Imbalanced Data Detection Kernel Method in Closed Systems

Youli-Lu
I.W.R department, Naval Command College
Network Security Office
Nanjing, China
zyqs1981@163.com

Jun-Luo
I.W.R department, Naval Command College
Network Security Office
Nanjing, China
lyl7384@tom.com

***Abstract*— Under the study of Kernel Methods, this paper put forward two improved algorithm which called R-SVM & I-SVDD in order to cope with the imbalanced data sets in closed systems. R-SVM used K-means algorithm clustering space samples while I-SVDD improved the performance of original SVDD by imbalanced sample training. Experiment of two sets of system call data set shows that these two algorithms are more effectively and R-SVM has a lower complexity.**

Keywords-component; Kernel Method; SVDD; Imbalaced Classification;

I. INTRODUCTION

The Closed System is a thermodynamic concept, which refers to an exchange of energy with the outside world only (for power or heat) without exchange of the quality of the system. If one has only one kind of particles (atoms or molecules) of the system for performing chemical reactions, process all kinds of particles can be generated or destroyed. However, the number of atoms of elements within the closed system will be conserved. In computer science, the Closed System is equivalent to the isolated information system such as military system, banking system and electricity management system etc., in which the condition (including applications & communications) seldom change.

In the Closed Systems, it will be very easy to gather the 'normal data' and hard to find the 'abnormal' that result in a great difference in the number of two types, this is considered highly skewed or 'Imbalanced Data Environment (IDE)'. For example, the proportion of sick and normal in medical care, the proportion of military targets and civilian targets in satellite photographs [1].

The minority class sample is often more important. Take the example of satellite, although they put a lot of manpower to verify military objectives, however, if the camera photographs of military targets mistakenly classified as normal photographs, the loss must be much greater than on a case; Similarly, the false alarm of military information system terminal will give staff bring some unnecessary trouble, but the real intrusion, the damage will be more unthinkable. Therefore, how to effectively improve existing kernel methods to improve this chapter focused on solving the problem of the detection rate of minority class.

II. IMPROVEMENT OF KERNEL METHOD

The lack of minority class samples will lead to a decline in SVM is SVDD itself as a single-classification algorithm cannot take advantage of the minority class sample. In the same time, the noise data and be trained to have a greater impact. In this regard, this paper presents 3 ways to improves:

A. Re-dividing samples in Hilbert 's space

The so-called Hilbert space sample weight divided is the use of the basic principles of kernel methods; training samples are mapped to the Hilbert space re-training sample set is divided into K sub. The specific method is the majority class samples roughly equally divided into K roughly each number of minority class samples, respectively, and the minority class sample of the new sub-set of training samples. Thus, the training samples in each sub-set, the balance of the two types of the number of samples [2]. Finally, SVM training sample set of N sub training will be the K classifier using the majority vote manner judgment category.

B. Improve the svdd algorithm by using abnormal samples

Specific ways to improve it is to generate the Hilbert space SSVD generates sphere contains all minority classes as much as possible. The system calls imbalance classification only when the test sample falling into the majority class (normal class) hypersphere of training, while in the minority (abnormal) training outside the hypersphere before judgment sample is normal, and otherwise abnormal. This method will be described in Figure 1.

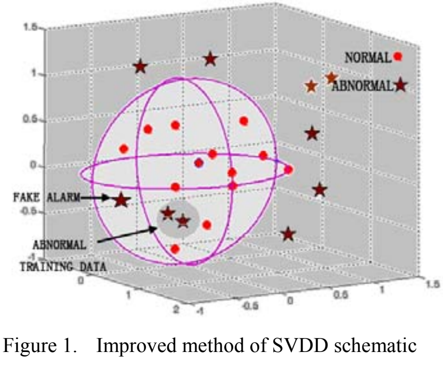

Figure 1. Improved method of SVDD schematic

C. Using different penalty parameters

To overcome the noise minority class samples, while in imbalanced classification problems, the minority class is often the focus of the classification. In this case, the minority class samples are more valuable than the identification of the majority class sample correctly identified [3]. Conversely, misclassified minority class samples need to pay a higher price. For imbalanced classification based on kernel methods, specific change measures are using two types of samples different penalties parameters.

