Support Vector Machine (SVM) in the anomaly and outlier detection

Student: Ulvi Bajarani
Student ID: 20539914
e-mail: ulvi.bajarani01@utrgv.edu
Class: CSCI 6366 - Data Mining and Warehousing
Semester: Spring 2021

April 5, 2021

Abstract

Support Vector Machine (SVM) algorithms is one of the powerful algorithms to classify the items, which also makes it useful in the anomaly and outlier detection. However, as other algorithms, it has its own advantages and disadvantages. In this term paper, the author provides the algorithms, and the available extensions of SVM, and their advantages and disadvantages. In the final section, the author provides his own method.

1 Introduction to Support Vector Machine (SVM) methods

1.1 The basic idea of the methods

The idea behind the Support Vector Machine methods is to transform training data $\mathbf{X} = \{x_1, x_2, ..., x_n\}$ to the high dimensional space by the function $\phi(\mathbf{X})$ such that the evaluation of the data members could be done by some kernel $\kappa(\mathbf{X}, x_i) = \phi(\mathbf{X}) \cdot \phi(x_i)$ where $i \leq n$. Despite the existence of various kernels, this is usually done by the Gaussian kernel:

$$\kappa(\mathbf{X}, x_i) = e^{-\frac{||\mathbf{X}-x_i||^2}{2\sigma^2}} \tag{1}$$

For the hyperplane $\mathbf{w}^T \cdot \mathbf{X} + b = 0$ and the classes $y_i \in \{-1, 1\}$ finding the optimal plane problem might be described as the Quadratic Program Problem[1]:

$$\min_{\mathbf{w},\ \xi} \frac{||\mathbf{w}||^2}{2} - C \sum_{i=1}^{n} \xi_i \tag{2}$$

$$\textbf{subject to } y_i((\mathbf{w}^T \phi(x_i + b)) \geq 1 - \xi_i \tag{3}$$

Here, $\mathbf{w}$ is the normal vector of $\phi(\mathbf{X})$, b is the bias, n is the number of training instances, C is the smoothness constant deciding the training error (compromise between the number of training data within the margin and the margin maximisation) with $C > 0$, ξ is the slackness variable with $\xi \geq 0$

With Lagrange multipliers α_i, the decision about the data is done by the formula below

$$f(x) = \textbf{sign}(\sum_{i=1}^{n} \kappa(\mathbf{X}, x_i)\alpha_i y_i + b) \tag{4}$$

1.2 Available implementations

- The first and the widespread implementation of SVM is One-Class SVM. This method separates the normal data from the outliers. According to [2], the optimization problem is described as

$$\min_{\mathbf{w},\ \rho,\ \xi} \frac{||\mathbf{w}||^2}{2} - \rho + \frac{1}{n\nu} \sum_{i=1}^{n} \xi_i \tag{5}$$

$$\textbf{subject to } (\mathbf{w} \cdot \phi(x_i)) \geq \rho - \xi \tag{6}$$

Here, $\rho = \mathbf{w} \cdot \phi(\mathbf{X})$, ν is the hyper-parameter with $\nu \in (0, 1]$. With Lagrange multipliers a_i, $\mathbf{w}$ could be calculated by

$$\mathbf{w} = \sum_{i=1}^{n} a_i \phi(x_i) \tag{7}$$

While the original implementation finds the plane, it is also possible to find the sphere [3][4] and quarter-sphere [5] where the normal values are inside of the sphere. The idea behind them is to minimize the sphere size with radius R and center a:

$$\mathbf{min}\ R^2 + C \sum_{i=1}^{n} \xi_i \tag{8}$$

$$\mathbf{subject\ to}\ x_i^{'} - a^{'} \leq R^2 + \xi_i \tag{9}$$

Here, $x^{'}$ and $a^{'}$ are normalized vectors.

- In the article [6], Hao successfully managed to implement Fuzzy One-Class SVM with the μ_i fuzzy membership coefficient by changing $C \sum_{i=1}^{n} \xi_i$ to $C \sum_{i=1}^{n} \mu_i \xi_i$ in Equation 2
- The successful combination of Bayesian Networks with One-Class SVM is also implemented [7]. The idea behind it is the calculation the posterior probability of the data constructed by One-class SVM with class labels.
- There are successful usage of ensemble SVM methods using SVM classifiers. While [8] used Simplified Mahalanobis Distance, [9] used clustering based approach.

