On Binary Reduction of Large-scale Multiclass Classification Problems

Bikash Joshi[†], Massih-Reza Amini[†], Ioannis Partalas[‡], Liva Ralaivola[♭],
Nicolas Usunier[⋆], Eric Gaussier[†]

[†]University of Grenoble Alpes
Grenoble Informatics Laboratory
{name.surname}@imag.fr

[‡]VISEO
R.&D. department
ioannis.partalas@viseo.com

[♭]Université Aix-Marseille
Fundamental Informatics Laboratory
liva.ralaivola@lif.univ-mrs.fr

[⋆]Université Technologique de Compiègne
Heudiasyc
nicolas.usunier@hds.utc.fr

Abstract. In the context of large-scale problems, traditional multiclass classification approaches have to deal with class imbalancement and complexity issues which make them inoperative in some extreme cases. In this paper we study a transformation that reduces the initial multiclass classification of examples into a binary classification of pairs of examples and classes. We present generalization error bounds that exhibit the interdependency between the pairs of examples and which recover known results on binary classification with i.i.d. data. We show the efficiency of the deduced algorithm compared to state-of-the-art multiclass classification strategies on two large-scale document collections especially in the interesting case where the number of classes becomes very large.

1 Introduction

The overwhelming growth of textual and visual data contents on the Web raises the issue of automatically structuring these collections into large, open-domain taxonomies. These taxonomies contain categories organized in a hierarchical structure such as a tree or a directed acyclic graph. The open directory project, maintained by roughly $90,000$ human editors, is an example of such taxonomies: it lists about 4 million websites distributed among more than 1 million classes. In that context, large-scale multiclass classification consists in assigning one class label to each document from the set of leaf nodes of the hierarchy.

In these Web-scale datasets, the classes exhibit a long-tailed distribution [1] in the sense that most of them contain very few examples. As most state-of-the-art multiclass classification approaches learn one scoring function for each class, they do not scale well to large number of classes in terms of training time, and, more importantly, they struggle with under-represented classes that tend to be never predicted. Ultimately, the predictions would be unchanged if most of the least represented classes are ignored.

In this paper, we present a new approach for multiclass classification that can deal with large number of classes with very few representative examples. The approach hinges on a theoretical analysis of algorithms that optimize ranking criteria for multiclass classification, such as those proposed in [22,17]. We provide a generalization error bound based on the Rademacher complexity for interdependent data that provides guarantees for the multiclass classification strategy based on a reduction to binary classification of pairs of couples (instance, class). The analysis suggests that the guarantees in terms of generalization performance degrades linearly with the number of classes for previous approaches that learn one parameter vector per class. To avoid this undesirable scaling of the sample complexity with respect to the number of classes, we present a new approach based on learning a combination of similarity features between instances and classes, where the similarities are computed by identifying a class with the set of its representative examples. Further, the reduction framework described above allows us to learn a single parameter vector with a dimension that does not depend on the number of classes. We empirically demonstrate that our approach is competitive with state-of-the-art multiclass classification approaches, in particular in terms of the macro F-measure, which gives higher emphasis to the correct prediction of rare classes than the classification accuracy. In addition, the number of parameters we learn is of order 10^7 times less than conventional multiclass classification models, which makes the approach appealing for large-scale classification.

In Section 2, we position our work with respect to the literature. Section 3 presents our theoretical analysis and our proposed classification strategy. The design of the features and the experimental results are in Section 4.

2 Related work

Several techniques exist to reduce multiclass problems with K classes into binary classification problems. The most popular approaches include the well-known one-versus-one (OVO), one-versus-all (OVA) [10], and Error Correcting Output Codes (ECOC) approaches. In OVO, a binary problem is created for each pair of classes of the initial problem, leading to $K(K-1)/2$ binary problems and, therefore, to as many binary classifiers. The prediction for a new instance is the class which receives the majority of the votes. In OVA, K binary problems are created, each one being associated to a specific class seen as the positive class and the other as forming the negative class. Given real-valued predictors $g_1, \ldots, g_K$, the predicted class for an instance x is given by $\arg\max_y g_y(x)$.

In the ECOC-based approach, a binary code $\mathbf{c}_k$ of length L is assigned to each class k, giving rise to L binary classification problems. One binary predictor is learned for each of the L induced binary problems and, at prediction time, inference is performed by selecting the class that minimizes the Hamming distance between its code and the predicted code. Methods to speed up prediction or training with ECOC have recently been proposed: for example, only a subset of the classifiers may be used at inference time without loss of accuracy [13]; in another direction, a Naive Bayes approach that only requires a single pass over the data for training has proved effective [14].

Methods that achieve logarithmic-time prediction or training have been proposed in [2,3]: they rest on binary tree structures where each leaf corresponds to a class and inference is performed by traversing the tree from top to bottom, a binary classifier being used at each node to determine the child node to develop.

Ranking approaches to multiclass classification [22,17], or the constraint classification framework of [6] can be seen as a reduction using binary classifications of pairs of classes (given an example), similar to ours. The proposed reduction strategy allows to obtain new generalization error bounds, and lead to a different algorithm. While state-of-the-art approaches learn one scoring function per class, and thus have similar computational and sample complexities similar to the `OVA` approach, we design similarity features between classes and examples allowing to learn a single parameter vector for the whole problem.

