Using a dynamically selective support vector data description model to discover novelties in the control system of electric arc furnace

Jiong Zhang1,2, Yue Wang3, Qian Li4 and Biao Wang5

Abstract
As increasing data-driven control strategies are applied in electric arc furnace systems, the problem of novelty detection has drawn more attentions than before. The presence of outliers should be the main obstacle in practical applications for these advanced control techniques. To this end, this paper proposes a dynamically selective support vector data description model to discover novelties in electric arc furnace. In this model, support vector data description plays the role of base detector. Artificial outliers are generated with two objectives, one is to assist the dynamic selection, and the other is to optimize two parameters of support vector data description. Then clustering technique is used to determine the validation set for each test point. Finally, a probabilistic method is used to compute the competence of base detectors. In contrast to other novelty ensembles that have parallel structures, our ensemble model has a dynamic selection mechanism that could facilitate the mining of the potential of base detectors. Three synthetic and three real-world data-sets are used to validate the effectiveness of the proposed detection model. Experimental results have approved our method by comparing it with several competitors.

Keywords
Electric arc furnace, novelty detection, support vector data description, ensemble learning

Introduction
Novelty detection or outlier/anomaly detection techniques have been applied in many practical domains, such as fraud detection for credit cards, intrusion detection for cyber-security, faulty detection for industrial systems, to name but a few. However, developing dedicated novelty detectors for industrial control systems has rarely been taken into account, let alone for an electric arc furnace (EAF) control system. During recent years, many industrial systems (including EAF) have introduced data-driven techniques to facilitate the modeling and control processes as more and more process data could be collected and stored. Inspired by this, novelty detection is drawing increasing attentions in industrial systems, because anomalous observations have adverse impact on both the modeling and the control process that use any data-driven technique. In the control system of EAF, outliers are referred to as observations that cannot reflect the normal system states.

Recently, several advanced control strategies such as adaptive control and model predictive control have been proposed to apply to EAF systems, in order to improve the control performance and save energy. We have noticed that in these control methods, some machine learning algorithms like neural network have been used to establish the process model of EAF. It is well known that these data-driven models are often very sensitive to outliers in training sets and the resultant control performance will deteriorated when they have been included. (This should be the main reason why these advanced data-driven control strategies have not been extensively used in practical EAF control.

1School of Mechanical Engineering, Shandong University, Jinan, China
2School of Information, Shandong Institute of Commerce & Technology, Jinan, China
3Department of Fundamental Teaching, Liaoning Vocational College of Ecological Engineering, Shenyang, China
4School of Medical Information and Engineering, Shenyang Medical College, Shenyang, China
5School of Automation, Shenyang Aerospace University, Shenyang, China

Corresponding author:
Biao Wang, School of Automation, Shenyang Aerospace University, Shenyang 110136, China.
Email: wangbiao_neu@126.com

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
In this situation, implementing an efficient novelty detector on the EAF control system may be beneficial. It is noteworthy that this novelty detector has a subtle distinction with a process monitoring model. Here, we are only interested in the controlled variables or variables that would be used by data-driven control strategies. In contrast, process monitoring concerns more about the state variables of systems.

In spite that rare dedicated detectors in literature have been proposed for EAF, there still many existing techniques in machine learning and data mining that can be used here. According to the availability of supervision, these novelty detectors are often categorized into three groups, that is, supervised detectors, unsupervised detectors, and semi-supervised detectors. Supervised detectors use labeled training data to learn conventional classifiers, such as support vector machine (SVM) and decision tree, that can separate normal observations from anomalous ones. A crucial drawback of this type of detectors resides on the need of labeled training data, which is of great challenges for most practical applications including EAF, because labeling process data is a time- and human-consuming work. In contrast, unsupervised detectors can use some similarity criteria like distance and density to mine potential outliers in databases. Therefore, these detectors are usually used in off-line ways and most representative methods should be distance-based detector and local outlier factor (LOF). Semi-supervised detectors are also referred to as one-class (OC) classifiers and data description techniques. The basic idea of this type of detectors is that a normal pattern can be learnt since all training samples are assumed from the target class. In this paper, we will often use the term OC classifiers to denote semi-supervised detectors. The most representative OC classifiers should be support vector data description (SVDD), which aims to enclose all training data with a hypersphere whose volume should be as small as possible. Based on the mechanisms of these detectors, those semi-supervised ones are more appropriate for novelty detection in an online manner. Then in order to improve the performance of single detectors, several ensemble models have been proposed. By combining diverse base detectors, the final detection performance could be enhanced. Note that a notable limitation of these ensemble models is that the used base detectors should be accurate and diverse simultaneously. This assumption can hardly be satisfied when we have none data labels. In this paper, we propose a dynamic selective model that uses SVDD as base detectors. Our detector can also be deemed an ensemble model as several base learners are necessary. In contrast to above ensemble detectors where a fix model structure will be used for all test points, our detector will select the most competent base detector for each test point dynamically. In order to facilitate the selective procedure, we use a trick to generate artificial outliers. Moreover, these outliers will also be used to obtain an optimal parameters of algorithm SVDD in our detector. We have noted that this problem has always been ignored by many researches that use SVDD. In addition, clustering technique is also used in the selective process to determine validation set for each test point.

Here, we conclude the contributions as follows:

- A dynamic novelty detection model is proposed for EAF control system.
- Artificial outlier examples are generated to determine validation sets and optimize parameters of SVDD.
- Datasets from real-world EAF system are used to verify the effectiveness of our detection model.