III. RESEARCH ON R-SVM

The way to deal with complex issues often be broken down into multiple sub-problems. This section is mainly to study Re-dividing samples in Hilbert space method to solve the problem of SVM training dataset imbalance. This method will original imbalanced classification problem into classification sub-problems of the plurality of balance, classified data is balanced in these sub-problems. Finally, the use of the majority vote manner to the judgment result of the processing of the sub-problems, [3] the improved algorithm referred to as R-SVM. Specific descriptions are as follows:

A. Description of R-SVM

Assumption given training data set, defined collection T:

$$T = \{X, Y, \lambda\}$$

$$\lambda = count(X^{+}) \big/ count(X^{-}) \,! \qquad (1)$$

X is the training sample, the few classes for training samples for the majority class in the training sample; Y as a feature vector; imbalance, i.e. majority class number of samples and the ratio of the number of samples of the few classes;

Mapping through kernel function, the training data set is mapped to the Hilbert space, Hilbert space training data collection:

$$\Phi(T) = \{\phi(X), Y, \lambda\} \qquad (2)$$

The majority class samples are divided in the data collection in the Hilbert space K region (K is odd), the minority class sample new sub-set of training samples:

$$\Phi(T_i^{'}) = \{\phi(x_i), Y_i, \lambda_i\} \qquad i \in [1 \quad K] \qquad (3)$$

Hilbert space sub training sample collection, the majority class samples between disjoint;

$$\phi(x_n^{+}) \bigcap \phi(x_m^{+}) = \varnothing \quad n \neq m \quad m, n \in [1 \quad K] \qquad (4)$$

Most number of samples and the number of samples the minority class remained largely balanced, $\lambda_i \in (0.8 \quad 1.2)$ in general. For each sub training sample set SVM training sub-SVM classifier, the R-SVM objective function as follows:

$$\sum_{i=1}^{k} \phi_i(\omega_i, \xi_i) = \sum_{i}^{k} \frac{1}{2} \langle \omega_i, \omega_i \rangle + C_i \xi_i \quad i \in [1 \quad K] \; (5)$$

The punishment parameter of hyperspace vector and relaxation variables are: $C_i \; w_i \; \xi_i$.

SVM classifier is trained in each sub-sample, the majority class roughly the same number of samples and the number of samples of the few classes, so each sub-classifier effect is desirable.

$$f_i(x) = \text{sgn}[< \omega_i, \phi_i(x) > + b_i] \qquad i \in [1 \quad K] \qquad (6)$$

As shown in Figure 2, when the measured samples for testing, K SVM classifier its output test results, the final majority vote to give the final verdict.

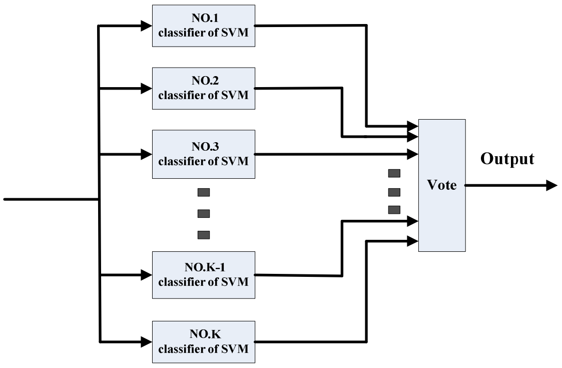

Figure 2. R-SVM classifier structure diagram

B. Samples space division based on the K-means

The core idea of the R-SVM To Hilbert space data collection in the majority class sample is divided into K zones (K is odd), then the sample with a small number of classes in each region a new sub-set of training samples [4]. In order to ensure that R-SVM has good classification effect, the division of space must be such that the majority class sample within each region required having similar characteristics. Based on this consideration, this paper uses the K-means algorithm for a regional breakdown of the majority class samples.

K-means clustering allows each cluster area as compact as possible, while among the regions as separate as possible, so that is conducive to each sub-SVM classifier.