A similar approach for learning representations of classes was also proposed in [21]. The latter learns a projection of examples and classes into a low dimensional space, which reduces both training and inference time. In contrast to our approach, the aforementioned learns one parameter vector per class, while we use joint features of classes and examples allowing to reduce the number of vector parameters to one.

3 Multiclass to Binary reduction

3.1 Framework

We consider monolabel multiclass classification problems defined on a joint space $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \subseteq \mathbb{R}^d$ is the *input space* and $\mathcal{Y} = \{1, \ldots, K\}$ the *output space*, made of K class labels. Elements of $\mathcal{X} \times \mathcal{Y}$ are denoted as $\mathbf{x}^y = (x, y)$. Furthermore, we assume the training set $\mathcal{S} = (\mathbf{x}_i^{y_i})_{i=1}^m$ is made of i.i.d pairs distributed according to a fixed but unknown probability distribution $\mathcal{D}$, and we consider a class of functions $\mathcal{G} = \{g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}\}$ as our predictors. We define the instantaneous loss of $g \in \mathcal{G}$ on an example $\mathbf{x}^y$ as:

$$e(g, \mathbf{x}^y) = \frac{1}{K-1} \sum_{y' \in \mathcal{Y} \setminus \{y\}} \mathbb{1}_{g(\mathbf{x}^y) \leq g(\mathbf{x}^{y'})}, \tag{1}$$

where $\mathbb{1}_\pi$ is the indicator function that is equal to 1 if the predicate π is true and 0 otherwise. Compared to the classical multiclass error:

$$e'(g, \mathbf{x}^y) = \mathbb{1}_{y \neq \text{argmax}_{y' \in \mathcal{Y}} g(\mathbf{x}^{y'})},$$

the loss of (1) estimates the average number of classes, given any input data, that get a greater scoring by g than the correct class. The loss (1) is hence a *ranking* criterion, and the multiclass `SVM` of [22] and AdaBoost.MR [17] optimize convex surrogate functions of this loss. This is also used in label ranking [7], where the task is to predict a ranking of all labels instead of predicting a single label y given an instance x. The multiclass classification problem we are going to study is that of finding a function $g \in \mathcal{G}$ using the labeled training set $\mathcal{S}$ with small generalization error $L(g)$:

$$L(g) = \mathbb{E}_{\mathbf{x}^y \sim \mathcal{D}} \left[e(g, \mathbf{x}^y) \right]. \tag{2}$$

Accordingly, the empirical error of $g \in \mathcal{G}$ over $\mathcal{S}$ is

$$\hat{L}_m(g, \mathcal{S}) = \frac{1}{m} \sum_{i=1}^m \underbrace{(\frac{1}{K-1} \sum_{y' \in \mathcal{Y} \setminus \{y\}} \mathbb{1}_{g(\mathbf{x}_i^y) \leq g(\mathbf{x}_i^{y'})})}_{e(g, \mathbf{x}_i^{y_i})} \tag{3}$$

3.2 Reduction Strategy

We further work out the empirical loss of Equation (3) in order to *i*) have it ressemble a more usual binary classification loss with, in particular, a single sum running over only one index, *ii*) make apparent the need of dealing with non-i.i.d. random variables and *iii*) after a theoretical introduction, set the stage for our practical binary reduction approach.

A first step to reshape the empirical loss is to see that the instantaneous loss (1) can be rewritten as

$$e(g, \mathbf{x}^y) = \frac{1}{K-1} \sum_{y' \in \mathcal{Y} \setminus \{y\}} \mathbb{1}_{\tilde{y} h(\mathbf{x}^y, \mathbf{x}^{y'}) \leq 0},$$

where h is defined as $h(\mathbf{x}^y, \mathbf{x}^{y'}) = g(\mathbf{x}^y) - g(\mathbf{x}^{y'})$. This bears strong resemblance with a binary-classification-loss-based risk, a resemblance that can be strengthened by introducing the transformed set $T(\mathcal{S})$ of size $n = m(K-1)$ defined as

$$T(\mathcal{S}) = \left\{ (\boldsymbol{Z}_j, \tilde{y}_j) : j = 1, \ldots, n \right\}, \tag{4}$$

where each $\boldsymbol{Z}_j$ is one of the pairs $(\mathbf{x}_i^y, \mathbf{x}_i^{y'})$, and $\tilde{y}_j = 1$ if the first observation in $\boldsymbol{Z}_j$ is constituted by an example $\mathbf{x}_i$ and its true class in $\mathcal{S}$ (i.e. $y = y_i$) and the second observation is constituted by the same example and any other of the $K-1$ classes; and $\tilde{y}_j = -1$ otherwise (i.e. if the order is reverse). This allows us to rewrite the empirical loss of (3) as

$$L_n^T(h, T(\mathcal{S})) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\tilde{y}_j h(\boldsymbol{Z}_j) \leq 0}. \tag{5}$$

With these definitions at hand, it is clear that the selection of a hypothesis in $\mathcal{G}$ minimizing the empirical risk of (3) over the training set $\mathcal{S}$, is equivalent to the search of a hypothesis in $\mathcal{H} = \{h : h(\mathbf{x}^y, \mathbf{x}^{y'}) = g(\mathbf{x}^y) - g(\mathbf{x}^{y'}), g \in \mathcal{G}\}$ minimizing the empirical risk of (5) over $T(\mathcal{S})$. However, even if the examples in $\mathcal{S}$ are i.i.d., the examples in $T(\mathcal{S})$ are no longer independent since the same observations $\mathbf{x}^y \in \mathcal{S}$ are involved in different pairs of $T(\mathcal{S})$. Thus, in order to obtain generalization error bounds $L_n^T(h, T(\mathcal{S}))$ we need to address the issue of learning with interdependent data.