The rest of paper is organized as follows. Some related works and preliminaries will be presented in section “Related works and preliminaries.” The proposed method will be introduced in section “Methodology,” followed by the experiments in section “Experiments and analysis.” Finally, some conclusions will be drawn in section “Conclusion.”

**Related works and preliminaries**

Several related works regarding novelty detection in EAF systems will be introduced in spite of their sparseness. Then some necessary preliminaries will also be presented simply.

**Related work**

In Liu et al., a model-based novelty detection model is proposed for the process control system of EAF. In this model, an improved Radial Basis Function (RBF) neural network is first used to establish the process model of EAF. Then hidden Markov model (HMM) is used to analyze the residuals between the true measurements and the results of this process model. From our point of view, the main drawback of this method is that it can only be used for univariate dataset. For multivariate datasets, several such models may be necessary. Then it heavily depends on the predictive model, and the detection performance will deteriorate much when the predictions are biased. In Wang and Mao, a clustering-based ensemble detector is proposed for EAF control system. In this method, a clustering algorithm is used first to separate the training set into several subsets, in each of which a single detector is established. Then any test point will be labeled as an outlier if it is rejected by all base detectors. In Wang and Mao, technique Random Subspace (RS) is used to develop an ensemble detector. In this method, RS is used first to divide the feature space into several subspaces, then all training points will be projected onto these subspaces to generate several training subsets, on which corresponding base detectors can be trained. Then a combination rule is used to derive the ultimate
result for each test point. As we have mentioned previously that a main drawback of these ensemble detectors is that generating accurate and diverse base detectors may be difficult for some situations.

In fields of machine learning and data mining, novelty detection is always a hot topic since outliers often indicate interesting data patterns. In general, existing novelty detection methods can be categorized into probabilistic detection models, distance-based detection models, reconstruction-based detection models, and domain-based detection models. For probabilistic detection models, Gaussian mixture model (GMM) should be one of the most popular parametric ones, and HMM and Kalman filter are another two commonly used parametric ones. Kernel density estimation should be the most popular non-parametric detection method. For distance-based detection models, methods based on nearest neighbors and clustering algorithms are the most common used. Gaussian mixture model (GMM) should be one of the most popular non-parametric ones. The SVM is introduced. In addition, kernel trick $\phi(x)$ is also used to make the description more flexible. The minimization problem becomes hence the following

$$\min_{R,a,\xi} R^2 + C \sum_{i=1}^{n} \xi_i$$

subject to

$$||\phi(x_i) - a||^2 \leq R^2 + \xi_i$$

where the parameter $C$ controls the balance between the volume of the hypersphere and errors on the target class. Larger value of $C$ implies that less training points should be inside the hypersphere. Note that the above convex optimization problem can be solved through its dual form

$$L(R,a,\alpha_i,\gamma_i,\xi_i) = R^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i (R^2 + \xi_i - ||\phi(x_i) - a||^2) - \sum_{i=1}^{n} \gamma_i \xi_i$$

By setting partial derivations to zero, the dual optimization problem has changed into

$$\min_{\alpha_i} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j (\phi(x_i) \cdot \phi(x_j)) - \sum_{i=1}^{n} \alpha_i (\phi(x_i) \cdot \phi(x_i))$$

subject to

$$\sum_{i=1}^{n} \alpha_i = 1$$

$$0 \leq \alpha_i \leq C, i = 1, 2, \ldots, n$$

Solution of this dual optimization problem are a set of values of $\alpha_i$. Objects $x_i$ with $\alpha_i > 0$ are referred to as support vectors (SVs), with which the center $a$ can be calculated. Furthermore, SVs on the boundary can be used to calculate the radius

$$R^2 = \frac{1}{|SV < C|} \sum_{x_i \in SV < C} ||\phi(x_i) - a||^2$$

where $SV < C$ is the set of SVs with $0 < \alpha_i < C$.

For any test point $x$, the distance from $x$ to the center $a$ is derived by

$$d = \sqrt{1 - 2 \sum_{i=1}^{n} \alpha_i K(x_i, x) + \sum_{i=j=1}^{n} \alpha_i \alpha_j K(x_i, x_j)}$$

If $d$ is larger than $R$, sample $x$ is labeled as an outlier or a fault.

**Preliminaries**

Basic concepts concerning SVDD and ensemble learning will be introduced simply.

**SVDD.** Algorithm SVDD defines a model by using a hypersphere to give a closed boundary around all observations in the training set. This hypersphere can be characterized by center $a$ and radius $a$ in original feature space. In order to enlarge the possibility of outliers in the training set, a slack variable $\xi$ is introduced. In addition, kernel trick $\phi(x_i)$ is also used to make the description more flexible. The minimization problem becomes hence the following

$$\min_{R,a,\xi} R^2 + C \sum_{i=1}^{n} \xi_i$$

subject to

$$||\phi(x_i) - a||^2 \leq R^2 + \xi_i$$

where the parameter $C$ controls the balance between the volume of the hypersphere and errors on the target class. Larger value of $C$ implies that less training points should be inside the hypersphere. Note that the above convex optimization problem can be solved through its dual form

$$L(R,a,\alpha_i,\gamma_i,\xi_i) = R^2 + C \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i (R^2 + \xi_i - ||\phi(x_i) - a||^2) - \sum_{i=1}^{n} \gamma_i \xi_i$$

By setting partial derivations to zero, the dual optimization problem has changed into

$$\min_{\alpha_i} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j (\phi(x_i) \cdot \phi(x_j)) - \sum_{i=1}^{n} \alpha_i (\phi(x_i) \cdot \phi(x_i))$$

subject to

$$\sum_{i=1}^{n} \alpha_i = 1$$

$$0 \leq \alpha_i \leq C, i = 1, 2, \ldots, n$$

Solution of this dual optimization problem are a set of values of $\alpha_i$. Objects $x_i$ with $\alpha_i > 0$ are referred to as support vectors (SVs), with which the center $a$ can be calculated. Furthermore, SVs on the boundary can be used to calculate the radius

$$R^2 = \frac{1}{|SV < C|} \sum_{x_i \in SV < C} ||\phi(x_i) - a||^2$$

where $SV < C$ is the set of SVs with $0 < \alpha_i < C$.

