Pattern Recognition Letters 80 (2016) 107–112

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Cost-Sensitive Large margin Distribution Machine for classification of imbalanced data☆

Fanyong Cheng[a,b,*], Jing Zhang[a], Cuihong Wen[a]

[a] College of Electrical and Information Engineering, Hunan University, Changsha 410082, China
[b] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Department of Computer Science, Minjiang University, Fuzhou 350121, China

ARTICLE INFO

Article history:
Received 9 November 2015
Available online 21 June 2016

Keywords:
Minimum margin
Margin distribution
Imbalanced training data
Cost-sensitive learning
Balanced detection rate

ABSTRACT

This paper proposes a new method to design a balanced classifier on imbalanced training data based on margin distribution theory. Recently, Large margin Distribution Machine (LDM) is put forward and it obtains superior classification performance compared with Support Vector Machine (SVM) and many state-of-the-art methods. However, one of the deficiencies of LDM is that it easily leads to the lower detection rate of the minority class than that of the majority class on imbalanced data which contradicts to the needs of high detection rate of the minority class in the real application. In this paper, Cost-Sensitive Large margin Distribution Machine (CS-LDM) is brought forward to improve the detection rate of the minority class by introducing cost-sensitive margin mean and cost-sensitive penalty. Theoretical and experimental results show that CS-LDM can gradually improve the detection rate of the minority class with the increasing of the cost parameter and obtain a balanced classifier when the cost parameter increases to a certain value. CS-LDM is superior to some popular cost-sensitive methods and can be used in many applications.

1. Introduction

As one of the most popular classification algorithms, Support Vector Machine (SVM) can be seen as a learning approach trying to maximize the minimum margin on the training data. It acquires a good theoretical foundation for the generalization performance from margin theory [8]. What is noteworthy for the margin theory is that it not only plays a guidance role for SVM, but also has been extended to interpret the strong generalization performance of many other learning approaches [4]. However, according to margin theory, Breiman [2] designed Arc-gv, which is able to maximize the minimum margin but has a weak generalization performance. Therefore, he doubted about the margin theory, and this almost sentenced margin theory to death. After ignoring the issue for years, Reyzin and Schapire [3] found that although Arc-gv directly maximizes the minimum margin, it suffers from a weak margin distribution. Thus, they conjectured that the margin distribution is more crucial than the minimum margin to the generalization performance. Gao and Zhou [1] proved this conjecture and revealed that rather than simply considering a single margin, margin distribution is really significant for the generalization performance. Consequently, Large margin Distribution Machine (LDM) which maximizes the margin mean and minimizes the margin variance is proposed by Zhang and Zhou [6,7]. Though LDM is superior to SVM and many state-of-the-art methods whether on theoretical results or many experimental results [7], it is not satisfactory when the training data is imbalanced.

We find that LDM generally makes the separator incline to the minority class to obtain a larger margin mean, and this leads to the problem that the minority class examples are more easily misclassified than the majority class examples. Generally speaking, real-world data sets are composed of the majority class examples with only a small percentage of the minority class examples. The minority class is usually more interesting or costly. For example, the mammography data set which might contain 98% normal examples and 2% abnormal examples. A simple default strategy of guessing the majority class would give a predictive accuracy of 98%, but miss all abnormal examples. However, the nature of the application requires a high detection rate of the minority class allowing a certain error rate in the majority class. Therefore, Cost-Sensitive Large margin Distribution Machine (CS-LDM) is proposed to address the imbalanced detection rate between two classes by adjusting margin weight for each class in the margin mean. CS-LDM increases

☆ "This paper has been recommended Ananda S Chowdhury."

* Corresponding author at: College of Electrical and Information Engineering, Hunan University, Changsha 410082, China. Tel.: +86 18673197062.
E-mail address: b12090031@hnu.edu.cn (F. Cheng).

http://dx.doi.org/10.1016/j.patrec.2016.06.009
0167-8655/

the margin weights of the minority class examples in the margin mean to force the separator to move towards the majority class examples, while increasing the misclassification penalty of the minority class. It is trained and tested on some UCI data sets, and experimental results show that CS-LDM can effectively improve the detection rate of the minority class to achieve a balanced detection rate, while has a strong generalization performance.

2. Related works

To deal with the imbalanced detection rate, various techniques have been proposed. These techniques can be mainly divided into three basic types: data-preprocessing, algorithmic approach, and boosting approach. The first type tries to increase the number of the minority examples (over-sampling) [22] or decrease the number of the majority class examples (under-sampling) [23] in different ways. Batista [24] combined these two methods to acquire good detection results for the minority class. The second type adjusts the cost of error or decision threshold in classification on the imbalanced data and tries to control the detection rate of the minority class. For example, Elkan [17], Seiffert et al. [25] decreased the proportion of the majority examples in training data set to train an optimal classification decision, and many methods were proposed to improve the prediction performance by adjusting the weight (cost) for each class [5,18]. Huang et al. [19] proposed the BMPM algorithm to build up biased classifier. The third type uses the cost of misclassifications to update the training distribution on successive boosting rounds [20,21,26]. In addition, there are some other combined ways to address this problem. For example, RUSBoost [25] was proposed by combining data sampling and fast boosting. Although these algorithms mentioned above obtained good results, there is a lack of consideration of the cost of the margin distribution which is crucial to generalization performance. In the following, LDM and CS-LDM will be introduced in detail and compared.

