On Multiclass Active Learning with Support Vector Machines

Klaus Brinker[1]

Abstract. In supervised machine learning, a training set of examples which are assigned to the correct target labels is a necessary prerequisite. However, in many applications, the task of assigning target labels cannot be conducted in an automatic manner, but involves human decisions and is therefore time-consuming and expensive. In the case of classification learning, the active learning framework has been considered to address this problem. While most research on active learning in the field of kernel machines has focused on binary problems, less attention has been given to the problem of learning classifiers in the case of multiple classes. We consider three common decomposition methods to express multiclass problems in terms of sets of binary classification problems and propose novel active learning heuristics in order to reduce the labeling effort. Various experiments conducted on real-world datasets demonstrate the merits of our approach in comparison to previous research.

1 INTRODUCTION

In machine learning research, the field of kernel machines has produced learning methods which have a strong theoretical foundation and yield state-of-the-art results. Among the most popular methods are support vector machines. With increasing computational power being available today, optimized training algorithms are able to cope with large-scale classification and regression problems involving tens of thousands of training examples. However, advances in computational speed and more efficient training algorithms do not solve the inherent problem that conventional supervised machine learning relies on a set of examples which have to be assigned to the correct target labels. In many applications, the task of assigning target labels cannot be conducted in an automatic manner, but involves human decisions and is therefore time-consuming and expensive. Hence, assuming the availability of *a priori* given *labeled* training sets disregards the labeling effort that is necessary in many cases.

More precisely, we consider the problem of learning a classifier in the case of an arbitrary number of class labels in a supervised learning scenario. Since many learning algorithms are restricted to binary classification problems, there exist several approaches to express multiclass problems in terms of a set of binary classification problems. The one-versus-one approach [2] decomposes multiclass problems into binary classification problems by considering all pairwise decisions between two class labels. Each pairwise decision problem is treated independently as a binary classification problem and predictions are made by means of a voting procedure. In DAG multiclass classification [5], multiclass problems are decomposed in an analogous manner, however, instead of evaluating aggregated votes, binary classifiers are considered as nodes in a decision directed acyclic graph (DDAG). An alternative approach to the expression of multiclass problem is the one-versus-all approach [7] which trains a separate classifier for each possible class against the rest of classes and predicts class labels according to the maximum output[2] among all binary classifiers.

2 MULTICLASS ACTIVE LEARNING

The superordinate concept of *active learning* refers to a collection of approaches which aim at reducing the labeling effort in supervised machine learning. We consider the *pool-based active learning model*[3] [4]: Starting with only a small amount of labeled examples, the learning algorithm sequentially selects new examples from a finite set of unlabeled examples and requests the corresponding class labels. The crucial point is that by selecting only the most informative examples to be labeled, in many applications, it is possible to learn a model by using fewer labeled examples without a significant loss of generalization accuracy in comparison to conventional batch learning based on the entire set of labeled examples.

In the field of kernel machines, active learning has been successfully applied to classification problems to reduce the labeling effort [1, 8, 6]. Both [1] and [8] are restricted to binary classification whereas [6] additionally considers active learning of multiclass classifiers using the one-versus-all approach. The latter approach considers optimization of a global measure on the volume reduction in the so-called version space model as the selection criterion. We propose a novel extension of active learning to multiclass problems (BINARYMIN strategy) which aims at maximizing the worst-case version space volume reduction on a single binary classifier among the set of classifiers. Considering one-versus-one decomposition, DAG multiclass classification and the one-versus-all technique, we propose heuristic strategies to the selection of new training examples. These strategies are compared to [6] (GLOBAL strategy) in the case of one-versus-all decomposition and a straightforward generalization of this approach to both pairwise decomposition techniques. Additionally, we consider random selection (RANDOM strategy) of new examples as a baseline strategy.

3 EXPERIMENTAL SETTING

To compare the efficiency of all the considered selection strategies, we have conducted several experiments on multiclass datasets[4] that

[1] International Graduate School of Dynamic Intelligent Systems, University of Paderborn, 33098 Paderborn, Germany, email: kbrinker@upb.de

[2] In the following, we consider real-valued classifiers which are thresholded at zero to make binary predictions $\{-1, +1\}$.

[3] We refer to *pool-based active learning* as *active learning* herein after.

[4] We selected all datasets with cardinality > 500 from a recent study on multiclass classification [3].

Table 1. This table shows the estimated final classification accuracy results and the corresponding standard error of the mean estimates. For each decomposition technique, the best result is indicated in bold face.