For any test point $x$, the distance from $x$ to the center $a$ is derived by

$$d = \sqrt{1 - 2 \sum_{i=1}^{n} \alpha_i K(x_i, x) + \sum_{i=j=1}^{n} \alpha_i \alpha_j K(x_i, x_j)}$$

If $d$ is larger than $R$, sample $x$ is labeled as an outlier or a fault.

**Ensemble of OC classifiers.** Designing parallel ensembles of OC classifiers is easier than sequential ones because those techniques used in conventional classification problem can be directly used here. The rationale of parallel ensemble resides in reducing the variance by inducing diverse base detectors or OC classifiers and aggregating them. Several strategies could be used to enhance this diversity. The most commonly used technique should be Bagging that uses a bootstrap sampling. By fusing individuals learnt on different training subsets, Bagging is expected to obtain more robust result. Another well-known strategy of enhancing diversity is to use different feature subsets. RS and feature bagging (FB) should be the most commonly used techniques of this type. Note that subspace-based outlier ensembles are more efficient on high-dimensional datasets since outliers could be easily masked there and may be exposed in certain subspace. Apart from these two strategies, using different model parameters (or initializations) and even different algorithms have also been proposed to enhance diversity of outlier ensembles.

**EAF.** The EAF is a highly energy-intensive process used to convert scrap metal into molten steel. EAFs range in capacity from a few tons to as many as 400 tons. Figure 1 gives a simple description of the EAF operation. The graphite electrodes that connect to the
electrical supply could convert the electrical energy into thermal energy via the electric arcs between electrodes and the steel scrap surface. In addition, natural gas and oxygen would also been injected into the furnace so that the releasing chemical energy could be converted into the thermal energy. The scrap keeps melting by absorbing extensive thermal energy. Therefore, the scrap surface becomes irregular as parts of the scrap melt and removed, leading to the contours of the scrap surface. The corresponding disturbances will occur for the arc length. Then the electrode regulate system will response to these disturbances by adjusting the distance from the electrodes to the scrap surface so that the optimal arc length can be obtained. We should also note that when sufficient space is available inside the furnace, another scrap charge will be added. Then the melting process will proceed until a flat batch of molten steel is formed at the end of the batch.

Methodology

Base detectors

As discussed previously that our detection model can also be deemed an ensemble model, the generation of base detectors is thus indispensable. In order to train diverse and accurate SVDD models in our ensemble, a subspace-based ensemble technique called FB17 is used. It is observed that subspace-based ensemble techniques are often more efficient than subsampling techniques in unsupervised learning, even the dimension of the given data is not very high.18 The basic steps of FB can be described as follows:

1. Sample an integer \( r \) from \( d/2 \) to \( d - 1 \);
2. Select \( r \) dimensions from the data \( D \) randomly to create an \( r \)-dimensional projection;
3. Use base detector on projected representation;
4. Repeat the above steps until \( M \) iterations.

The rationale is to sample an integer \( r \) between \( d/2 \) and \( d - 1 \), and \( r \) dimensions are then sampled from the dataset. The data are finally scored in this projection.

Artificial outliers

Before introducing the generation of artificial outliers, we first give a simple description of dynamic ensemble, by which the goal of artificial outliers can hence be explained. A general training and test process of dynamic ensemble learning can be demonstrated in Figure 2.

Once base detectors have been trained, a selective procedure will be implemented for each test point. We can see from Figure 2 that a validation set is necessary to complete the selection. The objective of this validation set is to provide reference examples in order to identify the competence of all base detectors, with respect to this test point. Then the most competent base detector(s) will be selected according to the result of competence calculations.

From the above description of dynamic ensemble, we can find that the role of validation set is very critical as it is the premise of competence calculation. However, we have none labeled training examples in novelty detection to constitute this validation set. As a result, we have to generate some artificial ones instead. A simple strategy of generating artificial outliers is by sampling examples from a bounded uniform distribution. Another strategy is to assume that outlier examples locate in sparse regions of the target domain, that is, regions where the target data are either absent or isolated from the rest of the data.19 Fan et al.20 propose to generate outliers close to the target data by constraining the learning algorithm to form an accurate boundary between known classes and anomalies. The value of one feature of a target point is changed randomly while leaving other features unchanged. Major drawbacks of these two methods are the impossibility to generate a sufficient amount of outlier examples in high-dimensional situations due to the curse of dimensionality. To this end, Tax and Duin21 propose to generate a uniform hyper-spherical outlier distribution that might fit tighter around the target class than a hyper-box distribution. However, such an approach to generate artificial outliers is mainly used to optimize parameters of OC classifiers, and the resultant outlier...
class heavily covers the target class. Consequently, it is not appropriate for providing reference validation sets in dynamic ensembles.

