Support Vector Clustering for Outlier Detection

Hai-Lei Wang[1,2]
[1]the Institute of Intelligence Machines,
Chinese Academy of Sciences
[2]the department of Automation, the University of
Science and Technology of China
Hefei, Anhui, China
e-mail: hlwang@iim.ac.cn

Wen-Bo Li and Bing-Yu Sun
the Institute of Intelligence Machines,
Chinese Academy of Sciences
Hefei, Anhui, China
e-mail: wbli@iim.ac.cn; bysun@ustc.edu

***Abstract*—In this paper a novel Support vector clustering(SVC) method for outlier detection is proposed. Outlier detection algorithms have application in several tasks such as data mining, data preprocessing, data filter-cleaner, time series analysis and so on. Traditionally outlier detection methods are mostly based on modeling data based on its statistical properties and these approaches are only preferred when large scale set is available. To solve this problem, in this paper we focus on establishing the context of support vector clustering approach for outlier detection. Compared to traditional outlier detection methods , the performance of the SVC is not sensitive to the selection of needed parameters. The experiment results proved the efficiency of our method.**

Keywords-Support vector clustering; Outlier detection; Nearest Distance.

I. INTRODUCTION

Given a set of data points drawn from some probability distribution, an outlier with respect to that distribution is an unlikely point. Studies from the field of statistics have typically considered outliers to be an observation that deviates so much from other observations as to arouse suspicion that it was generated by different mechanism. Outlier detection, as a familiar problem in machine learning, is to determine which data points are outliers when the underlying distribution is unknown. Outlier detection algorithms have application in several tasks such as data mining, data preprocessing, data filter-cleaner, time series analysis and so on. Moreover, several applications require the classifier to act as a detector rather as a classifier, that is, the requirement is to detect whether an input is part of the data that the classifier was trained on or it is in fact unknown.

The fist most commonly used approaches are statistical approaches. Statistical approaches are mostly based on modeling data based on its statistical properties and using this information to estimate whether a samples comes from the same distribution or not. So the key of this kind of approaches is the estimation of the underlying distribution of the data. Two main strategies exist to this estimation: parametric and non-parametric methods [1][6]. The parametric approach assumes that the data comes from a family of known distributions, such as the normal distribution and certain parameters are calculated to fit this distribution. However, in most real world situations the underlying distribution of the data is not known therefore such techniques have little practical importance. In non-parametric methods the overall form of the density function is derived from the data as well as the parameters of the model. As a result non-parametric methods give greater flexibility in general systems. In these approaches, however, the drawback is that in general (in higher dimensional feature spaces) a large number of samples is required. So these approaches are only preferred when large scale set is available.

The second most commonly used approaches are neural network approaches. V.Vapnik argued that in order to solve a problem, one should not try to solve a more general problem as an intermediate step. The estimation of the complete density instead of computing the boundary around a data set might require too much data and could result in bad descriptions. So in order to get the boundaries of a data set, neural network can be used. Unfortunately these methods inherit the weak points in neural network training, i.e. the choice of the size of the network, weight initialization, the stopping criterion, etc. Therefore, the major issue of this type of approaches is the avoidance of the local optimal during the training of the neural network.

In this paper we focus on establishing the context of support vector clustering approach for outlier detection. Just like the Support vector machine(SVM), the support vector clustering (SVC) also maps the data to a high dimension feature space by a kernel function [2]; but the SVC looks for a smallest sphere in the feature space that encloses the image of the data, then this sphere is mapped back to the data space. As a result, a convex quadratic program problem is needed to derive the SVC model. So compared to neural network, the undesirable local minima of the training processes can be avoided easily.

As we known, the SVC has been proposed for unsupervised clustering in [2] and data description in [3][5]. In this paper we will use the SVC algorithm for solving outlier detection problems. Then we will compare the performance of different algorithms.

The rest of this paper is organized as follows. In section 2 we review some basic notions of SVC and introduce how to use SVC to do outlier detection problems. The simulate results are given in section 3 and finally in section 4 we will draw the conclusions.

Published by Atlantis Press, Paris, France.

