Nonlinear Dyn (2011) 64: 127–130
DOI 10.1007/s11071-010-9851-y

ORIGINAL PAPER

Local density one-class support vector machines for anomaly detection

Jiang Tian · Hong Gu · Chiyang Gao · Jie Lian

Received: 22 June 2010 / Accepted: 17 September 2010 / Published online: 8 October 2010

Abstract Anomalies are patterns in data that do not conform to a well-defined notion of normal behavior. One-class Support Vector Machines calculate a hyperplane in the feature space to distinguish anomalies, however, it may not identify the ideal hyperplane especially when the support vectors do not have the overall characteristics of the target data. So, we propose a new local density OCSVM by incorporating distance measurements based on local density degree to reflect the distribution of a given data set. Experimental results on UCI data sets show that the proposed method can achieve better performance than other one class learning schemes.

Keywords One class support vector machine · Anomaly detection · Local density degree

1 Introduction

Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. These nonconforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities, or contaminants in different application domains. With the development of information technology, vast quantities of data are captured and stored. The capacity, dimensions, and complexity of database have grown rapidly, but usually the real data set could not be used directly for data mining due to ignorance, human errors, rounding errors, transcription factor, instrument malfunction, and biases. Anomaly detection is used to find the objects that do not comply with the general behavior of the data and then lead to potentially useful information. Anomalies can be translated to significant or critical actionable information. Anomaly detection can be applied to many fields, such as credit card fraud detection, security systems, medical diagnosis, network intrusion detection, and information recovery [1, 2].

One-class classification based anomaly detection techniques assume that all training instances have only one class label. Such techniques learn a discriminative boundary around the normal instances using a one-class classification algorithm. Any test instance that does not fall within the learned boundary is declared as anomalies. Support Vector Machines (SVMs) have been applied to anomaly detection in the one-class setting, for example, One-class SVMs [3, 4] find a hyperplane in feature space which has maximal margin to the origin and a prespecified fraction of the training examples lay beyond it. A variant of the basic technique named support vector data description (SVDD) [5, 6] finds the smallest hypersphere in the kernel

J. Tian (✉) · C. Gao
China Software Testing Center, China Center for Information Industry Development, Beijing 100048, China
e-mail: tianjiang@gmail.com

H. Gu · J. Lian
School of Control Science and Engineering, Dalian University of Technology, Dalian 116023, China

space, which contains all training instances, then decides which side of that hypersphere a test instance lies. They differ slightly in spirit and geometric notion, in particular, the OCSVM uses the margin arguments. Many one-class SVM-based techniques have been proposed for anomaly detection in medical diagnosis [7], novelty detection in power generation plants [8], seizure analysis from intracranial electroencephalogram [9], protein location prediction [10], and intrusion detection [11].

The unsupervised SVMs are promising in detecting new anomalies, however, the conventional OCSVM has limits to reflect overall characteristics of a target data set on its density distribution. In the OCSVMs, the small portion of samples called support vectors fully decide the hyperplane in the feature space, whereas all the nonsupport vectors have no influence on the hyperplane, regardless of the density distribution. But the region around a nonsupport vector with higher density degree should be included rather than other regions to more correctly identify the decision hyperplane. Therefore, the solution solely based on the support vectors, without considering the density distribution, can miss the optimal solution. To address this issue and find a more robust solution, we propose a new One-class SVM to reflect the different local density of a target data set by introducing the notion of a local density for each data point. We refer to the proposed method as a local density One-class SVM (LD-OCSVM).

2 Local density OCSVM method

2.1 Local density degree

Lee et al. [10] proposed a method to extract a local density degree for each data point from a target data set using a nearest neighborhood approach. Calculate a local density degree δ_i for a target data sample x_i. Let $d(x_i, x_i^K)$ be the distance between x_i and x_i^K, where x_i^K is the Kth nearest neighborhood of x_i. Let M^K represents the mean distance of Kth nearest neighborhoods of all target data. The local density δ_i for x_i is defined by

$$\delta_i = \exp\left(\frac{M^K}{d(x_i, x_i^K)}\right), \quad i = 1, \ldots, l, \tag{1}$$

where $M^K = \frac{1}{\ell}\sum_i d(x_i, x_i^K)$, ℓ is the number of data samples. It is obvious that (1) reports a higher local density degree δ_i for a sample in higher density region.

2.2 LD-OCSVM

The conventional OCSVM [3, 4] was proposed for estimating the support of a data distribution instead of the full density. It estimates a hyperplane in feature space such that a prespecified fraction of the training examples will lie beyond that hyperplane, while the hyperplane has maximal margin to the origin. However, the support vectors fully decide the hyperplane in the kernel space, whereas all the nonsupport vectors have no influence on the hyperplane. Therefore, to find the ideal separating hyperplane, the support vectors should be calculated with considering density distribution of a target data set.