There exist several ways to tackle this problem among which two settings received particular attention in the literature. The first one deals with learning from mixing processes, where the dependency between random variables decreases over time [12,18]. The second direction, on which the present work is based on, is developed around the idea of graph coloring that divides a graph, representing the relations between random variables, into sets of independent random variables called proper cover of the graph [8].

A proper cover of $T(\mathcal{S})$ is constituted of $K-1$ disjoint sets $(C_k)_{k=1}^{K-1}$ each containing m independent examples. For all $k \in \{1, \ldots, K-1\}$ it is defined as

$$C_k = \{(\boldsymbol{Z}_{k+j(K-1)}, \tilde{y}_{k+j(K-1)}); j \in \{0, \ldots, m-1\}\}$$

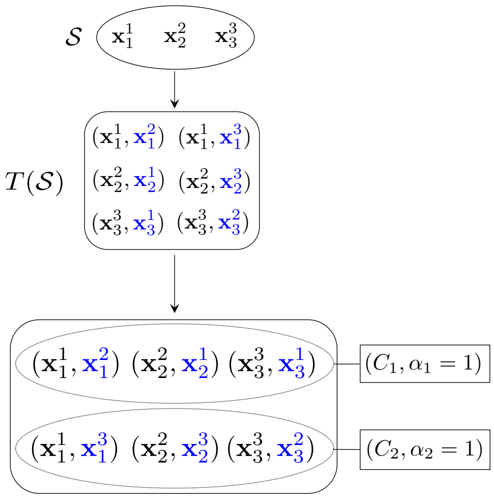
$\mathcal{S}$
$\mathbf{x}_1^1$ $\mathbf{x}_2^2$ $\mathbf{x}_3^3$
$T(\mathcal{S})$
$(\mathbf{x}_1^1, \mathbf{x}_1^2)$ $(\mathbf{x}_1^1, \mathbf{x}_1^3)$
$(\mathbf{x}_2^2, \mathbf{x}_2^1)$ $(\mathbf{x}_2^2, \mathbf{x}_2^3)$
$(\mathbf{x}_3^3, \mathbf{x}_3^1)$ $(\mathbf{x}_3^3, \mathbf{x}_3^2)$
$(\mathbf{x}_1^1, \mathbf{x}_1^2)$ $(\mathbf{x}_2^2, \mathbf{x}_2^1)$ $(\mathbf{x}_3^3, \mathbf{x}_3^1)$
$(C_1, \alpha_1 = 1)$
$(\mathbf{x}_1^1, \mathbf{x}_1^3)$ $(\mathbf{x}_2^2, \mathbf{x}_2^3)$ $(\mathbf{x}_3^3, \mathbf{x}_3^2)$
$(C_2, \alpha_2 = 1)$

Fig. 1. The proper exact fractional cover of the set $T(\mathcal{S})$ obtained after transformation of the training set $\mathcal{S} = \{\mathbf{x}_1^1, \mathbf{x}_2^2, \mathbf{x}_3^3\}$. For the sake of clarity, the class labels of pairs of examples are omitted. The fractional chromatic number of T is in this case $\chi_T^* = 2$.

Moreover, $(C_k, \alpha_k)_{k=1}^{K-1}$ is said to be a proper exact fractional cover of $T(\mathcal{S})$, if $(C_k)_{k=1}^{K-1}$ is a proper cover of $T(\mathcal{S})$ and if $\forall k, \alpha_k > 0$ and

$$\forall i \in \{1, \ldots, n\}, \sum_{k=1} \alpha_k \mathbb{1}_{(\boldsymbol{Z}_i, \tilde{y}_i) \in C_k} = 1.$$

The fractional chromatic number of T, denoted as χ_T^* is then the minimum sum of weights, or the minimum number of sets containing each independent random variables, which for the proposed transformation is equal to $K - 1$. Figure 1 depicts the transformation and its associated proper exact fractional on a toy problem.

Using graph coloring arguments, [8] extended Hoeffding's inequality to sums of interdependent random variables and based on that result, different studies proposed new generalization error bounds for learning with interdependent data, thus proving the consistency of the ERM principle for this case [20,16]. Here we build on [20] who proposed a generalization of [11] concentration inequality to the case of interdependent random variables.