II. SUPPORT VECTOR CLUSTERING

Let $\{\mathbf{X}_i\}$ be a data set of N points in the data space R^d , SVC finds a sphere of radius R and center $\mathbf{a}$ containing all these data. To get the smallest such sphere, we can define the error function to minimize:

$$\min J(R,\mathbf{a}) = R^2 \qquad (1)$$

$$\text{s.t. } \|\mathbf{x}_i - \mathbf{a}\|^2 \le R^2, \forall i$$

To allow the possibility of outliers in the training set, the distance from $\mathbf{X}_i$ to the center $\mathbf{a}$ should not be strictly smaller than R^2 , but larger distances should be penalized. Therefore we introduce slack variables $\zeta_i \ge 0$ and the minimization problem changes into:

$$\min J(R,\mathbf{a}) = R^2 + C\sum_{i=1}^{N}\zeta_i \qquad (2)$$

$$\text{s.t. } \|\mathbf{x}_i - \mathbf{a}\|^2 \le R^2 + \zeta_i, \zeta_i \ge 0, \forall i$$

The parameter C controls the trade-off between the volume and the errors.The optimal solutions of Eq.(2) can be achieved by a variational calculation applied to the Lagrangian:

$$L(R,\mathbf{a},\alpha_i,\gamma_i,\zeta_i) = R^2 + C\sum_{i=1}^{N}\zeta_i - \sum_{i=1}^{N}\alpha_i\left\{R^2 + \zeta_i - \left(\|\mathbf{x}_i\|^2 - 2\mathbf{a}\cdot\mathbf{x}_i + \mathbf{a}^2\right)\right\} - \sum_{i=1}^{N}\gamma_i\zeta_i \qquad (3)$$

with the Lagrange multipliers $\alpha_i \ge 0, \gamma_i \ge 0$

Setting partial derivatives to zero gives the constraints:

$$\frac{\partial L}{\partial R} = 0 \to \sum_{i=1}^{N}\alpha_i = 1 \qquad (4)$$

$$\frac{\partial L}{\partial \mathbf{a}} = 0 \to \mathbf{a} = \frac{\sum_{i=1}^{N}\alpha_i\mathbf{x}_i}{\sum_{i=1}^{N}\alpha_i} = \sum_{i=1}^{N}\alpha_i\mathbf{x}_i \qquad (5)$$

$$\frac{\partial L}{\partial \zeta_i} = 0 \to C - \alpha_i - \gamma_i = 0 \qquad (6)$$

To Eq.(), because $\alpha_i \ge 0, \gamma_i \ge 0$, Lagrange multipliers γ_i can be removed when we demand that:

$$0 \le \alpha_i \le C \qquad (7)$$

Resubstituting (4)–(6) into (3) results in:

$$L = \sum_{i=1}^{N}\alpha_i(\mathbf{x}_i\cdot\mathbf{x}_i) - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j(\mathbf{x}_i\cdot\mathbf{x}_j) \text{。} \qquad (8)$$

subject to constraints (4) and (7). Maximizing (8) gives a set α_i .

To test an object $\mathbf{x}$, the distance to the center of the sphere has to be calculated:

$$\|\mathbf{x}-\mathbf{a}\|^2 = (\mathbf{x}\cdot\mathbf{x}) - 2\sum_{i=1}^{N}\alpha_i(\mathbf{x}\cdot\mathbf{x}_i) + \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j(\mathbf{x}_i\cdot\mathbf{x}_i) \le R^2 \quad (9)$$

When instead of the rigid hypersphere a more flexible clustering boundary is required, just as SVM, we can replace the inner product $(\mathbf{x}\cdot\mathbf{x})$ by a kernel function $K(\mathbf{x}_i\cdot\mathbf{x}_i) = \varphi(\mathbf{x}_i)\cdot\varphi(\mathbf{x}_i)$,where φ is a function mapping the data into another (possibly high dimensional) feature space. As a result, Eq.(8) can be rewrite as:

$$L = \sum_{i=1}^{N}\alpha_i K(\mathbf{x}_i\cdot\mathbf{x}_i) - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j K(\mathbf{x}_i\cdot\mathbf{x}_i) \qquad (10)$$

The center of the hypersphere in feature space can be expressed as follows:

$$\mathbf{a} = \sum_{i=1}^{N}\alpha_i\varphi(\mathbf{x}_i) \qquad （11）$$

and the distance of the test object to can be calculated with following equation:

$$\|\mathbf{x}-\mathbf{a}\|^2 = (\mathbf{x}\cdot\mathbf{x}) - 2\sum_{i=1}^{N}\alpha_i(\mathbf{x}\cdot\mathbf{x}_i) + \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j(\mathbf{x}_i\cdot\mathbf{x}_i) \le R^2 \quad (12).$$

From the above description it can be known that the training samples are divvied into 3 types by SVC:

1) $\mathbf{x}_i$ lie in the boundary. In this case $\alpha_i = 0$ and $\mathbf{x}_i$ is regarded belonging to target set;

2) $\mathbf{x}_i$ lie on the boundary. In this case $\alpha_i = 0$ and $\mathbf{x}_i$ is SV. This type of instances is also regarded belonging to target set;

3) $\mathbf{x}_i$ lie out the boundary. In this case $\alpha_i > 0$ and $\mathbf{x}_i$ is regarded is outlier data;

Figure.1 shows the experimental results of outlier detection using SVC. The data used in this figure is the same as that in figure but several outliers are added. In this figure points A, B and C are regarded as outliers for their lie out the clustering boundary.