Given training data $x_1, \ldots, x_\ell \in \mathcal{X}$, where $\ell \in \mathbb{N}$ is the number of observations, and $\mathcal{X}$ is some data set. After incorporating the local density of a training data set, we solve the following quadratic problem:

$$\min \frac{1}{2}\|w\|^2 + \frac{1}{\nu\ell}\sum_i \delta_i\xi_i - \rho, \tag{2}$$

subject to

$$\begin{aligned} &\big(w \cdot \phi(x_i)\big) \geq \rho - \xi_i, \\ &\xi_i \geq 0. \end{aligned} \tag{3}$$

Here, $\nu \in (0, 1]$ is a parameter controlling this tradeoff, δ_i is the density degree in (1), ρ is the offset, nonzero slack variables $\delta_i\xi_i$ are penalized in the objective function. As described in the former section, a sample in higher density region has a higher local density degree δ_i, so the corresponding slack variable is higher. After introducing Lagrange multipliers $\alpha_i, \beta_i \geq 0$, we build the Lagrangian:

$$\begin{aligned} L(w, \xi, \rho, \alpha, \beta) = \frac{1}{2}\|w\|^2 + \frac{1}{\nu\ell}\sum_i \delta_i\xi_i - \rho \\ - \sum_i \alpha_i\big(\big(w \cdot \phi(x_i)\big) \\ - \rho + \xi_i\big) - \sum_i \beta_i\xi_i, \end{aligned} \tag{4}$$

and set the derivatives with respect to the primal variables w, ξ, ρ equal to zero, yielding

$$\begin{aligned} w &= \sum_i \alpha_i \phi(x_i) \\ \alpha_i &= \frac{\delta_i}{\nu\ell} - \beta_i \leq \frac{\delta_i}{\nu\ell} \\ \sum_i \alpha_i &= 1. \end{aligned} \tag{5}$$

Replacing the corresponding terms in (4) by those in (5), we obtain the dual objective of LD-OCSVM:

$$\min \frac{1}{2} \sum_{ij} \alpha_i \alpha_j K(x_i, x_j), \tag{6}$$

subject to

$$\begin{aligned} 0 \leq \alpha_i &\leq \frac{\delta_i}{\nu\ell}, \\ \sum_i \alpha_i &= 1, \end{aligned} \tag{7}$$

where α is the Lagrangian parameter, $K(x, y) = (\phi(x) \cdot \phi(y))$ represents a kernel function, and the point x_i with the corresponding $\alpha_i > 0$ is called a support vector (SV). After deriving the dual problem, the decision function can be shown to have a SV explanation:

$$f(x) = \text{sgn}\left(\sum_i \alpha_i K(x_i, x) - \rho \right). \tag{8}$$

The offset ρ can be recovered by exploiting that for any α_i which is not at the upper or lower bound, the corresponding pattern x_i satisfies

$$\rho = \left(w \cdot \phi(x_i)\right) = \sum_j \alpha_j k(x_j, x_i). \tag{9}$$

When we apply the proposed algorithm to anomaly detection applications, training examples with decision values greater than zero are considered as normal, while the others are detected as anomalies.

3 Experiments

3.1 Data sets

We validated our method on five different data sets from the UCI repository for machine learning. These data sets belong to conventional classification problems with multiple classes. To make them suitable for anomaly detection applications, the data sets were converted to binary class problems. We selected one of the classes to represent the anomalies, while the remaining classes were gathered to form the normal class. Table 1 lists the data sets, along with some fitting statistics. The training data sets include only normal patterns, while the test data sets include both normal and anomaly patterns. Because the information about the UCI data sets is public, we do not report it here.

Table 1 Five data sets from the UCI repository

Data set	Outlier class	Training samples	Test samples	
			Normal	Anomalies
Balance	B	357	219	49
Glass	headlamps	114	71	29
Hypothyroid	Positive	2136	1442	194
Optdigits	6	3023	2039	558
Sick	Sick	2105	1436	231

3.2 Results and discussion

We compared our method with three other One-class SVM based anomaly detection methods, the conventional OCSVM [4] , SVDD [6], and D-SVDD [10], a polynomial kernel, and a Gaussian RBF kernel were used. We used a cross validation strategy to estimate the optimizing parameters of the one-class classifiers. Our program was developed based on Libsvm [12]. These methods were compared using respective area under receiver operating characteristic curves (AUCs) [13].

