Large margin multiple hyperplane classification for content-based multimedia retrieval

Serhiy Kosinov SERHIY.KOSINOV@CUI.UNIGE.CH
Ivan Titov IVAN.TITOV@CUI.UNIGE.CH
Stéphane Marchand-Maillet MARCHAND@CUI.UNIGE.CH
University of Geneva, 24 Rue du General-Dufour, CH-1211, Geneva, Switzerland

Abstract

This introductory note considers an application of content-based multimedia retrieval, where a semantic concept of a user query must be learned from only a few documents, provided as relevance feedback, that are vastly outnumbered by the irrelevant items of the collection. Formally, the problem in question is situated in the context of asymmetric classification where due to substantial imbalance, different classes are not treated equally. In contrast to the popular optimal separating hyperplane techniques that use only one hyperplane, an attempt is made to further exploit the asymmetric problem setting by incorporating multiple hyperplanes in a classifier so as to favor the under-represented class. Although the introduced modification leads to a more difficult optimization problem, a preliminary empirical evaluation of such a method in the asymmetric "one-against-all" classification setting provides encouraging results, which warrants further investigation.

1. Introduction

In this note, we consider the asymmetric classification problem setting, often encountered in content-based multimedia retrieval performed as a "one-against-all" classification scheme. The essence of the proposed technique is to increase the number of hyperplanes used in an optimal separating hyperplane classifier, so as to favor the under-represented class. Such a distinction that singles out a certain target class from the rest of the data, when modeled explicitly, has been previously shown to improve classification accuracy for undersampled and unbalanced data sets, (Akbani et al., 2004; Veropoulos et al., 1999; Zhou et al., 2004). While being applicable in the general classification scenario, the proposed method is designed to further exploit the asymmetry of the classification problem at hand.

The intuition behind the idea of introducing one or more extra hyperplanes in a classifier is exemplified in Figure 1, where it is shown how an additional hyperplane may improve the class separation margin, and thus have the potential to reduce the classification error rate. The following section details the formulation of the multiple-hyperplane (MH) classification, considers its generalization properties and presents preliminary experimental results.

Appearing in *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

2. Multiple-hyperplane classification

2.1. Problem formulation

The standard 2-class optimal separating hyperplane problem setting can be extended trivially in order to accommodate more than one hyperplane:

$$\min_{\omega_1,\ldots,\omega_{N_H}} \quad ||\omega_1||^2 \tag{1}$$

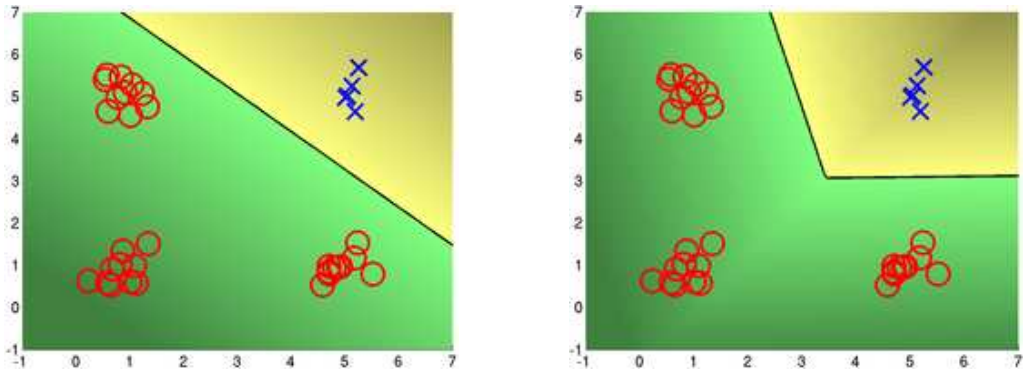
$$\text{subject to:} \quad y_i \min_{j=1\ldots N_H} (\omega_j^T x_i) \geqslant 1, \tag{2}$$

