Research on abnormal detection of one-class support vector machine based on ensemble cooperative semi-supervised learning

Dinghai Wu¹, Chunhua Liu², Hongbo Fan¹, Bin Song¹

¹Shijiazhuang Campus, Army Engineering University, Shijiazhuang 050003, China
²Information center, Hebei Provincial Public Security Department, Shijiazhuang 050000, China
wudh81@163.com

Abstract. In order to promote the accuracy of anomaly detection model under the condition of only a small number of labeled samples and large number of unlabeled samples, abnormal detection of One-class Support Vector Machine(SVM) based on ensemble cooperative Semi-supervised Learning is proposed. A kind of One-class SVM model which bring supervision with a small number of abnormal samples can classify samples with max interval. The semi-supervised learning methods easily suffer from the low accuracy because the mistake labeled sample are chosen as training sample set. Refer to the semi-supervision method of Tri-training, the K-Nearest Neighbour(KNN) and Naive Bayes classifier are used to assist the One-class SVM based on ensemble cooperative Semi-supervised learning method which can classify the large number of unlabeled samples as accurate as possible. The weight is also given after ensemble cooperative Semi-supervised Learning. Then the proposed semi-supervised One-class SVM would be trained with the result and used to classify test samples. The experimental results on UCI dataset show that the proposed algorithm achieves higher classification accuracy with less labeled samples and it improves generalization performance and reduces the labelling cost.

1. Introduction
In the process of fault diagnosis, there are problems such as unbalanced data samples and incomplete fault modes, resulting in poor timeliness and low accuracy of equipment condition monitoring and fault diagnosis. Therefore, it is of great significance to identified and alarm the equipment operating state in time[1-4].Support vector machine(SVM) is based on statistical learning theory and has better learning performance and generalization ability than traditional learning methods. The one-class SVM proposed by Tax and Duin could obtain the minimum hyper sphere which is containing normal sample data through training and could distinguish between normal data and abnormal data[5,6]. The one-class SVM provides an effective way for anomaly detection.

However, the classical Support Vector Domain Description(SVDD) is based on a large number of supervised learning with labeled data. In practical engineering applications, it is obviously difficult to obtain a large number of labeled data samples due to many factors such as experimental conditions, system operating environment and labor costs.

How to use a large number of unidentified samples to improve the performance of the classifier is one of the most concerned issues in current machine learning research[7-8]. Semi-supervised learning is
a learning model between supervised learning and unsupervised learning. It can simultaneously use both labeled and unlabeled sample information for better classification performance. In the semi-supervised training mode, the semi-supervised learning methods with typical representation include self-training learning mode, Co-Training algorithm and Tri-training algorithm\cite{9,10}. Among them, the Tri-training algorithm expands the Co-Training algorithm and uses three classifiers for training and learning. It does not require a sufficient redundant view, nor does it require a typical cooperative training semi-supervised learning algorithm without a type classifier, and it use integrated learning to improve generalization capabilities. The semi-supervised learning method is to select the sample of the un-labeled sample from which the current learner has high confidence, and then add to the training sample set to improve the learning performance. Some scholars have proposed to combine the active learning method which uses the field experts to mark the most uncertain and most influential classifier training samples and then adds them to the training sample set to reduce the classification error\cite{11}.

Traditional one-class support vector machines lack supervision of abnormal samples and are sensitive to model parameters\cite{12}. A new model is proposed in this paper which uses the supervision with small sample of anomaly data, and expands the maximum interval one-class SVM to semi-supervised mode. The cumbersome and costly problem of obtaining labeled data is solved by integrated cooperative semi-supervising training and marking the unlabeled data samples based on k-nearest neighbor (KNN) and naive Bayes. The model determines the weights of the sensitive data samples based on the historical discriminant precision. Then the detection performance of the model is improved in a semi-supervised manner by retraining with the obtained labeled data samples and weight information.

