. SVDD aims to minimize the volume of the sphere as well as the training errors $\xi_i$ for objects falling outside the boundary, which is equivalent to

$$\min \quad \mathcal{L}_{SVDD} = R^2 + C \sum_{i=1}^{N} \xi_i^2$$

s.t. $\|x_i - a\|^2 \leq R^2 + \xi_i, \quad i = 1, \ldots, N$

$$\xi_i \geq 0, \quad \forall i, \quad (9)$$

where $R$ and $a$ are the hypersphere’s radius and center, respectively. Parameter $C$ controls the tradeoff between the volume and the errors.

The corresponding function of the primal SVDD optimization (9) is

$$L_{SVDD} = R^2 + C \sum_{i=1}^{N} (R^2 + \xi_i - (\|x_i\|^2 - \|a\|^2) - \sum_{i=1}^{N} \alpha_i \xi_i) - \sum_{i=1}^{N} \beta_i \xi_i \quad (10)$$

with the Lagrange variables $\alpha_i \geq 0$ and $\beta_i \geq 0$. $L_{SVDD}$ should be minimized with respect to $R$, $a$, $\xi_i$ and maximized with respect to $\alpha_i$, $\beta_i$.

Based on the Karush-Kuhn-Tucker (KKT) theorem [22], to get the optimal solutions of (10), we should have

$$\frac{\partial L_{SVDD}}{R} = 0 \implies \sum_{i=1}^{N} \alpha_i = 1, \quad (11a)$$
\[
\frac{\partial L_{SVDD}}{a} = 0 \implies a = \sum_{i=1}^{N} \alpha_i x_i, \quad (11b)
\]
\[
\frac{\partial L_{SVDD}}{\xi_i} = 0 \implies C = \alpha_i + \beta_i, \quad (11c)
\]
From (11c) and \(\alpha_i \geq 0\) and \(\beta_i \geq 0\), \(\beta_i\) can be removed and \(\alpha_i\) can be further limited to the interval \([0, C]\):
\[
0 \leq \alpha_i \leq C. \quad (12)
\]
Substituting (11a)–(11c) into (10), the dual optimization function can be derived as
\[
L_{D_{SVDD}} = \sum_{i=1}^{N} \alpha_i (x_i \cdot x_i) - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j (x_i \cdot x_j), \quad (13)
\]
subject to constraints (11a) and (12). To constitute a flexible data description model, kernel function \(K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)\), with an implicit feature mapping \(\phi\) of the data into a higher dimensional feature space, can be adopted to replace the inner product \(x_i \cdot x_j\). In this case, the corresponding dual optimization function is changed to
\[
L_{D_{SVDD}} = \sum_{i=1}^{N} x_i K(x_i, x_i) - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j K(x_i, x_j). \quad (14)
\]
The KKT conditions of the target functions are
\[
\alpha_i (R^2 + \xi_i - \|x_i - a\|^2) = 0,
\]
\[
\beta_i \xi_i = 0. \quad (15)
\]
The constraints have to be enforced and we have three cases as follows:

(1) \(\alpha_i = 0\)
\[
\beta_i = C \implies \xi_i = 0,
\alpha_i = 0 \implies R^2 \geq \|x_i - a\|^2, \quad (16)
\]

(2) \(0 < \alpha_i < C\)
\[
\beta_i > 0 \implies \xi_i = 0,
0 < \alpha_i < C \implies R^2 = \|x_i - a\|^2, \quad (17)
\]

(3) \(\alpha_i = C\)
\[
\beta_i = 0 \implies \xi_i = 0,
\alpha_i = C \implies R^2 \leq \|x_i - a\|^2. \quad (18)
\]

Only a small ratio of objects with \(\alpha_i > 0\) are called the support vectors. The dual optimization functions (13) and (14) are standard Quadratic Programming (QP) problems and the Lagrange variables \(\alpha_i\) can be obtained using some optimization methods such as SMO algorithm [39]. To test a new object \(z\), its distance to the center of the sphere is calculated. The classifier will accept the object if the distance is less than or equal to the radius:
\[
\|z - a\|^2 = (z \cdot z) - 2\sum_{i=1}^{N} \alpha_i (z \cdot x_i) + \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j (x_i \cdot x_j) \leq R^2. \quad (19)
\]

In addition to the batch learning model of SVDD, incremental learning methods [40] of SVM are extended to SVDD algorithm. Yin et al. [28] show an online fault diagnosis process through a hybrid model of incremental SVDD (ISVDD) and ELM with incremental output structure (IOELM). They used the ISVDD to detect the unknown failure model, and the output nodes of IOELM are adaptively increased to recognize the new failure mode.

