Classifier Combination Using Random Walks on the Space of Concepts

Jorge Sánchez[1,2] and Javier Redolfi[2]

[1] CIEM-CONICET, Universidad Nacional de Córdoba
[2] CIII, Universidad Tecnológica Nacional, Fac. Reg. Córdoba
jsanchez@scdt.frc.utn.edu.ar

Abstract. We propose a novel approach for the combination of classifiers based on two commonly adopted strategies in multiclass classification: one-vs-all and one-vs-one. The method relies on establishing the relevance of nodes in a graph defined in the space of concepts. Following a similar approach as in the ranking of websites, the relative strength of the nodes is given by the stationary distribution of a Markov chain defined on that graph. The proposed approach do not requires the base classifiers to provide calibrated probabilities. Experiments on the challenging problem of multiclass image classification show the potentiality of our approach.

Keywords: multiclass, classification, random walks, image classification, Fisher vectors.

1 Introduction

Multiclass classification is a fundamental problem in pattern recognition. Here, the task is to assign a given sample to one or more instances from a predefined set of concepts or classes. According to whether a sample can belong either to just one or to several of such concepts, the classification problem can be further characterized as a multiclass single-label (MCSL) or a multiclass multi-label (MCML) task. In what follows, we assume the availabily of a training set consisting of a fair amount of manually annotated samples of each class.

Although a large number of methods exists aiming to solve the multiclass problem as a whole, the most common approach is to decompose the classification task into a set of binary subproblems and to solve them independently. This class of methods have been shown to perform on par with more elaborated techniques when used properly [15,4]. Let $\mathcal{C} = \{1, \ldots, C\}$ denote the set of classes. A common binarization strategy, known as *one-vs-all* (OVA) or *one-vs-the-rest*, is to generate a set of C binary classifiers trained by using as positives the samples form each class and as negatives those from the others, i.e. each model is trained to separate one class from the rest. Given a new sample, each classifier provides a score s_i, $1 \leq i \leq C$, reflecting its confidence in assigning the input sample to the class $i \in \mathcal{C}$. The final decision regarding class membership is generally made using the "argmax" rule (MCSL), i.e. $\hat{i} = \arg\max_i s_i$, or via

L. Alvarez et al. (Eds.): CIARP 2012, LNCS 7441, pp. 789–796, 2012.

a simple threshold (MCML), i.e. the input sample belongs to class $\hat{i} \subset \mathcal{C} \iff s_i > threshold$.

Another strategy, known as *one-vs-one* classification (OVO), consists in training a set of $\binom{C}{2}$ binary classifiers to discriminate between every pairs of classes. Let r_{ij} be the output of the classifier trained with samples of the ith and $j$the classes as positives and negatives respectively. In order to decide which class the input sample belongs, a common approach is to compute a weighted vote, e.g. $\sum_j r_{ij}$, followed by the application of one of the above assignment strategies.

As the OVA and OVO schemes use different subsets of data for learning the classifiers, they are likely to provide complementary information about the structure of the feature space they act on. Based on this hypothesis, we propose a novel approach for combining the scores of OVA and OVO classifiers based on the stationary distribution of a Markov chain defined in the space of concepts. The approach does not requires the base classifiers to provide calibrated probabilities, nevertheless the combined scores do. We demonstrate the effectiveness and potentiality of the approach on the challenging problem of multiclass image classification, for both the single- and multi-label settings.

Related Work. Next, we provide a summary of the methods most closely related to our work in the context of multiclass classification.