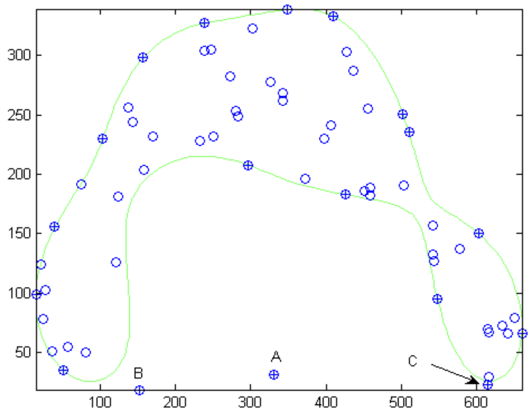

Fig.1 the outlier detection using SVC

III. EXPERIMENTAL RESULTS

The IRIS data set is used to verify our proposed approaches of outlier detection. In our experiments two outlier detection methods, i.e., SVC and NND (Nearest Neighbor Data Description) [4]are used.

The IRIS data set is one of the well known databases existing in pattern recognition literature. According to [7], the original data consists of three species of plants named as IRIS, which are setosa, versicolor and virginica, respectively. This data set respectively measured in their sepal lengths, sepal widths, petal lengths and petal widths, includes 150 samples (each species contains 50 samples) with the dimension size of 4. Moreover, the setosa is obviously linearly separable from the other two and the remaining two species,i.e., versicolor and virginica, have significant overlapping.

In our experiments the instances of versicolor are regarded as target class and the instances from setosa and virginica are selected randomly to serve as outliers. Accordingly the training data set is composed of all the instances from versicolor and the instances selected randomly from setosa and virginica. All the experiments are conducted by averaging 10 experimental results, so they are authentic.

The experimental results are shown in Table 1. From this table it can be found that performance of SVC algorithms is better than NDD algorithm.

Table 1 The experimental results of outlier detection using different algorithms

Outliers	# of outlier that is correctly detected		# of object that is correctly classified	
	SVC	NDD	SVC	NDD
2	1.8	1.5	49.7	44
4	2.6	2	48.2	44
6	5.5	1.4	45.3	44
8	7.1	1.8	46.1	44
10	8	1.2	44.2	44

IV. CONCLUSION

In this paper a novel Support vector clustering(SVC) method for outlier detection is proposed. Outlier detection algorithms have application in several tasks such as data mining, data preprocessing, data filter-cleaner, time series analysis and so on. Traditionally outlier detection methods are mostly based on modeling data based on its statistical properties and these approaches are only preferred when large scale set is available. To solve this problem, in this paper we focus on establishing the context of support vector clustering approach for outlier detection. Compared to traditional outlier detection methods , the performance of the SVC is not sensitive to the selection of needed parameters. The experiment results proved the efficiency of our method..

ACKNOWLEDGMENT

The work of this paper has been supported by the National Natural Science Foundation of China (No 91024008).

REFERENCES

[1] David, “Clustering via Hilbert Space”, Physica A, 302(2001),pp70-79.

[2] M.J.Tax Dvaid, “Support Vector Data Description”, Machine Learning ,,54,45-66,2004

[3] B.Scholkopf ,A.J.Smolad, “New Support Vector Algorithmsm”, Technical Report ,GMD first and Australian National University, 1998

[4] E.Knorr, V.Tuckkov, "Distance-based outliers: algorithm and applications," VLDB journal,8(3):pp.237-253,2000.

[5] B. Sch¨olkopf_ , R.Williamson_ ,Alex Smola_ , John Shawe-Taylor , J.Platt " Support Vector Method for Novelty Detection" S.A. Solla, T.K. Leen and K.-R. M¨ uller (eds.), 582–588MIT Press, 2000.

[6] J.Han, M.Kamber. Data Mining : Concepts and Techniques. Morgan Kaufmann Publishers.Inc, 2001.

[7] C. Blake, UCI repository of machine learning databases, Department of Information and Computer Science, University of California, Irvine, CA, http://www.ics.uci.edu/~mlearn/MLRepository.html

Published by Atlantis Press, Paris, France.