2. Semi-supervised maximum interval one-class SVM

Given data set $X = \{x_1, x_2, \cdots, x_n, x_{n+1}, \cdots, x_{n+m}\}$, The first n are unlabeled data, and the last m are labeled data. Tag category $Y = \{1, -1\}$, among, +1 is the mark of normal samples, -1 is the mark of abnormal samples. The data set contain $m_1$ normal samples and $m_2$ abnormal samples with the conditions of $m_1 + m_2 = m$, and the model of Semi-supervised maximum interval one-class SVM is proposed as follow\cite{13}: 

$$
\min_{d, a, \xi} R^2 - (2d)^2 + C_1 \sum_{i=1}^{n} \xi_i + C_2 \sum_{j=n+1}^{n+m} \xi_j + C_3 \sum_{k=n+m+1}^{n+2m} \xi_k \\
\text{s.t.} \left\| \phi(x_i) - a \right\| \leq R^2 - d^2 + \xi_i, \quad 1 \leq i \leq n \\
\left\| \phi(x_j) - a \right\| \leq R^2 - d^2 + \xi_j, \quad n + 1 \leq j \leq n + m, \quad \xi_j \geq 0 \\
\left\| \phi(x_k) - a \right\| \leq R^2 + d^2 + \xi_k, \quad n + m + 1 \leq k \leq n + m, \quad \xi_k \geq 0
$$

(1)

Where $a$ denotes the Center of sphere, $\phi(\cdot)$ is mapping function, $R$ denotes the optimal classification surface with maximum interval and $d$ denotes classification interval, $C_i$ is a trade-off parameter of different types of data, $\xi_i$ is the relaxed variable introduced to optimize problem constraints. $C_1$ reflects the constraint of unmarked data, $C_2$ and $C_3$ denote the type and guiding role of the labeled samples. Usually set $C_2 < C_3$.

Unlabeled data is divided into two categories, which contain normal samples $n_1$ and abnormal samples $n_2$ with $n_1 + n_2 = n$. Then the model have four constraint conditions which makes it too complicated to find the global optimal solution.

For the purpose of minimizing the effect of misidentified training samples, weights are introduced in the semi-supervised model for unlabeled samples, range as (0,1]. and $w_i = \{1, \cdots, w_{n_1+1}, \cdots, w_{n_1+n_2}\}$
are weights of normal samples and abnormal samples, \( W_2 = \{1, 1, \cdots, w_{n_2 + 1}, \cdots, w_{n_2 + m_2}\} \). The proposed model takes the control proportional parameters of Positive and negative samples \( v_1, v_2 \) as a substitute of trade-off parameter \( C \). Then the optimization goal of the model is simplified as follow:

\[
\min_{\alpha, u, \xi} R^2 - (2d)^2 + \frac{W_1}{v_1(n_1 + m_1)} \sum_{i=1}^{n_1 + m_1} \xi_i + \frac{W_2}{v_2(n_2 + m_2)} \sum_{j=n_1 + m_1 + 1}^{n_1 + m_1 + m_2} \xi_j \\
\text{s.t.} \|\phi(x_i) - a\|^2 \leq R^2 - d^2 + \xi_i, \quad 1 \leq i \leq n_1 + m_1, \quad \xi_i \geq 0 \\
\|\phi(x_j) - a\|^2 \leq R^2 + d^2 - \xi_j, \quad n_1 + m_1 + 1 \leq j \leq n + m, \quad \xi_j \geq 0 \\
\xi_i \geq 0, 1 \leq k \leq n + m
\]

(2)

Use the dual principle to turn the above optimization problem into

\[
\min_{\alpha} \sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n_1+m_1} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{n_1} \alpha_i y_i K(x_i, x_i)
\]

s.t. \( 0 \leq \alpha_i \leq \frac{w_1}{v_1 m_1}, \quad 1 \leq i \leq n_1 + m_1 \)

\( 0 \leq \alpha_j \leq \frac{w_2}{v_2 m_2}, \quad n_1 + m_1 + 1 \leq j \leq n + m \)

\( \sum_{i=1}^{n_1} \alpha_i y_i = 1 \)

\( \sum_{i=1}^{n_1} \alpha_i = \frac{w_1}{v_1} \)

(3)

Each classification surface is ultimately determined by the supported support vector. The support vectors set belonging to the positive domain \( H_+(a, R) \) and negative domain \( H_-(a, R) \) could be solved by Lagrange \( \alpha_i \), as follow formulas

\[
SV_+ = \{x_i \mid 0 < \alpha_i \leq \frac{w_1}{v_1 m_1}\}
\]

