Indicative Support Vector Clustering with its Application on Anomaly Detection

Huang Xiao
Chair of IT Security
Computer Science Department
Technical University of Munich
Boltzmannstr.3 Garching (Germany)
Email: xiaohu@in.tum.de

Claudia Eckert
Chair of IT Security
Computer Science Department
Technical University of Munich
Boltzmannstr.3 Garching (Germany)
Email: claudia.eckert@in.tum.de

***Abstract*—In many learning scenarios, supervised learning is hardly applicable due to the unavailability of a complete set of data labels, while unsupervised model overlooks valuable user feedback in an interactive system setting. In this paper, a novel semi-supervised support vector clustering algorithm is presented, where a small number of user indicated labels are available as supervised information. We apply the clustering algorithm in the anomaly detection area, and show that the given labels significantly improve the recognition of anomalies. Moreover, the partially labeled data proliferates the information without extra computation but strengthening the robustness to anomalies.**

I. Introduction

In recent years, kernel and spectral clustering methods [1] have invoked immense interest of researchers due to its non-parametric characteristics. Ben-Hur et.al [2] developed the support vector clustering (SVC) algorithm that discovers the smallest sphere in the feature space enclosing all the data points. With a delicate selection of parameters, the support vector clustering can naturally separate data samples into various classes. This support vector descriptive model shares the common core idea with the one-class support vector machine [3]. Unfortunately, the SVC has not yet been adapted to the semi-supervised mode. In many systems, we observe that some feedback can be contributed by users at a very low expense, which encourages the raise of the semi-supervised learning [4]. Especially in anomaly detection, it is highly expected that a least effort from users could improve the performance at a significant amount. He et al. [5] introduces the semantic anomaly factor to measure the deviation an outlier behaves from the majority of its cluster members by inquiring their labels. Another well studied semi-supervised learning method is the semi-supervised support vector machine [6], which prevents the decision boundary from passing through high density area. A graph based semi-supervised method leverages the non-negative matrix factorization [7] to cluster data by minimizing distances of same labeled samples while maximizing distances of different labeled samples.

In this paper, we present the semi-supervised version of support vector clustering, named as indicative support vector clustering (iSVC). Given a limited number of binary labels as normal or abnormal, the support vector clustering algorithm can be guided to produce a more reliable and accurate boundary separating normal instances from anomalies. The iSVC reweighs part of the input data points utilizing the supervised information given by users, and then constructs the boundary alleviating the impact of outliers on the hypersphere. To our knowledge, this is the first semi-supervised support vector clustering method.

II. Support Vector Clustering

Given a data set of N points $\{x_i\}_{i=1}^N \subseteq \mathcal{X}$, it forms a d-dimensional real-valued input space with $\mathcal{X} \subseteq \mathbb{R}^d$. The support vector clustering looks for a smallest sphere $\mathcal{S} = \langle R, g \rangle$ in $\mathcal{H}$ enclosing all the mapped data points, where R is the radius of the sphere and g is the center in $\mathcal{H}$. It is formulated as Tikhonov regularization problem, and the Wolfe dual form to it is as follows:

$$W = \sum_j K(x_j, x_j)\beta_j - \sum_{i,j} \beta_j \beta_j K(x_i, x_j),$$

where β_i is the Lagrangian multiplier and $K(\cdot, \cdot)$ is the kernel function. Solving this quadratic problem, we obtain the variables $\{\beta_j\}$ assigning the input data into three disjoint sets $\{$SVs, BSVs, Insiders$\}$ [2]. And the cluster boundary is determined by the radius R and mapped back to the input space to form the corresponding contour.

Cluster assignment

The SVC algorithm searches the smallest sphere enclosing all the points, but not classifies each point into its containing cluster. To assign the cluster labels, it requires a further geometrical analysis of the input data. However, in anomaly detection, the cluster labeling process is not a necessity when anomalies are already identified. This characterizes the SVC as an ideal choice for anomaly detection. Throughout this paper, we neglect the cluster assignment of SVC, therefore convert it to an one-class classifier separating outliers from normal instances.

