Neurocomputing 73 (2009) 397–411

Contents lists available at ScienceDirect

Neurocomputing

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

Margin calibration in SVM class-imbalanced learning

Chan-Yun Yang [a,*], Jr-Syu Yang [b], Jian-Jun Wang [c]

[a] Department of Mechanical Engineering, Technology and Science Institute of Northern Taiwan, No. 2 Xue-Yuan Road, Beitou, Taipei 11202, Taiwan, ROC
[b] Department of Mechanical and Electro-Mechanical Engineering, Tamkang University, No. 151 Ying-Chuan Road, Tamsui, Taipei County 25137, Taiwan, ROC
[c] School of Mathematics and Statistics, Southwest University, Chongqing 400715, PR China

ARTICLE INFO

Article history:
Received 13 December 2008
Received in revised form
1 June 2009
Accepted 1 August 2009
Communicated by T. Heskes
Available online 20 August 2009

Keywords:
Margin
Cost-sensitive learning
Class-imbalanced learning
Support vector machines
Classification

ABSTRACT

Imbalanced dataset learning is an important practical issue in machine learning, even in support vector machines (SVMs). In this study, a well known reference model for solving the problem proposed by Veropoulos et al., is first studied. From the aspect of loss function, the reference cost sensitive prototype is identified as a penalty-regularized model. Intuitively, the loss function can change not only the penalty but also the margin to recover the biased decision boundary. This study focuses mainly on the effect from the margin and then extends the model to a more general modification. As proposed in the prototype, the modification first adopts an inversed proportional regularized penalty to re-weight the imbalanced classes. In addition to the penalty regularization, the modification then employs a margin compensation to lead the margin to be lopsided, which enables the decision boundary drift. Two regularization factors, the penalty and margin, are hence suggested for achieving an unbiased classification. The margin compensation, associating with the penalty regularization, is here utilized to calibrate and refine the biased decision boundary to further reduce the bias. With the area under the receiver operating characteristic curve (*AuROC*) for examining the performance, the modification shows relative higher scores than the reference model, even though the optimal performance is achieved by the reference model. Some useful characteristics found empirically are also included, which may be convenient for the future applications. All the theoretical descriptions and experimental validations show the proposed model's potential to compete for highly unbiased accuracy in a complex imbalanced dataset.

1. Introductions

One frequent difficulty encountered in the practical application of machine learning is that the learning classes are imbalanced, and thus due to that, topics of class-imbalanced learning are worth further research. In the past few years, there have been many studies on reducing the bias of decision boundaries, which is generally produced by the class imbalance [1–4]. As recognized, the purpose of a common learning machine, such as a Bayesian classifier, decision tree, neural network, or support vector machine, acting as a categorical classification function, is to produce an assessment with the maximal accuracy of classification. But the purpose is difficult to fulfill when faced with imbalanced classes in the learning set. In the case, the more common machines tend to be overwhelmed by the large classes and therefore ignore the small ones. To solve this problem, modifications to machines have been proposed to generate a hypothesis which is robust to the majority overwhelming [2–7]. According to actions in this process, the resolving modifications can mainly be categorized into two levels, the data and algorithmic levels [2]. Both of these levels of modification adopt strategies prior to or in the process of balancing the problem.

At the data level, strategies of up-sampling and down-sampling are often used to deal with imbalanced datasets [2,4]. Down-sampling eliminates samples in the majority class, whereas while up-sampling duplicates those in minority-classes. Both techniques reduce the degree of imbalance. Common justifications, like "discard useful majority samples when down-sampling" or "do not really helpful for minority class when employing exact copies of samples for up-sampling", restrict the use of the techniques and have led force researchers to develop advanced sampling strategies [3]. Furthermore, at the algorithmic level, researchers have introduced cost-sensitive learning as a solution for class-imbalanced learning since it gives minority-class samples a higher learning cost in order to reduce the degree of imbalance [5–7]. A general practice is to exploit the misclassification cost from samples in the majority-class to outweigh those in the minority-class. Generally, the re-weighing scheme is merged with the common edition of classification algorithms [6]. In addition to the main approaches, there are also some

* Corresponding author. Tel.: +886 2 28927154x8012; fax: +886 2 28473721.
E-mail address: cyyang.research@gmail.com (C.-Y. Yang).

0925-2312/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2009.08.006

interesting strategies for managing imbalance, for example "learning only the rare class", "segmenting the data", "iterative boosting," and others. Some good comments for selecting a typical strategy to deal with a particular imbalanced learning problem can be found in Ref. [3].

For this study, we narrow the study of class-imbalanced problem to support vector machines (SVMs) [8–13], and examine relevant paper surveys. The surveys show the development of class-imbalanced learning in SVMs actually has the same technical merits which we have mentioned above. Known as a popular approach, variant techniques of up-sampling and down-sampling have been employed in some real applications, such as automatic detection of pulmonary nodules in thoracic computer tomography [14], protein homology prediction in bioinformatics [15], and identifying the respondents of a mailing campaign in customer relationship management [16]. On the other hand, Akbani et al. [17] have explained why the down-sampling strategy is not the best choice for SVM. They developed a method combining a well-known synthetic minority up-sampling technique (SMOTE) [18] with different error costs algorithm [19] to push the biased decision boundary away from the minority-class. Ertekin et al. [20] proposed an efficient active learning strategy for down-sampling. This method iteratively selects the closest instance to the separating hyperplane from the unseen training data and adds it to the training set to retrain the classifier. With an early stopping criterion, the method can decrease significantly the training time in the large scale imbalanced dataset. In addition, a backward pruning technique, identified as one type of down-sampling, is also employed to deal with the class-imbalanced classification [21].