(4)

\[
SV_- = \{x_j \mid 0 < \alpha_j \leq \frac{w_2}{v_2 (n - m_1)}\}
\]

After solving, the final discriminant function is

\[
f(x) = \text{sign}(R^2 - \|\phi(x) - a\|^2)
\]

(5)

Where \( R \) can be solved by

\[
R^2 = \frac{\max \|\phi(x_{n_2}) - a\|^2 + \min \|\phi(x_{n_2}) - a\|^2}{2}
\]

(6)

3. Labels prediction and weight assignment based on semi-supervised collaborative training model

3.1. Semi-supervised collaborative training framework

According to the above semi-supervised model, it is necessary to predict the categories of unlabeled samples and give corresponding weights. In order to ensure the diversity of classifiers, the proposed model combined with the advantages of tri-training collaborative training and selects the nearest neighbor KNN and naive Bayes as auxiliary training classifiers in this paper as shown in Fig.1. The KNN classifier has the parameter \( k=1 \) and Naive Bayes does have parameters, which avoiding parameters optimization problem of the initial classifiers in the cooperative training. At the same time, the Boosting method is used to group the unlabeled samples, and the initial classifier is established and performed. Semi-supervised learning, obtaining pseudo-labels of unlabeled samples, recording the discriminating results of each iteration, and then combining the discriminating results to perform statistics and assigning corresponding weights to these unlabeled samples.
3.2. Category tag and weight assignment

The key to selecting which unlabeled samples to perform semi-supervised learning is to calculate the class marker confidence of the sample. In the process of collaborative training, the two classifiers of KNN and Naive Bayes are used to assist the one-class SVM for semi-supervised learning. Then unlabeled samples are identified. Drawing on the idea of tri-training, the sample is considered to belong to the category when two classifiers identify the same result. However, if the data point is located near the boundary of the data set, it is obvious that the data category is difficult to discriminate, and it is inevitable that the classification error will occur, and the pseudo-marking confidence of the data point is considered to be low.

The inclusion of mislabeled data samples in the subsequent semi-supervised learning of the model will inevitably lead to a decrease in the accuracy of the model. Therefore, it is necessary to establish category tag weights. Obviously, the data sample is easy to classified when located inside the boundary of the data set, and the classification error rate will be significantly increased when the data sample is around the boundary. Therefore, a class weight determination method based on classification correct rate is proposed in this paper.

Make \( k_i \text{Count} \) denotes the identification results of the data set of group \( i \) in the iteration loop \( k \). The operation can be expressed by

\[
\text{Count}_i^k = \begin{cases} 
1 & \text{correct identify} \\
0 & \text{error identify}
\end{cases}
\]

After the integration of semi-supervised collaborative training, the category weights of each group of data samples are determined according to the previous classification and recognition process. It is given by

\[
\text{weight}_i = \frac{\text{Count}_i}{\text{Iter}_{\text{max}}}
\]

The specific steps of the semi-supervised one-class SVM anomaly detection algorithm based on collaborative training are as follows:

1. Train KNN, Naive Bayes and Maximum Interval one-class SVM with the identified training samples to obtain three initial classifiers.
2. Set the maximum number of loop iterations. Repeat uses the boosting method to select test data set from the unlabeled data set, and uses above three classifiers to distinguish them.
3. The discriminant result is analyzed as follow. When the results of any two classifiers are consistent, the unlabeled data sample is marked as the category and added to the pseudo-label data sample set and record as \( \text{Count}_i^k \).
4. The training set of each classifier is obtained by Boosting sampling from the pseudo-labeled data set. The amount of training data sets increases with the iterations increase, which combined the original training samples to form a new training data set.
5. Train the model with a new data set and then jump to step (2) until iteration loops reaches the maximum number.

6. The label of final training data set is given by the last identify result and the category weight is calculated based on the correct rate of previous recognition results.

7. The semi-supervised maximum interval one-class SVM is trained with the final training data set and weight, and the model after training learning is used for detecting the new test sample.