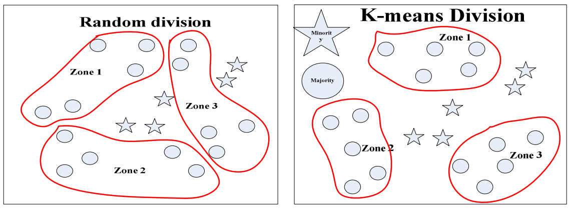

Figure 3. Comparison of sample space division method

As shown in Figure 3, the circular sample the majority class samples, the pentagram sample minority class samples. The two types of samples presented imbalance distribution. In the left side of the figure, the majority class samples were randomly divided into three areas; right, the majority class samples through the K-means algorithm is divided into three regions [5]. It is clear that compared randomly divided the sample space, the same area of the sample with a higher degree of similarity, based on the K-means algorithm divided the sample space, so more use of R-SVM classification results.

Published by Atlantis Press, Paris, France.

C. Sequence of division

Specific sample space division process is as follows:

Step1: choose from the N data samples K samples as the initial cluster centers;

Step2: cycle Step3 to Step4 change date not happen again until each cluster (i.e. the objective function convergence).

Step3: Calculate the mean of the cluster center of the cluster region, calculation of the Euclidean distance of the other samples, and these cluster center; corresponding sample divided according to the minimum distance re;

Step4: the recalculated mean area (change) clustering to identify new cluster centers.

IV. IMBALANCED SAMPLE TRAINING OF SVDD

A. The I-SVDD algorithm analysis

According to the location and types of test samples over the sphere of distribution of sample classification, the improved algorithm referred to SVDD was called the I-SVDD. Assume the majority of class training samples as X+, a small number of class training samples as X-, Penalty parameter is C. In I-SVDD algorithm two types of training samples are mapped to the Hilbert space of two balls in vivo using kernel methods [6].

By adjusting the transformation, the structure of the error in the original SVDD becomes:

$$\varepsilon_{struct}(R,a,\xi_i,\xi_l) = R^2 + \sum_i \xi_i c + \sum_l \xi_l c \tag{7}$$

Similarly adding noise, Lagrange function becomes:

$$\begin{aligned} L(R,a,\xi_i,\xi_l,\alpha_i,\alpha_l,\gamma_i,\gamma_l) &= R^2 + \sum_i \xi_i c + \sum_l \xi_l c \\ &- \sum_i \alpha_+ \left\{ R^2 + \xi_i - (x_i \cdot x_i - 2a \cdot x_i + a \cdot a) \right\} \\ &- \sum_l \alpha_- \left\{ R^2 + \xi_i - (x_l \cdot x_l - 2a \cdot x_l + a \cdot a) \right\} - \sum_i \gamma_i \xi_i - \sum_l \gamma_l \xi_l \end{aligned} \tag{8}$$

Do partial derivatives to a, R, γ_i, γ_l, so that partial derivatives to zero, we get:

$$\begin{aligned} L = &\sum_i \alpha_i (x_i \cdot x_i) - \sum_l \alpha_l (x_l \cdot x_l) - \sum_{i,j} \alpha_i \alpha_j (x_i \cdot x_j) \\ &- \sum_{l,m} \alpha_l \alpha_m (x_l \cdot x_m) + 2 \sum_{i,l} \alpha_i \alpha_l (x_i \cdot x_l) \end{aligned} \tag{9}$$

Constraints are $0 \le \alpha_i \le C$,$0 \le \alpha_l \le C$,$\forall i,l$

Introduction of kernel function, the final decision condition becomes:

$$\mathrm{f}_{I-\mathrm{SVDD}}(\mathrm{z};\alpha_i,\alpha_l,R_i,R_l) = \mathrm{I}(\|\mathrm{f(z)} - \mathrm{f}(\alpha_i)\|^2 \le R_+{}^2, \|\mathrm{f(z)} - \mathrm{f}(\alpha_l)\|^2 \ge R_-{}^2) \tag{10}$$

This discriminant function I is defined as:

$$\mathrm{I(A)} = \begin{cases} 1 & \text{if A is true} \\ -1 & \text{otherwise} \end{cases} \tag{11}$$

B. Hilbert space distribution of the minorityclass

As there were two hydro spheres, there may be representative of the position include, intersect, are away from the three cases, this section will detail these three situations described:

Case1：Minority hydrosphere within the majority

Training samples generated minority class hypersphere contains the body hypersphere in most categories. Circular sample of majority class samples, the pentagram sample minority class samples, black diamond for the test sample. The two types of training samples render the imbalance distribution.