Garcia-Pedrajas and Ortiz-Boyer [5] proposed a method for the combination of OVO and OVA classifiers. The method is a two-stage approach in which the best two scoring classes of an OVA scheme are used as hypothesis for OVO classification. The method relies on the following observations: *i)* in many cases, when an OVA scheme using the "argmax" rule fails, the correct class is given by the second best performing classifier; and *ii)* most of the errors in OVO classification are due to *incompetent classifiers*, i.e. those classifiers that have not been trained using the correct class of the query sample. Our method differs from [5] in that we take into account not only the second but all the scores provided by the pool of OVA classifiers in a principled way, avoiding early decisions that may affect the final classification. Reid [14] proposed to weight each pairwise (OVO) prediction by an estimate of the probability that the sample belongs to that pair. The method is very computationally demanding as it involves the training and evaluation of $C(C-1)$ classifiers, half of which must be learned using all available sample instances. Moreover, an additional calibration step must be performed in order to use state-of-the-art classifiers, e.g. Support Vector Machines (SVM). Also close to our work is the first of the methods proposed by Wu *et al.* [16]. The authors formulate an optimization problem involving all pairwise (OVO) estimates and the unknown class-probabilities. The solution to this problem is shown to be the stationary distribution of an irreductible Markov chain (cf. Sec. 2) whose transition matrix involves the set of (calibrated) pairwise predictions. Our method differs from [16] in the following: *a)* we go beyond simple OVO classification, *b)* we do not require the base classifiers to provide calibrated probabilities and *c)* we do not assume the training data to be balanced.

This paper is organized as follows: we first give a brief introduction to the theory of random walks on graphs and its application to the node ranking problem (Sec. 2). In Sec. 3 we formalize our approach for classifier combination. In Sec. 4 we give a detailed explanation of the experimental setup. Results of our experiments are shown in Sec. 5. Finally, we draw some conclusions in Sec. 6.

2 Preliminaries

Let $G = (\mathcal{V}, \mathcal{E}, \mathrm{A})$ be a weighted directed graph with nodes $\mathcal{V} = \{1, \ldots, n\}$ and edges $\mathcal{E} = \{(i,j) | i \in \mathcal{V}, j \in \mathcal{V}\}$. The $n \times n$ adjacency matrix $\mathrm{A} = [a_{ij}]$ is defined such that $a_{ij} > 0 \iff (i,j) \in \mathcal{E}$ and 0 otherwise. Let us now consider the following random walk on G: starting from an arbitrary node, if at time t the walker is at node i, it makes a jumps to node j with probability $p_{ij} := \hat{a}_{ij} = a_{ij} / \sum_{j=1}^{n} a_{ij}$ (independent of t). Each "step" of the process can be associated with a random variable X_t taking values on $\mathcal{V}$. The sequence $X_1, X_2, \ldots, X_t, \ldots$ corresponds to a Markov chain defined on the space of nodes and $P(X_{t+1} = j | X_t = i) = p_{ij}$. *Thus, a random walk on G is a Markov chain with states in $\mathcal{V}$ and transition matrix* $\mathrm{P} = [\hat{a}_{ij}]$. The distribution $\mathbf{\Pi}$ is said to be stationary if

$$\mathbf{\Pi}^T = \mathbf{\Pi}^T \mathrm{P} \ . \tag{1}$$

It can be shown that such a distribution exists if the Markov chain encoded by P is *irreductible* (any state must be reachable from any other state in a finite number of steps) and *aperiodic* (returning to state i can occur at irregular number of steps). Given P, the stationary distribution $\mathbf{\Pi}$ can be found by solving the eigenvalue problem (1) with the constraint $\mathbf{\Pi}^T \mathbf{e} = 1$. Here, $\mathbf{e}$ denotes the n-dimensional vector whose elements are all equal to 1. The solution to this problem can be found numerically, e.g. by the power-method.

PageRank and the Relevance of Nodes in a Graph. PageRank [10] was proposed as a model to determine the relevance of web-pages. The model considers the hyperlink structure of the web as a directed graph, on which a random walker located at node i can jump to any of the nodes linked by i with equal probability, i.e. $p_{ij} = 1 / \sum_k a_{ik}$. Here, $a_{ij} = 1$ if $(i,j) \in \mathcal{E}$ and 0 otherwise. A particularity of this structure is the presence of nodes with no out-going links ("dangling" links). For these nodes, the corresponding row of the transition matrix contains only zeros. Beeing non-stochastic, the resulting P do not corresponds to a valid transition matrix. Page *et al.* proposed the following definition for P:

$$\tilde{P} = \alpha P + (1 - \alpha) \frac{\mathbf{e}\mathbf{e}^T}{n} \tag{2}$$

where $0 \leq \alpha \leq 1$. Here, the convex combination of P with the perturbation matrix $E = \frac{\mathbf{e}\mathbf{e}^T}{n}$ ensures $\tilde{P}$ to be irreductible by definition[1] [7]. The intuition behind this approach is to model the behaviour of a "random surfer" that with