Our theoretical result is the following theorem which provides data-dependent bound on the generalization error of the multiclass classifier (Eq. 2). This result is at the basis of the algorithm for the binary classification of pairs of examples that we expose in the next section. We consider here kernel-based hypotheses with $\kappa : \mathcal{Z} \to \mathbb{R}$ a *positive semidefinite* (PSD) kernel and $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{H}$ its associated feature mapping function, defined as:

$$\mathcal{G}_B = \{\mathbf{x}^y \in \mathcal{X} \times \mathcal{Y} \mapsto \langle \boldsymbol{w}, \Phi(\mathbf{x}^y) \rangle \mid ||\boldsymbol{w}|| \leq B\} \tag{6}$$

where $\boldsymbol{w}$ is the weight vector defining the kernel-based hypotheses and $\langle\cdot,\cdot\rangle$ denotes the dot product. We further define the following associated function class:

$$\mathcal{H}_B = \{(\mathbf{x}^y, \mathbf{x}'^{y'}) \in \mathcal{Z} \mapsto g_w(\mathbf{x}^y) - g_w(\mathbf{x}'^{y'}) \mid g_w \in \mathcal{G}_B\}.$$

Theorem 1 *Let $\mathcal{S} = (\mathbf{x}_i^{y_i})_{i=1}^m \in (\mathcal{X} \times \mathcal{Y})^m$ be a dataset of m examples drawn i.i.d. according to a probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and $T(\mathcal{S}) = ((\boldsymbol{Z}_i, \tilde{y}_i))_{i=1}^n \in (\mathcal{Z} \times \{-1,1\})^n$ the transformed set obtained with the transformation function T defined above. Further let $\kappa : \mathcal{Z} \to \mathbb{R}$ be a PDS kernel, and let $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{H}$ be the associated feature mapping function. Then for all $1 > \delta > 0$ with probability at least $(1-\delta)$ over $T(\mathcal{S})$ the following generalization bound holds for all $h_w \in \mathcal{H}_B$:*

$$L^T(h_w) \leq \epsilon_n^T(h_w, T(\mathcal{S})) + \frac{2B\mathfrak{G}(T(\mathcal{S}))}{m\sqrt{K-1}} + 3\sqrt{\frac{\ln(\frac{2}{\delta})}{2m}} \tag{7}$$

where $\epsilon_n^T(h, T(\mathcal{S})) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\tilde{y}_i h_w(\boldsymbol{Z}_i))$ with the surrogate Hinge loss $\mathcal{L} : t \mapsto \min(1, \max(1-t, 0))$, $L^T(h_w) = \mathbb{E}_{T(\mathcal{S})}[L_n^T(h_w, T(\mathcal{S}))]$ and $\mathfrak{G}(T(\mathcal{S})) = \sqrt{\sum_{i=1}^n d_\kappa(\boldsymbol{Z}_i)}$ with

$$d_\kappa(\mathbf{x}^y, \mathbf{x}^{y'}) = \kappa(\mathbf{x}^y, \mathbf{x}^y) + \kappa(\mathbf{x}^{y'}, \mathbf{x}^{y'}) - 2\kappa(\mathbf{x}^y, \mathbf{x}^{y'})$$

Proof. Exploiting the fact that $\mathcal{L}$ dominates the $0/1$ loss and using the fractional Rademacher data-dependent generalization bound proposed for interdependent data in Theorem 4 of [20] one has

$$L^T(h_w) \leq \epsilon^T(h_w) \leq \hat{\epsilon}_n^T(h_w, T(\mathcal{S})) + \hat{\mathcal{R}}_n^T(\mathcal{L} \circ \mathcal{H}_B, \mathcal{S}) + 3\sqrt{\frac{\chi_T^* \ln(\frac{2}{\delta})}{2n}}$$

Where $\epsilon^T(h_w) = \mathbb{E}_{T(\mathcal{S})}[\hat{\epsilon}_n^T(h_w, T(\mathcal{S}))]$ and $\hat{\mathcal{R}}_n^T(\mathcal{L} \circ \mathcal{H}_B, \mathcal{S})$ is the empirical fractional Rademacher complexity of $\mathcal{L} \circ \mathcal{H}_B$ on $T(\mathcal{S})$. Further, as $\mathcal{L}$ is 1-Lipschitz, so

$$\hat{\mathcal{R}}_n^T(\mathcal{L} \circ \mathcal{H}_B, \mathcal{S}) \leq \hat{\mathcal{R}}_n^T(\mathcal{H}_B, \mathcal{S})$$

where

$$\hat{\mathcal{R}}_n^T(\mathcal{H}_B, \mathcal{S}) = \sum_{k=1}^{K-1} \frac{2\alpha_k}{M} \mathbb{E}_\sigma \sup_{h \in \mathcal{H}_B} \sum_{j=0}^{m-1} \sigma_j h_w(\boldsymbol{Z}_{k+j(K-1)})$$