3.2. The ELM Based One-Class Classifier. When data only from the target class are available, the one-class classifier is trained to accept target objects and reject objects that deviate significantly from the target class. In the training phase, the one-class classifier, which defines a distance function \(d\) between the objects and the target class, takes in the training set \(X\) to build the classification model. In general, the classification model contains two important parameters to be determined: threshold \(\theta\) and modal parameter \(\lambda\). A generic test sample \(z\) is accepted by the classifier if \(d(z \mid X, \lambda) < \theta\).

In the training phase, not all the training samples are to be accepted by the one-class classifier due to the presence of outliers or noisy data contained in the training set. Otherwise, the trained classification model may generalize poor to unknown test set when the training set includes abnormal data samples. Usually, threshold \(\theta\) is determined such that a user-specified fraction \(\mu\) of training samples most deviant from the target class are rejected. For instance, if one is told five percent of training samples are mislabeled, setting \(\mu = 0.05\) makes the classifier more robust. Even when all the samples are correctly labeled, rejecting a small fraction of training samples helps the classifier to learn the most representative model from the training samples.

Any one-class classifier has model parameters which influence the model complexity (flexibility), for example, the number of hidden nodes in autoencoder neural networks or the tradeoff parameter \(C\) of SVDD. Minimizing the errors of both the target and outlier classes on a cross-validation set is no longer available since there is no data from the outlier class. Fortunately, several model selection criteria [2] have been proposed. Assuming the uniform distribution of the outlier class, consistency-based model selection [41] method is one of the most effective methods used to select the model parameters. The basic idea is that the complexity of the classifier can be increased as long as it still fits the target data. The more complex the model, the smaller the volume of the classifier in the object space and the less the probability of outlier objects falling inside the domain of the classifier. In practice, one can make an ordering of the potential model parameters such that the latter parameter
always yields the more complex classifier and chooses the most complex classifier without overfitting the target data.

The compactness hypothesis [42] is the basis for object recognition. It states that similar real-world objects have to be close in the feature space. Therefore, for similar objects from the target class, the target outputs should be the same:

\[ t_i = y, \quad \forall x_i \in X, \quad (20) \]

where \( y \) is a real number. All the training samples’ target outputs are set to the same value \( y \). Then, the desired target output vector is \( T = [t_1, \ldots, t_N] = [y, \ldots, y]^T \). Training the samples from the target class can directly use the optimization function (2). For a new test sample \( z \), the distance function between the sample object and the target class is defined as

\[
d_{\text{ELM}}(z \mid X, \lambda) = \left| h(z)^T \beta - y \right| = \left| h(z)^T \mathbf{H}^T \left( \frac{1}{C} + \mathbf{H} \mathbf{H}^T \right)^{-1} \mathbf{T} - y \right|. \quad (21)
\]

The decision whether \( z \) belongs to the target class or not is based on threshold \( \theta \). Recall that \( \theta \) is optimized to reject a small fraction \( \mu \) of training samples to avoid overfitting. The distances of the training samples to the target class can be directly determined using (21) and the constraint of (2)

\[
d_{\text{ELM}}(x_i \mid X, \lambda) = \left| h(x_i)^T \beta - y \right| = \left| \xi \right|. \quad (22)
\]

From (22), we find the distances are \( \left| \xi \right| \) and the larger \( \left| \xi \right| \) means the more deviant of the training sample \( x_i \) from the target class. Hence, we derive threshold \( \theta \) based on a quantile function to reject the most deviant training samples. Denote the sorted sequence of the distances of training samples by \( d = [d_1, \ldots, d_N] \) such that \( d_1 \geq \cdots \geq d_N \). Here, \( d_1 \) and \( d_N \) represent the most and the least deviant samples. The function determining \( \theta \) can be written as

\[
\theta = d_{\text{floor}}(\mu N), \quad (23)
\]

where \( \text{floor}(a) \) returns the largest integer not greater than \( a \).

