SVM-OD: SVM Method to Detect Outliers[1]

Jiaqi Wang[1], Chengqi Zhang[1], Xindong Wu[2], Hongwei Qi[3], Jue Wang[3]

[1] Faculty of Information Technology, University of Technology, Sydney, Australia

[2] Department of Computer Science, University of Vermont, USA

[3] Laboratory of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, China

ABSTRACT

Outlier detection is an important task in data mining because outliers can be either useful knowledge or noise. Many statistical methods have been applied to detect outliers, but they usually assume a given distribution of data and it is difficult to deal with high dimensional data. The Statistical Learning Theory (SLT) established by Vapnik et al. provides a new way to overcome these drawbacks. According to SLT Schölkopf et al. proposed a ν-Support Vector Machine (ν-SVM) and applied it to detect outliers. However, it is still difficult for data mining users to decide one key parameter in ν-SVM. This paper proposes a new SVM method to detect outliers, SVM-OD, which can avoid this parameter. We provide the theoretical analysis based on SLT as well as experiments to verify the effectiveness of our method. Moreover, an experiment on synthetic data shows that SVM-OD can detect some local outliers near the cluster with some distribution while ν-SVM cannot do that.

[1] This research is partly supported by the National Key Project for Basic Research in China (G1998030508).

1 Introduction

Outliers are abnormal observations from the main group, and are either noise or new knowledge hidden in the data. Researchers always wish to remove noise in the data during the pre-processing of data mining because noise may prevent many data mining tasks. In addition, data mining users are interested in new knowledge or unusual behaviors hidden behind data such as the fraud behavior of credit cards. Therefore, outlier detection is one of the important data mining tasks.

Statistics has been an important tool for outlier detection [1], and many researchers have tried to define outliers using statistical terms. Ferguson pointed out [4] that, "In a sample of moderate size taken from a certain population it appears that one or two values are surprisingly far away from the main group." Barnett et al. gave another definition [1], "An outlier in a set of data is an observation (or a subset of observations) which appears to be inconsistent with the remainder of that set of data." Hawkins characterized an outlier in a quite intuitive way [7], "An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism."

All these definitions imply that outliers in a given data set are events with a very low probability or even those generated by the different distribution from most data. Although statistical methods have been applied to detect outliers, usually they need to assume some distribution of data. It is also difficult for statistical methods to deal with high dimensional data [3, 6].

To some extent, the Statistical Learning Theory (SLT) established by Vapnik et al. and the corresponding algorithms can overcome these drawbacks [12]. According to this theory Schölkopf et al. proposed a ν-Support Vector Machine (ν-SVM) to estimate the support of a high dimensional distribution of data and applied it to detect outliers [11]. As pointed out by Schölkopf et al., a practical method has not been provided to decide the key parameter ν in ν-SVM though Theorem 7 in [11] gives the confidence that ν is a proper parameter to adjust. Therefore, it is still difficult for data mining users to decide this parameter.

In fact, we find in some experiments that this parameter can be avoided if another strategy is adopted. This strategy consists of two components: (1) a geometric method is applied to solve ν-SVM without the penalty term and (2) the support vector with the maximal coefficient is selected as the outlier. This method is called SVM-OD in this paper.

Vapnik et al. originally provided a standard SVM without the penalty term to solve the classification problem and then added the penalty term to deal with noise and nonlinear separability in the feature space [12]. In this paper, SVM-OD tries to detect outliers by solving ν-SVM without the penalty term. Although removing the penalty term from ν-SVM may drastically change the classification model, we can theoretically analyze why the strategy adopted in SVM-OD is reasonable for outlier detection based on SLT and the popular definition of outliers.

Some experiments on toy and real-world data show the effectiveness of SVM-OD. The experiment on a synthetic data set shows that when the kernel parameter is given, SVM-OD can detect some local outliers near the cluster with some distribution (e.g. *O1* and *O2* in Fig. 1) while ν-SVM cannot do that. The details about local outliers can be found in [3]. Another interesting phenomenon is found in the experiment on stock data that SVM-OD is insensitive for some values of the kernel parameter compared with ν-SVM though this still needs to be verified by the theoretical analysis and more experiments.

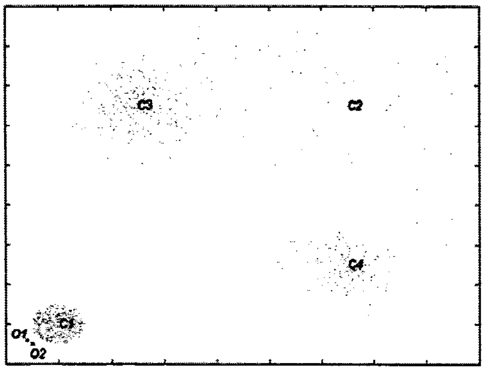


Fig. 1. Four clusters with different distributions and two local outliers (bigger points *O1, O2*)

The other sections in this paper are organized as follows. SVM-OD and ν-SVM are introduced in Section 2 and the theoretical analysis of SVM-OD is given in Section 3. Experiments are provided to illustrate the effectiveness of SVM-OD in Section 4. A discussion and conclusions are given in Sections 5 and 6. The notations used in this paper are shown in Table 1.