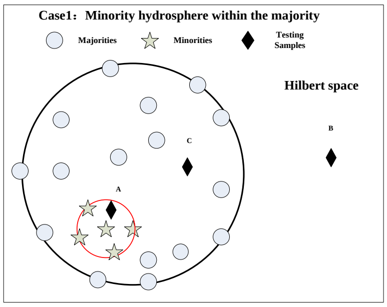

Figure 4. The within Case

Seen in most categories over the sphere contains a small class of hydrosphere case, I-SVDD class for a few samples of the introduction of success will be mistaken for the original test sample A majority of the class to determine a small number of classes, effectively raising the I-SVDD minority class samples for the detection rate.

(2) Case 2: Intersection of the hydrosphere

As Shown in Figure 2, training samples were generated into two types of hydrosphere intersection in Hilbert space At this time, the test samples to determine also divided into three categories: one is when the test samples A, B hydrosphere into a few categories, will test samples A, B sub-class for the few; the second is when to be Most fall into the category D test samples over the ball when in vitro, the test sample for a few sub-category D; Third, when the majority of test samples fall into category C, when the body over the ball and the ball in a few classes over the body, then the test verdict Sample C is the majority class.

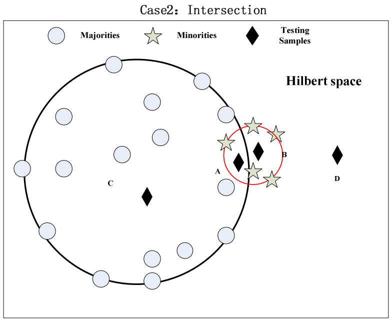

Figure 5. The intersection case

Comparing I-SVDD and SVDD in this case, only if the test sample containing a similar concentration of such a test samples A, I-SVDD will improve the detection rate.

By analyzing the above three cases, a small number of classes of Hilbert space distribution of the sample, the following conclusions:

- Comparing with SVDD, based on the imbalanced sample training, I-SVDD can increase the detection rate of a few classes, but appear only when the two hydrosphere intersection contains or may occur only when the relationship.

- The tendency to use way of discrimination and detection of minority class samples, and different types of classification algorithms (such as two-way classification of discrimination.

C. Penalty parameters optimation

For most classification and a small number of samples used in the same class, the penalty parameters C, for imbalanced data of the environment, makes the I-SVDD a large improvement which sampled the minority class to impose greater penalties parameters, for most classes of small samples of the punishment imposed parameters. If there are two types of super-ball contains between different punishments parameters conducive to small class of super balls shrink as much as possible, so as to separate the two types of training samples[5], will help reduce the I-SVDD of the samples misclassified.

V. EXPERIMENTS AND PERFORMANCE COMPARISONS

As to the traditional classification algorithm, the most commonly used evaluation criteria are the recognition rate of the test sample. However, in the case of imbalanced data classification accuracy to the minority class sample assessment is biased. For most classes, the minority class sample has little effect on the recognition rate. Therefore need to adopt a new assessment methods and indicators to describe the performance of the imbalance classification algorithm classification.

A. Evaluation Criteria

The system calls anomaly detection, each process data set for the system call, the result of the determination of the following four:

TP (True Positive), this belongs to a small number of categories (abnormal process) is judged to be the number of minority class (abnormal process).

FP (False Positive), this belongs to the majority class (normal process) is judged to be the number of minority class (abnormal process).

FN (False Negative), this belongs to a minority class (the abnormal process) is judged to be the number of the majority class (normal process).

TN (True Negative) of this belongs to the majority class (normal process) is judged to be the number of minority class (normal process).