[1] Note that by adding E we are effectively creating an arc between every pair of nodes.

probability $(1-\alpha)$ gets bored and makes a jump to an arbitrary site. An extension to this model –known as *personalization*– consists on replacing $\mathbf{e}^T/n$ by $\mathbf{v}^T$: a distribution over states reflecting the preferences of each particular user [6].

3 Random Walks for Classifier Combination

Let G^C be a graph with nodes $\mathcal{V} = \mathcal{C}$, i.e. a graph defined on the space of concepts. Let us consider a random walk on G^C with a transition matrix defined as the convex combination of two terms, as follows:

$$\tilde{P} = \alpha P_O + (1-\alpha) P_A \ , \tag{3}$$

where $0 \leq \alpha \leq 1$. Let us also define the matrix $A = [a_{ij}]$, with elements:

$$a_{ij} = \begin{cases} \sigma(\beta r_{ij}), & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases} \ , \tag{4}$$

where $\beta > 0$ corresponds to a tuning parameter and $\sigma(x) = (1 + \exp(-x))^{-1}$ is the logistic function. The matrix A can be seen as the adjacency matrix of the graph corresponding to the first term in (3). The $C \times C$ matrix P_O is defined as the the row-normalized version of the adjacency matrix (4). The matrix P_A is defined as $P_A = \mathbf{e}\mathbf{q}^T$, where the "personalization" vector $\mathbf{q} = (q_1, \ldots, q_C)^T$ takes the form:

$$q_i = \frac{\sigma(\beta s_i)}{\sum_{k=1}^{C} \sigma(\beta s_k)} \ . \tag{5}$$

Using (4) and (5) in the definition of $\tilde{P}$ makes it a valid transition matrix[2]. It comprises two terms: the first, reflecting all pairwise relations between nodes; the second, modelling the behaviour of a "random surfer" which prefers those nodes with a high one-vs-all classification score. The trade-off between these terms is controlled by the parameter α.

Given a new sample $\mathbf{x}$ and a set of trained OVO and OVA classifiers, we define the *classification score w.r.t. the ith class as the corresponding element of the stationary distribution vector of the Markov chain having $\tilde{P}$ as transition matrix.*

The computation of $\tilde{P}$ involves the evaluation of $C(C+1)/2$ classifiers. It is interesting to see that in the case of $\alpha = 0$, i.e. when considering only the OVA-terms, the stationary distribution is $\mathbf{\Pi} = \mathbf{q}$. From the definition of q_i in eq. (5), it follows that the "argmax" rule will make the same prediction as with a traditional OVA scheme.

4 Experimental Setup

We evaluate our approach in the context of multiclass image classification. The evaluation was performed using two challenging image datasets: PASCAL

[2] It corresponds to a fully connected graph, as $q_i > 0$, $\forall i$.

VOC2007 [3] and MIT Indoor Scenes [13]. The image representation we used was the improved Fisher Vector (FV) but without spatial pyramids (cf. [12]). Before going into details regarding the experimental procedure, we give a brief overview of this state-of-the-art image signature. Details can be found in [11,12].