Then, we can get the decision function for \( z \) to the target class:

\[
\psi_{\text{ELM}}(z) = \text{sign} (\theta - d_{\text{ELM}}(z \mid X, \lambda)) = \begin{cases} 
1 & z \text{ is classified as a target} \\
-1 & z \text{ is classified as an outlier}. 
\end{cases} \quad (24)
\]

Remark 1. The target output \( y \) can be assigned to arbitrary real number except 0. When \( y = 0 \), seen from (6), the output weights between the hidden layer and the output layer become 0 (\( \beta = 0 \), \( 0 \) is the \( L \)-dimensional zero vector). Therefore, the decision value of any sample \( z \) is 0 using the proposed classifier. It is obvious that, in such case, the one-class classifier cannot distinguish between the target class and the outlier class. When \( y \neq 0 \), as there are infinite possible \( y \), there seem to exist infinite ELM based one-class classifiers.
get a universal ELM based one-class classifier, we normalize the distance function (21) by dividing the target output $y$

$$d_{\text{NORM-ELM}} (z \mid X, \lambda) = \frac{h(z)^T (1/C + HH^T)^{-1} T - y}{y}$$

$$= \frac{h(z)^T (1/C + HH^T)^{-1} e_y - y}{y}$$

$$= \frac{h(z)^T H^T \left( \frac{1}{C} + HH^T \right)^{-1} e}{y},$$

(25)

where $e$ is the $N$-dimensional unit vector. The normalization formula (25) is to eliminate the possible bias introduced by the target output $y$. In practice, one can set the target output $y = 1$ such that (21) is equivalent to (25) and the normalization step is implicitly done.

**Remark 2.** Both random feature mappings and kernels can be used for the proposed one-class classifier. When nonlinear piecewise continuous functions satisfying ELM universal approximation capability theorems [20, 21] are used as the activation function, the ELM network can approximate any target continuous function as long as the number of hidden nodes $L$ is large enough. When the feature mapping is unknown, kernel methods can be adopted as shown in (8a) and (8b). Huang et al. [17] have shown ELM, the unified solution for regression, binary, and multiclass classifications. Since the same optimization formula (2) is used in the proposed one-class classifier, this paper also shows ELM, the unified learning mode for one-class classification.

Figure 1 shows the decision boundaries (black curves) of the classifier with incremental hidden nodes using sigmoid function as the activation function. The dataset (blue points) is composed of 100 samples in the plane. Threshold $\theta$ is determined such that $\mu = 0.01$ and the model parameter $C$ is automatically determined by the consistency-based model selection method. When the number of hidden nodes is small ($L = 50$), the classifier fails to approximate the target region and some unexpected "holes" without any targets can be seen from the leftmost picture of Figure 1. The weakness alleviates as more hidden nodes are added. When the number of hidden nodes $L$ gets large enough, the classifier can be close enough to describe the target class well. This is consistent with ELM universal approximation capability theorems [20, 21].