Segment For analyte process four determination result, it is imbalanced classification commonly used in the assessment standards are given in Table I

TABLE I. FOUR RESULTS OF CLASSIFIER DECISION

	Recognized Minority	Recognized Majority
Actual Minority	TP	FN
Actual Majority	FP	TN

Currently many for imbalance classification performance metrics currently used metric the Sensitivity measure and Specificity metrics. Sensitivity (SE) metric is $TP/(TP+FN)$ which is used to evaluate samples (exception process), the recognition rate of the minority class; Specificity (SP) metric is $TN/(TN+FP)$ which is used to evaluate the recognition rate of the majority class samples (normal process).G-means used herein, now widely used in the imbalanced classification metric, defined as:

$$G = \sqrt{\text{Sensitivity*Specificity}} \qquad (14)$$

B. Experimental data and Parameter Selection

In order to fully evaluate SVDD, R-SVM, I-SVDD anomaly detection performance in Closed Systems, a total of two sets of experiments, respectively, using the University of New Mexico (University of New Mexico, UNM) website provides the Sendmail process data and MIT lpr process data. Sendmail data sets and MIT lpr data sets as an international standard data by researchers widely used [7], and has a certain representation on the sample size and imbalance.

TABLE II. DESCRIPTION OF DATASETS

Description		Category	SSC No	P No
Sendmail	Normal	Sendmail	19526	151
	Abnormal	Syslog-local-1	1516	6
		Syslog-local-2	1574	6
		Syslog-remote-1	1861	7
		Syslog-remote-2	1553	4
MIT lpr	Normal	Normal	2912133	2704
	Abnormal	Abnormal	163246	1001

For Sendmail datasets, experiments selected 40 normal process and the three abnormal process as a training sample to K = 6, the sliding window segmentation training samples, the normal non-repetitive system calls short sequence 255; selected three abnormal process the most representative system calls 15 of the short sequence as the minority class samples, the degree of imbalance of the two types of training samples of 17:1. Gaussian RBF kernel is selected as SVDD, R-SVM and the kernel function of the I-SVDD, SVDD parameters, I-SVDD parameters; R-SVM training integrated 17 subsets.

MIT lpr data sets, experiments selected 80 normal processes and 20 abnormal process as a training sample to K = 6, the sliding window segmentation training samples, the normal non-repetitive system calls short sequence 512; selected 20 abnormal the process most representative of the system calls a short sequence of 13 as a minority class samples [8].

TABLE III. SELECTED EXPERIMENTAL DATA

Description		Process for Training	SSC For Training	Experiment data	Imbal-ance degree
Send mail	Norm	40	255	110	17：1
	Abnor m	3	15	23	
MIT lpr	Norm	600	512	2704	32：1
	Abnor m	30	13	1001	

C. *Results*

Use SE, SP and G-means to evaluate these methods. Two sets of experimental results are shown in Table IV.

TABLE IV. RESULTS OF TWO EXPERIMENTAL DATASETS

Dataset	**Sendmail**			**MIT lpr**		
Evaluation criteria	SVDD	R-SVM	I-SVDD	SVDD	R-SVM	I-SVDD
SE	0.739	0.826	0.913	0.953	0.991	0.994
SP	0.963	0.954	0.963	0.948	0.975	0.976
G-means	0.844	0.881	0.938	0.951	0.983	0.985

The results reflect the R-SVM proposed in this paper can be very effective I-SVDD handling imbalance datasets, the best detection performance of the I-SVDD 17:1 Sendmail datasets; imbalance degree, I- the SVDD significantly better than the R-SVM, the imbalance degree less than 32:1 MIT lpr data sets, I-SVDD R-SVM detection performance.

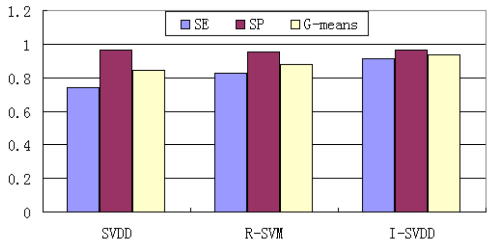

Figure 6. Comparaision of three measure results to Sendmail

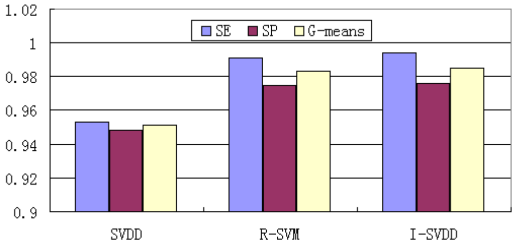

Figure 7. Comparaision of three measure results to MIT lpr

In this paper, the ROC curve drawing method: First, in the 0-1 interval setting a certain number of threshold, and then obtain three kernel methods in each threshold FFR (false alarm rate) and Tampered Failure Rate (detection rate) as the coordinates of the ROC curve, and finally these coordinates are arranged in accordance with the order of small to large connection and then generating a ROC curve. Figure 8 for the Sendmail and MIT lpr ROC curve:

