Cost-sensitive probabilistic predictions for support vector machines

Sandra Benítez Peña, Rafael Blanquero, Emilio Carrizosa and Pepa Ramírez-Cobo

Abstract

Support vector machines (SVMs) are widely used and constitute one of the best examined and used machine learning models for two-class classification. Classification in SVM is based on a score procedure, yielding a deterministic classification rule, which can be transformed into a probabilistic rule (as implemented in off-the-shelf SVM libraries), but is not probabilistic in nature.

In this paper we propose a novel approach to generate probabilistic outputs for the SVM. The new method has the following three properties. First, it is designed to be cost-sensitive, and thus the different importance of sensitivity (or true positive rate, TPR) and specificity (true negative rate, TNR) is readily accommodated in the model. As a result, the model can deal with imbalanced data. Second, the SVM is embedded in an ensemble method to improve its performance, making use of the valuable information generated in the parameters tuning process. Finally, the probabilities estimation is done via bootstrap estimates, avoiding the use of parametric models as competing approaches.

Keywords: Support Vector Machines; Probabilistic Classification; Cost-Sensitive Classification

1 Introduction

Supervised classification is one of the most relevant tasks in Data Science. We are given a set Ω of individuals. Each element $i \in \Omega$ is represented by a pair (x_i, y_i), where $x_i \in \mathbb{R}^n$ is the attribute vector, and $y_i \in \mathcal{C}$ is the class membership of object i. We only have class information in $T \subset \Omega$, which is called the *training sample*. In its most basic version, the one considered in this paper, supervised classification addresses two-class problems, that is to say, $\mathcal{C} = \{-1, +1\}$.

Support Vector Machine (SVM) is a powerful and state-of-the-art method in supervised classification, that aims, in the simpler case of linear SVM, at separating both classes by means of a linear classifier, $\omega^\top x_i + \beta$. SVM can be addressed by solving a convex quadratic programming (QP) formulation with linear constraints. It is usual to consider its dual formulation, which allows us to use the so-called kernel

trick, and is given by

$$\begin{array}{ll} \max_{\lambda} & -\frac{1}{2}\sum_{j\in T}\sum_{k\in T}\lambda_j\lambda_k y_j y_k K(x_j,x_k)+\sum_{l\in T}\lambda_l \\ s.t. & \sum_{i\in T}\lambda_i y_i = 0 \\ & 0\le\lambda_i\le C, \qquad\qquad i\in T \end{array} \tag{1}$$

where λ are the usual variables of the dual SVM formulation, $C>0$ is a *regularization parameter* to be tuned, which controls the trade-off between margin minimization and misclassification errors, and K is a *kernel* such that $K(x_j,x_k)=\phi(x_j)^\top\phi(x_k)$ (where ϕ is a mapping function that embeds the dataset into a higher dimensional space). The kernel function $K(x_j,x_k)$ may include other parameters, such as the σ parameter in the Radial Basis Function (RBF) kernel. Such parameters have to be also tuned, in a grid Θ.

Given an object with attribute vector x_0, the SVM algorithm produces a hard labeling in such a way that this instance is classified in the positive or the negative class according to the sign of $f(x_0)$, where $f(x)=\sum_{i\in T}\lambda_i y_i K(x,x_i)+\beta$ is the *score function*. When an attribute vector x_0 is given, the value $f(x_0)$ is called the *score value* of x_0. However, the SVM method does not result in probabilistic outputs as posterior probabilities $P(y=+1\mid x)$, which are of interest if a measure of confidence in the predictions is sought.

2 State of the art

Several attempts to obtain the posterior probabilities $P(y=+1\mid x)$ for SVM have been already carried out previously. One of them is based on assigning posterior class probabilities assuming a specific parametric family for the posterior probability. For example, [15] proposed a logistic link function.

Also, [14] suggested to estimate $P(y=+1\mid x)$ in terms of a series of the trigonometric functions, where the coefficients of the trigonometric expansion minimizes a regularized function. Another considered option has been to fit Gaussians to the class-conditional densities $P(f(x)\mid y=+1)$ and $P(f(x)\mid y=-1)$, as proposed in [5]. From such a choice, the posterior probability $P(y=+1\mid f(x))$ is assumed to be a sigmoid, whose slope is determined by the tied variance. One of the best-known heuristics to obtain probabilities is due to [10], which considers $f(x)$ as the log-odds ratio $\log\dfrac{P(y=+1\mid x)}{P(y=-1\mid x)}$. Although SVM is designed for binary classification, there are several extensions for multiclass problems, e.g. [1], and also some attempts to construct class probabilities are found in the literature. In particular, multiclass versions of Platt's approach can be found in [8] and have been implemented in software packages like LIBSVM ([2]). However, it has been criticized for failing to provide insight and for interpreting $f(x)$ as a log-odds ratio, which may be not accurate for some datasets, see [9].