**Remark 3.** Autoencoder is one of the most effective neural networks approaches for one-class classification, which has been applied by Manevitz and Yousef for document retrieval [43]. Constrain the number of output nodes that must be equal to the number of input nodes ($m = n$). The hidden layer in such a network actually acts as a bottleneck, where $L < n$. The idea is that while the bottleneck prevents learning the full identity function on $n$-space, the identity on the small set of examples is in fact learnable. Traditional learning algorithms like BP are used to train the network. Several challenging issues such as local minimum, trivial human intervention,
Table 2: The value of \( F_1 \) measure with standard deviations (in parentheses) for a number of one-class classifiers. Twenty trials have been conducted for each dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classifier</th>
<th>SPECTF heart</th>
<th>Arrhythmia</th>
<th>Sonar</th>
<th>Liver</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>41.7 (4.2)</td>
<td>61.8 (1.1)</td>
<td>46.8 (2.2)</td>
<td>41.5 (0.9)</td>
<td></td>
</tr>
<tr>
<td>Naive Parzen</td>
<td>39.3 (1.7)</td>
<td>63.7 (1.2)</td>
<td>49.8 (2.9)</td>
<td>40.7 (1.4)</td>
<td></td>
</tr>
<tr>
<td>Parzen</td>
<td>38.3 (4.7)</td>
<td>63.7 (1.7)</td>
<td>53.2 (3.2)</td>
<td>41.7 (1.4)</td>
<td></td>
</tr>
<tr>
<td>k-means</td>
<td>31.8 (2.6)</td>
<td>59.2 (1.5)</td>
<td>60.4 (2.2)</td>
<td>41.3 (1.3)</td>
<td></td>
</tr>
<tr>
<td>1-NN</td>
<td>34.7 (1.2)</td>
<td>62.4 (0.9)</td>
<td>55.3 (1.3)</td>
<td>42.0 (1.2)</td>
<td></td>
</tr>
<tr>
<td>k-NN</td>
<td>39.3 (3.4)</td>
<td>64.8 (1.6)</td>
<td>50.6 (2.4)</td>
<td>42.2 (1.7)</td>
<td></td>
</tr>
<tr>
<td>Autoencoder</td>
<td>36.3 (15.2)</td>
<td>62.8 (1.2)</td>
<td>53.3 (2.3)</td>
<td>41.6 (1.3)</td>
<td></td>
</tr>
<tr>
<td>PCA</td>
<td>38.9 (4.7)</td>
<td>60.5 (4.8)</td>
<td>51.2 (5.8)</td>
<td>40.6 (1.3)</td>
<td></td>
</tr>
<tr>
<td>MST</td>
<td>31.1 (8.7)</td>
<td>51.9 (5.0)</td>
<td>44.6 (6.3)</td>
<td>40.7 (2.0)</td>
<td></td>
</tr>
<tr>
<td>k-centers</td>
<td>38.3 (3.9)</td>
<td>63.8 (2.0)</td>
<td>52.2 (4.2)</td>
<td>40.7 (1.6)</td>
<td></td>
</tr>
<tr>
<td>SVDD</td>
<td>31.8 (6.4)</td>
<td>63.4 (1.9)</td>
<td>53.6 (3.1)</td>
<td>40.5 (2.4)</td>
<td></td>
</tr>
<tr>
<td>Autoencoder</td>
<td>38.3 (15.4)</td>
<td>65.5 (1.9)</td>
<td>37.9 (10.9)</td>
<td>48.7 (2.7)</td>
<td></td>
</tr>
<tr>
<td>PCA</td>
<td>33.7 (15.4)</td>
<td>65.3 (1.9)</td>
<td>31.1 (1.0)</td>
<td>46.0 (0.5)</td>
<td></td>
</tr>
<tr>
<td>MST</td>
<td>36.3 (12.9)</td>
<td>67.5 (0.7)</td>
<td>34.4 (3.4)</td>
<td>473 (0.9)</td>
<td></td>
</tr>
<tr>
<td>k-centers</td>
<td>38.8 (6.5)</td>
<td>67.7 (1.1)</td>
<td>49.4 (22.9)</td>
<td>42.5 (2.6)</td>
<td></td>
</tr>
<tr>
<td>SVDD</td>
<td>50.9 (10.6)</td>
<td>61.1 (2.5)</td>
<td>68.0 (9.1)</td>
<td>44.5 (3.3)</td>
<td></td>
</tr>
<tr>
<td>MPM</td>
<td>38.8 (11.8)</td>
<td>63.5 (1.7)</td>
<td>71.7 (6.5)</td>
<td>44.9 (1.9)</td>
<td></td>
</tr>
<tr>
<td>LPDD</td>
<td>67.8 (11.0)</td>
<td>66.6 (0.8)</td>
<td>79.6 (7.5)</td>
<td>45.8 (2.0)</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>57.2 (12.8)</td>
<td>66.6 (1.1)</td>
<td>83.1 (3.0)</td>
<td>46.2 (1.2)</td>
<td></td>
</tr>
<tr>
<td>ELM</td>
<td>42.6 (1.8)</td>
<td>63.6 (1.6)</td>
<td>54.2 (3.5)</td>
<td>43.0 (1.6)</td>
<td></td>
</tr>
</tbody>
</table>

\( ^1 \)None of target data is recalled.