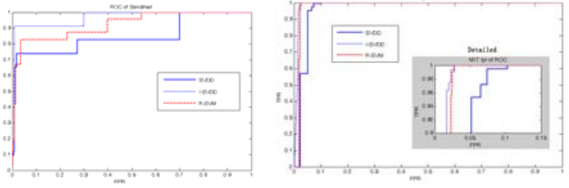

Figure 8. ROC curves of three classifiers to Sendmai & MIT lpr

ROC curve of SVDD is more close to the upper left corner, and a greater AUC area can be found by observation of the ROC curve of the two data sets. By quantitative calculation for Sendmail datasets, SVDD I-SVDD, R-SVM AUC were 0.8481,0.9666,0.9267; MIT lpr data sets SVDD, I-SVDD, R-SVM AUC is for 0.9586,0.9896,0.9795.

The evaluation results and by AUC quantitative comparison, R-SVM and I-SVDD this paper to solve the traditional SVM can not deal with the imbalance classification problem is easy to know, and the performance is better than single classification algorithm SVDD. In addition, R-SVM time complexity is lower than SVM and SVDD.

VI. CONCLUSIONS

The study of kernel methods proposed R-SVM and the I-SVDD focused on the imbalance data environment, existing kernel methods to improve, to make it more effective anomaly detection applied to the system call in the military environment. R-SVM, which is a Hilbert space sample weight-based approach of the original imbalanced classification algorithm, divided the classification problem into sub-problems and finally solved he result of the processing by using a majority vote decision, which can be divided by using the K-means algorithm clustering space samples; I-SVDD which is considered as an improved SVDD on the imbalanced sample training, can be used to solve the defects of the original the algorithm.

REFERENCES

[1] Tandom G, Chan P. Learning Useful System Call Attributes for Anomaly Detection[A]. In : Proc 18th Intl FLAIRS Conf, 2005.

[2] Amoroso, E.: Intrusion detection: an introduction to internet surveillance, correlation, trace back,traps, and response, 1st edn. Intrusion NetBooks ,1999, pp. 24-27.

[3] GhasemiGol, M., Monsefi, R., Sadoghi-Yazdi, H.: Ellipse Support Vector Data Description. EANN 2009, Springer, CCIS 43, pp. 257–268 (2009)

Article in a journal:

[4] Banerjee A, Burlina P, Diehl C. A support vector method for anomaly detection in hyperspectral imagery. IEEE Transactions on geoscience and remote sensing, 2008, 44(8) : 2282-2291. .

Article in a conference proceedings:

[5] Agarwal C (2005) An empirical bayes approach to detect anomalies in dynamic multidimen-sionalarrays. In: Proceedings of the 5th IEEE international conference on data mining. IEEE Computer Society,Washington, DC, USA, pp 26–33.

[6] Liu, Y., Gururajan, S., Cukic, B., Menzies, T., Napolitano, M.: Validating an online adaptive system using SVDD. In: Proceedings of the 15th IEEE international conference on tools with artificial intelligence (ICTAI'03), pp. 384–388. Sacramento, California, USA, 3–5 Nov 2003.

[7] Parmer Gabriel, West Richard, Hijack : Taking Control of COTS Systems for Real-Time User-Level Services. In : Proceedings of 13th IEEE on Real Time and Embedded Technology and Applications Symposium. April 2007, 133-146.

[8] Ji, R., Liu, D., Wu, M., Liu, J.: The application of SVDD in gene expression data clustering.In: Proceedings of the 2nd international conference on bioinformatics and biomedical engineering (ICBBE'08), pp. 371–374. Shanghai, China, 16–18 May 2008.