multilabel strategies: SVM binary minimum

On Active Learning in Multi-label Classification

Klaus Brinker

Data and Knowledge Engineering
Faculty of Computer Science
Otto-von-Guericke-University Magdeburg
D-39106 Magdeburg, Germany

Abstract. In conventional multiclass classification learning, we seek to induce a prediction function from the domain of input patterns to a mutually exclusive set of class labels. As a straightforward generalization of this category of learning problems, so-called multi-label classification allows for input patterns to be associated with multiple class labels simultaneously. Text categorization is a domain of particular relevance which can be viewed as an instance of this setting. While the process of labeling input patterns for generating training sets already constitutes a major issue in conventional classification learning, it becomes an even more substantial matter of relevance in the more complex multi-label classification setting. We propose a novel active learning strategy for reducing the labeling effort and conduct an experimental study on the well-known Reuters-21578 text categorization benchmark dataset to demonstrate the efficiency of our approach.

1 Introduction

The conventional multiclass classification setting assumes each input pattern to be associated with one specific class label, i.e., the target space consists of a mutually exclusive finite set of class labels. However, many applications require a more flexible setting which allows for input patterns to be associated with multiple class labels simultaneously. We refer to this setting as *multi-label classification*. Semantic image classification [2] and text categorization [6] form learning problems of particular relevance in practice where each input patterns is potentially associated with multiple class labels, such as *building, beach, animal* and *sports, economy, politics*, respectively. Thus, these learning problems can be cast as instances of multi-label classification in a straightforward manner. We are particularly interested in the latter domain of text categorization and consider the well-studied Reuters-21578 text categorization benchmark dataset.

Many machine learning algorithms are inherently restricted to binary classification learning. In multiclass classification, one principal means for the generalization of binary algorithms is to construct and combine several binary classifiers using a one-versus-all decomposition scheme [4]. The one-versus-all technique trains a separate classifier for each possible class against the rest of classes with the training examples being binary relabeled in a suitable manner and predicts class labels according to the maximum output[1] among all binary classifiers (*MAX-wins*). This technique can be generalized to the multi-label setting by a modification of the decomposition and prediction step, where input patterns are submitted as positive examples to all binary problems corresponding to the associated (relevant) class labels and submitted to the remaining problems as negative examples. Moreover, the *MAX-wins* prediction technique has to be replaced by an *ALL-positive* procedure.

The effort necessary to construct training sets of labeled examples in a supervised machine learning scenario is often disregarded, though in many applications, it is a time-consuming and expensive procedure. While this process already constitutes a major issue in classification learning, it becomes an even more substantial matter of relevance in multi-label learning, which considers the more complex target space of a set of *not* necessarily mutually exclusive class labels.

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, which was originally introduced by [7] in the context of text classification learning. If not noted otherwise, we refer to the *pool-based active learning model* as *active learning* herein after to simplify our presentation. In contrast to conventional supervised learning, pool-based active learning considers an extended learning model in which the learning algorithm is granted access to a set of initially unlabeled examples and the algorithm is provided with the ability to determine the order of assigning target objects, i.e., associated subsets of class labels. The essential idea behind active learning is to select promising unlabeled examples with the objective of attaining a high level of accuracy without requesting the complete set of corresponding target objects.

In particular, text categorization is a characteristic learning problem which is amenable to the active learning approach [12, 8, 10]: While a relatively cheap source of unlabeled examples is available, acquiring the associated sets of target labels is an expensive procedure. More precisely, large corpora of text documents are readily available in many domains. However, assigning given text documents to target categories to generate labeled training sets is a time-consuming task as it requires human decisions. This general pattern is not only characteristic for text categorization problems, but also arises in many other domains.

We propose a novel generalization of pool-based active learning to reduce the labeling effort based on the one-versus-all technique for representing multi-label classifiers. The remainder of this paper is organized as follows: The subsequent section establishes the notational basis and reviews the aforementioned binary decomposition approach to multi-label classification.

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

In Section 3, we discuss active learning in the context of multi-label classification and propose our novel generalization. Experimental results on the Reuters-21578 text categorization benchmark dataset which demonstrate the efficiency of our approach are discussed in Section 4.

2 Multi-label Classification

Assume we are given a nonempty input space $\mathcal{X}$ and a finite set $\{1, \ldots, d\}$ of class labels. Then, the target space $\mathcal{Y}$ in multi-label classification is defined as $\mathcal{Y} \overset{\text{def}}{=} \mathfrak{P}(\{1, \ldots, d\})$ where $\mathfrak{P}(A)$ denotes the power set of a given set A. The fundamental learning task consists in inducing a prediction function $f : \mathcal{X} \to \mathcal{Y}$ based on a given training set of labeled examples

$$L = \{(x_1, Y_1), \ldots, (x_m, Y_m)\} \subset \mathcal{X} \times \mathcal{Y}. \tag{1}$$

We consider support vector machines as the binary base learning algorithm as they have demonstrated excellent generalization ability in the domain of text categorization [6].

