Rare Class Classification by Support Vector Machine

He He
Department of Electronic and Information Engineering
The Hong Kong Polytechnic University
Hong Kong, China
07821020d@polyu.edu.hk

Ali Ghodsi
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, Canada
aghodsib@uwaterloo.ca

***Abstract*—The problem of classification on highly imbalanced datasets has been studied extensively in the literature. Most classifiers show significant deterioration in performance when dealing with skewed datasets. In this paper, we first examine the underlying reasons for SVM's deterioration on imbalanced datasets. We then propose two modifications for the soft margin SVM, where we change or add constraints to the optimization problem. The proposed methods are compared with regular SVM, cost-sensitive SVM and two re-sampling methods. Our experimental results demonstrate that this constrained SVM can consistently outperform the other associated methods.**

***Keywords*-Support vector machines; Classification; Novelty detection;**

I. Introduction

The problem of rare class classification (also known as outlier analysis, anomaly detection, etc.) arises in various areas including drug discovery, fraud detection and cancer diagnosis, where the target class is heavily under-represented in a sample. In real world applications, these datasets can often have an imbalance ratio ranging from 10:1 to 100:1. Usually the rare events or objects are of major interest and are very costly if misclassified.

Most classifiers perform poorly on imbalanced datasets since they are intended to maximize the accuracy. This results in a simple decision: classify all data points to the negative (common) class; this decision will still maintain high accuracy. The two main approaches to address this problem are to rebalance the dataset, or modify the algorithm. The first approach operates on the data level by either oversampling the minority class or undersampling the majority class. However, these methods make distributional assumptions about the data, and they could easily eliminate important samples and introduce noise. Alternatively, Chawla et al. [1] have developed a popular sampling method: the Synthetic Minority Oversampling Technique (SMOTE). It inserts new samples in between the identified neighbors and the instance to be oversampled. At the algorithm level, major efforts have been put on cost-sensitive learning, where a higher penalty is assigned to misclassified positive data points. This method indeed biases the classifier toward the positive class to improve the detection rate of the rare objects.

In this paper, we modify the constraints of the optimization problem of support vector machines (SVM) in two ways. For the first approach, we derive an extreme case of 2C-SVM [2]. The second approach adds one constraint to the soft margin SVM in order to forbid the positive data points from crossing the decision boundary.

The rest of the paper is organized as follows. Section 2 gives a brief overview of SVMs and investigates the reason for its degradation on imbalanced datasets. Section 3 presents the proposed modification of SVM and Section 4 shows the experimental results. Section 5 concludes the paper with a summary.

II. Background

A. Soft Margin SVM

Support Vector Machines are considered the state-of-the-art classification method. According to the Structural Risk Minimization Principle [3], SVM aims to find the best separating hyperplane of the data. This results in the following optimization problem, where x_i is the vector of the i-th training sample, y_i is its class label, β and α are hyperplane parameters, C is the regularization parameter, K represents the kernel function, and ξ_i is the slack variable which allows for some points to be on the wrong side of the margin:

Primal:

$$\begin{aligned} \min_{\beta,\beta_0,\xi_i} \quad & \frac{1}{2}\boldsymbol{\beta^T\beta} + C\sum_{i=1}^{n}\xi_i \\ s.t. \quad & y_i(\boldsymbol{\beta^T x_i} + \boldsymbol{\beta_0}) \geq 1 - \xi_i \quad \forall_i \\ & \xi_i \geq 0 \end{aligned}$$

Dual:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x_i}, \boldsymbol{x_j}) - \sum_{i=1}^{n}\alpha_i \\ s.t. \quad & \sum_{i=1}^{n}\alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned}$$