4.1 Image Signature

Let $X = \{x_t, t = 1 \dots T\}$ be the set of D-dimensional local descriptors extracted from a given image. Let $u_\lambda : \mathbb{R}^D \to \mathbb{R}_+$ be a pdf with parameters λ modelling the generation process of low-level descriptors in *any* image. Here, u_λ is defined to be a mixture of N Gaussians with diagonal covariances: $u_\lambda(x) = \sum_{i=1}^{N} w_i u_i(x)$, $\lambda = \{w_i, \mu_i, \sigma_i, i = 1 \dots N\}$. w_i, μ_i and σ_i^2 denote, respectively, the mixing weight, mean and variance vectors corresponding to the ith component of the mixture. The FV is defined as $\mathcal{G}_\lambda^X = L_\lambda G_\lambda^X$, where G_λ^X corresponds to the gradient of the (average) log-likelihood of X, i.e. $\frac{1}{T}\sum_{t=1}^{T} \nabla_\lambda \log u_\lambda(x_t)$ and L_λ a diagonal normalizer. The image signature is the concatenation of normalized partial derivatives, resulting in a vector of dimensionality $2ND$. Following [12], we apply the transformation $f(z) = sign(z)\sqrt{|z|}$ on each dimension and L_2-normalize the resulting vector as it was shown to improve classification accuracy.

Low-Level Features. We used 128-dimensional SIFT descriptors [9] extracted from image patches of 32×32 pixels uniformly distributed on the image (from the nodes of a regular grid with a step size of 8 pixels). We did not perform any normalization on the image patches before computations. The dimensionality of the resulting descriptors were further reduced to 80 by Principal Components Analysis (PCA). To account for variations in scale, we extracted patches at 5 different resolutions using a scale factor of 0.707 between levels.

Generative Model. We trained a GMM under a Maximum Likelihood (ML) criterion using the Expectation-Maximization (EM) algorithm. We used $1M$ random samples from the training set of PASCAL VOC2007. We initialized the EM iterations by running k-means and using the statistics of cluster assignments (relative count, mean and variance vectors) as initial estimates.

4.2 Base Classifiers

As base classifiers we used linear SVMs trained on the primal using Stochastic Gradient Descent (SGD) [1], i.e. minimizing the L_2 regularized hinge-loss in a sample-by-sample basis. The regularization parameter λ was chosen by cross-validation on the training set.

4.3 Datasets

PASCAL VOC2007. This dataset contain images of 20 object categories. The set of images for each class exhibits a large degree of intra-class variation, including changes in viewpoint, illumination, scale, partial occlusions, etc.. Images

from this dataset are split into three groups: *train*, *val* and *test*. We followed the recommended procedure of tuning parameters on the *train* set while using the *val* set for testing. Once the best choice for the parameters have been selected, the system was re-trained using the *train+val* sets. Classification performance is measured using the mean Average Precision (mAP) computed on the *test* set.

MIT Indoor Scenes. This dataset consists on more than $15K$ images depicting 67 different indoor environments. We created 10 different train/test splits by randomly selecting 50% of the images for training and 50% for testing. In order to adjust model parameters, we ran a 5-fold cross validation on the training set of the first of such splits. The best configuration was used in all runs. Classification performance was measured using the *multiclass prediction accuracy* (MPA), i.e. the mean over the diagonal elements of the confusion matrix [13]. We report the mean as well as standard deviation over runs.

5 Results

We observed that finely tuning the parameter β has little effect on performance. For the hyperparameter α, we found that a value of 0.6 was the optimal choice in most situations. We set $\alpha = 0.6$ and $\beta = 2$ in all our experiments.

Table 1 show classification performances obtained on PASCAL VOC 2007 as a function of the model complexity (the number of Gaussian components, N) for two classification schemes: one-vs-all (OVA) and our RW based approach (RWC). We compare only against OVA because it is the best performing method on this dataset[3]. It can be observed that the gain brought by our method decreases as the model complexity increases. For instance, our approach achieves a better score on 16, 15, 14, 13, 10 and 7 classes out of 20 for model complexities of $N = 8$, 16, 32, 64, 128 and 256 respectively. This seems to indicate that the proposed approach helps to ameliorate –in the final stage of the classification pipeline– the use of representations with less expressive power. For this dataset, the feature space induced by models with more than 64 Gaussians makes OVA classification a good multiclass scheme, provided this particular representation. From this point, a better performance can be expected due to a more descriptive (complex) model and not to the capabilities of the system on solving possible ambiguities between concepts.