Table 1. Notations and their meanings

Notations	Meaning	Notations	Meaning
χ	Sample space	$\varphi(x)$	Points in kernel space
X	Sample set of size l	$(\bullet)$	Inner product
x_o	Outlier	Ω	Index set of sample
$K(x_i,x_j)$	Kernel function	Λ	Index set of SV

2 SVM-OD for Outlier Detection

In 2001, Schölkopf et al. presented ν-SVM to estimate the support of a high dimensional distribution of data and applied it to detect outliers [11]. ν-SVM solves the following optimization problem in the kernel space:

$$min\ 0.5\times\|w\|^2-\rho+\textstyle\sum_i\xi_i/\nu l \quad s.t.\ (w\bullet\varphi(x_i))\geq\rho-\xi_i,\ \xi_i\geq 0,\ i=1\ldots l. \qquad (1)$$

The geometric description of ν-SVM is shown in Fig. 2:

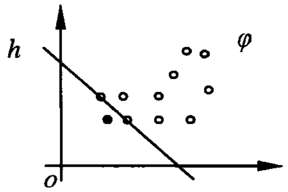


Fig. 2. o is the origin, φ is the sample set, h is the hyper-plane. h separates φ and o

Sequential Minimization Optimization (SMO), an optimization method proposed by Platt to solve classification problems [10], is extended to solve the optimization problem in (1). After the decision function f is obtained by solving (1), ν-SVM selects the samples x_i whose function value $f(x_i)<0$ as outliers. A key parameter ν in ν-SVM needs to be decided. However, it is not easy for data mining users to decide this parameter as implied in [11]. Thus SVM-OD, which can avoid ν, is introduced as follows.

Firstly, the Gaussian Radius Basis Function (RBF) $K(x,y) = exp\{-\|x-y\|^2/2\sigma^2\}$ is used as the kernel function in SVM-OD. The three properties of the RBF kernel, which are implied in [11] and will lead to the theorems in this paper, are discussed here.

Property 1 $\forall\ x\in X,\ K(x,x)=1.$

Property 2 For any $x,y\in X$ and $x\neq y,\ 0<K(x,y)<1.$

Property 3 In the kernel space spanned by RBF, the origin o and the mapped point set φ are linearly separable.

Proof According to Properties 1 and 2,

$$\exists\ x_i\in X,\ w=\varphi(x_i),\ \forall\ x_j\in X,\ (w\bullet\varphi(x_j))>0\ .$$

Let $0<\varepsilon<min(w\bullet\varphi(x_j))\ (x_j\in X)$, then $\forall\ x_j\in X,\ (w\bullet\varphi(x_j))\ -\ \varepsilon\ >\ 0$ and $(w\bullet o)-\varepsilon<0$, where o is the origin in the kernel space. Therefore the hyperplane $(w\bullet\varphi(x))-\varepsilon>0$ can separate φ and the origin o in the kernel space. ■

Secondly, SVM-OD solves the following optimization problem in the kernel space:

$$min\ 0.5\times\|w\|^2-\rho \quad s.t.\ (w\bullet\varphi(x_i))\geq\rho,\ i=1\ldots l. \qquad (2)$$

Although the optimization problem in (2) removes the penalty term in (1), the solution of (2) always exists for any given data because Property 3 guarantees that the mapped point set φ and the origin o in the kernel space are linearly separable. It can be proved that the optimization problem in (2) is equivalent to finding the shortest distance from the origin o to the convex hull C (see Fig. 3). This is the special case of the nearest point problem of two convex hulls and the proof can be found in [8].

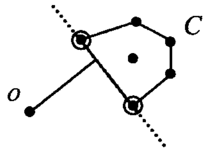


Fig. 3. Minimal normal problem (the shortest distance from the origin o to the convex hull C)

The corresponding optimization problem is as follows:

$$min\ \|\textstyle\sum_{i=1}^{l}\beta_i x_i\|^2 \ \ s.t.\ \sum_i \beta_i = 1,\ \beta_i \geq 0,\ i=1\ldots l. \tag{3}$$

Many geometric methods have been designed to solve (3), e.g. the Gilbert algorithm [5], and the MDM algorithm [9]. Since program developers can implement these geometric algorithms very easily, we combine them with the kernel function to solve the optimization problems in (2) or (3) in the kernel space. In this paper, the Gilbert algorithm is used to solve the problems in (2) or (3).

