Multi-class to Binary reduction of Large-scale classification Problems

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

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

Large-scale multi-class classification problems have gained increased popularity in recent time mainly because of the overwhelming growth of textual and visual data in the web. However, this is a challenging task for many reasons. The main challenges in Large-scale classification problems are: scalability, complexity of model and class imbalance problem. In this work, we present an algorithm for binary reduction of multi-class classification problems, which aims at addressing the above-mentioned challenges.

Let us consider *input space* $\mathcal{X}$, *output space* $\mathcal{Y}$, class labels K, class of predictor functions $\mathcal{G}$ and $\mathcal{S} = (\mathbf{x}_i^{y_i})_{i=1}^m$ training examples made of i.i.d pairs. We define the instantaneous loss of $g \in \mathcal{G}$ on an example $\mathbf{x}^y$ as:

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

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

$$\hat{L}_m(g, \mathcal{S}) = \frac{1}{m} \sum_{i=1}^{m} e(g, \mathbf{x}_i^{y_i}) \tag{2a}$$

$$= \frac{1}{m(K-1)} \sum_{i=1}^{m} \sum_{y' \in \mathcal{Y} \setminus \{y_i\}} \mathbb{1}_{g(\mathbf{x}_i^{y}) \leq g(\mathbf{x}_i^{y'})} \tag{2b}$$

This can be rewritten as:

$$\hat{L}_m(g, \mathcal{S}) = \frac{1}{m(K-1)} \sum_{i=1}^{m} \sum_{y' \in \mathcal{Y} \setminus \{y_i\}} \mathbb{1}_{\tilde{y} h(\mathbf{x}^y, \mathbf{x}^{y'}) \leq 0} \tag{3}$$

Where, h is defined as $h(\mathbf{x}^y, \mathbf{x}^{y'}) = g(\mathbf{x}^y) - g(\mathbf{x}^{y'})$. and $\tilde{y}_j = 1$

	DMOZ-7500			Wikipedia-7500		
	Accuracy	$\mathtt{MaF_1}$	N_c	Accuracy	$\mathtt{MaF_1}$	N_c
mRb	$0.499^{\downarrow}{}_{\pm.011}$	**0.352** $\pm$ .009	0.495	$0.467^{\downarrow}{}_{\pm.023}$	**0.378** $\pm$ .012	0.551
OVA	**0.549**$_{\pm.036}$	$0.282^{\downarrow}{}_{\pm.018}$	0.379	**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

Table 1. Accuracy, $\mathtt{MaF_1}$ 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.

Equation 3 resembles binary classification loss-based risk or in other words selecting hypothesis G minimizing risk over S is equivalent to search a hypothesis in H minimizing risk over a transformed set $T(\mathcal{S})$ of size $n = m(K-1)$. This forms the main foundation of our proposed reduction algorithm.

Also, we discuss the generalization error bound based on Rademacher complexity for interdependent data. The analysis of the Rademacher complexity based bound shows linear degradation of generalization performance for traditional approaches which learn one parameter vector for each class. Whereas in our work, we learn combination of similarity features between instances and classes, which results in data-dependent bound for our proposed algorithm. This non-trivial feature representation remains same for any number of classes. So the goal of learning is to combine these features using same parameter vector for all classes. Empirically, we evaluate the performance of our proposed algorithm on multi-class document classification. The datasets used in the experiments are DMOZ and Wikipedia of Large Scale Hierarchical Text Classification Challenge (LSHTC 2011) [3]. Further, we randomly drew samples from both datasets with increasing number of classes: 100, 500, 1000, 3000, 5000 and 7000. We adequately chose 10 similarity features between documents and class of documents. SVM with linear kernel was used as our binary classification algorithm. The results of the proposed algorithm were compared with hierarchical reduction approach (LogT) [1] and LibLinear package [2] implementation of One Vs. All, One Vs. One and Multiclass SVM (M-SVM) [4]. To compare the results Accuracy, Macro F-Measure and training time were used as evaluation measures. Additionally, N_c was used to evaluate the proportion of classes that were covered in the classification result.

The results for maximum number of class case (i.e. 7500) is shown in Table 1. Algorithms M-SVM and OVO are not presented in the table as they were not able to scale up for high number of classes because of their high complexity. As can be seen in the result mRb (proposed reduction algorithm) significantly outperforms OVA and LogT in terms of Macro F-Measure which is intuitively better evaluation measure than Accuracy for large-scale multi-class classification problems.

In conclusion, this work presents binary reduction of multi-class classification problems. Analysis based on Rademacher complexity shows that learning single scoring function for all classes using similarity features helps to avoid linear degradation of generalization bound in contrast to traditional methods. Also the use of joint feature representation facilitated in better scalability and covering of rare classes as compared to state of the art methods.

References

1. Choromanska, A., Langford, J.: Logarithmic time online multiclass prediction. CoRR abs/1406.1822 (2014)
2. 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)
3. 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)
4. 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)