III. SVC Algorithm on Anomaly Detection

The SVC lends its advantages to anomaly detection when it becomes an one-class classifier. It forms an enclosed hypersphere separating the data into three sets, or generally

speaking, two sets: $\{\mathcal{X}^+, \mathcal{X}^-\}$. We denote the positive set $\mathcal{X}^+$ for anomalies and the negative set $\mathcal{X}^-$ for normalities. The bounded support vectors (BSVs) correspond the set $\mathcal{X}^+$ which are far away from the sphere center, and all the other points belong to the set $\mathcal{X}^-$. According to the properties of SVC, the number of anomalies is upper bounded by $1/C$, where C is the penalty coefficient to the regularization problem. That is, the probability of anomalies in the input data P_a is bounded by $\frac{1}{NC}$, Therefore, the penalty C and the sample size N determine a tolerance level for the outliers in the input data. Besides, the distance of an outlier x_l to the center g can be computed as:

$$d_{x_l} = K(x_l, x_l) - 2\sum \beta_j K(x_l, x_j) + \sum \beta_i \beta_j K(x_i, x_j)$$

For a hypersphere $\mathcal{S} = \langle R, g \rangle$ obtained by SVC, the third term in d_{x_l} is fixed, and for Gaussian kernel $K(x_l, x_l) = 1$. When the outlier x_l is distant from the center g, it implies a small value for $\sum \beta_j K(x_l, x_j)$, which is a weighted average of the Gaussian kernels on all the samples. When all the points are regarded as anomalies, i.e., $P_a = 1$ and $C = 1/N$, we have an approximated Parzen Window Estimate,

$$P(x_l) = \frac{1}{N}\sum K(x_l, x_j)$$

Thus, given a SVC hypersphere $\mathcal{S} = \langle R, g \rangle$, the further a point lies away from the center g, the lower its density estimate could be in comparison with others. This forms the fundamental idea of using SVC for anomaly detection.

IV. Indicative Support Vector Clustering

A study in Section III reveals two problems of the SVC in anomaly detection applications. First, the sparsity of anomalies does not always hold. Most anomaly detection algorithms are not employable, if the anomalies form cluster themselves. Second, the existence of anomalies will tamper with the center and radius of the hypersphere, the robustness of SVC algorithm is thus questionable when the negative impact of the outliers is overlooked. In this section, we propose a novel semi-supervised version of the SVC algorithm by integrating user given labels.

A. Weighted Regularization for Robustness

The SVC generalizes the solution as a smallest sphere in the Hilbert space $\mathcal{H}$ by allowing a portion of points to be ruled out, but simultaneously penalizing their outlying. In accordance with the properties of SVC, the optimal center g is a weighted mean of all the SVs and BSVs under feature mapping Φ. Suppose the Lagrangian multipliers $\{\beta_j\}$ can be fragmented as $\{\beta_{\mathrm{sv}}, \beta_{\mathrm{bsv}}, \beta_{\mathrm{in}}\}$ in corresponds to $\{$SVs, BSVs, Insiders$\}$ respectively:

$$g = \sum \beta_{\mathrm{sv}} \Phi(x_{\mathrm{sv}}) + \sum \beta_{\mathrm{bsv}} \Phi(x_{\mathrm{bsv}}) + \sum \beta_{\mathrm{in}} \Phi(x_{\mathrm{in}}),$$

where $0 < \beta_{\mathrm{sv}} < C$, $\beta_{\mathrm{bsv}} = C$ and $\beta_{\mathrm{in}} = 0$. The bounded support vectors contribute even more than the support vectors on positioning the center. However, in anomaly detection or other noise-aware applications, BSVs are recognized as outliers or noises, to which a robust learning model should be more resilient. For robustness, a weighted regularization term to the original SVC objective function is proposed, the Lagrangian is now reformed as:

$$\begin{aligned} L = R^2 &- \sum \left(R^2 + \xi_j - \|\Phi(x_j) - g\|^2\right) \beta_j \\ &- \sum \xi_j \mu_j + \sum c_j \xi_j, \end{aligned} \tag{1}$$

Instead of equivalently penalizing all the input data with a constant C, the weighted regularization term $\sum c_j \xi_j$ in Eq. (1) treats each point individually. The objective is that, the larger the regularization coefficient c_j is, the less possible the point x_j would be driven away from the hypersphere, and vice versa. In this way, the BSVs (anomalies) should have lower values of β_j, and the SVs or insiders are optimized with higher values. Consequently, we alleviate the negative effect of the BSVs on the formation of the hypersphere $\mathcal{S}$, and the SVs are the main factors leading to the decision boundary.

B. Integration of Indicative Labels

Assumption 1: Anomalies are the patterns found to behave distinctly from the normal patterns, and similarly behaving instances are more likely hosted in the same cluster.