Finally, the function obtained by solving the optimization problem in (3) in the kernel space is as follows:

$$f(x)=\textstyle\sum_i \alpha_i K(x,x_i)-\rho,\ i\in \Lambda. \tag{4}$$

where $\rho=\sum_{i,j}\alpha_i K(x_j,x_i)$ and $i,j\in \Lambda$.

According to the function in (4), we define a decision function class as follows:

$$\{f_k|\ f_k(x)=\textstyle\sum_i \alpha_i K(x,x_i)-\rho+\alpha_k-\varepsilon,\ i\in \Lambda\text{-}\{k\},\ k\in \Lambda\}. \tag{5}$$

where $\varepsilon=min_i \alpha_i(1-K),\ i\in \Lambda,\ K=max_{i,j}K(x_i,x_j),\ i,j\in \Omega,\ i\neq j$.

The decision region of a decision function $f_k(x)$ in the decision function class in (5) is $R_{k,0}=\{x: f_k(x)\geq 0\}$. The following property holds true for the decision function class in (5).

Property 4 For each decision function in the decision function class in (5) $f_k(x)$,

$$f_k(x_k)<0 \text{ and } \forall\ i\in \Omega\text{-}\{k\}, f_k(x_i)\geq 0.$$

Proof $f_k(x)=\sum_i \alpha_i K(x,x_i)-\rho+\alpha_k(1-K(x,x_k))-\varepsilon$ $(i\in \Lambda)$ and according to Property 2, $0<\varepsilon<1$, $0<K<1$.

(1) $k\in \Lambda$, that is, x_k is a support vector, therefore $\sum_i \alpha_i K(x_k,x_i)-\rho=0$ $(i\in \Lambda)$. According to Property 1, $f_k(x_k)=\alpha_k(1-K(x_k,x_k))-\varepsilon=-\varepsilon<0$.

(2) $\forall\ j\in \Omega -\{k\}$, $\sum_i \alpha_i K(x_j,x_i)-\rho\geq 0$ $(i\in \Lambda)$. According to the definition of ε,

$$f_k(x_j)\geq \alpha_k(1-K(x_j,x_k))-\varepsilon\geq 0.$$

Therefore Property 4 holds true. ■

Then SVM-OD selects a function $f_o(x)$ in the decision function class in (5), where o is the index of α_o and $\alpha_o=max_i \alpha_i$ $(i\in \Lambda)$, as the decision function and the support vector x_o with the maximal coefficient α_o as the outlier. After removing the outlier x_o from the given data set, according to the same strategy we re-train the data and select the other outlier. The theoretical analysis for this strategy will be given in the next section.

The steps of SVM-OD are described below:

Step 1 Use the Gilbert algorithm to solve the optimization problem in (2) or (3) in the kernel space.

Step 2 Select the support vector with the maximal coefficient as an outlier.

Step 3 Remove this outlier from the given data set and go to Step 1.

3 Theoretical Analysis

According to statistics, a sample x_o will be regarded as an outlier if it falls in the region with a very low probability compared with other samples in the given set. The statistical methods to detect outliers usually assume some given distribution of data and then detect outliers according to the estimated probability. However, the real distribution of data is often unknown. When analyzing ν-SVM according to SLT, Schölkopf et al. provided the bound of the probability of the non-decision region, where outliers fall.

Definition 1 (Definition 6 in [11]) Suppose that f is a real-valued function on χ. Fix $\theta\in R$. For $x\in\chi$, let $d(X,f,\theta)=max\{0,\theta-f(x)\}$. And for a given data set X, we define $D(X,f,\theta)=\sum_x d(X,f,\theta)$ $(x\in X)$.

Then the following two theorems can be proved according to the results in [11], Definition 1 and four properties discussed in Section 2.

Theorem 2 Suppose that an independent identically distributed sample set X of size l is generated from an unknown distribution P that does not contain discrete components. For a decision function $f_k(x)$ in the decision function class in (5), the corresponding decision region is $R_{k,0}=\{x: f_k(x)\geq 0\}$, $k\in \Lambda$. Then with probability $1-\delta$ over randomly drawn training sequences X of size l, for all $\gamma>0$ and any $k\in \Lambda$,

$$P\{x: x\notin R_{k,\cdot}\}\le 2\times(k+log(l^2/2\delta))/l$$

where $c_1=4c^2$, $c_2=ln(2)/c^2$, $c=103$, $\hat{\gamma}=\gamma/\|w_k\|$, and $\|w_k\|=(\sum_{i1,i2}\alpha_{i1}\alpha_{i2}K(x_{i1},x_{i2}))^{1/2}$ $(i1,i2\in\Lambda-\{k\})$, $k=(c_1/\hat{\gamma}^2)log(c_2l\hat{\gamma}^2)+(\varepsilon/\hat{\gamma})log(e((2l-1)\hat{\gamma}/\varepsilon+1))+2$.