Table 2 show the performance obtained by the OVA and RWC systems on a problem involving a larger number of classes (MIT Indoor Scenes). As before, it can be seen that the gain in performance is greater for systems based on less complex representations. In this particular case, the RWC approach allows a model with a small number of Gaussians to achieve a performance comparable to that achieved by a model using twice as many components.

[3] In preliminary experiments we also considered the use of OVO classification with voting, but its performance was consistently lower compared to the simpler and more usual OVA strategy.

Table 1. PASCAL VOC2007. Classification performance for one-vs-all (OVA) and the proposed approach (RWC), for increasing model complexity (number of Gaussians, N).

	N=8		N=16		N=32		N=64		N=128		N=256	
Class	OVA	RWC	OVA	RWC	OVA	RWC	OVA	RWC	OVA	RWC	OVA	RWC
aeroplane	69.3	**72.4**	71.0	**72.1**	72.6	**73.9**	**75.8**	75.3	**78.2**	77.5	**78.4**	77.2
bicycle	50.8	**53.7**	53.0	**54.9**	59.0	**60.4**	61.2	**62.2**	**64.7**	64.5	**65.9**	65.8
bird	31.8	**35.6**	39.2	**39.5**	43.9	**46.5**	46.0	**47.3**	45.4	**46.7**	**48.3**	47.8
boat	62.0	**63.9**	64.6	**66.1**	65.1	**66.4**	68.2	**69.0**	**68.7**	68.4	69.6	**69.7**
bottle	24.2	**24.9**	28.3	**29.0**	**32.3**	31.3	**31.1**	30.9	**31.9**	**31.9**	**33.6**	32.2
bus	54.9	**55.3**	56.5	**59.1**	59.2	**61.0**	63.7	**65.0**	64.0	**64.8**	**64.7**	**64.7**
car	70.5	**73.4**	73.2	**76.1**	75.9	**77.6**	76.7	**78.9**	78.1	**79.2**	79.9	**80.5**
cat	46.5	**46.2**	**51.0**	49.5	**54.0**	53.9	55.8	**56.8**	**56.8**	55.4	**58.6**	57.2
chair	45.3	**46.6**	43.2	**46.4**	46.6	**48.6**	48.5	**50.1**	48.4	**49.0**	49.8	**51.1**
cow	**30.0**	26.8	**35.2**	32.5	35.3	**36.6**	**41.4**	41.1	**42.5**	39.8	**45.2**	42.5
diningtable	41.0	**41.6**	42.8	**43.0**	48.0	**48.2**	50.7	**51.4**	53.7	**54.8**	**55.3**	54.7
dog	31.0	**41.9**	36.3	**44.8**	40.4	**45.1**	40.2	**46.6**	41.5	**46.1**	45.5	**48.3**
horse	**68.9**	67.9	72.8	**73.6**	**74.5**	74.3	**75.1**	74.7	**76.4**	75.6	**77.6**	76.8
motorbike	**51.3**	50.4	**57.5**	56.7	**62.3**	60.9	**64.0**	63.6	**66.5**	65.2	**65.7**	65.4
person	**76.6**	74.7	**79.2**	78.1	**81.1**	80.8	**81.5**	80.9	**82.2**	81.5	**82.6**	82.3
pottedplant	14.1	**15.7**	21.9	**25.3**	24.2	**26.5**	27.5	**28.9**	30.9	**32.0**	30.1	**31.2**
sheep	30.6	**34.4**	**38.4**	36.8	**40.6**	39.8	**40.2**	38.5	**38.9**	38.0	**43.3**	40.2
sofa	43.7	**45.1**	43.5	**45.5**	45.8	**47.4**	49.9	**51.1**	49.2	**49.8**	**51.9**	51.7
train	68.2	**70.2**	72.0	**73.6**	74.2	**76.1**	75.1	**76.2**	77.6	**77.8**	**79.2**	79.1
tvmonitor	44.6	**46.8**	46.0	**48.9**	47.6	**50.1**	49.8	**51.9**	50.8	**54.4**	51.8	**54.1**
average	47.8	**49.4**	51.3	**52.6**	54.1	**55.3**	56.1	**57.0**	57.3	**57.6**	**58.8**	58.6
gain		+1.6		+1.3		+1.2		+0.9		+0.3		-0.2