Decomposition	Strategy	vowel	vehicle	segment	dna	satimage	letter	shuttle
One-vs-All	Random	0.670 ±0.004	0.741 ±0.002	0.893 ±0.001	0.861 ±0.003	0.828 ±0.002	**0.590** ±0.003	**0.951** ±0.003
	Global	0.705 ±0.003	0.737 ±0.003	0.876 ±0.003	**0.908** ±0.002	0.829 ±0.004	0.524 ±0.011	0.938 ±0.006
	BinaryMin	**0.753** ±0.003	**0.767** ±0.002	**0.919** ±0.001	0.904 ±0.002	**0.851** ±0.001	0.516 ±0.008	0.936 ±0.005
One-vs-One	Random	0.732 ±0.004	0.735 ±0.003	0.904 ±0.002	0.866 ±0.002	0.836 ±0.002	0.606 ±0.006	0.960 ±0.003
	Global	0.744 ±0.005	0.720 ±0.003	0.881 ±0.003	**0.907** ±0.002	0.827 ±0.006	0.506 ±0.009	0.864 ±0.026
	BinaryMin	**0.844** ±0.003	**0.745** ±0.003	**0.941** ±0.001	0.906 ±0.002	**0.866** ±0.002	**0.609** ±0.006	**0.993** ±0.002
DAG	Random	0.739 ±0.004	0.734 ±0.002	0.900 ±0.002	0.870 ±0.002	0.834 ±0.002	**0.610** ±0.005	0.961 ±0.002
	Global	0.738 ±0.004	0.718 ±0.003	0.880 ±0.003	**0.904** ±0.001	0.827 ±0.007	0.474 ±0.011	0.881 ±0.022
	BinaryMin	**0.848** ±0.003	**0.749** ±0.003	**0.941** ±0.001	**0.904** ±0.002	**0.854** ±0.002	0.609 ±0.004	**0.995** ±0.001

are publicly available from the UCI repository of machine learning databases and from the Statlog collection.

For all problems which include a separate test set (dna, satimage, letter, shuttle), the selection strategies were initialized using a randomly drawn set of 20 examples (30 for the letter dataset which contains 26 classes) from the training set with at least one example from each class. While new training examples were selected from the remaining training sets, the classification accuracy was estimated on the test sets after every 10 selection steps. We fixed the final number of labeled examples to 200 and averaged the results over 20 runs of random initialization. Each of the remaining datasets (without separate test sets) was randomly split 100 times into a training and a test set of equal size. Analogously, we used initial sets of cardinality 20 which were drawn from the training sets. Furthermore, we fixed the number of labeled examples to 100 for these smaller datasets and averaged the results over all 100 runs of random initialization and separation into training and test sets.

In our experiments, we used a modified version of libsvm [3] that learns support vector machines without bias and L2-loss to stay consistent with the theoretical motivation. We make use of RBF-kernels with the default value of $\gamma = \frac{1}{\text{\# input features}}$ and $C = 100$.

4 EXPERIMENTAL RESULTS

Note that the kernel has not been optimized with respect to the given problems and multiclass techniques. Therefore, we compare the selection strategies separately for each approach and do not focus on a quantitative comparison between different multiclass techniques. Due to space restrictions, we refrain from a detailed exposition of every learning curve. Instead, we focus our presentation of the experimental results on the average classification accuracy, i.e. the proportion of correctly classified test examples, at the end of the experiments with 100 and 200 labeled examples respectively, to summarize the efficiency of a given selection strategy. Table 1 shows average classification results and associated standard error of the mean estimates. In the case of the one-versus-one and the DAG pairwise technique, the BinaryMin strategy achieves the best result for 6 out of 7 problems (with one tie for DAG and the dna problem). For the only dataset where BinaryMin is outperformed, its average accuracy is very close to the winning strategy. In the case of one-versus-all classification, BinaryMin is the winner for 4 out of 7 problems, while Random and Global achieve the best result on 2 and 1 problem respectively.

In our experiments, both the original Global strategy and its generalization were outperformed by the Random selection strategy on most of the problems. In contrast to this, the BinaryMin strategy based on all decomposition techniques clearly indicates that is possible to reduce the labeling effort in multiclass classification learning.

5 CONCLUSION

We have introduced a novel extension of pool-based active learning to multiclass classification based on both the one-versus-all classification technique and two pairwise decomposition methods. Experimental results clearly indicate that our approach to active learning yields a significant reduction of the labeling effort in multiclass learning and outperforms previous approaches.

REFERENCES

[1] C. Campbell, N. Cristianini, and A. Smola, ‘Query learning with large margin classifiers’, in *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pp. 111–118, (2000).

[2] J. H. Friedman, ‘Another approach to polychotomous classification’, Technical report, Department of Statistics, Stanford University, Stanford, CA, (1996).

[3] C.-W. Hsu and C.-J. Lin, ‘A comparison of methods for multi-class support vector machines’, *IEEE Transactions on Neural Networks*, **13**, 415–425, (2002).

[4] David D. Lewis and William A. Gale, ‘A sequential algorithm for training text classifiers’, in *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, eds., W. Bruce Croft and Cornelis J. van Rijsbergen, pp. 3–12, Dublin, IE, (1994). Springer Verlag, Heidelberg, DE.

[5] J. Platt, N. Cristianini, and J. Shawe-Taylor, ‘Large margin dags for multiclass classification’, in *Advances in Neural Information Processing Systems 12 (NIPS)*, eds., S.A. Solla, T.K. Leen, and K.-R. Mueller, pp. 547–553, (2000).

[6] S. Tong, *Active Learning: Theory and Applications*, Ph.D. dissertation, Stanford University, 2001.

[7] V. Vapnik, *Statistical Learning Theory*, John Wiley, N.Y., 1998.

[8] M. K. Warmuth, G. Rätsch, M. Mathieson, J. Liao, and C. Lemmen, ‘Active learning in the drug discovery process’, in *Advances in Neural information processing systems*, eds., T.G. Dietterich, S. Becker, and Z. Ghahramani, volume 14, pp. 1449–1456, (2002).