Table 3: The value of \( F_1 \) measure with standard deviations (in parentheses) for a number of one-class classifiers. Twenty trials have been conducted for each dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>E. coli</th>
<th>Diabetes</th>
<th>Breast</th>
<th>Abalone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>71.7 (7.0)</td>
<td>68.7 (1.1)</td>
<td>82.4 (3.2)</td>
<td>51.8 (0.3)</td>
</tr>
<tr>
<td>Naive Parzen</td>
<td>75.1 (5.7)</td>
<td>67.7 (1.4)</td>
<td>80.1 (7.7)</td>
<td>49.0 (1.0)</td>
</tr>
<tr>
<td>Parzen</td>
<td>54.6 (15.2)</td>
<td>68.9 (1.1)</td>
<td>58.8 (17.2)</td>
<td>45.2 (1.3)</td>
</tr>
<tr>
<td>k-means</td>
<td>21.2 (3.9)</td>
<td>64.8 (0.9)</td>
<td>35.3 (5.8)</td>
<td>35.7 (1.0)</td>
</tr>
<tr>
<td>1-NN</td>
<td>43.9 (14.2)</td>
<td>68.8 (1.0)</td>
<td>34.9 (7.5)</td>
<td>49.8 (1.4)</td>
</tr>
<tr>
<td>k-NN</td>
<td>53.5 (15.8)</td>
<td>66.9 (1.3)</td>
<td>37.9 (10.9)</td>
<td>48.7 (2.7)</td>
</tr>
<tr>
<td>Autoencoder</td>
<td>33.7 (15.4)</td>
<td>65.5 (1.9)</td>
<td>31.1 (1.0)</td>
<td>46.0 (0.5)</td>
</tr>
<tr>
<td>PCA</td>
<td>36.3 (12.9)</td>
<td>67.5 (0.7)</td>
<td>34.4 (3.4)</td>
<td>473 (0.9)</td>
</tr>
<tr>
<td>MST</td>
<td>38.8 (6.5)</td>
<td>67.7 (1.1)</td>
<td>49.4 (22.9)</td>
<td>42.5 (2.6)</td>
</tr>
<tr>
<td>k-centers</td>
<td>50.9 (10.6)</td>
<td>61.1 (2.5)</td>
<td>68.0 (9.1)</td>
<td>44.5 (3.3)</td>
</tr>
<tr>
<td>SVDD</td>
<td>38.8 (11.8)</td>
<td>63.5 (1.7)</td>
<td>71.7 (6.5)</td>
<td>44.9 (1.9)</td>
</tr>
<tr>
<td>MPM</td>
<td>67.8 (11.0)</td>
<td>66.6 (0.8)</td>
<td>79.6 (7.5)</td>
<td>45.8 (2.0)</td>
</tr>
<tr>
<td>LPDD</td>
<td>57.2 (12.8)</td>
<td>66.6 (1.1)</td>
<td>83.1 (3.0)</td>
<td>46.2 (1.2)</td>
</tr>
<tr>
<td>SVM</td>
<td>77.1 (4.8)</td>
<td>69.1 (1.1)</td>
<td>80.1 (5.1)</td>
<td>53.0 (0.7)</td>
</tr>
</tbody>
</table>

and time consuming in learning stage discourage people who are not familiar in the field to use it, while the ELM based one-class classifier can approximate the target class well as long as the dimensionality of the feature mappings is large enough (cf. Figure 2).

4. Experiments

4.1. Artificial Datasets. First, we illustrate the proposed method with both random feature mappings and kernels on three specific designed artificial datasets, which all contain 100 samples created in a 2D feature space. The first dataset contains four Gaussian distributions (each has 25 samples) with the same unit covariance matrix but with different mean vectors. It is set to test the classifier’s sensitivity to multimodality. The second dataset contains one Gaussian distribution with the first feature with a variance of 1 and the second feature with a variance of 40. Moreover, the two features are rotated over 45 degrees to construct a strong correlation. The third banana-shaped dataset, which has been shown in Section 3, contains one uniform distribution along an arc curve with some small position offsets. It is to test the influence of convexity. In Figure 3, the datasets (blue points) together with the decision boundaries (black curves) in the feature space are illustrated. Sigmoid function acts as
the activation function for the method with random feature mappings \( L \text{ large enough} \) and Gaussian kernel \( K(x_i, x_j) = \exp(-\|x_i - x_j\|^2/\sigma^2) \) is used for the method with kernels. All the thresholds are determined such that \( \mu = 0.01 \). The pictures show that methods using both random feature mappings and kernels give reasonable results. However, the method with kernels tends to be superior to the method with random feature mappings since the boundary captures the distribution more precisely, while in Figure 3(a) some small “holes” still exist in the upper left and lower right regions for the method with random feature mappings.