To obtain the regularization weights $\{c_j\}_{j=1}^N$, the supervised information from user given labels can be integrated in addition to the input data based on the assumption 1, that is, similar data patterns are more closely located to each other. Given a data set $\mathcal{X} \subseteq \mathbb{R}^d$, two supervised data sets are indicated by users,

$$\begin{aligned} \mathcal{X}^+ &= \{(x_l, y_l) \,|\, x_l \in \mathcal{X}, y_l = 1)\} \\ \mathcal{X}^- &= \{(x_r, y_r) \,|\, x_r \in \mathcal{X}, y_r = -1)\} \\ & l, r \in \{1, \ldots, N\} \end{aligned}$$

$\mathcal{X}^+$ is a labeled subset of $\mathcal{X}$ with only anomalies, while $\mathcal{X}^-$ is likewise a labeled subset with normal samples. On the Assumption 1, a given label of an instance indicates the similarity of its neighborhood and can broadcast the supervised information over its neighbors. More precisely, a given anomaly increases the probabilities of its neighborhood of being anomalies, while a given normal instance implies that its neighborhood are more confident of being normal.

To leverage the favorable supervised information to obtain the regularization weights, an impact function f is defined in the input space $\mathcal{X}$ given the labeled sets $\mathcal{X}^+$ and $\mathcal{X}^-$.

$$f(x_i) = \frac{\sum_{x_l \in \mathcal{X}^+} K(x_l, x_i)}{\sum_l \mathbf{1}^+(x_l, x_i)} + \frac{\sum_{x_r \in \mathcal{X}^-} K(x_r, x_i)}{\sum_l \mathbf{1}^-(x_r, x_i)} \tag{2}$$

And $\mathbf{1}^+$ and $\mathbf{1}^-$ are both indicator functions defined as

$$\mathbf{1}^+(x_l, x_i) = \begin{cases} 1 & \|x_l - x_i\| \le 2h \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{1}^-(x_r, x_i) = \begin{cases} -1 & \|x_r - x_i\| \le 2h \\ 0 & \text{otherwise} \end{cases}$$

We define the radius of the affected neighborhood as $2h$ for simplicity, which means that the impact of the given labeled

data is bounded. The bounds of the indicator functions $\mathbf{1}^+$ and $\mathbf{1}^-$ can also be configured individually so that we have flexibility on controlling the degree of impact by anomalies or normal samples. Besides, the similarity measurement $K(x, x_i)$ follows a Gaussian distribution $\mathcal{N}(x, h)$ with respect to a normalization factor. Therefore, the impact function $f(x_i)$ represents the probability of x_i being an anomaly affected by $\mathcal{X}^+$ or a normal instance affected by $\mathcal{X}^-$.

By use of the impact function $f(x_j)$, the regularization weights c_j can be computed

$$c_j = \begin{cases} c_0 \cdot \frac{1-f(x_j)}{1-\exp(-2)} + \frac{1}{N} \cdot \frac{f(x_j)-\exp(-2)}{1-\exp(-2)} & \text{if} \quad f(x_j) > 0 \\ c_0 \cdot \frac{1-|f(x_j)|}{1-\exp(-2)} + \frac{|f(x_j)|-\exp(-2)}{1-\exp(-2)} & \text{if} \quad f(x_j) < 0 \end{cases} \tag{3}$$

where c_0 is the initial value of c_j and N is the sample size. Together with the constraints on the labeled samples, the Wolfe dual form becomes:

$$\begin{aligned} W = \sum_j K(x_j,x_j)\beta_j - \sum_{i,j} \beta_j \beta_j K(x_i, x_j) \qquad (4) \\ \text{subject to} \quad 0 \le \beta_j \le c_j \\ \beta_{\mathcal{X}^+} = c_{\mathcal{X}^+} \\ 0 \le \beta_{\mathcal{X}^-} < c_{\mathcal{X}^-} \end{aligned}$$

The weighted regularization term and the additional constraints do not change the convexity. Solving this dual problem by maximizing W, we obtain the robust SVC boundary providing a more reliable separation of anomalies from normal instances. Note that we only need to compute the impact function for each point, the additional complexity is then $\mathcal{O}(N)$. Nevertheless, the problem can still be solved in polynomial time without noticeable computational burden.

V. Experiments

In this section, we evaluated the iSVC algorithm both on synthetic data and real-world data. The required parameters for iSVC involve the kernel bandwidth h, an initial penalty constant c_0 and user given positive and negative label sets that correspond to anomalies and normal instances respectively. The bandwidth h controls the smoothness of the cluster boundary and also the number of the support vectors. For simplicity, we assume an optimal h without explicitly evaluating it in experiments. Note that this can be done by the cross-validation. For the initial value of c_0, a good estimate could be derived from the proportion of the outliers in input data set, namely, $c_0 = 1/n_{bsv}$, where n_{bsv} is the number of outliers in data set. In the end, only a small number of labels are supposed to be given by users and we show that the involvement of the supervised information indeed improves the performance of support vector clustering significantly on anomaly detection.