4. Test Results and Discussions

In order to verify the effectiveness of the proposed algorithm, the standard UCI test data set is used for experimental verification. The false alarm rate and missed detection rate are also introduced to evaluate the detection accuracy of the proposed model. Make $TP$ is the number of positive samples and identified as positive, $FN$ denotes the number of positive samples but identified as negative samples, $TN$ denotes the number of positive samples and identified as negative samples, $FP$ denotes the number of negative samples and identified as positive samples. Then the Evaluation index of false alarm rate, missed detection rate and accuracy rate are defined as follows:

\[
\text{False alarm rate} = \frac{FN}{TP + TN + FN + FP} \quad (9)
\]

\[
\text{Missed detection rate} = \frac{FP}{TP + TN + FN + FP} \quad (10)
\]

\[
\text{Accuracy rate} = \frac{TP + TN}{TP + TN + FN + FP} \quad (11)
\]

In order to verify the effectiveness of the proposed model, the ionosphere data set was used for testing. The training data sample contain 10 positive samples, 25 negative samples and 50 unlabeled samples (positive sample). The parameter kernel function $\sigma$ is 8, control proportion parameter is $[0.02, 0.03]$. The unlabeled data samples are identified by the Initial classifier and the corresponding weights is given. Set the number of predicted sample errors of the data set samples from 0:5:50, then the proposed model, maximum interval one-class SVM and support vector description (SVDD) are trained and used to identify the test samples. As show in Fig.2, the established model uses the information of weights which weakens the influence of the error identification samples in the process of training. As the number of errors labeled samples increases, the classification accuracy rate decreases but relatively stable. The above results show that the model has a good training effect. For the other two models, the mislabeled samples have a greater impact on the of training process and the learning results. As the number of mislabeled increases, the classification performance of the model deteriorates and decreases rapidly, and there is a large fluctuation.

![Fig.2 Influence of ensemble semi-supervised learning result](image-url)
The UCI test data set was used to test the performance of the proposed model in this paper. The attributes of the five UCI test data set samples used are shown in Tab.1.

| name          | Initial label sample | Number of unlabeled samples | Number of test samples | data category | Data dimension | Target data |
|---------------|----------------------|-----------------------------|------------------------|---------------|---------------|-------------|
| Banana        | 10/10                | 400                         | 200                    | 2             | 2             | 1           |
| Iris          | 15/5                 | 65/35                       | 100/50                 | 3             | 4             | 1,2         |
| wine          | 20/10                | 30/80                       | 59/119                 | 3             | 13            | 1           |
| Sonar         | 30/10                | 60/50                       | 111/97                 | 2             | 60            | 2           |
| Ionosphere    | 30/10                | 120/80                      | 225/126                | 2             | 34            | 0           |

Only a small number of initial labeled samples are extracted from UCI data set, and the maximum interval one-class SVM is trained to form an initial detection classifier. From the accuracy rate, false alarm rate and missed detection rate, as shown in Table 2, the classification performance of the initial classifier is not good. The integrated cooperative semi-supervised training method is used to identify the unlabeled samples. The number of loop iterations is set to 30. It can be seen from the results in Table 2 that the data sets under different dimensions have been well recognized, but the accuracy cannot be very high when completely rely on traditional machine learning. Combined with the corresponding prediction categories and weights, the above recognized results are used to train the model established in this paper which base on semi-supervised. Then the data set is tested. Compared with the initial classifier detection performance of the data set, the detection accuracy of the model after semi-supervised learning is significantly improved, and it has better generalization performance, which indicating the effectiveness of the algorithm.

| Data set Samples name | Parameters of the Semi-supervised model | Test results of Initial classifier | Unlabeled sample prediction accuracy | The proposed model test results after semi-supervised learning |
|-----------------------|----------------------------------------|-----------------------------------|-------------------------------------|-------------------------------------------------------------|
|                       | \( \sigma \) \( \nu \) | Accuracy rate | False alarm rate | Missed detection rate | Accuracy | False alarm rate | Missed detection rate |
| Banana                | 5 [0.01,0.01] | 0.8200 | 0.1600 | 0.0200 | 0.9775 | 0.9800 | 0.0150 | 0.0050 |
| Iris                  | 3 [0.02,0.01] | 0.9067 | 0.0867 | 0.0067 | 0.9700 | 0.9800 | 0.0200 | 0 |
| wine                  | 4 [0.02,0.01] | 0.8596 | 0.1404 | 0 | 0.9600 | 0.9733 | 0.0267 | 0 |
| Sonar                 | 1 [0.02,0.01] | 0.7981 | 0.1250 | 0.0769 | 0.8636 | 0.8894 | 0.0625 | 0.0481 |
| Ionosphere            | 8 [0.01,0.05] | 0.7806 | 0.1709 | 0.0456 | 0.8300 | 0.8689 | 0.0883 | 0.0427 |