$$||\omega_1||^2 = \ldots = ||\omega_{N_H}||^2, \tag{3}$$

where $(x_i, y_i) \in \mathbb{R}^n \times \{\pm 1\}$ are data samples with their respective class labels, and N_H is the number of hyperplanes, each of which is defined by ω_j. Here, labels $+1$ and -1 correspond to under-represented and over-represented classes respectively. Additionally, we require that the sum of distances to compound border be less or equal to the sum of signed distances to the average hyperplane $\bar{\omega}$:

$$\sum_i y_i \min_{j=1\ldots N_H} (\omega_j^T x_i) \leqslant \sum_i y_i \bar{\omega}^T x_i, \tag{4}$$

where $\bar{\omega} = \frac{1}{N_H} \sum_{j=1\ldots N_H} \omega_j$. This condition ensures some degree of flatness of the compound border avoid-



(a) SVM, $\gamma = 1.14$ (b) MH, $\gamma = 1.55$

Figure 1. SVM vs. Multiple-hyperplane (MH) method on a toy problem in 2D: an additional hyperplane leads to a better separation margin γ (both methods use linear kernels).

ing overfitting. There is guaranteed to be at least one set of hyperplanes that meets this requirement. The actual role of this average signed distance constraint, however, will be clarified in greater detail in the following section.

A disadvantage of the proposed formulation is that the above optimization problem may be quite diffcult due to the use of non-differentiable min-function, which necessitates the use of auxiliary numerical strategies for attaining differentiability via smoothing of the loss function and avoiding unacceptable local minima via annealed penalty terms. Its advantage, on the other hand, is that (1-4) are expressed in terms of dot products, and thus are easily extended to nonlinear cases via kernel trick.

2.2. Generalization performance assessment

The following result, which we state without a detailed proof due to space limitations, establishes the generalization properties of the proposed technique.

Proposition 1. *Consider thresholding a class* $\mathbf{F}$ *of functions* $\min_{j=1\ldots N_H}(\omega_j^T t)$ *with unit weight vectors on an inner product space* $\mathcal{X}$ *and fix* $\gamma \in \Re^+$. *For any probability distribution* $\mathcal{D}$ *on* $\mathcal{X} \times \{-1, 1\}$ *with support in a ball of radius* R *around the origin, with probability* $1-\delta$ *over* l *random examples* S, *any hypothesis* $f \in \mathbf{F}$ *that has margin* $m_S(f) \geq \gamma$ *on* S *has error no more than*

$$\varepsilon(l, \mathbf{F}, \delta, \gamma) = \frac{2}{l}\left(\frac{64R^2}{\gamma^2}\log\frac{el\gamma}{4R}\log\frac{128lR^2}{\gamma^2} + \log\frac{4}{\delta}\right), \tag{5}$$

provided $l > 2/\varepsilon$ *and* $64R^2/\gamma^2 < l$.

Note that the error bound is absoultely the same as presented in (Cristianini & Shawe-Taylor, 2000) for a single hyperplane case. In order to clarify the intuition behind this result, we observe that the proof of a standard result on fat-shattering dimension, $fat_{\mathbf{F}}$, of an optimal hyperplane classifier (Schölkopf & Smola, 2002; Bartlett & Shawe-Taylor, 1999; Vapnik, 1982) is applicable in the multiple-hyperplane setting (1-4). That is, proceeding in a manner similar to the original proof and explicitly taking constraint (4) into account leads to an identical bound on $fat_{\mathbf{F}}$:

$$r^2\gamma^2 N_H \leqslant \left\|\sum_i^r y_i S^T x_i\right\|^2 \leqslant N_H r R^2$$
$$\Rightarrow \quad fat_{\mathbf{F}}(\gamma) \leqslant r \leqslant \left(\frac{R}{\gamma}\right)^2. \tag{6}$$

Then, result (5) naturally follows, once (6) is substituted into the theoretical result that establishes the link between the fat-shattering dimension and generalization error (Bartlett & Shawe-Taylor, 1999; Vapnik, 1982). In equation (6) above, r is the number of observations x_i, R is the radius of the smallest sphere containing all x_i, γ is the separation margin, and $S = \mathbf{1}^T \otimes I$ for a vector $\mathbf{1}$ of all ones of length N_H.