[12] considers a different probabilistic framework for SVM classification, based on Bayesian theory.

Again, this method as the previously commented approaches make modeling assumptions that might not be satisfied by the data. Finally, other procedures seeking probabilistic outputs are found in the literature, as [11, 7].

None of the previously mentioned works produce cost-sensitive models, which are of crucial importance in many managerial decision-making problems.

Cost-sensitivity is closely related to the problem of imbalancedness in datasets. Imbalancedness may produce unnacurate classification rates for the minority class that is often the most critical one, [3]. Several attempts in the literature have considered probabilistic outputs for the SVM in a context of imbalancedness. For example, [13] propose robust SVM that turn out insensitive to the class imbalancedness. Their approach, the *posterior probability support vector machine* (PPSVM), is distribution-free and weighs imbalanced training samples. A multiclass approach based on the method in [13] is proposed by [4]. Also, a more sophisticated and computationally expensive alternative is proposed by [6], which combines layers of SVM with class probability output networks (CPONs), in which strong statistical assumptions are imposed.

3 Main Contributions

In this paper we have proposed a procedure to obtain probabilistic outputs for the Support Vector Machines, through point estimates. Contrary to existing proposals, we present a method that is distribution-free and cost-sensitive. Also, it makes use of not only a single classifier but a weighted average of the scores corresponding to the different classifiers built for the different parameters of the SVM, obtaining more accurate results. The method turns out advantageous for operational business processes as credit scoring or churn prediction, where the class of interest may suffer from imbalancedness.

Our proposal is compared to some benchmark methodologies. The results show that our approach is comparable or better than such approaches if the focus is on point estimates. Two cost-sensitive alternatives are proposed here. The first one is based on changing the way the probabilities are estimated and the second one proposes to modify the original classifier by a cost-sensitive version.

References

[1] Emilio Carrizosa, Belen Martin-Barragan, and Dolores Romero Morales. Multi-group Support Vector Machines with Measurement Costs: A Biobjective Approach. *Discrete Applied Mathematics*, 156(6):950–966, 2008.

[2] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3), May 2011.

[3] Nazeeh Ghatasheh, Hossam Faris, Ismail AlTaharwa, Yousra Harb, and Ayman Harb. Business analytics in telemarketing: Cost-sensitive analysis of bank campaigns using artificial neural networks. *Applied Sciences*, 10(7), 2020.

[4] M. Gonen, A. G. Tanugur, and E. Alpaydin. Multiclass posterior probability support vector machines. *IEEE Transactions on Neural Networks*, 19(1):130–139, 2008.

[5] Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. In *Advances in neural information processing systems*, pages 507–513, 1998.

[6] Sangwook Kim, Zhibin Yu, Rhee Man Kil, and Minho Lee. Deep learning of support vector machines with class probability output networks. *Neural Networks*, 64:19 – 28, 2015. Special Issue on "Deep Learning of Representations".

[7] James Tin-Yau Kwok. Integrating the evidence framework and the support vector machine. In *ESANN*, volume 99, pages 177–182, 1998.

[8] J. Milgram, Mohamed Cheriet, and R. Sabourin. Estimating accurate multi-class probabilities with support vector machines. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 3, pages 1906–1911 vol. 3, 2005.

[9] Kevin P. Murphy. *Machine learning, a probabilistic perspective*. The MIT Press, 2012.

[10] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*, pages 61–74. MIT Press, 2000.

[11] Matthias Seeger. Bayesian model selection for support vector machines, gaussian processes and other kernel classifiers. In *Advances in neural information processing systems*, pages 603–609, 2000.

[12] Peter Sollich. Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Machine learning*, 46(1):21–52, 2002.

[13] Qing Tao, Gao-Wei Wu, Fei-Yue Wang, and Jue Wang. Posterior probability support vector machines for unbalanced data. *IEEE Transactions on Neural Networks*, 16(6):1561–1573, 2005.

[14] Vladimir Naumovich Vapnik and Vlamimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[15] Grace Wahba. Multivariate function and operator estimation, based on smoothing splines and reproducing kernels. In *Santa Fe Institute Studies in the Sciences of Complexity-Proceedings Volume-*, volume 12, pages 95–95. Addison-Wesley Publishing Co, 1992.