In Désir et al., randomization principles of ensemble learning methods to subsample the number of features and the number of training target instances are used to generate artificial outliers from a computational point of view. Random subspace method (RSM) and Bagging are used to subsample features and training set, respectively. Then the amount of required outliers has been reduced much more than the original version. Furthermore, sparsity information extracted from the original training set is also used to make the artificial outlier class complementary to the target class. Experimental results on several benchmark datasets have shown the superiority of this method. Inspired by this, we also use such a strategy to generate outliers in this paper. But some adjustments are made here. In particular, RSM and Bagging will not be used to sample features and training set, and only FB is used instead since it is also used to train base detectors.

Score normalization

Before the selective procedure, a normalization procedure for all base detectors is of great necessity in order to achieve an unbiased selection. As has been proved that even when using the same method as base detector and identical parameterization, outlier scores obtained from different subspaces could vary considerably, if some subspaces have largely different scales. For techniques concerning outlier score normalization, those converting outlier scores of different base detectors into probability estimates are more acknowledged. As claimed by Gao and Tan, there are many advantages to transforming outlier scores into well-calibrated probability estimates, and a dominant one is that the probability estimates are more appropriate for developing an ensemble outlier detection framework. Sigmoid function and mixture modeling are accordingly used to fit outlier scores into probability values in this study. While in Kriegel et al., a more general framework of outlier score normalization is provided. The fundamental motivation of this framework is to establish sufficient contrast between outlier scores and inlier scores so that outliers could be easily separated from inliers. This seems more practical than only the interpretation of outlier scores because we actually need to pick out outliers in some applications. However, we may encounter a problem if we directly use these normalization methods. Note that normalization methods in Gao and Tan and Kriegel et al. are used for mining outliers in given databases. The normalization or scaling procedures are implemented with samples only in the given database. When using these procedures for unseen test samples, the probability of an observation being an outlier may be out of the range of \([0, 1]\), leading to the loss of interpretability of probability estimates.

To address this problem, we first do some adjustments on outputs of base detectors before converting to probabilistic estimates. Generally, normalizing them to Z-scores is a good choice, and its effectiveness has been verified in Aggarwal and Sathe and Nguyen et al. However, space of improvement still exists like that using Z-score for outlier detection for univariate data ("3σ edit rule"). As we all know that Z-scores are sensitive to outliers because the presence of outliers tends to inflate the variance estimate. Z-scores of all normal data would therefore move toward those of outliers, making the contrast between outliers and inliers smaller. As we have emphasized that sufficient contrast between outliers and inliers is preferred by most scenarios of outlier or novelty detection. In light of this, we propose to adjust the calculation of Z-score just like the improvement by "Hample Identifier," which replaces the outlier-sensitive mean and standard deviation estimates with the outlier-resistant median and median absolute deviation from the median (MAD), respectively.

Dynamic selection

Validation set. In general, algorithm K-nearest neighborhood (KNN) is always used to determine the validation set of any test point in dynamic ensembles. However, calculating the distances to all data points in the training set is necessary to find its nearest K neighbors. The cost of computation is too expensive for online applications sometimes. While for the clustering-based method only the determination of its belonging cluster is
necessary so long as all data points have been divided into several clusters at the training phase. As a result, we prefer to use a clustering algorithm to determine the validation set for each test point.

Here, we choose three representative clustering algorithms as candidates and do some quantitative comparisons. The first one is the classical K-means clustering, the second one is GMM, the third one is a density-based algorithm named DBSCAN (density-based spatial clustering of applications with noise). K-means should be the most popular clustering algorithm due to its simple theory and implementation. Its drawbacks mainly reside on its sensitivity to initial values, noise, and outliers. Correspondingly, several improved versions have also been proposed. GMMs are among the most statistically mature methods for clustering. Each cluster is represented by a Gaussian distribution. The clustering process thereby turns to estimate the parameters of the Gaussian mixture, usually by the Expectation-Maximization algorithm. Its probabilistic form of output may be an advantage, which make the learning models more smoothly and naturally. But its drawbacks lie on its probabilistic assumption, which also requests more on the size of samples and representativeness. The largest advantage of DBSCAN should reside on its ability of discovering clusters of arbitrary shape. It also requires less input parameter, including the number of clusters. While it becomes unstable when detecting border objects of adjacent clusters.

The quantitative criterion we use is Calinski–Harabasz (CH). This quantity is defined as

$$CH_k = \frac{SSB}{SSW} \times \frac{(N - k)}{k - 1}$$

(10)

where $SSB$ denotes the overall between-cluster variance, $SSW$ is the overall within-cluster variance, $k$ is the number of clusters, and $N$ is the number of observations. Then $SSB$ and $SSW$ are defined as

$$SSB = \sum_{i=1}^{k} n_i \| m_i - m \|^2$$

(11)

$$SSW = \sum_{i=1}^{k} \sum_{x \in C_i} \| x - m_i \|^2$$

(12)

where $n_i$ indicates the number of observations in cluster $i$, $m_i$ is the centroid of cluster $i$, $m$ is the overall mean of the sample data, $\| m_i - m \|^2$ is the $L^2$ norm between $m_i$ and $m$, and $\| x - m_i \|^2$ is the $L^2$ norm between $x$ and $m_i$. Note that the CH index can be deemed the ratio of between-cluster variance and within-cluster variance, and larger CH value indicates better data partition. The optimal number of clusters can be determined by maximizing $CH_k$ with respect to $k$.

When classifying a test point, we should decide its belonging cluster first. Then all data points in that cluster constitute the validation set with respect to this test point.