Now, for all $k \in \{1, .., K-1\}$ and $j \in \{0, .., m-1\}$, let z_{kj} and z'_{kj} be the first and the second pair of $\boldsymbol{Z}_{k+j(K-1)}$, then from the bilinearity of dot product and the Cauchy-Schwartz inequality, $\hat{\mathcal{R}}_n^T(\mathcal{H}_B, \mathcal{S})$ is upper-bounded by

$$\begin{aligned} &\sum_{k=1}^{K-1} \frac{2\alpha_k}{n} \mathbb{E}_\sigma \sup_{h_w \in \mathcal{H}_B} \left\langle \boldsymbol{w}, \sum_{j=0}^{m-1} \sigma_j (\Phi(z_{kj}) - \Phi(z'_{kj})) \right\rangle \\ &\qquad \leq \sum_{k=1}^{K-1} \frac{2B\alpha_k}{n} \mathbb{E}_\sigma \left\| \sum_{j=0}^{m-1} \sigma_j (\Phi(z_{kj}) - \Phi(z'_{kj})) \right\| \end{aligned}$$

Further, for all $i, j \in \{0, \dots, m-1\}^2, i \neq j$, we have $\mathbb{E}_\sigma[\sigma_i \sigma_j] = 0$ so

$$\begin{aligned} \hat{\mathcal{R}}_n^T(\mathcal{H}_B, \mathcal{S}) &\leq \sum_{k=1}^{K-1} \frac{2B\alpha_k}{n} \sqrt{\sum_{j=0}^{m-1} d_\kappa(z_{kj}, z'_{kj})} \\ &= \frac{2B\chi_T^*}{n} \sum_{k=1}^{K-1} \frac{\alpha_k}{\chi_T^*} \sqrt{\sum_{j=0}^{m-1} d_\kappa(z_{kj}, z'_{kj})} \end{aligned}$$

Now as $\sum_{k=1}^{K-1} \frac{\alpha_k}{\chi_T^*} = 1$ and that $t \mapsto \sqrt{t}$ is concave, from Jensen inequality we have

$$\hat{\mathcal{R}}_m^T(\mathcal{H}_B, \mathcal{S}) \leq \frac{2B\chi_T^*}{n} \sqrt{\sum_{k=1}^{K-1} \frac{\alpha_k}{\chi_T^*} \sum_{j=0}^{m-1} d_\kappa(z_{kj}, z'_{kj})}$$

The result follows from rearranging the examples and the equalities $\chi_T^* = K - 1$, and $n = (K-1)m$.

3.3 Favoring Low-Dimensional Feature Maps

Our reduction relies on the joint representation $\Phi(\mathbf{x}^y)$ of features and classes. Such feature maps are at the basis of algorithms such as structured SVM (see e.g. [19]), to account for features encoding properties of structures such as sequences or trees. However, in multiclass classification, the output space is unstructured and these algorithms are then applied by taking a "‘trivial"’ feature map such that even if a single parameter vector is used, it is in fact the concatenation of one parameter vector per class. In that case, $\Phi(\mathbf{x}^k) \in \mathbb{R}^{dK}$ (with $\mathbf{x} \in \mathbb{R}^d$) is a vector where all entries are zero except those with indices in the range $[1+(k-1)d; kd]$, which are equal to $\mathbf{x}$. The reduction of multiclass classification to constraint classification of [6] follows the same idea. With this kind of joint (instance, class) representation, the natural regularization is to constrain each parameter vector to have a norm smaller than some B. The whole vector $\boldsymbol{w}$ would then have a norm about KB, leading the capacity term of Theorem 1, $\mathfrak{G}(T(\mathcal{S}))$, to grow linearly with K. To avoid this linear detorioration of the generalization performance guarantees, we might choose to put a stronger regularization on some classes, e.g. the rare classes. But then these heavily regularized classes would be penalized because the magnitude of their predicted scores would be smaller: they would rarely or never be predicted. We propose to give an alternative answer to avoid the dependence of the penalty term on K. We advocate the design of a non-trival joint feature representation $\Phi(\mathbf{x}^y)$ by using a small number of adequatley chosen similarity features between examples and classes, so that this joint feature space is the same for any number of classes. The goal of learning is then to combine these features, using the same parameter vector for all classes. Then, the natural scaling of the penalty term of Theorem 1 should remain constant, and the detrimental effect of having stronger regularization on certain classes disappears. The proposed approach denoted by `mRb`, for multiclass reduced to binary classification, is hence depicted in Algorithm 1. As the learned classifier from the function class $\mathcal{G}_B$ is linear in the feature space, the output of a function $h \in \mathcal{H}_B$ over

Algorithm 1: Multiclass reduced to binary classification (mRb)

```
Input: Labeled training set $\mathcal{S} = (\mathbf{x}_i^{y_i})_{i=1}^{m}$ ;
A binary classifier $\mathcal{A}$ ;
Initialize
$T(S) \leftarrow \emptyset$ ;
for $i = 1..m$ do
    for $k = 1..K$ do
        if $y_i > k$ then
            $T(S) \leftarrow \{(\Phi(\mathbf{x}_i^{y_i}) - \Phi(\mathbf{x}_i^{k}), +1)\}$;
        end
        if $y_i < k$ then
            $T(S) \leftarrow \{(\Phi(\mathbf{x}_i^{k}) - \Phi(\mathbf{x}_i^{y_i}), -1)\}$;
        end
    end
end
Learn $\mathcal{A}$ on $T(S)$;
```

an example $(\mathbf{x}^{y}, \mathbf{x}^{y'})$ can be computed as the dot product between the learned weight vector, $\boldsymbol{w}$, and the difference between the vector representations $\Phi(\mathbf{x}^{y}) - \Phi(\mathbf{x}^{y'})$. For testing a new example $\mathbf{x}'$, we estimate $\Phi(\mathbf{x}'^{y})$ for all $\mathbf{x}'^{y}$ pairs. Given the learned weight vector $\boldsymbol{w}$, the predicted class is the one which maximizes the dot product $\langle \boldsymbol{w}, \Phi(\mathbf{x}'^{y}) \rangle$.