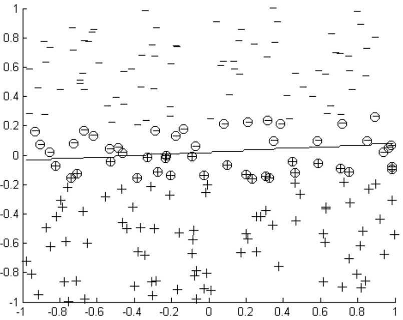

Figure 1. SVM on the dataset with balance ratio 1:1

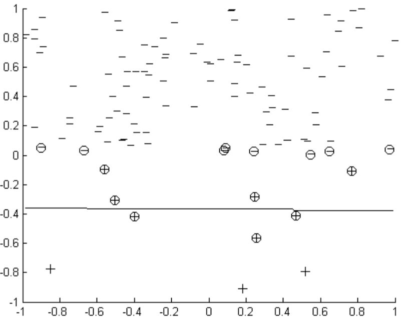

Figure 2. SVM on the dataset with imbalance ratio 10:1

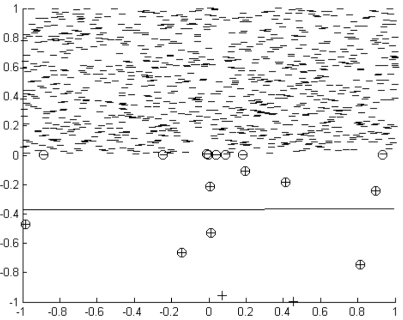

Figure 3. SVM on the dataset with imbalance ratio 100:1

B. SVM on Imbalanced Datasets

The fact that the SVM solution only depends on a few support vectors makes it relatively robust to noise and moderate imbalance. The simple experiment below gives some insight into SVM's behavior on skewed datasets. The positive sample is uniformly distributed between $y = -1$ and $y = 0$ and the negative samples are between $y = 1$ and $y = 0$. Obviously, the ideal boundary is the x axis. Figure 1 to Figure 3 show the classification boundary for datasets with increasing imbalance ratios. Support vectors are shown as circled points.

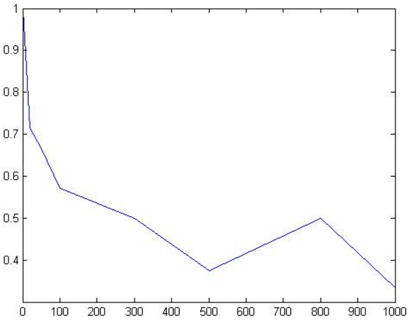

Figure 4. Ratio of number of support vectors vs. Imbalance ratio

Several observations from these experimental results are summarized as follows:

- When the distribution is highly skewed, the number of support vectors in the positive class is less than the number in the negative class.
- Examples from the positive class tend to reside farther from the real boundary than those from the negative class.
- As the imbalance becomes more severe, the predicted decision boundary is pressed towards the rare class.

The equality constraint $\sum_{i=1}^{n} \alpha_i y_i = 0$ indicates that if there are more negative samples than positive samples, i.e. if more y_i equal -1 than +1, then the positive class will have higher α_i values in order to guarantee a zero sum. In the decision function $y = sign(\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + \beta_0)$, α_i can be regarded as the weight of each example; thus larger α_i values essentially increase the influence of the minority class, which automatically rebalances the skewed dataset.

However, if the size of the negative class keeps growing, its support vectors can be much more than the positive class', making some α_i of the positive instances equals C. According to the complementary slackness, when $\alpha_i = C$, $y_i(\boldsymbol{\beta^T x_i} + \boldsymbol{\beta_0}) = 1 - \xi_i < 1$. This indicates that the data point enters the margin or even crosses the decision boundary, which explains why SVM has a degrading performance on highly imbalanced datasets.

C. 2C-SVM

In standard soft margin SVM, the penalty parameter C is the same for both classes. As a result, even if all the positive examples are ignored, this loss is still acceptable since there are so few of them. To address this problem, the cost-sensitive version of SVM is proposed in [2]. It essentially re-weights (denoted by C_+ and C_-) the examples to make the error from the rare class more obvious to the classifier. In the optimization problem, 2C-SVM has the penalty term $C\gamma \sum_{i \in I^+} \xi_i + C(1-\gamma) \sum_{i \in I^-} \xi_i$. To balance the dataset, γ is usually set to be the ratio of the number of negative instances to the number of positive instances.

III. Modification of SVM

A. Special Case of 2C-SVM

As mentioned before, in highly imbalanced datasets, positive examples end up entering the margin or crossing the decision boundary. Thus, we can intuitively change the constraint $0 \leq \alpha \leq C$ to $0 \leq \alpha < C$ for the rare class. Recalling the complementary slackness condition, for $\alpha_i = 0$ or $0 < \alpha_i < C$, we have $y_i(\boldsymbol{\beta^T x_i} + \boldsymbol{\beta_0}) \geq 1$. This is equivalent to the hard margin SVM constraint; thus the modification can be further illustrated as the 2C-SVM when $\gamma = 0$.