### Competence calculation

It is desired that we could select the most competent base detector by computing their competences through the validation sets. With the artificial outliers in the training set, we can employ selection mechanisms proposed traditional classification problem. More than just estimating the classifier accuracy on the basis of a simple percentage of corrected classified samples, here we use a probabilistic measure to select the most competent classifier.

Let $V = \{ X_1, \ldots, X_N \}$ denote the validation set of test pattern $X_t$. For each base classifier $C_j$, the probability of correct classification of any test pattern can be estimated by (if class labels of training patterns are available)

$$p(\text{correct}_j) = \frac{N_j}{K} j = 1, \ldots, L$$

(13)

where $N_j$ is the number of neighbor patterns that are correctly classified by classifier $C_j$. Assume that a neighbor pattern $X_l \in \omega_i, l \in \{O, M\}$ ($O, M$ denotes the outlier and normal class, respectively), then $p(\omega_i|X_l)$ provided by classifier $C_j$ can be deemed its measure of competence for $X_l$. As a result, the competence of classifier $C_j$ on pattern $X_t$ can be derived by averaging competences on all neighbor patterns

$$p(\text{correct}_j) = \frac{1}{K} \sum_{i=1}^{K} p(\omega_i|X_l) W_l$$

(14)

where $W_l = 1/d_i$, and $d_i$ is the distance from pattern $X_l$ to $X_t$.

As outputs of base classifiers have been transformed into posterior probability estimates, we can exploit this information to measure classifier competence

$$p(\text{correct}_j|C(X_l) = \omega_i) = p(X_t \in \omega_i|C_j(X_l) = \omega_i)$$

$$= \frac{N_{jl}}{\sum_{j \in \{M, O\}} N_{jl}}$$

(16)

where $N_{jl}$ is the number of neighbor patterns that have been correctly classified by $C_j$ to class $\omega_l$, and $\sum_{j \in \{M, O\}} N_{jl}$ is the total number of neighbor patterns that have been classified to class $\omega_l$ by classifiers $C_j$.

Then we exploit the posterior probabilities of neighbor patterns according to the Bayesian theory

$$p(X_t \in \omega_i|C_j(X_l) = \omega_i) = \frac{p(C_j(X_l) = \omega_i|X_t \in \omega_i) p(\omega_i)}{\sum_{j \in \{M, O\}} p(C_j(X_l) = \omega_i|X_t \in \omega_i) p(\omega_i)}$$

(17)
The term $\hat{p}(C_j|X_t) = \omega_j|X_t \in \omega_j$ indicates the probability that patterns belonging to class $\omega_j$ are correctly classified. It can be estimated as

$$\hat{p}(C_j|X_t) = \omega_j|X_t \in \omega_j = \frac{1}{\sum_{l=1}^{M} N_{l}j} \sum_{X_t \in \omega_j} p'(\omega_j|X_t')$$

(18)

The term $\hat{p}(\omega_j)$ denotes the prior probability of class $\omega_j$, which can be estimated as

$$\hat{p}(\omega_j) = \frac{\sum_{l=1}^{M} N_{lj}}{K}$$

(19)

Then a weight is also assigned to each neighbor pattern to reduce the uncertainty triggered by the neighbor size. Finally, competence of classifier $C_j$ on test pattern $X_t$ can be obtained

$$\hat{p}(\text{correct}|C_j|X_t) = \omega_j = \frac{\sum_{X_t \in \omega_j} p'(\omega_j|X_t') \cdot W_l}{\sum_{l=1}^{M} \sum_{X_t \in \omega_j} p'(\omega_j|X_t') \cdot W_l}$$

(20)

We conclude this procedure in Algorithm 1.

**Algorithm 1.**

**Input:** the pool of base detectors $C_j$, training set $Tr$ and test set $Te$; the number of samples in validation set $K$;  
**Output** $c^*_t$, the most promising classifier for each unknown sample $t$ in the $Te$;  

for each testing sample $t$ in $Te$ do

Find $\Psi$ as the $K$ samples in the region of competence of $t$ in $Te$;  

for each classifier $c_j$ in $C$ do  

Compute $p(\text{correct})$ on $\Psi$;  

if $p(\text{correct}) > 0.5$ then  

$C_j = CS_i$;  

end if  

end for  

$p(\text{correct}) = \max(p(\text{correct}));$  

$c_m = \arg \max(p(\text{correct}));$  

selected $= \text{TRUE}$;  

for each classifier $c_j$ in $CS$ do  

d $= p(\text{correct}_m) - p(\text{correct})$;  

if $(\forall j \neq m \text{ and } d < \text{ Threshold})$ then  

selected $= \text{FALSE}$;  

end if  

end for  

if (selected $= \text{TRUE}$) then  

$c^*_t = c_m$;  

else  

$c^*_t$ is a classifier randomly selected from $CS$, with $d < \text{ threshold}$;  

end if  

end for  

Use the classifier $c^*_t$ to classify $t$.