Table 2. MIT Indoor Scenes. Multiclass prediction accuracy (in %). OVA vs. RWC (*left*). Comparison with the state-of-the-art (*right*).

	N=16		N=32		N=64		N=128		N=256	
	OVA	RWC	OVA	RWC	OVA	RWC	OVA	RWC	OVA	RWC
avg.	46.3	**48.7**	48.9	**50.8**	51.2	**52.6**	52.9	**53.9**	53.6	**54.4**
s.d.	0.6	0.7	0.6	0.6	0.5	0.5	0.6	0.6	0.6	0.6
gain		+2.4		+1.9		+1.4		+1.0		+0.8

Method	MPA
OB [8]	37.6
NNbMF [2]	47.0
OVA	50.7
RWC	52.3

As a final comparison, we ran experiments using the same train/test as in [13]. We compare the OVA and RWC schemes based on Fisher vectors ($N = 128$) and simple linear classifiers against the *Object Bank* (OB) approach of Li *et al.* [8] and the *Nearest-Neighbor based Metric Functions* (NNbMF) of Çakir *et al.* [2]. Results are shown in Table 2 (right). It can be observed that the system based on FVs and linear OVA classification outperforms the state-of-the-art on this dataset and that even such a powerful representation can benefit from the proposed classifier combination scheme.

6 Conclusions and Future Work

We proposed a method to combine the scores of two common multiclass classification schemes: one-vs-all and one-vs-one. The approach is based on the stationary distribution of a Markov chain defined in the space of concepts. Results on the challenging problem of image classification showed the potentiality of our approach. In a future work we will investigate other types of graph connectivity structures, specially those leading to sparse transition matrices.

References

1. Bottou, L.: SGD, http://leon.bottou.org/projects/sgd
2. Çakir, F., Güdükbay, U., Ulusoy, Ö.: Nearest-neighbor based metric functions for indoor scene recognition. Computer Vision and Image Understanding 115(11), 1483–1492 (2011)
3. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC 2007) Results, http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html
4. Galar, M., Fernández, A., Tartas, E.B., Sola, H.B., Herrera, F.: An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. Pattern Recognition 44(8), 1761–1776 (2011)
5. García-Pedrajas, N., Ortiz-Boyer, D.: Improving multiclass pattern recognition by the combination of two strategies. IEEE Tr. on Pattern Analysis and Machine Intelligence 28(6), 1001–1006 (2006)
6. Haveliwala, T.H.: Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. IEEE Tr. on Knowledge and Data Eng. 15(4), 784–796 (2003)
7. Langville, A.N., Meyer, C.D.: Deeper inside PageRank. Internet Mathematics 1(3), 335–400 (2004)
8. Li, L.-J., Su, H., Xing, E.P., Li, F.-F.: Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: Proc. NIPS (2010)
9. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Intl. Jrnl. on Computer Vision 60(2) (2004)
10. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab (1999)
11. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: Proc. CVPR (2007)
12. Perronnin, F., Sánchez, J., Mensink, T.: Improving the Fisher Kernel for Large-Scale Image Classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010)
13. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: Proc. CVPR (2009)
14. Reid, S.R.: Model Combination in Multiclass Classification. Ph.D. thesis, Univ. of Colorado (2010)
15. Rifkin, R.M., Klautau, A.: In: defense of one-vs-all classification. Jrnl. of Machine Learning Research 5, 101–141 (2004)
16. Wu, T.F., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. Jrnl. of Machine Learning Research 5, 975–1005 (2004)