B. Constraint on the Slack Variable

In reality, misclassification of a positive example usually costs more than that of a negative example. To ensure that no positive example is left out, we add one constraint to the slack variables of the positive class and solve the following optimization problem:

Primal:

$$\begin{aligned}
\min_{\beta,\beta_0,\xi_i} \quad & \frac{1}{2}\boldsymbol{\beta^T\beta} + C\sum_{i=1}^{n}\xi_i \\
s.t. \quad & y_i(\boldsymbol{\beta^T x_i} + \boldsymbol{\beta_0}) \geq 1 - \xi_i \\
& \xi_i \geq 0 \quad for\ i \in I^- \\
& 0 \leq \xi_i \leq 1 \quad for\ i \in I^+
\end{aligned}$$

Dual:

$$\begin{aligned}
\min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x_i}, \boldsymbol{x_j}) \\
& -\sum_{i=1}^{n}\alpha_i + \sum_{i \in I^+}\mu_i \\
s.t. \quad & \sum_{i=1}^{n}\alpha_i y_i = 0 \\
& 0 \leq \alpha_i \leq C \quad \text{for } i \in I^- \\
& \begin{cases} \alpha_i \geq 0 & \\ \mu_i \geq 0 & \text{for } i \in I^+ \\ \alpha_i - \mu_i \leq C & \end{cases}
\end{aligned}$$

Here, μ_i is the lagrange multiplier of the added constraint.

Since ξ_i actually measures the distance of the point from the margin, making $0 \leq \xi_i \leq 1$ for the positive class can forbid the point from crossing the decision boundary ($y_i(\boldsymbol{\beta^T x_i} + \boldsymbol{\beta_0}) \geq 1 - \xi_i \geq 0$), although it can still enter the margin. As shown in the previous experiment, the decision boundary tends to press on the positive class. Here it is pushed back to the negative class.

However, one potential problem is that the extreme constraint can make the classifier sensitive to noise. Even one outlier can pull the decision boundary far into the negative class. To avoid the affect of noise, we can properly extend the upper bound on ξ_i so that the new constraint is $0 \leq \xi_i \leq \theta$ for $i \in I^+$ where θ is the variable controlling how far the point can cross the decision boundary. In our experiments, we have tried θ from 1 to 1.3.

IV. Experiment

A. Evaluation Metrics

Since the accuracy over the whole dataset cannot objectively indicate the true performance of the classifier on imbalanced datasets, *sensitivity* (true positive rate) and *specitivity* (true negative rate) are defined to address the accuracy of both classes. Subsequently, *G-mean* [4] is proposed as a means of combining the two measurements:

$$G - mean = \sqrt{sensitivity \cdot specificity} \tag{1}$$

To examine the classifier's performance mainly on the rare class, the *F-measure* [5] is defined as the harmonic mean of *recall* and *precision*:

$$F - measure = \frac{2 \cdot recall \cdot precision}{recall + precison} \tag{2}$$

Another important evaluation metric is the area under ROC curve (AUR) [6], which is unaffected by the class distribution and different error costs.

B. Results

In the experiment, we compare the performance of the proposed approach with regular SVM, 2C-SVM, and undersampling and oversampling techniques. Undersampling is implemented by random sampling [7] which is empirically shown to be simple but effective. Oversampling is implemented by SMOTE [1]. Each training dataset is normalized before classification. A Gaussian kernel is used for all the datasets and 10-fold cross validation is utilized to select the best model. In 2C-SVM, γ is set to be the percentage of negative instances.

We use five UCI datasets with different imbalance ratios to test the classifiers as shown in Table I. Each dataset is randomly divided into a training set and test set. All the measurements are computed by the *perf* code provided by http://kodiak.cs.cornell.edu/kddcup/software.html. The results are shown in Table II.

C. Analysis

The results show that SVM ($C_+ = 0$) achieves high score on G-mean and ROC-curve; however, it is obviously lower than the other algorithms in precision which results in a low F-measure. This can be explained by the hard constraints $C_+ = 0$. While the recall rate is guaranteed by including most of the positive examples, it will inevitably include more negative examples. The second approach, SVM ($\xi_+ \leq \theta$), has a relatively better and more stable performance over all three metrics. In addition, although 2C-SVM has been

Table I
EXPERIMENT DATASETS

Dataset	Imbalanced Rate	Training Set	Test Set
Glass(5)	6.07%	108	106
Abalone(4)	1.36%	417	3760
Car	3.76%	345	1383
Segment(1)	14.29%	231	2079
Segment(3)	14.29%	231	2079
Yeast(ME2)	3.44%	297	1187
Yeast(ME1)	2.96%	297	1187