The term $\hat{p}(C_j|X_t) = \omega_j|X_t \in \omega_j$ indicates the probability that patterns belonging to class $\omega_j$ are correctly classified. It can be estimated as

$$\hat{p}(C_j|X_t) = \omega_j|X_t \in \omega_j = \frac{1}{\sum_{l=1}^{M} N_{l}j} \sum_{X_t \in \omega_j} p'(\omega_j|X_t')$$

(18)

The term $\hat{p}(\omega_j)$ denotes the prior probability of class $\omega_j$, which can be estimated as

$$\hat{p}(\omega_j) = \frac{\sum_{l=1}^{M} N_{lj}}{K}$$

(19)

Then a weight is also assigned to each neighbor pattern to reduce the uncertainty triggered by the neighbor size. Finally, competence of classifier $C_j$ on test pattern $X_t$ can be obtained

$$\hat{p}(\text{correct}|C_j|X_t) = \omega_j = \frac{\sum_{X_t \in \omega_j} p'(\omega_j|X_t') \cdot W_l}{\sum_{l=1}^{M} \sum_{X_t \in \omega_j} p'(\omega_j|X_t') \cdot W_l}$$

(20)

We conclude this procedure in Algorithm 1.

**Optimization of SVDD**

In algorithm SVDD, two parameters $\nu$ and $s$ need to be determined a prior. Parameter $\nu$ is defined as

$$\nu = \frac{1}{NC}$$

(21)

where $N$ is the number of training data, and $C$ indicates the trade-off parameter. It has been proved in Tax and Duin\textsuperscript{11} that parameter $\nu$ is an upper bound for the fraction of target class objects outside the description. In addition, the fraction of objects which become SV is a leave-one-out estimate of the error on the target set

$$E[P(\text{error target set})] = \frac{\#SV}{N}$$

(22)

where $\#SV$ indicates the number of SVs. Using this equation, one parameter ($\nu$ or $s$) can be optimized. Unfortunately, this will not uniquely define both parameters. Minimizing just the error on the target set is not sufficient. To estimate the outlier acceptance rate without outlier examples, we have to assume an outlier distribution. In section “Artificial outliers,” some artificial outliers have been generated. With these outlier examples, we can optimize both parameters simultaneously by minimizing the following error term

$$E(\nu, s) = \lambda \frac{\#SV'}{N} + (1 - \lambda) f_o^s$$

(23)

where $f_o^s$ denotes the fraction of accepted outlier examples. Therefore, parameter $\lambda$ indicates a trade-off between target error and outlier error. When we take $\lambda = 1/2$, then errors on the fraction of target and outlier observations are weighted equally.

**Experiments and analysis**

**Datasets**

In EAF control systems, three secondary current and three secondary voltage are often used by data-driven control strategies. For example, in Li and Mao,\textsuperscript{6} these six variables are used to identify the process model of EAF. In this paper, we also use these variables to constitute the training and test sets. Totally, we will use six datasets, including three synthetic ones by the simulation model and three real-world ones. A simple description of these datasets is shown in Table 1.

| Dataset | # Examples | % Outliers | # Attributes |
|---------|------------|------------|--------------|
| 1 | 10,000 | 10 | 6 |
| 2 | 10,000 | 10 | 6 |
| 3 | 10,000 | 10 | 6 |
| 4 | 5594 | 9 | 6 |
| 5 | 6422 | 11 | 6 |
| 6 | 7826 | 7 | 6 |

Table 1. Description of datasets.

The first three datasets are generated using the simulation model, and different faults are simulated in different datasets. The last three datasets are collected from real-world EAF control systems. In each dataset, 70% normal data is randomly selected to constitute the training set, and all remaining samples constitute the test set. This process will be repeated 10 times, and the averaging values will be used as the final results.
Table 2. Confusion matrix of two-class classification problem.

| Predicted label | Target class | Negative class |
|-----------------|--------------|----------------|
| Target class    | True positive (TP) | False positive (FP) |
| Negative class  | False negative (FN) | True negative (TN) |

Competitors and metrics

In order to put the experimental results of our method into context, we compare it with several competitors:

- Random subspace SVDD (RS-SVDD) proposed in Wang and Mao.\textsuperscript{13} In this detection model, technique RSM is used to develop a parallel ensemble model.
- Bagging SVDD (BA-SVDD) proposed in Ge and Song.\textsuperscript{33} In this detection model, technique Bagging is used to develop a parallel ensemble model.
- Clustering-based SVDD (C-SVDD) proposed in Wang and Mao.\textsuperscript{12} In this detection model, clustering technique is used to develop a parallel ensemble model.

Here, we refer to our method as Dynamic selection SVDD (DS-SVDD) as it is a dynamic selective SVDD model.

In this paper, we use three metrics, that is, \( G\)-\textit{mean}, \( F\)-\textit{measure}, and \( \text{ROC-curve} \) (receiver operating characteristic curve). A representation of classification performance is formulated by a confusion matrix as illustrated in Table 2.

Then we can formulate \( G\)-\textit{mean} as follows

\[
G\text{-\textit{mean}} = \sqrt{\frac{TP}{TP + FN}} \times \frac{TN}{TN + FP} \tag{24}
\]

This metric evaluates the degree of inductive bias in terms of a ratio of positive accuracy and negative accuracy.

\( F\)-\textit{measure} can be formulated as follows

\[
F\text{-\textit{measure}} = \frac{(1 + \beta^2) \cdot \text{Recall} \cdot \text{Precision}}{\beta^2 \cdot \text{Recall} + \text{Precision}} \tag{25}
\]

where \( \beta \) is a coefficient to adjust the relative importance of precision versus recall (usually, \( \beta = 1 \)).

\[
\text{Recall} = \frac{TP}{TP + FN}, \quad \text{Precision} = \frac{TP}{TP + FP} \tag{26}
\]

\( F\)-\textit{measure} combining recall and precision as a measure could provide more insight into the functionality of a classifier than the accuracy metric.