4 Experiments

We use non-trivial joint feature representation, which is popularly used in text classification domain. So, we evaluate the proposed method for multi-class classification in a large-scale scenario using DMOZ and Wikipedia datasets of the Large Scale Hierarchical Text Classification challenge (LSHTC 2011) [15]. These datasets contain 27875 and 36504 categories respectively for DMOZ and Wikipedia and they are provided in a pre-processed format using stop-word removal and stemming. The dimension of the vectorial space (d), the size of the training set (m) and the test set are respectively 594158, 394756 and 104263 for DMOZ and 346299, 456886 and 81262 for Wikipedia. For each of these datasets we randomly draw several samples with increasing number of classes: 100, 500, 1000, 3000, 5000 and 7500 and by keeping the same proportion of examples in the training and the test sets than in the initial collections. For the feature mapping, we used the following features in the vector representation of $\Phi(\mathbf{x}^{y})$ (table 1) by considering a class y as a mega-document, constituted by the concatenation of all of the documents in the training set belonging to it. Almost all the features, except 9 and 10, are classical features employed in learning to rank by assimilating a class and a document to respectively a document and a query. The former two are the distance of the example x to its two nearest neighbours in class y. Since the absolute values of each feature for the documents are different and not comparable, we normalize them such that the feature values are confined within the range of 0 to 1. Following our theoretical result, we used SVM with linear kernel as our binary classification

Table 1. Let x_t represent the term frequency of term t in document x, and $\mathcal{V}$ the set of distinct terms within $\mathcal{S}$, then $y_t = \sum_{x \in y} x_t$, $|y| = \sum_{t \in \mathcal{V}} y_t$, $\mathcal{S}_t = \sum_{x \in \mathcal{S}} x_t$, $l_{\mathcal{S}} = \sum_{t \in \mathcal{V}} \mathcal{S}_t$. I_t is the inverse document frequency of term t, and $d_1(\mathbf{x}^y)$ and $d_2(\mathbf{x}^y)$ are the distances of x to its two nearest neighbours in class y.

Features in the vector representation of $\Phi(\mathbf{x}^y)$.			
1. $\sum_{t \in y \cap x} \ln(1 + y_t)$	**2.** $\sum_{t \in y \cap x} \ln(1 + \frac{l_{\mathcal{S}}}{\mathcal{S}_t})$	**3.** $\sum_{t \in y \cap x} I_t$	**4.** $\sum_{t \in y \cap x} \frac{y_t}{\|y\|}.I_t$
5. $\sum_{t \in y \cap x} \ln(1 + \frac{y_t}{\|y\|})$	**6.** $\sum_{t \in y \cap x} \ln(1 + \frac{y_t}{\|y\|}.I_t)$	**7.** $\sum_{t \in y \cap x} \ln(1 + \frac{y_t}{\|y\|}.\frac{l_{\mathcal{S}}}{\mathcal{S}_t})$	**8.** $\sum_{t \in y \cap x} 1$
9. $d_1(\mathbf{x}^y)$	**10.** $d_2(\mathbf{x}^y)$		

algorithm. The value of the hyperparameter C is chosen from a range of values from 10^{-3} to 10^3 by cross-validation. We compared the proposed approach, `mRb` (Figure 1), with the hierarchical reduction approach (`LogT`) proposed by [3] and the following multiclass classification techniques using the LibLinear package [5] that implements them all: One Vs. All (`OVA`), One Vs. One (`OVO`) and Multiclass `SVM` (`M-SVM`) proposed by [4]. For all of these methods we adopted the *tfidf* encoding of features as it provided the best performance. Results are evaluated over the test set using the accuracy and the macro F1 measure ($\texttt{MaF}_1$), which is the harmonic average of macro precision and macro recall. The reported performance is averaged over 50 random (train/test) sets of the initial collection for every fixed number of classes we considered. In all of our experiments, we used a server with an intel Xenon 1.8HGz processor and 16GB of RAM.

Table 2. Accuracy, $\texttt{MaF}_1$ of methods that could be trained with 7500 classes of `DMOZ` and `Wikipedia` collections. N_c is the proportion of classes that are covered. Statistics are given over 50 random samples of training/test sets.