Proof According to Property 4, $f_k(x_k)=-\varepsilon$ and $\forall\ i\in\Omega-\{k\}$, $f_k(x_i)\ge 0$. So $D=D(X,f_k,0)=\varepsilon$, where D is defined in Definition 1. Theorem 2 holds true according to Theorem 17 in [11]. ■

Theorem 3 For any decision function in the decision function class in (5) $f_k(x)$, $(k\in\Lambda)$, $\|w_k\|^2=\sum_{i,j}\alpha_i\alpha_jK(x_i,x_j)$, $(i,j\in\Lambda-\{k\}, k\in\Lambda)$. Then

$$\|w_o\|^2=min_k\|w_k\|^2 \text{ iff } \alpha_o=max_k\alpha_k\ (k\in\Lambda).$$

Proof For $o\in\Lambda$ and any $k\in\Lambda\ k\neq o$, $\alpha_k+\sum_i\alpha_iK(x_k,x_i)=\alpha_o+\sum_j\alpha_jK(x_o,x_j)=\rho$ where $i\in\Lambda-\{k\}$, $j\in\Lambda-\{o\}$ and ρ is defined in the function in (4). So $\sum_i\alpha_iK(x_k,x_i)=\alpha_o-\alpha_k+\sum_j\alpha_jK(x_o,x_j)\ (i\in\Lambda-\{k\}, j\in\Lambda-\{o\})$.

$\|w_o\|^2=\sum_{i,j}\alpha_i\alpha_jK(x_i,x_j)(i,j\in\Lambda-\{o\})=\sum_{i,j}\alpha_i\alpha_jK(x_i,x_j)(i,j\in\Lambda)-2\alpha_o\sum_i\alpha_iK(x_o,x_i)(i\in\Lambda-\{o\})-\alpha_o^2$

$\|w_k\|^2=\sum_{i,j}\alpha_i\alpha_jK(x_i,x_j)(i,j\in\Lambda-\{k\})=\sum_{i,j}\alpha_i\alpha_jK(x_i,x_j)(i,j\in\Lambda)-2\alpha_k\sum_i\alpha_iK(x_k,x_i)(i\in\Lambda-\{k\})-\alpha_k^2$

So $\|w_o\|^2-\|w_k\|^2=2\alpha_k\sum_i\alpha_iK(x_k,x_i)\ (i\in\Lambda-\{k\})-2\alpha_o\sum_i\alpha_iK(x_o,x_i)\ (i\in\Lambda-\{o\})+\alpha_k^2-\alpha_o^2$

$$\begin{aligned}
&=2\alpha_k(\alpha_o-\alpha_k+\textstyle\sum_i\alpha_iK(x_o,x_i))-2\alpha_o\sum_i\alpha_iK(x_o,x_i))+\alpha_k^2-\alpha_o^2\ (i\in\Lambda-\{o\})\\
&=2\alpha_k\alpha_o-\alpha_k^2-\alpha_o^2+2(\alpha_k-\alpha_o)\textstyle\sum_i\alpha_iK(x_o,x_i)\ (i\in\Lambda-\{o\})\\
&=-(\alpha_o-\alpha_k)^2-2(\alpha_o-\alpha_k)\textstyle\sum_i\alpha_iK(x_o,x_i)\ (i\in\Lambda-\{o\})\\
&=-(\alpha_o-\alpha_k)(\alpha_o-\alpha_k+2\textstyle\sum_i\alpha_iK(x_o,x_i))\ (i\in\Lambda-\{o\})\\
&=-(\alpha_o-\alpha_k)(\textstyle\sum_i\alpha_iK(x_o,x_i)+\sum_j\alpha_jK(x_k,x_j))\ (i\in\Lambda-\{o\}, j\in\Lambda-\{k\}).
\end{aligned}$$

According to Properties 1 and 2 of RBF, the following inequality always holds true:

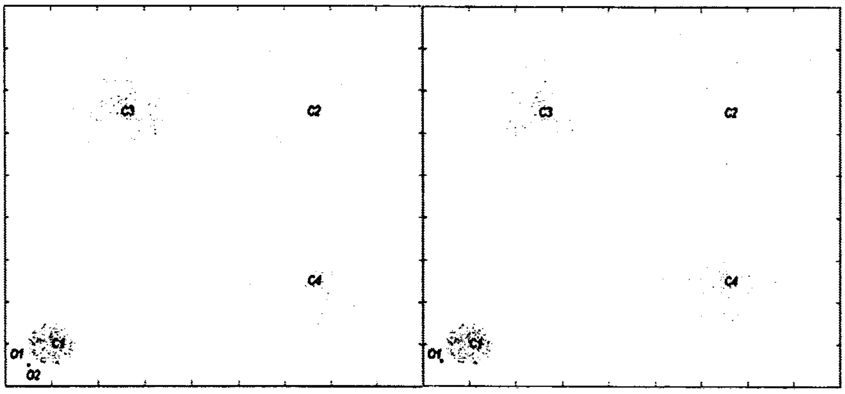
$$\forall\ i,j\in\Omega,\ K(x_i,x_j)>0.$$