To validate the effectiveness of our method, the experiment had been run for ten times, in each cycle the data sets were separated randomly into two parts. The average AUCs of different methods are given in Table 2. For the balance data set, the AUCs of OCSVM are 0.762 and 0.785. For the same data set, the AUCs of SVDD are 0.728 and 0.731, and the AUCs of D-SVDD are 0.791 and 0.791. The LD-OCSVM method, however, shows AUCs of 0.811 and 0.893, which are better than the other classifiers. Of all the results on the Balance data set, the proposed method with a RBF kernel has the best performance. For the other data sets, the OCSVM and SVDD show similar performances, while the proposed LD-OCSVM shows best performance. The proposed LD-OCSVM outperforms the

Table 2 Experimental results

AUC	OCSVM		SVDD		D-SVDD		LD-OCSVM	
	Poly-3	RBF	Poly-3	RBF	Poly-3	RBF	Poly-3	RBF
Balance	0.762	0.785	0.728	0.731	0.755	0.791	0.811	0.893
Glass	0.665	0.710	0.683	0.702	0.698	0.741	0.718	0.766
Hypothyroid	0.775	0.811	0.783	0.822	0.801	0.842	0.821	0.872
Optdigits	0.691	0.706	0.681	0.698	0.689	0.721	0.698	0.734
Sick	0.811	0.852	0.824	0.855	0.832	0.873	0.842	0.879

other methods, which means it may identify the optimal solution of the target description and provide a best solution for detecting anomalies. Furthermore, better results are obtained by classifiers with RBF kernels.

From these results, we drew a conclusion that the proposed method showed better performance than the OCSVM, SVDD, and D-SVDD for all the data sets used regardless of the type of kernel functions. Moreover, the best performance was obtained when the LD-OCSVM with a RBF kernel was used. For our method, it can be inferred that the anomalies can be effectively detected by the optimizing hyperplane, which incorporating the density distribution of a given data set. In addition, it is important to mention that we have also validated our method on data sets using other classes as anomaly class. The results were similar to those reported in this section.

4 Conclusion

In this study, we proposed a novel One-class SVM for identifying the anomalies in the feature space. To reflect the overall characteristics of a target data set on its density distribution, each data point was associated with a relative density degree. Using the distance measurements, we developed a local density OCSVM whose support vectors could better reflect the overall characteristics of the target data. Experimental results showed that the proposed method outperformed the OCSVM, SVDD and D-SVDD for all test data sets.

References

1. Hodge, V., Austin, J.: A survey of outlier detection methodologies. Artif. Intell. Rev. **22**(2), 85–126 (2004)
2. Filzmoser, P., Maronna, R., Werner, M.: Outlier identification in high dimensions. Comput. Stat. Data Anal. **52**(3), 1694–1711 (2008)
3. Scholkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J., Platt, J.: Support vector method for novelty detection. Adv. Neural Inf. Process. Syst. **12**(3), 582–588 (2000)
4. Scholkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Comput. **13**(7), 1443–1471 (2001)
5. Tax, D.M.J., Duin, R.P.W.: Support vector domain description. Pattern Recognit. Lett. **20**(11–13), 1191–1199 (1999)
6. Tax, D.M.J., Duin, R.P.W.: Support vector data description. Mach. Learn. **54**(1), 45–66 (2004)
7. Tian, J., Gu, H.: Anomaly detection combining one-class SVMs and particle swarm optimization algorithms. Nonlinear Dyn. **61**(1), 303–310 (2010)
8. King, S.P., King, D.M., Astley, K., Tarassenko, L., Hayton, P., Utete, S.: The use of novelty detection techniques for monitoring high-integrity plant. In: Proceedings of the 2002 International Conference on Control Applications, vol. 1, Anchorage, AK, pp. 221–226 (2002)
9. Gardner, A.B., Krieger, A.M., Vachtsevanos, G., Litt, B.: One-class novelty detection for seizure analysis from intracranial EEG. J. Mach. Learn. Res. **7**(7), 1025–1044 (2006)
10. Lee, K., Kim, D.W., Lee, K.H., Lee, D.: Density-induced support vector data description. IEEE Trans. Neural Netw. **18**(1), 284–289 (2007)
11. Giacinto, G., Perdisci, R., Del Rio, M., Roli, F.: Intrusion detection in computer networks by a modular ensemble of one-class classifiers. Inf. Fusion **9**(1), 69–82 (2008)
12. Chang, C.C., Lin, C.J. Libsvm: a library for support vector machines, software available at http://www.csie.ntu.edu.tw/cjlin/libsvm [cp] (2001)
13. Fawcett, T.: An introduction to ROC analysis. Pattern Recognit. Lett. **27**(8), 861–874 (2006)