3. Large margin Distribution Machine

We denote sample space by $\boldsymbol{x} \in R^d$ (d is the dimension of each example or feature), and the label set by $y = \{+1, -1\}$. A training set of size m $S = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots (\boldsymbol{x_m}, y_m)\}$ is drawn identically and independently according to an unknown distribution Γ over $\boldsymbol{x} \times y$. The final goal is to learn a linear function $f(\boldsymbol{x}) = \boldsymbol{\omega}^T \phi(\boldsymbol{x})$ with strong generalization performance to predict an unlabeled example, where $\boldsymbol{\omega}$ is a weight vector and $\phi(\boldsymbol{x})$ is a feature transform of x induced by a kernel k, i.e., $k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \phi(\boldsymbol{x_i})^T \phi(\boldsymbol{x_j})$. In the light of [8–11], the margin of example $(\boldsymbol{x_i}, y_i)$ is formulated as

$$\gamma_i = y_i \boldsymbol{\omega}^T \phi(\boldsymbol{x_i}), \forall i = 1, \ldots, m.$$

The margin mean is formulated as

$$\bar{\gamma} = \sum_{i=1}^{m} y_i \boldsymbol{\omega}^T \phi(\boldsymbol{x_i}) = \frac{1}{m} (X\boldsymbol{y})^T \boldsymbol{\omega},$$

where $X = [\phi(\boldsymbol{x_1}), \ldots, \phi(\boldsymbol{x_m})]$ represents a matrix ($\phi(\boldsymbol{x_i})$ is the ith column of the matrix) and $\boldsymbol{y} = [y_1, \ldots, y_m]^T$ is denoted as a vector. The margin variance is formulated as

$$\hat{\gamma} = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} (y_i \boldsymbol{\omega}^T \phi(\boldsymbol{x_i}) - y_j \boldsymbol{\omega}^T \phi(\boldsymbol{x_j})^2$$
$$= \frac{2}{m^2} (m \boldsymbol{\omega}^T X X^T \boldsymbol{\omega} - \boldsymbol{\omega}^T X \boldsymbol{y} \boldsymbol{y}^T X^T \boldsymbol{\omega}) \quad (1)$$

LDM maximizes the margin mean and minimizes the margin variance simultaneously. The optimal object function of LDM is formulated as

$$\min_{w,\xi} \frac{1}{2} \boldsymbol{\omega}^T \boldsymbol{\omega} + \lambda_1 \hat{\gamma} - \lambda_2 \bar{\gamma} + C \sum_{i=1}^{m} \xi_i.$$
$$s.t. \quad y_i \boldsymbol{\omega}^T \phi(\boldsymbol{x_i}) \geq 1 - \xi_i,$$
$$\xi_i \geq 0, i = 1, \ldots, m.$$

Where $C \geq 0$, $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are real trade-off parameters for the training error, the margin variance, the margin mean and the model complexity.

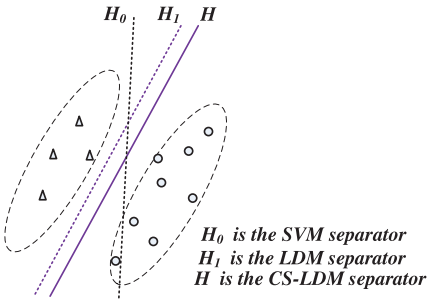

Fig. 1. Simple illustration of the separators with different algorithms. Triangles represent positive examples; circles represent negative examples.

4. Cost-Sensitive Large margin Distribution Machine

The idea is that we increase the margin weight of the minority class in the margin mean and the misclassification penalty of the minority class to address the problem that the separator of LDM (H_1 in Fig. 1) inclines to the minority class to get a larger margin mean as shown in Fig. 1.

4.1. Formulation

Considering that the positive training example size is m_+, and the negative training example size is m_- ($m_+ < m_-$), the separator of LDM (H_1 in Fig. 1) inclines to positive examples to increase the margin mean as shown in Fig. 1. To resist the trend of LDM, we introduce cost-sensitive learning method. We define the cost of margin as

$$\theta_i = \begin{cases} (m_-/m_+)^\rho & i \in I_+, \\ (m_+/m_-)^\rho & i \in I_-, \end{cases} \quad (2)$$

where $\rho \geq 0$ is a real cost parameter, $I_+ \equiv \{i | y_i = 1\}$, and $I_- \equiv \{i | y_i = -1\}$. We can find that the cost of the positive class (the minority class) is more than or equal to that of the negative class (the majority class) for all possible value of ρ. The cost-sensitive margin mean is formulated as

$$\bar{\bar{\gamma}} = \sum_{i=1}^{m} \theta_i y_i \boldsymbol{\omega}^T \phi(\boldsymbol{x_i}) = \frac{1}{m} (XD\boldsymbol{y})^T \boldsymbol{\omega} \quad (3)$$

where D is a diagonal matrix with diagonal elements $\theta_1, \ldots, \theta_m$. For simplicity, the separable case without outliers is considered. If ρ increases, the cost-sensitive margin mean of the minority class will increase and the cost-sensitive margin mean of the majority class will decrease. This induces the CS-LDM separator (H in Fig. 1) incline to the majority class examples (circles) to get a larger cost-sensitive margin mean as shown in Fig. 1.

To prevent the separator towards one class immoderately, the margin variance is also introduced into CS-LDM. Besides that, Fig. 2 shows a more complicated case where there are outliers or noisy examples. If we only optimize the minimum margin, the separator may be dominated by the outlier or noisy example. However, if