	`DMOZ`-7500			`Wikipedia`-7500		
	Accuracy	$\texttt{MaF}_1$	N_c	Accuracy	$\texttt{MaF}_1$	N_c
`mRb`	$0.499^{\downarrow}\pm.011$	$\mathbf{0.352} \pm .009$	0.495	$0.467^{\downarrow}\pm.023$	$\mathbf{0.378} \pm .012$	0.551
`OVA`	$\mathbf{0.549}\pm.036$	$0.282^{\downarrow}\pm.018$	0.379	$\mathbf{0.484}\pm.029$	$0.348^{\downarrow}\pm.017$	0.489
`LogT`	$0.311^{\downarrow}\pm.034$	$0.096^{\downarrow}\pm.029$	0.194	$0.231^{\downarrow}\pm.035$	$0.151^{\downarrow}\pm.021$	0.287

We start our evaluation by analyzing the performance measures of different approaches on the setting with the largest number of classes we considered in our experiments ($K = 7500$). Table 2 summarizes results obtained by `mRb`, `OVA` and `LogT`, as the corresponding training processes of `M-SVM` and `OVO` were killed by the system and did not pass the scale. Results are averaged over 50 random splits of tests sets. We use bold face to indicate the highest performance rates, and the symbol ↓ indicates that performance is significantly worse than the best result, according to a Wilcoxon rank sum test used

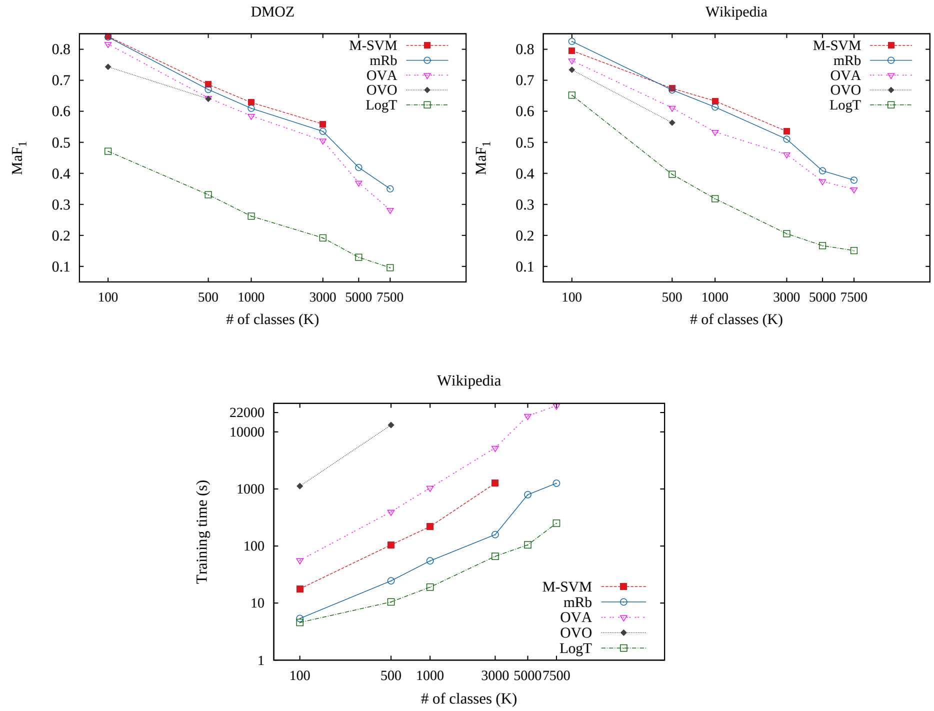

Fig. 2. MaF$_1$ of all methods with respect to the number of classes for DMOZ (top left) and Wikipedia (top right). Training time in seconds of all methods with respect to the number of classes for Wikipedia (bottom).

at a p-value threshold of 0.01 [9]. The competitive methods are OVA and mRb with a discrepancy over their accuracy and MaF$_1$ measures on both collections. To analyze this divergence we estimated the proportion of classes that have been covered, or for which at least one true positive document was found. It comes out that mRb covers 6% to 12% more classes than OVA (that is 465 to 900 more classes on both datasets). The reason here is that OVA is affected by the class imbalance problem especially in the extreme case where classes contain very few documents. For the large scale scenario this problem is accentuated as the class distribution is long-tailed, as for example in DMOZ-7500, more than half of the classes contain less than 5 documents (Figure 3). We also analyze the behavior of the various algorithms for increasing number of classes. Figure 2 (top) illustrates this by showing the MaF$_1$ measures on DMOZ and Wikipedia with respect to the number of classes. As expected all performance curves decrease monotonically with respect to an increasing number of classes. The breaking points beyond which OVO and M−SVM cannot be trained, happen at the same time on both collections for respectively $K = 500$ and $K = 3000$ classes. The performance of mRb are in between of those of OVA and M−SVM before the breaking point, with a slight advantage for M−SVM, while mRb uniformly outperforms OVA with a larger gap on Wikipedia. We notice

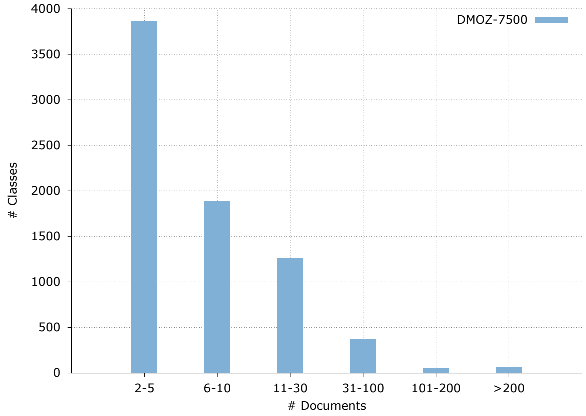

Fig. 3. Distribution of classes with respect to the number of documents they contain for DMOZ-7500.