5. Conclusion
Combining the advantages of collaborative training and semi-supervised learning, a maximum interval one-class SVM with integrated cooperative semi-supervised learning method is proposed in this paper. By integrating collaborative semi-supervised learning, high-accuracy unlabeled sample categories and weights are obtained with minimal cost, and input to the established model for extended learning. The UCI data set test results show that the proposed method in this paper effectively improves the performance of the anomaly detection model. In this paper, the parameters of the model have not been optimized in semi-supervised learning. The next step is to carry out parameter optimization and its application.

Acknowledgments
This work was financially supported by National Natural Science Foundation of China (Grant No.51305454).
References

[1] Li Zhaofei, Chai Yi, Li Huafeng, et al. Diagnosing faults in vibration signals by multi-fractal detrended fluctuation analysis [J]. J. Huazhong Univ. of Sci. & Tech. (Natural Science Edition) , 2016, 37(7): 1449-1456(in Chinese).

[2] Markou M, Singh S. Novelty detection: a review-part 1: statistical approaches[J]. Signal Processing, 2003,83(12):2481-2497.

[3] Tarassenko L, Clifton D A, Bannister P R, et al. Novelty Detection[C]. Proceedings of the International Conference in Encyclopaedia of Structural Health Monitoring, New York, 2008.

[4] Mohamed E S. Fault diagnosis of ICE valve train for abnormal clearance and valve head crack using vibration Signals[J]. Int. J. Vehicle Noise and Vibration, 2015,11(1):18-38.

[5] Tax D, Duin R. Support vector domain description[J]. Pattern Recognition Letters, 1999,20:1191-1199.

[6] Schlkopf B, Williamson RC, Smola A J, et al. Support vector method for novelty detection[J]. Advances in neural information processing systems, 2000,12(4):582-588.

[7] Kothari R, Jain V. Learining from Labeled and Unlabeled Data Using a Minimal Number of Queries[J]. IEEE Trans on Neural Networks, 2003,14(6):1496-1505.

[8] Mehdizadeh M, MacNish C. Semi-supervised neighborhood preserving discriminant embedding: a semi-supervised subspace learning algorithm[J]. Computer Science, 2011,6494:199-212.

[9] Blum A. Combining labeled and unlabeled data with co-training[C]// Proceedings of the eleventh annual conference on Computational learning theory. ACM, 2000:92-100.

[10] Zhou Zhihua, Li Ming. Tri-training exploiting unlabeled data using three classifiers.IEEE Trans on Knowledge and Data Engineering, 2005,17(11):1529-1540.

[11] Xu Hailong, Long Guangzheng, Bie Xiaofeng, et al. Active Learning Algorithm of SVM Combining Tri-training Semi-supervised Learning and Convex –Hull Vector[J].PR&AI, 2016,29(1):39-46(in Chinese).

[12] Qin Bin, Jiang Xuexiang, Zhou Hao, et al. SVM-based sliding mode control of pitch angle for direct driven PM wind turbine system[J]. Chinese Journal of Scientific Instrument, 2014, 35(7): 1525-1531(in Chinese).

[13] Liu Jianwei, Liu Yuan, Luo Xionglin. Semi-Supervised Learning Methods[J]. Chinese Journal of Computers, 2015, 38(8): 1592-1617(in Chinese).

[14] Xu Xiaodan,Yao Mhaiming, Liu Hluawen, et al. Pre-processing Method of Multi-label Classification Based on KNN[J]. Computer Science, 2015, 42(5): 106-108(in Chinese).

[15] Du Xuan. Research on weighted complement-based naive bayes text classification algorithm[J]. Computer Application and Software, 2014, 31(9): 253-255(in Chinese).