A. On Synthetic Data

To illustrate the mechanism and capability of iSVC algorithm, experiments were conducted firstly on a 2-d synthetic data set. It contains two normal classes of compacted Gaussians, each of which has 250 samples. Another set of 150 anomaly samples forms a sparse ring encircling the normal instances. The experiments were performed in three distinct configurations of given labels.

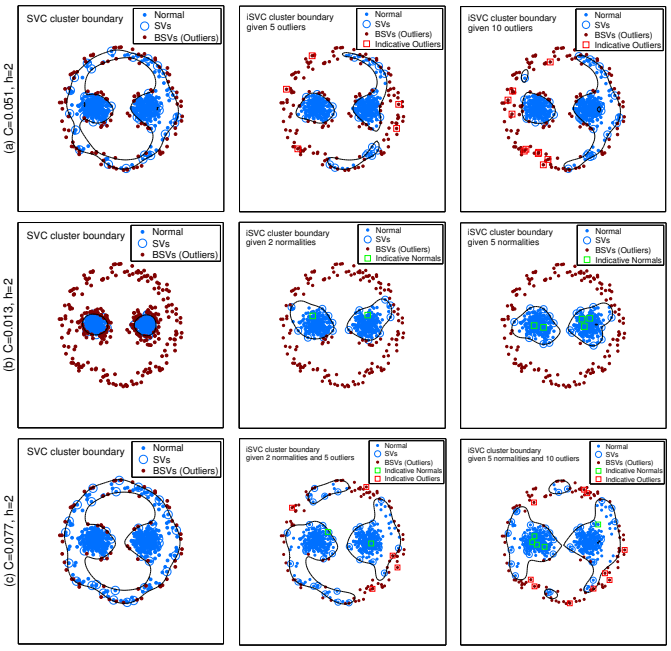


Fig. 1: Comparisons of SVC and iSVC by giving different configurations of sample labels on the synthetic data set.

In Figure 1, the iSVC clustering results given different label sets are shown in a 3x3 grid. Each row has the same bandwidth h and the initial c_0, and the first column of the results gives the cluster boundaries of original SVC algorithm on the synthetic data. In Figure 1a, a moderate value of C only discloses a minority of the actual anomalies, while a handful of normal instances are also recognized as anomalies due to their low densities. Clearly, the SVC algorithm fails when the densities of data samples are not explicitly separable between normalities and anomalies. In the middle of the first row, 5 given anomalies are evenly distributed along the ring. We observe that more anomalies are discovered under the supervised information. On the rightmost, more anomalies up to 10 are given, however, part of which are redundantly located on the left half ring. Consequently, the anomalies on the right half ring are completely unaffected. In the Figure 1b, a smaller c_0 produces an excessive discovery of anomalies with a significant false positive. Since the normal instances are more compacted in the center, supervised information can be passed on its neighborhood more efficiently. On the rightmost of the second row, only 5 given normal instances produce a perfect separation of the anomalies from the normal samples. In Figure 1c, it does not present a promising cluster boundary for anomalies, however, the normal instances are mostly detected under a small number of given normal labels. Additional anomaly labels are required to achieve a perfect separation.

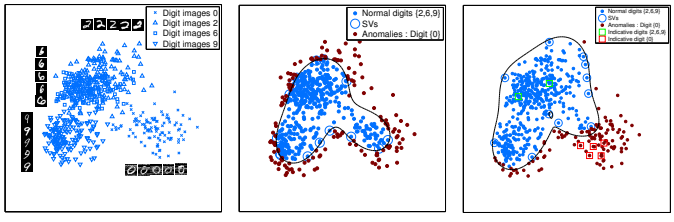
B. On Real-world Data

The evaluation of the iSVC algorithm was then performed on real-world data sets. Again, we assume an optimal kernel bandwidth h for each experiment session without explicitly estimating it. The experiments show that iSVC improves the SVC algorithm and outperforms other semi-supervised binary classifier, and also presents its potential to related practitioners.

MNIST Digit Images: We firstly selected four digit sets from the MNIST digit images data set [8]: $\{0, 2, 6, 9\}$. To fabricate a reliable anomaly-aware scenario, 100 images of digit $\{0\}$ were manually tampered with a couple of horizontal noisy lines. For digits $\{2, 6, 9\}$, we sampled 150 images out of each class as normal instances. For illustration, we ran the PCA first on the input data with respect to its first two principal components. In the lower dimension, the tampered digit images $\{0\}$ are relatively distant from other digits images. In Figure 2b, SVC algorithm generated several false positives and false negatives as well. Given 2 normalities and 5 anomalies in Figure 2c, iSVC gave a better estimate of the cluster boundary setting the anomaly digits $\{0\}$ apart from other normal digits.