And $\forall\ i\in\Lambda$, $\alpha_i>0$, therefore $\|w_o\|^2=min_k\|w_i\|^2$ iff $\alpha_o=max_k\alpha_k\ (k\in\Lambda)$. ■

Note that ε in Theorem 2 is a constant for any decision function in the decision function class in (5). Therefore Theorem 2 shows that the smaller value of $\|w\|^2$, the lower bound of probability of the non-decision region decided by the decision function. Furthermore, Theorem 3 shows that we can obtain the smaller value of $\|w\|^2$ if the function $f_o(x)$ in the decision function class in (5), where o is the index of α_o, is chosen as the decision function. This means that the probability of the non-decision region decided by $f_o(x)$ is low compared with others. In addition, according to Property 4, the following inequalities hold true: $\forall\ i\in\Omega-\{o\}$, $f_o(x_i)\ge 0$ and $f_o(x_o)<0$. This means that the support vector x_o with the maximal coefficient falls in the non-decision region decided by $f_o(x)$. Thus we select x_o as an outlier.

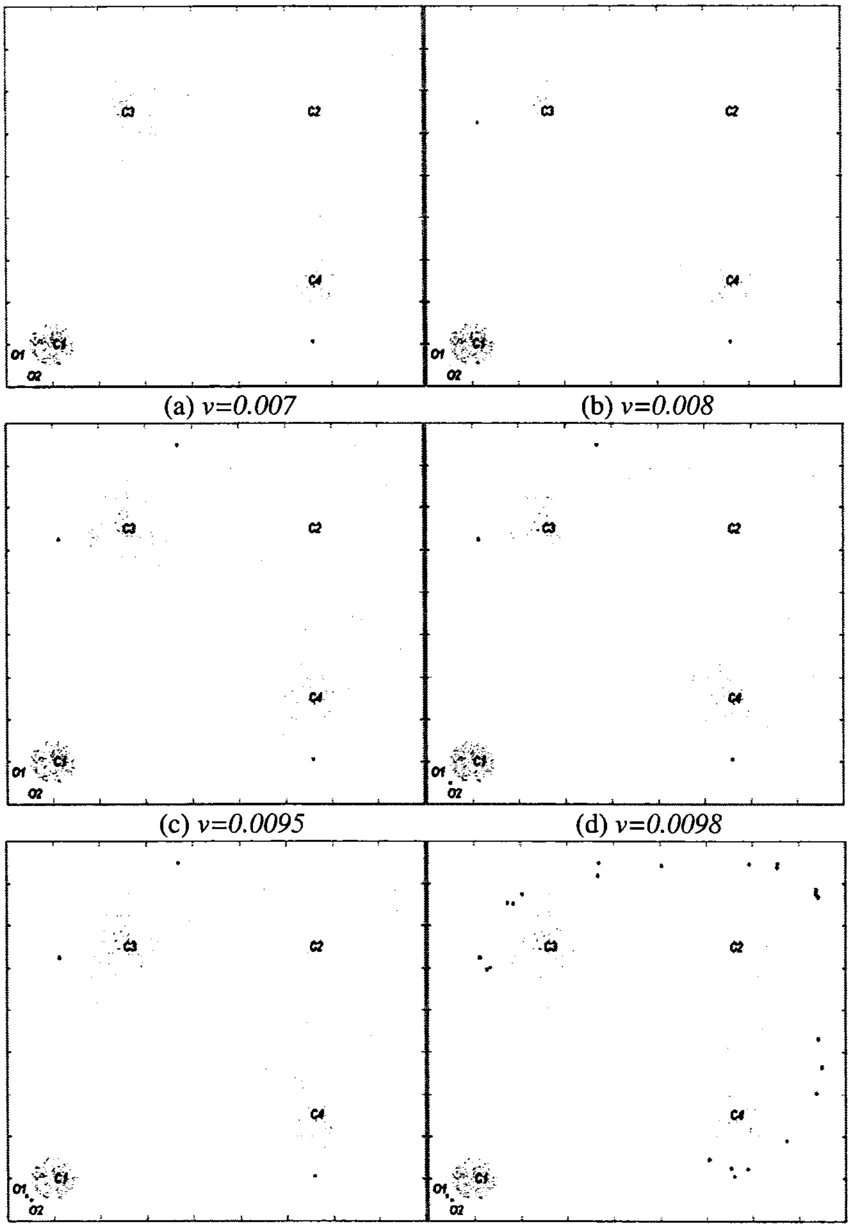
4 Experiments

Experiment 1. This experiment is performed on a synthetic data set including local outliers. There are respectively 400 and 100 points in the clusters *C1* and *C2* with the uniform distribution and respectively 300 and 200 points in the clusters *C3* and *C4* with the Gaussian distribution. In addition, there are two local outliers *O1* and *O2* near the cluster *C1* (see Fig. 1). The details about local outliers can be found in [3]. The goal of this experiment is to test whether ν-SVM and SVM-OD only detect these two local outliers since other points are regarded as the normal data from some distributions. σ=*5* is set as the kernel parameter value. In Fig. 1, bigger points are two local outliers. In Fig. 4 and Fig. 5, bigger points are some "outliers" detected by ν-SVM and SVM-OD. Comparing SVM-OD and ν-SVM, we find that SVM-OD can detect *O1* and *O2* after two loops. However, ν-SVM either does not detect both *O1* and *O2* or detect other normal data other than these two local outliers (see Fig. 4 and Fig. 5) when the different values of ν are tested. So in this experiment SVM-OD is more effective for detecting some kinds of local outliers than ν-SVM.



(a) first loop of SVM-OD (b) second loop of SVM-OD

Fig. 4. (a) only *O2* is detected as "outlier", (b) only *O1* is detected as "outlier"



(a) *v=0.007*

(b) *v=0.008*

(c) *v=0.0095*

(d) *v=0.0098*

(e) *v=0.01*

(f) *v=0.05*

Fig. 5. (a) 1 "outlier" is detected (without both *O1* and *O2*), (b) 2 "outliers" are detected (without both *O1* and *O2*), (c) 3 "outliers" are detected (without both *O1* and *O2*), (d) 4 "outliers" are detected (without *O1* and with *O2*), (e) 5 "outliers" are detected (with both *O1*, *O2* and others), (f) many "outliers" are detected (with both *O1*, *O2* and others)

Experiment 2. This experiment is performed on the first 5000 samples of the MNIST test set because there are fewer outliers in these samples than the last 5000. Data can be available from the website "yann.lecun.com/exdb/mnist/". In both SVM-OD and ν-SVM, $\sigma=8\times256$ is selected as the kernel parameter value. For ten hand-digits (0-9), the penalty factor ν in ν-SVM is *5.57%*, *5.47%*, *5.9%*, *5.71%*, *5.88%*, *7.3%*, *5.41%*, *6.51%*, *5.84%*, and *4.81%*, respectively. Classes (0-9) are labeled on the left side of Fig. 6 and Fig. 7. From both Fig. 6 and Fig. 7, a number of samples detected are either abnormal or mislabeled. The results of this experiment show that SVM-OD is effective for outlier detection. What is more important is that SVM-OD avoids adjusting the penalty factor ν, while this parameter is needed in ν-SVM.

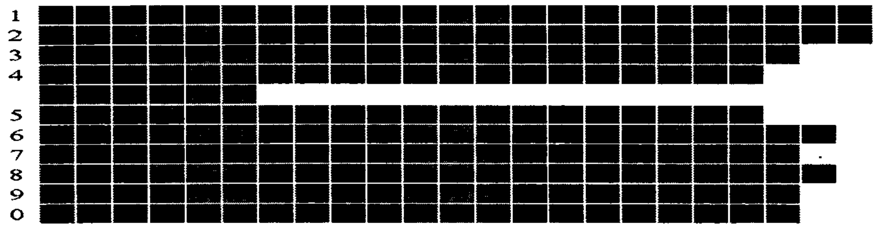


Fig. 6. ν-SVM detects some outliers in the first 5000 examples of MNIST (0-9) test set