1.3 Advantages and disadvantages

The SVM methods has its benefits and drawbacks[10][11]. The advantages of SVM methods:

- Comparing to other methods, SVM methods are effective in the high-dimensional spaces.
- SVM can handle non-linear data effectively.
- SVM methods are robust against noise data.
- SVM methods don't stuck on local optimum. Moreover, with optimal parameters, SVM methods provides the unique solution.

However, SVM methods also have the disadvantages:

- The methods are not suitable for the huge datasets, since the computationally costly.
- The results of methods are highly dependent on the hyper-parameters, especially for the one-class SVM that is dependent on ν.
- Number of the dimensions for each data point should be lower than the number of training samples, because it leads to underfitting. This could be solved by dimensionality reduction techniques such as PCA. [12]
- One-Class SVM methods are sensitive to both the training data and kernel choice.

2 The method of the author

Here, the author provides the method involving several iteration of the data and defining the best results:

1. Set clusters = [ ]
2. Set initial_outliers = [ ];
3. Set final_outliers = [ ];
4. For 1 to number of iterations:
5. Cluster the initial data into k clusters (For one-class SVM, k=2 is suitable);
6. Add each cluster to clusters.
7. For cluster in clusters:
8. Execute the One-Class SVM;
9. For data-point in cluster:
10. If the data is outlier:
11. Add data-point to the initial_outliers;
12. Count outlier-points in initial_outliers:
13. If count > threshold:
14. Add outlier-point to the final_outliers;
15. return final_outliers;

Such an approach has several advantages:

- Since the K-means is sensitive to the centroids, one iteration might provide the normal points as the outliers. However, if one point is considered as outlier again and again, it is possible to take it as the outlier. In other words, the accuracy increases;
- Each clusters are analysed separately, so the influence of outliers with high density reduces;
- It doesn't require the training data;
- Unlike ensemble methods, each cluster are evaluated with the same classifier.

However, it also has the disadvantages:

- It is computationally expensive, so it is not suitable for large data sets. The computational time might be reduced using appropriate data structures;
- K-means are sensitive for the centroid choice. This might be handled using both several iterations (as in the algorithm) and K-means++;
- The threshold should be chosen carefully.

References

[1] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[2] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[3] David MJ Tax and Robert PW Duin. Data domain description using support vectors. In *ESANN*, volume 99, pages 251–256. Citeseer, 1999.

[4] David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54(1):45–66, 2004.

[5] Pavel Laskov, Christin Schäfer, Igor Kotenko, and K-R Müller. Intrusion detection in unlabeled data with quarter-sphere support vector machines. *PIK-praxis der Informationsverarbeitung und Kommunikation*, 27(4):228–236, 2004.

[6] Pei-Yi Hao. Fuzzy one-class support vector machines. *Fuzzy Sets and Systems*, 159(18):2317–2336, 2008.

[7] Vasilis A Sotiris, W Tse Peter, and Michael G Pecht. Anomaly detection through a bayesian support vector machine. *IEEE Transactions on Reliability*, 59(2):277–286, 2010.

[8] Roberto Perdisci, Guofei Gu, and Wenke Lee. Using an ensemble of one-class svm classifiers to harden payload-based anomaly detection systems. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 488–498. IEEE, 2006.

[9] Jungsuk Song, Hiroki Takakura, Yasuo Okabe, and Yongjin Kwon. Unsupervised anomaly detection based on clustering and multiple one-class svm. *IEICE transactions on communications*, 92(6):1981–1990, 2009.

[10] Sasan Karamizadeh, Shahidan M Abdullah, Mehran Halimi, Jafar Shayan, and Mohammad javad Rajabi. Advantage and drawback of support vector machine functionality. In *2014 international conference on computer, communications, and control technology (I4CT)*, pages 63–65. IEEE, 2014.

[11] Dhiraj K. Top 4 advantages and disadvantages of Support Vector Machine or SVM. `https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107`, 2019. [Online; accessed 13-Jun-2019].

[12] Li Zhanchun, Li Zhitang, and Liu Bin. Anomaly detection system based on principal component analysis and support vector machine. *WuHan University Journal of Natural Sciences*, 11(6):1769–1772, 2006.