The ROC curve describes the trade-off between the true-positive rate and the false-positive rate. (Note that normal data are regarded as positive in this paper. So true-positive rate indicates the rate of correctly detected normal data.) It could thus evaluate the general performance rather than performance at only one working point. In practice, the area under the ROC curve (AUC) is used since comparing directly ROC curves of varying detectors is difficult. For a novelty detection task, the AUC value of a perfect algorithm equals 1, implying that all outliers have been identified and none misclassified normal data occur simultaneously. Algorithms with AUC values smaller than 0.5 are often deemed invalid since “random guessing” could obtain the AUC of 0.5. Here, we employ method in Huang and Ling\textsuperscript{34} to calculate the AUC.

Table 3. Comparative result with respect to \( G\)-\textit{mean} values.

| Dataset | SVDD | RS-SVDD | BA-SVDD | C-SVDD | DS-SVDD |
|---------|------|---------|---------|--------|---------|
| 1       | 0.855 | 0.881   | 0.874   | 0.893  | 0.904  |
| 2       | 0.847 | 0.863   | 0.839   | 0.852  | 0.895  |
| 3       | 0.811 | 0.798   | 0.826   | 0.837  | 0.830  |
| 4       | 0.907 | 0.921   | 0.918   | 0.902  | 0.948  |
| 5       | 0.944 | 0.959   | 0.949   | 0.949  | 0.975  |
| 6       | 0.967 | 0.976   | 0.972   | 0.980  | 0.994  |
| Ave     | 0.889 | 0.900   | 0.896   | 0.902  | 0.924  |

SVDD: support vector data description. The best result is in bold.

Table 4. Comparative result with respect to \( F\)-\textit{measure} values.

| Dataset | SVDD | RS-SVDD | BA-SVDD | C-SVDD | DS-SVDD |
|---------|------|---------|---------|--------|---------|
| 1       | 0.891 | 0.916   | 0.909   | 0.918  | 0.927  |
| 2       | 0.877 | 0.892   | 0.884   | 0.890  | 0.908  |
| 3       | 0.851 | 0.856   | 0.862   | 0.874  | 0.871  |
| 4       | 0.911 | 0.926   | 0.920   | 0.907  | 0.951  |
| 5       | 0.948 | 0.964   | 0.968   | 0.968  | 0.979  |
| 6       | 0.959 | 0.962   | 0.969   | 0.974  | 0.988  |
| Ave     | 0.906 | 0.919   | 0.919   | 0.922  | 0.937  |

SVDD: support vector data description. The best result is in bold.

Result and analysis

Results on all six datasets with respect to three metrics are shown in Tables 3–5, respectively. Apart from the values in terms of three metrics, we also provide the averaging values over all datasets so that we can have an insight into the general performance. The comparison of these averaging values can be understood clearly in Figure 3, from which we could find that our method (DS-SVDD) has achieved the best general result on all three metrics. Then we compare DS-SVDD with its competitor one by one:

- **DS-SVDD versus SVDD.** On all six datasets, DS-SVDD has outperformed SVDD in terms of three metrics. This result implies that the improvement provided by the dynamically selective procedure in DS-SVDD is very efficient.
Although the outlier examples in the training set are artificial, their role in the selection has effected much on the detection result. Certainly, other processes in DS-SVDD have also made great contributions to the final result.

- **DS-SVDD versus RS-SVDD.** The superiority of DS-SVDD over RS-SVDD has been shown on all datasets. The reason may be twofold from our point of view. FB technique used in DS-SVDD to generate base detectors may be more efficient than RS method. Actually, this conclusion has also been verified in other researches.

- **DS-SVDD versus BA-SVDD.** Before comparing these two models, we first compare the result of BA-SVDD with that of RS-SVDD. From Figure 3, we can see that RS-SVDD has achieved better general result than BA-SVDD. On four of five datasets, RS-SVDD outperforms BA-SVDD. This comparative result indicates that subspace-based novelty ensembles are often more efficient than subsampling-based ones. With this comparison, we can easily understand why DS-SVDD outperforms BA-SVDD on all datasets.

- **DS-SVDD versus C-SVDD.** An interesting point in this pair of comparison is that C-SVDD outperforms DS-SVDD on dataset 3. In order to find the reason behind this, we first check the performance of RS-SVDD. From results in terms of three metrics, we can see that RS-SVDD has obtained the worst result on this dataset. Although a little improvement has been achieved by DS-SVDD, but this improvement has not pushed DS-SVDD to outperform C-SVDD. But on other datasets, DS-SVDD still has better performance than its competitors.

**Conclusion**

To facilitate the development of advanced data-driven control strategies in EAF systems, this paper proposes a dedicated novelty detection model with the help of dynamic ensemble learning theory. In this detection model, SVDD plays the role of base detector. Artificial outliers are generated with two objectives, one is to complete the dynamic selection, and the other is to optimize two parameters of SVDD. Then clustering technique is used to determine the validation set for each test point. Finally, a probabilistic method is used to compute the competence of base detectors. In order to validation the proposed detection model, we compare it with four competitors on three synthetic and three real-world datasets. We compare results of all these methods and show the superiority of our method.

However, several issues regarding our method are still open to solve. For example, the procedure of generating artificial outliers may be not appropriate in some situation. When training set contains unknown outliers, the robustness of our method may be poor. These problems have not been considered in this paper, but they should be our future research directions.