4.2. UCI Datasets. This section compares the performance of the proposed method with a variety of one-class classification algorithms. The popular one-class classifiers to be compared include Parzen [7], Naïve Parzen, \( k \)-means [44], \( k \)-centers [45], 1-NN [46], \( k \)-NN [47], autoencoder, PCA [48], MST [14], MPM [13], SVDD [10], LPDD [11], and SVM [9]. The implementations for one-class SVM are carried out using compiled C-coded SVM packages: LIBSVM [49]. All the other algorithms are conducted with Matlab toolbox DDTOOLS [50]. Binary and multiclass classification datasets taken from UCI Machine Learning Repository [51] are used. The specifications of the datasets are shown in Table 1. The datasets are transformed for one-class classification by setting a chosen class as the target class and all the other classes as the outlier class.

In our experiments, all the inputs have been normalized into range \([0, 1]\). The samples from the target class are equally partitioned in two sets for training and testing, respectively. All one-class classifiers are trained on target data only and tested on both the remaining target data and all other nontarget data. To assess the performance, we use \( F_1 \) measure [52], which is defined as a combination of recall (R) and precision (P) with an equal weight in the following form:

\[
F_1 (R, P) = \frac{2RP}{(R + P)}.
\]

All the thresholds \( \theta \) are determined such that \( \mu = 0.1 \). The Gaussian kernel is used in Parzen, Naïve Parzen, MPM, SVDD, SVM, and ELM. The consistency-based model selection method is employed to select the model parameters. For

![Figure 3: The upper row shows the boundaries of the method with random feature mappings and the bottom row shows the boundaries of the method with Gaussian kernel. Parameter \( C \) of the method with random feature mapping from left to right is \( 2^{11}, 2^6 \), and \( 2^9 \). Parameters (\( C, \sigma \)) of the method with Gaussian kernel from left to right are \( (2^{11}, 5.76), (2^6, 2.81) \), and \( (2^9, 1.55) \).

### Table 4: The value of precision and recall for two datasets (arrhythmia and E. coli).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Arrhythmia</th>
<th>E. coli</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Naïve Parzen</td>
<td>45.6</td>
<td>96.0</td>
</tr>
<tr>
<td>Parzen</td>
<td>52.0</td>
<td>82.4</td>
</tr>
<tr>
<td>( k )-means</td>
<td>52.5</td>
<td>81.0</td>
</tr>
<tr>
<td>1-NN</td>
<td>44.5</td>
<td>88.6</td>
</tr>
<tr>
<td>( k )-NN</td>
<td>47.8</td>
<td>90.2</td>
</tr>
<tr>
<td>Autoencoder</td>
<td>52.6</td>
<td>84.7</td>
</tr>
<tr>
<td>PCA</td>
<td>79.1</td>
<td>15.9</td>
</tr>
<tr>
<td>MST</td>
<td>47.3</td>
<td>91.8</td>
</tr>
<tr>
<td>( k )-centers</td>
<td>50.4</td>
<td>83.4</td>
</tr>
<tr>
<td>SVDD</td>
<td>55.7</td>
<td>69.3</td>
</tr>
<tr>
<td>MPM</td>
<td>64.0</td>
<td>43.9</td>
</tr>
<tr>
<td>LPDD</td>
<td>51.7</td>
<td>83.3</td>
</tr>
<tr>
<td>SVM</td>
<td>52.4</td>
<td>80.5</td>
</tr>
<tr>
<td>ELM</td>
<td>52.8</td>
<td>80.1</td>
</tr>
</tbody>
</table>
Parzen, MPM, SVDD, SVM, and ELM, the kernel parameter $\sigma$ is chosen from 20 aliquots between the minimum and maximum pairwise object distances, so as the smoothing parameter of sigmoid transform function used in LPDD. For $k$-means, $k$-centers, parameter $k$ is selected from the range $\{1, 2, \ldots, 20\}$. For ELM, another parameter $C$ is chosen from the range $\{2^{-24}, 2^{-23}, \ldots, 2^{25}\}$ and we set $\sigma$ a higher priority than $C$; that is, when the parameter combinations, $(\sigma_1, C_1)$ and $(\sigma_2, C_2)$, both obtain consistent boundaries, we always choose a smaller $\sigma$ rather than a larger $C$. We try every possible parameter setting and find the most complex classifier as long as the classifier is consistent. For Naive Parzen and $k$-NN, the leave-one-out maximum likelihood estimation is used. One-class PCA retains 0.95 variance for the training set. For MST, the complete minimum spanning tree is used. The number of hidden nodes in autoencoder neural network is carefully chosen from a large range and the optimal number is selected.