2.3. Preliminary experimental results

For our content-based multimedia retrieval experiments we chose ETHZ80 collection (Leibe & Schiele, 2003), containing 3280 high-resolution color images of objects from 8 different semantic classes. The visual information for each image was represented by 286-dimensional feature vector containing 166 global color histogram and 120 Gabor filter texture descriptors.

Table 1. Classification accuracy (in %%) per class for ETHZ80 image collection

Method	**apple**	**car**	**cow**	**cup**	**dog**	**horse**	**pear**	**tomato**
MH classifier	97.12	88.44	89.75	95.41	92.37	88.44	95.19	98.38
N_H	6	2	5	5	6	4	5	2
SVM classifier	96.16	88.06	84.59	95.94	83.59	88.09	92.16	97.66
margin ratio (MH/SVM)	1.37	1.10	1.11	1.02	1.77	1.76	1.22	1.01

For each semantic class the training data comprised 80 images with an imbalance ratio of 10/70, and an overall training vs. testing data ratio was hence 80/3200. For each class, we compared the classification accuracy of the 2-class SVM (Cristianini & Shawe-Taylor, 2000; Vapnik, 1998) with a Gaussian kernel tuned by cross-validation to that of the MH classifier using the same kernel parameters, but letting the number of hyperplanes vary. The outcome of these experiments demonstrated that in most cases the performance of the SVM classifier is improved by introducing extra separating hyperplanes, while the ratio of the class separation margins achieved by the two methods indicated where such improvement was most likely. The summary of results is shown in Table 1.

3. Conclusion

We have presented a large margin classification method that exploits the asymmetric problem setting by increasing the number of hyperplanes used in an optimal separating hyperplane classifier. The performance of the proposed technique has been assessed theoretically by establishing a bound on generalization error, and practically by evaluating its performance in a content-based image retrieval task, providing encouraging results. Further research is warranted in order to gain a better insight into the method's theoretical properties via Rademacher complexity bounds (Bartlett & Mendelson, 2001; Koltchinskii & Panchenko, 2002), and to investigate its performance in related multimedia processing applications.

References

Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. *Proceedings of the 15th European Conference on Machine Learning (ECML'04)* (pp. 39–50).

Bartlett, P. L., & Mendelson, S. (2001). Rademacher and Gaussian complexities: Risk bounds and structural results. *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory and Fifth European Conference on Computational Learning Theory* (pp. 224–240).

Bartlett, P. L., & Shawe-Taylor, J. (1999). Generalization performance of support vector machines and other pattern classifiers. In B. Schölkopf, C. J. C. Burges and A. J. Smola (Eds.), *Advances in kernel methods – support vector learning*, 43–54. MIT Press.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press.

Koltchinskii, V., & Panchenko, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.*, *30*, 1–50.

Leibe, B., & Schiele, B. (2003). Analyzing appearance and contour based methods for object categorization. *International Conference on Computer Vision and Pattern Recognition (CVPR'03)* (pp. 409–415). Madison, Wisconsin.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA: MIT Press.

Vapnik, V. N. (1982). *Estimation of dependencies based on empirical data*. New York: Springer-Verlag.

Vapnik, V. N. (1998). *Statistical learning theory*. New-York: Wiley.

Veropoulos, K., Cristianini, N., & Campbell, C. (1999). Controlling the sensitivity of support vector machines. *Proceedings of the International Joint Conference on Artificial Intelligence, (IJCAI99)* (pp. 55–60). Stockholm, Sweden.

Zhou, X., Garg, A., & Huang, T. (2004). A discussion of nonlinear variants of biased discriminants for interactive image retrieval. *Proceedings of CIVR'04* (pp. 353–364). Dublin, Ireland.