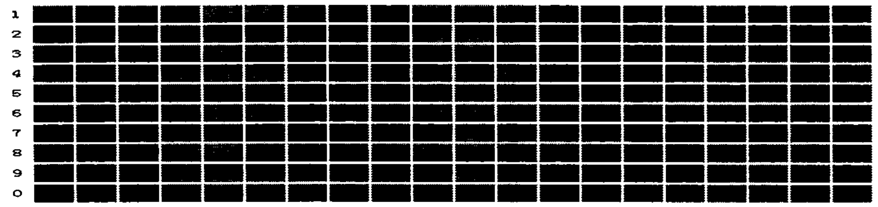
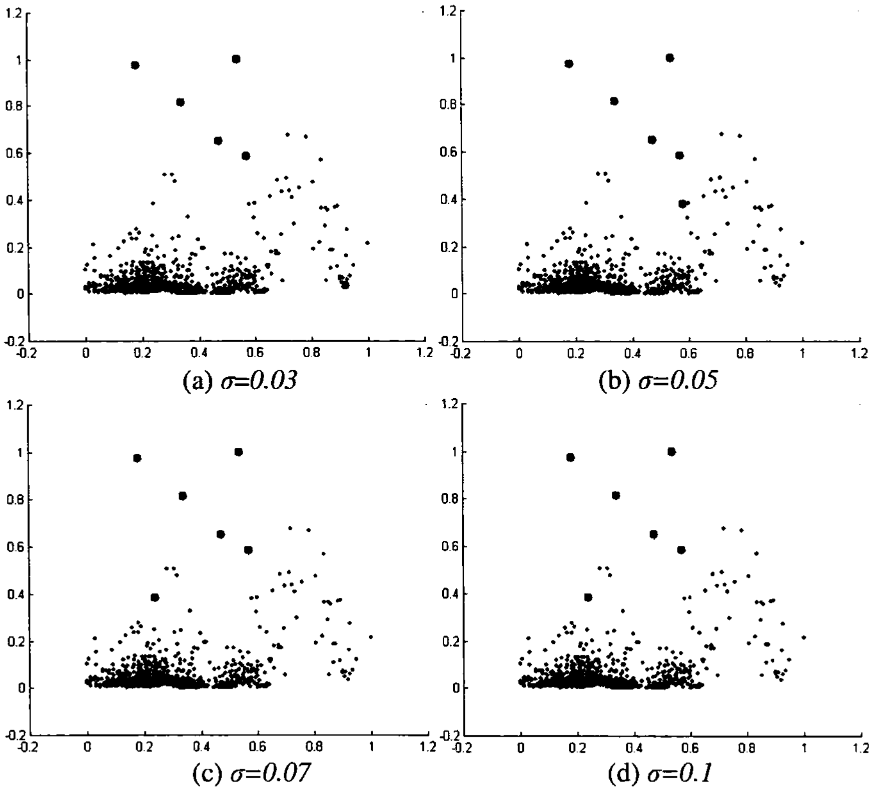


Fig. 7. SVM-OD detects some outliers in the first 5000 examples of MNIST (0-9) test set

Experiment 3. This experiment is conducted on the stock data including 909 daily samples from the beginning of 1998 to the end of 2001. The name of stock is not provided for the commercial reason. In Fig. 8 and Fig. 9, the horizontal coordinate refers to the stock return and the vertical coordinate the trading volume. The penalty factor ν in ν-SVM is *5.5%* and the kernel parameter σ in SVM-OD and ν-SVM are shown in Fig. 8 and Fig. 9. Bigger points in these two figures are outliers detected by SVM-OD and ν-SVM respectively. The goal of this experiment is to show that SVM-OD can also detect those points far away from the main group, though it is still necessary to verify whether or not outliers detected by SVM-OD and ν-SVM are unusual behaviors in the stock market. An interesting phenomenon is also found that SVM-OD is more insensitive for some values of the kernel parameter than ν-SVM though more experiments and the theoretical analysis are needed (see Fig. 8 and Fig. 9).