All the experiments are carried out in Matlab R2013a environment running in E5504 2GHz CPU, 4GB RAM. Twenty trials have been conducted for each dataset and the average $F_1$ and corresponding standard deviations are shown in Tables 2 and 3. The best results are shown in boldface. As an example, we give a detailed description for the diabetes experiment. First, all the samples from both the target class and the outlier class are normalized into range $[0, 1]$. Then, the 500 training target samples are randomly divided into two equal sets (250 samples for each set). One of the sets is used for training the one-class classifier and the other set, together with all the samples from the outlier class, is used for testing only. After that, the consistency-based model selection method is employed to select the model parameters for each classifier using only the training set. Finally, the other target set with the outlier set is judged by the trained classifier with precision and recall recorded. $F_1$ value is then derived as (26). The same procedure repeats for twenty times and the corresponding mean and deviation values are calculated. It can be seen that the generalization performance of ELM is the best in five of the eight experiments while in the other experiments, except for sonar dataset, the performance is comparable to the best classifier. Table 4 presents a detailed performance comparison of two datasets, including precision and recall.

Table 5 reports the execution time comparisons in seconds between the ELM, autoencoder, and SVDD classifiers for all the eight experiments. As observed from Table 5, the advantage of the ELM on training time is quite obvious. ELM can generally learn hundreds of times faster than autoencoder neural network due to the tuning-free mechanism. Besides, ELM also learns much faster than SVDD without solving a QP problem. For testing time, since autoencoder may obtain a more compact network and the parameters have been tuned in the training phase, the computational time depends on the specific task. The computational complexity of ELM mostly depends on the number of samples while autoencoder depends on both the number of samples and the number of dimensions. Thus, for datasets with relatively small size and high dimensions, such as arrhythmia dataset, ELM obtains a smaller testing time, while for datasets with relatively large size and low dimensions, such as abalone dataset, autoencoder reacts faster to the testing samples. However, ELM still tends to outperform autoencoder with respect to both training time and accuracy. It is obvious that ELM and SVDD obtain a similar testing time since both of them utilize a kernel function.

### 5. Conclusion

This paper presents a simple and efficient one-class classifier utilizing extreme learning machine, which also shows ELM, the unified learning mode for one-class classification. Both random feature mappings and kernels can be used for the proposed classifier while the method with kernels tends to be superior to the method with random feature mappings. Moreover, the proposed classifier with kernels achieves the best results on five of the eight UCI datasets, which suggests ELM being effective for one-class classification problem. We have also discussed the relationships and differences between autoencoder neural network and ELM network for one-class classification. Although autoencoder neural network has been successfully applied in many applications, the slow gradient based method is still used to tune all the parameters, which is far slower than required. On the other hand, the ELM based one-class classifier has an analytical solution which can obtain superior generalization performance at much faster learning speed. Possible future directions include the fusion of fuzzy logic and ELM for one-class classification, one-class classifier ensembles with ELM, and substituting autoencoder with the ELM based one-class classifier for deep learning.
Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is partially sponsored by Natural Science Foundation of China (nos. 6175115, 61272320, 61379100, and 61472388). The authors would like to thank the helpful discussions with Mr. Fan Wang and Dr. Laiyun Qing.

References


Submit your manuscripts at http://www.hindawi.com