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.
References

1. Aggarwal CC. Outlier analysis. 2nd ed. Cham: Springer, 2017.
2. Chandola V, Banerjee A and Kumar V. Anomaly detection: a survey. *ACM Comput Surv* 2009; 41(3): 1–58.
3. Hodge V and Austin J. A survey of outlier detection methodologies. *Artif Intell Rev* 2004; 22(2): 85–126.
4. Liu F, Mao Z and Su W. Outlier detection for process control data based on a non-linear auto-regression hidden Markov model method. *T Inst Meas Contr* 2012; 34(5): 527–538.
5. Jia M, Ping Y, Yan LI, et al. Model predictive control synthesis approach of electrode regulator system for electric arc furnace. *J Iron Steel Res* 2011; 18(11): 23–28.
6. Li L and Mao Z. A direct adaptive controller for EAF electrode regulator system using neural networks. *Neurocomputing* 2012; 82(4): 9198.
7. Li Q and Wu PH. Electric arc furnace based on fuzzy neural network control system and MATLAB simulation. *Appl Mech Mater* 2011; 48–49: 491–495.
8. Rashid MM, Mhaskar P and Swartz CLE. Multi-rate modeling and economic model predictive control of the electric arc furnace. *J Process Contr* 2016; 40: 50–61.
9. Breunig MM, Kriegel H-P, Ng RT, et al. LOF: identifying density-based local outliers. In: *Proceedings of the ACM SIGMOD 2000 international conference on management of data*, Dallas, TX, 16–18 May 2000. New York: ACM.
10. Knorr EM, Ng RT and Tucakov V. Distance-based outliers: algorithms and applications. *VLDB J* 2000; 8(3): 237–253.
11. Tax DMJ and Duin RPW. Support vector data description model for batch process monitoring. *T Inst Meas Contr*. Epub ahead of print 3 March 2020. DOI: 10.1177/0142331209058951.
12. Wang B and Mao Z. Detecting outliers in electric arc furnace under the condition of unlabeled, imbalanced, non-stationary and noisy data. *Meas Contr* 2018; 51(3–4): 83–93.
13. Wang B and Mao Z. One-class classifiers ensemble based anomaly detection scheme for process control systems. *T Inst Meas Contr* 2018; 40(12): 3466–3476.
14. Yuan P, Mao Z and Wang B. A pruned support vector data description-based outlier detection method: applied to robust process monitoring. *T Inst Meas Contr*. Epub ahead of print 3 March 2020. DOI: 10.1177/0142331209058951.
15. Pimentel MAF, Clifton DA, Clifton L, et al. Review: a review of novelty detection. *Sig Process* 2014; 99: 215–249.
16. Bird S, Ghobara YEM, Khater MM, et al. Modeling, optimization and estimation in electric arc furnace (EAF) operation. *Chem Eng* 2013, https://macsphere.mcmaster.ca/bitstream/11375/13345/1/fulltext.pdf.
17. Lazarevic A and Kumar V. Feature bagging for outlier detection. In: *Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining*, Chicago, IL, 21–24 August 2005. New York: ACM.
18. Ho TK. The random subspace method for constructing decision forests. *IEEE T Pattern Anal Mach Inte* 1998; 20(8): 832–844.
19. Hempstalk K, Frank E and Witten IH. One-class classification by combining density and class probability estimation. In: *Proceedings of the joint European conference on machine learning and knowledge discovery in databases*, Antwerp, 15–19 September 2008.
20. Fan W, Miller M and Miller M. Using artificial anomalies to detect unknown and known network intrusions. In: *Proceedings of the 2001 IEEE international conference on data mining*, San Jose, CA, 29 November–2 December 2001. New York: IEEE.
21. Tax DMJ and Duin RPW. Uniform object generation for optimizing one-class classifiers. *J Mach Learn Res* 2001; 2: 155–173.
22. Désir C, Bernard S, Petitjean C, et al. One class random forests. *Pattern Recogn* 2013; 46: 3490–3506.
23. Zimek A, Campello RJ and Sander J. Ensembles for unsupervised outlier detection: challenges and research questions. In: *Proceedings of the IEEE sixth international conference on data mining (ICDM’06)*, Hong Kong, China, 18–22 December 2006. New York: IEEE.
24. Gao J and Tan P-N. Converting output scores from outlier detection algorithms into probability estimates. In: *Proceedings of the 2011 SIAM international conference on data mining*, Mesa, AZ, 28–30 April 2011.
25. Aggarwal CC and Sathe S. Theoretical foundations and algorithms for outlier ensembles. *ACM SIGKDD Explor News* 2015; 17(1): 24–47.
26. Nguyen HV, Ang HH and Gopalkrishnan V. Mining outliers with ensemble of heterogeneous detectors on random subspaces. In: *Proceedings of the international conference on database systems for advanced applications*, Tsukuba, Japan, 1–4 April 2010.
27. Pearson RK. Exploring process data. *J Process Contr* 2001; 11(2): 179–194.
28. He X, Cai D, Shao Y, et al. Laplacian regularized Gaussian mixture model for data clustering. *IEEE T Knowl Data Eng* 2011; 23(9): 1406–1418.
29. Ester M, Kriegel H-P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of 2nd international conference on knowledge discovery and data mining*, Anchorage, AK, 4–8 August 1996.
30. Calinski T and Harabasz J. A dendrite method for cluster analysis. *Commun Stat: Theor Method* 1974; 3(1): 1–27.
31. Ge ZQ and Song ZH. Bagging support vector data description model for batch process monitoring. *J Process Contr* 2013; 23(8): 1090–1096.
32. Huang J and Ling CX. Using AUC and accuracy in evaluating learning algorithms. *IEEE T Knowl Data Eng* 2005; 17(3): 299–310.