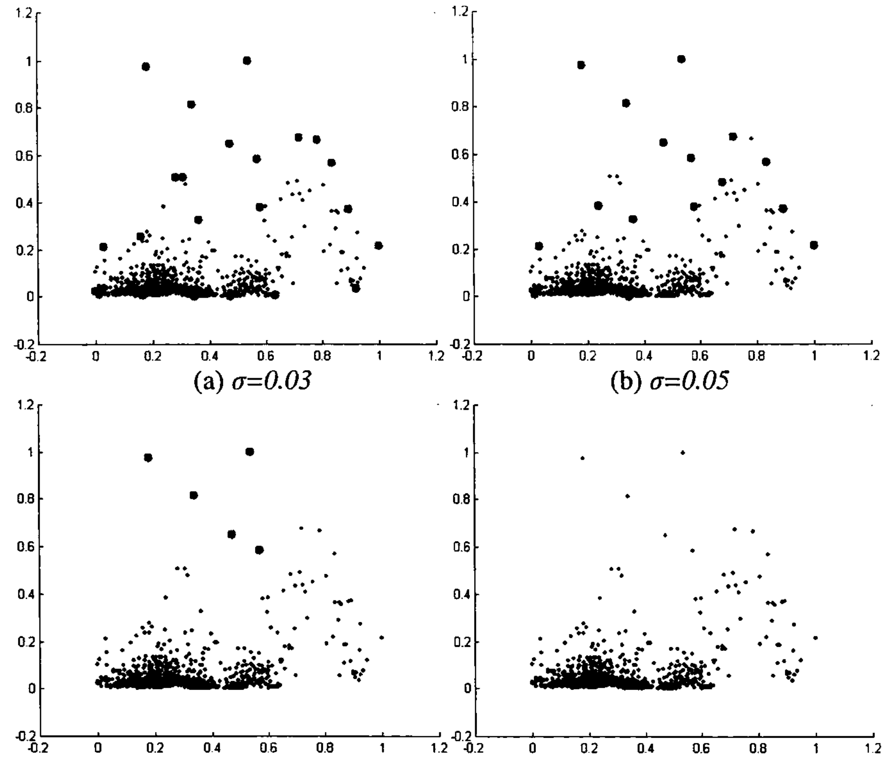
(a) σ=*0.03*

(b) σ=*0.05*

(c) σ=*0.07*

(d) σ=*0.1*

Fig. 8. Outliers (bigger points) detected by SVM-OD are shown in these four figures



(a) σ=0.03

(b) σ=0.05

(c) σ=0.07

(d) σ=0.1

Fig. 9. Outliers (bigger points) detected by ν-SVM are shown in these four figures

5 Discussion

While Section 3 explains why it is reasonable for SVM-OD to select the support vector with the maximal coefficient as an outlier, SVM-OD does not tell us the number of outliers in a given data set and the stopping criteria of the algorithm. In fact, the number of outliers depends on the user's prior knowledge about the fraction of outliers. Actually, the number of outliers is a more intuitive concept than the penalty factor ν in ν-SVM and the user can more easily know the approximate ratio of outliers compared with ν.

ν-SVM does not provide a relationship between the fraction of outliers and ν although ν is proved to be the upper bound of the fraction of outliers [11]. It is still difficult for the user to decide ν even if they know the fraction of outliers in advance. For example, Table 1 in [11] showed that when ν is

4%, the fraction of outliers is *0.6%* and when ν is *5%*, the fraction of outliers is *1.4%*. However, when the user knows the fraction of outliers (e.g. *1.4%*) before detecting outliers, which value of ν (*4%*, *5%*, or another upper bound of *1.4%*) should be selected to obtain *1.4%* samples as outliers? SVM-OD can avoid this problem when the user knows the approximate fraction of outliers. There is the similar problem in another work about detecting outliers based on support vector clustering [2].

In addition, Table 1 in [11] pointed out that the training time of ν-SVM increases as the fraction of outliers detected becomes more. There is a similar property for the training cost of SVM-OD. This paper does not compare the training costs of these two methods, which is a topic for the future work.

6 Conclusion

This paper has proposed a new method to detect outliers called SVM-OD. Compared to ν-SVM, SVM-OD can be used by data mining users more easily since it avoids the penalty factor ν required in ν-SVM. We have verified the effectiveness of SVM-OD according to the theoretical analysis based on SLT and some experiments on both toy and real-world data. The experiment on a synthetic data set shows that when the kernel parameter is fixed, SVM-OD can detect some local outliers while ν-SVM cannot do that. In the experiment on stock data, it is found that SVM-OD is insensitive for some values of the kernel parameter compared with ν-SVM. In the future work, we will try to give a theoretical explanation to this phenomenon and compare SVM-OD with more methods to detect outliers on more real-world data from the different sides, e.g. the training cost and the effectiveness.

References

1. Barnett V, Lewis T (1984) Outliers in statistical data. John Wiley & Sons, New York
2. Ben-Hur A, Horn D, Siegelmann HT, Vapnik V (2001) Support vector clustering. Journal of Machine Learning Research 2: 125-137
3. Breunig MM, Kriegel H-P, Ng RT, Sander J (2000) LOF: identifying density-based local outliers. Proceedings of ACM SIGMOD Conference
4. Ferguson TS (1961) On the rejection of outliers. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1: 253-287
5. Gilbert EG (1966) Minimizing the quadratic form on a convex set. SIAM Journal of Control 4: 61-79

6. Han J, Kamber M (2000) Data mining: concepts and techniques. Morgan Kaufmann Publishers, Inc.
7. Hawkins D (1980) Identification of outliers. Chapman and Hall, London
8. Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK (2000) A fast iterative nearest point algorithm for support vector machine classifier design. IEEE Transactions on Neural Networks 11(1): 124-136
9. Mitchell BF, Dem'yanov VF, Malizemov VN (1974) Finding the point of a polyhedron closet to the origin. SIAM Journal of Control 12: 19-26
10. Platt JC (1999) Fast training of support vector machines using sequential minimal optimization. In: Advances in kernel methods – support vector learning. MIT press, Cambridge, MA, pp 185-208
11. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC (2001) Estimating the support of a high-dimensional distribution. Neural Computation 13 443-1471
12. Vapnik V (1999) The nature of statistical learning theory (2nd edn). Springer-Verlag, New York