The other type of approach for dealing with class-imbalanced SVM learning is to modify the algorithms of an underlying machine, which can be done by cost-sensitive learning. Modifying the kernel function provides one solution for such cost-sensitive learning [22,23]. The kernel function can be conformally transformed according to the structure of the Riemannian geometry in the imbalanced dataset to improve the bias. The idea has been utilized in learning with a one-class SVM to handle a case of class imbalance [24].

One another crucial approach to solve this problem was proposed by Veropoulos et al. [19], who use different penalty constants for different classes to make errors on minority-class samples costlier than errors on majority-class samples. Different scales of penalization are applied to cost the errors in different classes. The penalty-regularized model deserves much more attention because its straightforward idea gives the model intrinsic coherence with its original prototype of SVM. Indeed, this remedy has broadly been applied and extended to deal with applications [17,25–28].

2. Generalized cost-sensitive learning in SVM

2.1. Penalty-regularized model

In order to reduce the majority overwhelming, the penalty-regularized model proposed by Veropoulos et al. [19] is closely inspected. The key idea of this model is to introduce uneven loss functions to re-weight the penalties of the samples in the imbalanced classes [29]. The loss function associated with a minority-class sample, also a positive sample, retains penalty higher than a function with a majority-class sample, also a negative sample, in the optimization. The lopsided penalties then re-configure the Lagrange multipliers to achieve the decision boundary low bias due to heavier misclassification cost. The model here can be started with a set $S=(\mathbf{X},\mathbf{Y})=\{(\mathbf{x}_i,y_i)|\ \mathbf{x}_i\in \Re^d,\ y_i\in\{+1,-1\},\ i=1,2,\ldots,l\}$ consisting of l^+ positive and l^- negative training samples in a d-dimensional input space $\Re^d$. In order to distinguish the samples in positive and negative classes, the class labels of training samples are first examined and partitioned into their exact categories:

$$I^+=\{i|y_i=+1,(\mathbf{x}_i,y_i)\in S,\ \text{ for }\ \forall i\}, \tag{1}$$

and

$$I^-=\{i|y_i=-1,(\mathbf{x}_i,y_i)\in S,\ \text{ for }\ \forall i\}, \tag{2}$$

where I^+ and I^- denote, respectively, the index set for both the positive and negative class. With set S and categorized indices I^+ and I^-, the Veropoulos model based on the prototype of soft-margin SVM [30] is presented as

$$\arg\min_{\mathbf{w},\boldsymbol{e}}\frac{1}{2}\mathbf{w}^T\mathbf{w}+\boldsymbol{C}^T\boldsymbol{e}, \tag{3}$$

subject to

$$y_i(\mathbf{w}^T\Lambda(\mathbf{x}_i)+b)\geq 1-e_i, \tag{4}$$

and

$$e_i\geq 0,\ \ \forall i, \tag{5}$$

to learn the target concept $f(\mathbf{x})=\mathbf{w}^T\Lambda(\mathbf{x})+b$. In these expressions, $\boldsymbol{C}$ denotes a column vector consisting of two types of penalty constants for both positive and negative classes:

$$\boldsymbol{C}=[C_1,C_2,\ldots,C_l]^T \text{where } C_i=C^+ \text{ if } i\in I^+,\ \text{or } C_i=C^- \text{ if } i\in I^-, \tag{6}$$

and $\boldsymbol{e}$ denotes a column vector of corresponding losses of the training samples, $\boldsymbol{e}=[e_1,e_2,\ldots,e_l]^T$. As is known, a map function, $\Lambda: S\rightarrow\Re^h$, mapping the learning set from the lower d-dimensional input space to a higher reproducing kernel Hilbert space (RKHS) $\Re^h$ can be introduced for solving the classification problem non-linearly [8,10–11]. The weight vector $\mathbf{w}$ is a d-dimensional transposed vector normal to the decision boundary, the bias b is a scalar for offsetting the decision boundary, and the slack variables e_i's measuring the losses are used to urge samples to satisfy the boundary constraints in the optimization. As mentioned above, the recovery of the biased decision boundary is sought to assign different cost to the misclassified samples in different classes. In general, the misclassifications in the positive class are costlier than those in negative class. When there are fewer positive samples, a higher misclassification cost should be assigned. However, with the recovered decision boundary, supplemental conditions from the KKT (Karush–Kuhn–Tucker) conditions [19] should still be satisfactory after optimization:

$$0\leq\alpha_i\leq C_i,\ \ \forall i, \tag{7}$$

and

$$\sum_{i\in I^+}\alpha_i=\sum_{i\in I^-}\alpha_i, \tag{8}$$

where α_i denotes the Lagrange multiplier of corresponding sample i. The details of α_i will be described in the following section.

2.2. From the viewpoint of a loss function

From the aspect of optimization, the prototype SVM can also be expressed as a risk regularized model [8,9]:

$$\arg\min_{f\in\Re^h}\frac{1}{2}\mathbf{w}^T\mathbf{w}+\lambda R_{\text{emp}}[f], \tag{9}$$

where $R_{\text{emp}}[f]$ denotes classification risk and $\Re^h$ denotes the RKHS. The model seeks to simultaneously maximize the margin and minimize the classification risk, and λ in the expression is a tradeoff factor for regularization. One of the major assumptions of the learning machine is that no information about the probability