Table II
EXPERIMENT RESULTS

Dataset	Methods	F-measure	G-mean	AUR
Glass(5)	SVM	0.6667	0.7071	0.7500
	2C-SVM	0.5714	0.8002	0.8135
	SVM($C_+ = 0$)	0.6667	**0.9674**	**0.9654**
	SVM($\xi_+ \leq \theta$)	**0.8333**	0.9083	0.9117
	SMOTE	**0.8333**	0.8452	0.8571
	undersmp	0.6111	0.9225	0.9224
Abolone(4)	SVM	0.3516	0.5530	0.6507
	2C-SVM	**0.4364**	0.9451	0.9454
	SVM($C_+ = 0$)	0.4221	0.8944	0.8975
	SVM($\xi_+ \leq \theta$)	0.4210	0.8653	0.8714
	SMOTE	0.3035	0.9503	0.9503
	undersmp	0.2750	**0.9539**	**0.9542**
Car	SVM	0.9039	0.9489	0.9500
	2C-SVM	**0.9541**	0.9533	**0.9981**
	SVM($C_+ = 0$)	0.7647	**0.9879**	0.9880
	SVM($\xi_+ \leq \theta$)	0.9039	0.9489	0.9500
	SMOTE	0.7161	0.9845	0.9846
	undersmp	0.6125	0.9748	0.9751
Segment(1)	SVM	**0.9898**	**0.9940**	0.9907
	2C-SVM	**0.9898**	0.9927	**0.9927**
	SVM($C_+ = 0$)	0.9882	0.9924	0.9924
	SVM($\xi_+ \leq \theta$)	**0.9898**	0.9927	**0.9927**
	SMOTE	0.9640	0.9713	0.9717
	undersmp	0.9370	0.9792	0.9792
Segment(3)	SVM	0.8493	0.9239	0.9248
	2C-SVM	0.8635	0.9405	0.9408
	SVM($C_+ = 0$)	**0.8663**	**0.9585**	**0.9585**
	SVM($\xi_+ \leq \theta$)	0.8653	0.9422	0.9425
	SMOTE	0.7702	0.9080	0.9083
	undersmp	0.7754	0.9425	0.9430
Yeast(ME2)	SVM	0.2319	0.4378	0.5888
	2C-SVM	0.3529	0.7410	0.7617
	SVM($C_+ = 0$)	0.2963	0.7688	0.7892
	SVM($\xi_+ \leq \theta$)	**0.3546**	0.7549	0.7722
	SMOTE	0.3545	0.8111	0.8172
	undersmp	0.2791	**0.8256**	**0.8264**
Yeast(ME1)	SVM	**0.6364**	0.7712	0.7957
	2C-SVM	0.5833	0.9781	0.9783
	SVM($C_+ = 0$)	0.5667	0.9635	0.9636
	SVM($\xi_+ \leq \theta$)	0.6153	**0.9803**	**0.9805**
	SMOTE	0.5909	0.9670	0.9671
	undersmp	0.5246	0.9584	0.9585

proposed for quite a long time, it is not given much attention in previous work for addressing the imbalance problem. However, in our experiments 2C-SVM demonstrates decent performance. As for the re-sampling techniques, SMOTE has better overall performance than random undersampling, but both techniques have low F-measure scores.

V. CONCLUSION

In this paper, we propose two modifications of Support Vector Machines to address the problem of classifying highly imbalanced datasets. Four versions of SVM as well as two re-sampling techniques are compared: regular SVM, cost-sensitive 2C-SVM, the modified 2C-SVM ($C_+ = 0$), SVM ($\xi_+ \leq \theta$), SMOTE oversampling, and random undersampling. We studied their behavior comprehensively using three comparison methods. The results show that the two proposed methods have a consistent improvement over SVM's performance. According to our experiments, 2C-SVM is comparable to other algorithms as well, which is neglected in the former studies. The re-sampling methods can mainly be criticized for changing the dataset and introducing unnecessary noise. We conclude that the proposed approaches are promising candidates for addressing the rare class problem.

REFERENCES

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[2] E. E. Osuna, R. Freund, and F. Girosi, "Support vector machines: Training and applications," Massachusetts Institute of Technology, Tech. Rep., 1997.

[3] V. N. Vapnik and A. Chervonenkis, *Theory of pattern recognition*. Nauka, Moscow, 1974.

[4] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proceedings of the Fourteenth International Conference on Machine Learning*. Morgan Kaufmann, 1997, pp. 179–186.

[5] C. J. V. Rijsbergen, *Information Retrieval*, 2nd ed. London: Butterworths, 1979.

[6] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, pp. 1145–1159, 1997.

[7] J. V. Hulse, T. M. Khoshgoftaar, and A. Napolitano, "An emprical comparison of repetitive undersampling techniques," 2009, pp. 29–34.