On the WDBC Data: Last experiments were conducted on the *Wisconsin Diagnostic Breast Cancer (WDBC)* data set from UCI repository [9] to illustrate the empirical advantages of iSVC. The data set contains in total 569 samples, out of which 212 are malignant and 357 are benign. For comparison, we employed the semi-supervised fast linear SVM solver [6] as a binary classifier, denoted as S3VM. Note that S3VM requires an additional parameter indicating the ratio of outliers in the unlabeled data set, and this is normally not available in practice. To approximate a value of the ratio, we estimated it from the labeled data instead of the unlabeled data. Suppose the labeled data are actively sampled according to a certain distribution, this reflects the underlying ratio of positive samples. The results are shown in Table I.

We started with a very small value of penalty constant C which should identify excessive anomalies by iSVC. By given some labels, iSVC outperforms the original SVC algorithm and behaves much more stable than S3VM. Especially when imbalanced labels are given, e.g., only giving 5% positive or 5% negative labels will completely exterminate the function of S3VM.



(a) Original distribution (b) SVC cluster boundary (c) iSVC cluster boundary

Fig. 2: MNIST digit images of $\{0, 2, 6, 9\}$ are demonstrated on its first two principals under PCA. Both SVC and iSVC are conducted on the reduced data set with the parameters $h = 2.5$ and $C = 0.012$.

TABLE I: Empirical results on the WDBC data set given different proportions of sample labels.

	WDBC data set with $\mathbf{h = 0.9, C = 0.001}$			
	Accuracy	F_1-measure	FPR	FNR
SVC	63.4%	50.5%	28.6%	50%
	5% *positive* and 0% *negative* labels			
iSVC	**90.5%**	**87.0%**	**6.4%**	14.6%
S3VM	39.2%	55.0%	96.9%	**0%**
	0% *positive* and 5% *negative* labels			
iSVC	**89.8%**	**87.7%**	14.8%	2.4%
S3VM	64.9%	10.7%	**0%**	94.3%
	5% *positive* and 5% *negative* labels			
iSVC	**92.3%**	**89.8%**	**7.0%**	8.9%
S3VM	88.4%	86.3%	17.4%	**1.9%**
	10% *positive* and 10% *negative* labels			
iSVC	**93.9%**	**91.9%**	**6.4%**	5.6%
S3VM	92.4%	90.6%	10.4%	**2.8%**

VI. Conclusion

The iSVC algorithm is proposed to circumvent the inherent deficiencies of SVC of integrating additional supervised information. Given a small number of sample labels, the iSVC exhibits its prominent performance on anomaly detection, more generally, on semi-supervised binary classification problem. In future, we will extend our work on the automatic estimate of an optimal bandwidth and also on the active selection of data samples, to which users are supposed to give labels, such that a minimal expenditure on the labeling process can improve the performance at the most.

Acknowledgment

This paper is funded by the project ARAMiS of BMBF[1] in Germany.

References

[1] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means: spectral clustering and normalized cuts," ser. KDD '04. New York, NY, USA: ACM, 2004, pp. 551–556.

[2] A. Ben-Hur, D. Horn, H. Siegelmann, and V. Vapnik, "A support vector clustering method," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, vol. 2, 2000, pp. 724–727 vol.2.

[3] B. Schlkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," 2000.

[4] X. Zhu, "Semi-supervised learning literature survey," 2006.

[5] Z. He, S. Deng, and X. Xu, "Outlier detection integrating semantic knowledge," in *Advances in Web-Age Information Management*, ser. LNCS. Springer Berlin Heidelberg, 2002, vol. 2419, pp. 126–131.

[6] V. Sindhwani and S. S. Keerthi, "Large scale semi-supervised linear svms," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '06. New York, NY, USA: ACM, 2006, pp. 477–484.

[7] N. Guan, X. Huang, L. Lan, Z. Luo, and X. Zhang, "Graph based semi-supervised non-negative matrix factorization for document clustering," in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, vol. 1, 2012, pp. 404–408.

[8] Y. Lecun and C. Cortes, "The mnist database of handwritten digits."

[9] K. Bache and M. Lichman, "UCI machine learning repository," 2013.

[1] BMBF: Bundesministerium für Bildung und Forschung