that on this collection, mRb achieves for 7500 classes $\mathtt{MaF}_1$ score comparable to the OVA's one for 5000 classes. Comparatively, for $K = 3000$, the numbers of parameters of these two models are roughly 5.4×10^8 to 6.5×10^8 on respectively Wikipedia and DMOZ collections which are $O(10^7)$ with respect to the fixed number of parameters of mRb we have. Figure 2 (bottom) summarizes the training time of all methods for an increasing number of classes on Wikipedia. mRb has the second fastest running time after LogT which together with its small number of parameters and its performance makes it appealing for classification in large-scale taxonomies.

5 Conclusion

We presented a new method for large-scale multiclass classification based on a reduction of multiclass classification to binary classification. The theoretical analysis based on the fractional Rademacher complexity shows that learning a single scoring function for all classes, instead of one scoring function per class, avoids the capacity term to grow linearly with the number of classes, contrarily to existing methods. In addition, to have better scalability than existing methods, the features that we designed to jointly represent classes and documents improved the covering of rare classes compared to its counterparts, which is also depicted on $\mathtt{MaF}_1$ score.

Acknowledgments

This work is partially supported by the LabEx PERSYVAL-Lab ANR-11-LABX-0025, and Titan CNRS-Mastodons.

References

1. Babbar, R., Metzig, C., Partalas, I., Gaussier, E., Amini, M.R.: On power law distributions in large-scale taxonomies. SIGKDD Explorations 16(1) (2014)
2. Beygelzimer, A., Langford, J., Ravikumar, P.: Error-correcting tournaments. In: Proceedings of the 20th International Conference on Algorithmic Learning Theory. pp. 247–262. ALT'09 (2009)
3. Choromanska, A., Langford, J.: Logarithmic time online multiclass prediction. CoRR abs/1406.1822 (2014)
4. Crammer, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. J. Mach. Learn. Res. 2, 265–292 (2002)
5. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. J. Mach. Learn. Res. 9, 1871–1874 (2008)
6. Har-peled, S., Roth, D., Zimak, D.: Constraint classification: A new approach to multiclass classification and ranking. In: In Advances in Neural Information Processing Systems 15. pp. 365–379 (2002)
7. Hüllermeier, E., Fürnkranz, J.: On minimizing the position error in label ranking. In: Machine Learning: ECML 2007, pp. 583–590. Springer (2007)
8. Janson, S.: Large deviations for sums of partly dependent random variables. Random Structures and Algorithms 24(3), 234–248 (2004)
9. Lehmann, E.: Nonparametric Statistical Methods Based on Ranks. McGraw-Hill, New York, USA (1975)
10. Lorena, A.C., Carvalho, A.C., Gama, J.a.M.: A review on the combination of binary classifiers in multiclass problems. Artif. Intell. Rev. 30(1-4), 19–37 (2008)
11. McDiarmid, C.: On the method of bounded differences. Survey in Combinatorics pp. 148–188 (1989)
12. Mohri, M., Rostamizadeh, A.: Rademacher complexity bounds for non-i.i.d. processes. In: Advances in Neural Information Processing Systems 21. pp. 1097–1104 (2009)
13. Park, S.H., Fürnkranz, J.: Efficient prediction algorithms for binary decomposition techniques. Data Mining and Knowledge Discovery 24(1), 40–77 (2012)
14. Park, S., Fürnkranz, J.: Efficient implementation of class-based decomposition schemes for naïve bayes. Machine Learning 96(3), 295–309 (2014)
15. Partalas, I., Kosmopoulos, A., Baskiotis, N., Artieres, T., Paliouras, G., Gaussier, E., Androutsopoulos, I., Amini, M.R., Galinari, P.: LSHTC: A Benchmark for Large-Scale Text Classification. ArXiv e-prints (Mar 2015)
16. Ralaivola, L., Szafranski, M., Stempfel, G.: Chromatic PAC-Bayes Bounds for Non-IID Data: Applications to Ranking and Stationary β-Mixing Processes. Journal of Machine Learning Research 11, 1927–1956 (2010)
17. Schapire, R.E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. Machine learning 37(3), 297–336 (1999)
18. Steinwart, I., Christmann, A.: Fast learning from non-i.i.d. observations. In: Advances in Neural Information Processing Systems 22. pp. 1768–1776 (2010)
19. Tsochantaridis, I., Hofmann, T., Joachims, T., Altun, Y.: Support vector machine learning for interdependent and structured output spaces. In: Proceedings of the twenty-first international conference on Machine learning. p. 104. ACM (2004)
20. Usunier, N., Amini, M.R., Gallinari, P.: Generalization error bounds for classifiers trained with interdependent data. In: Advances in Neural Information Processing Systems 18. pp. 1369–1376 (2006)
21. Weston, J., Bengio, S., Usunier, N.: Wsabie: Scaling up to large vocabulary image annotation. In: Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI (2011)

22. Weston, J., Watkins, C.: Multi-class support vector machines. Tech. rep., Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London (1998)