A common binary decomposition method for solving multi-label problems is to train a separate binary classifier $h_i : \mathcal{X} \to \{-1, +1\}$ for each of the d target classes against the remaining set of classes [2]. More precisely, all examples $(x, Y) \in L$ with $i \in Y$ are relabeled as positive examples in the process of training the binary classifier h_i, whereas the remaining examples are relabeled as negative examples. Target objects Y for unseen patterns x are predicted according to positive classification of the underlying set of binary classifiers (*ALL-positive*):

$$h : \mathcal{X} \to \mathcal{Y} \tag{2}$$

$$x \mapsto \underset{i=1,\ldots,d}{\operatorname{argpos}} h_i(x) = \big\{ i \in \{1, \ldots, d\} \mid h_i(x) = +1 \big\}. \tag{3}$$

3 Active Multi-label Learning

As mentioned in the preceding section, we employ support vector machines [13] as binary base learning algorithm. Support vector machines and, more generally, the class of kernel machines form linear learning algorithms which as a distinctive feature perform an implicit embedding of input patterns in a kernel-induced feature space $\mathcal{F}$. In the following, we will denote the given kernel by $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and the corresponding kernel feature map by $\phi : \mathcal{X} \to \mathcal{F}$. Given a weight vector $w \in \mathcal{F}$, the corresponding (binary) kernel classifier is defined as

$$h_w : \mathcal{X} \to \{-1, +1\} \tag{4}$$

$$x \mapsto \operatorname{sign}(\langle w, \phi(x) \rangle_{\mathcal{F}}) \tag{5}$$

where $\text{sign}(t) = +1$ for $t > 0$ and $\text{sign}(t) = -1$ otherwise.

Let us assume that we are given a linearly separable (in the feature space) binary classification training set $\{(x_1, y_1), \ldots, (x_m, y_m)\}$. This assumption can be relaxed using a suitable modification of the kernel matrix when using the quadratic-loss [11]. The nonempty set

$$\mathcal{V} \overset{\text{def}}{=} \{w \in \mathcal{F} \mid h_w(x_i) = y_i \ \text{ for } \ i = 1, \ldots, m, \ \ \|w\|_{\mathcal{F}} \leq 1\} \tag{6}$$

which consists of weight vectors in the unit hyperball corresponding to linear classifiers in the feature space which are consistent with the training set is called *version space* [9]. Learning can be viewed as a search process within the version space: Each labeled example (x_i, y_i) imposes a constraint on the version space because to correspond to a consistent classifier a weight vector has to satisfy $\text{sign}(\langle w, \phi(x_i)\rangle_{\mathcal{F}}) = y_i \ \Leftrightarrow \ y_i \langle w, \phi(x_i)\rangle_{\mathcal{F}} > 0$. In other words, consistent classifiers are restricted to a halfspace whose boundary is the hyperplane with normal vector $y_i \phi(x_i)$. For a fixed feature vector $\phi(x_i)$, the class label y_i determines the orientation of the halfspace. Furthermore, $\mathcal{V}$ is the intersection of m halfspaces (a convex polyhedral cone) with the unit hyperball in the feature space $\mathcal{F}$.

So far, we considered the conventional batch learning scenario where the completely labeled set of examples is required as a prerequisite for training. Moreover, the version space model provides the basis for active learning strategies which sequentially select the most promising unlabeled examples and then request the corresponding class label. From a theoretical perspective, there exists an appealing connection to the theory of convex set which provides further insides on appropriate active selection strategies: [5] showed that any halfspace containing the center of mass of a convex set contains a fraction of at least $1/e$ of the overall volume. Assume we are able to repeatedly select unlabeled examples which correspond to restricting hyperplanes passing exactly through the current center of mass of the version space w_{center}. Then, independently of the actual class label, the volume of the version space is reduced exponentially in terms of the number of labeled examples. For computational efficiency, the exact center of mass can be approximated by the center of maximum radius hyperball inscribable in the version space. In the case of normalized feature vectors[2], this center is given by the weight vector $w^{(\text{svm})}$ of the support vector machine trained on the labeled set of examples. As a consequence of the *finite* number of unlabeled examples which in general does not allow to satisfy this criterion, a common approach in pool-based active learning with kernel machines is to select the unlabeled example whose restricting hyperplane is closest to the center of the maximum radius hyperball, i.e., unlabeled examples minimizing $|\langle w^{(\text{svm})}, \phi(x)\rangle_{\mathcal{F}}|$ [12].

[2] Normalization can be achieved by a straightforward kernel modification: $k^{(\text{NORM})}(x, x') \overset{\text{def}}{=} \frac{k(x,x')}{\sqrt{k(x,x)k(x',x')}}$.

For generalizing this selection strategy from binary to multi-label classification, we have to take into account that instead of a single version space the aforementioned one-versus-all decomposition technique yields a set of d version spaces. In the case of label ranking learning where similar decomposition techniques are required, a best worst-case approach with respect to individual volume reduction was demonstrated to achieve a substantial reduction of the labeling effort [3]. We propose an analogous generalization for multi-label classification in the following.

For a *labeled* binary example, the (rescaled) margin $\frac{1+y\langle w^{(\mathrm{svm})},\phi(x)\rangle_{\mathcal{F}}}{2}$ can be viewed as a (coarse) measure of the reduction of the version space volume. Indeed, a straightforward derivation reveals that the above-defined selection strategy can be interpreted as measuring the volume reduction for the worst-case class label. For multi-label classification, the notion of worst-case can be generalized to the case of a set of binary classification problems by evaluating the minimum absolute distance $\min_{i=1,\ldots,d} |\langle w_i^{(\mathrm{svm})}, \phi(x)\rangle_{\mathcal{F}}|$ among all binary problems, where $w_i^{(\mathrm{svm})}$ denotes the weight vector of the support vector machine trained on the one-versus-i subproblem. From a different perspective, we aim at selecting an unlabeled multi-label example which maximizes the (binary) volume reduction with respect to the worst-case set of associated target class label. Denoting the set of labeled and unlabeled examples by L and U, respectively, the active selection strategy is formally given by

$$(U, L) \mapsto \operatorname*{argmin}_{x \in U} \Big(\min_{i=1,\ldots,d} |\langle w_i^{(\mathrm{svm})}, \phi(x)\rangle_{\mathcal{F}}| \Big). \tag{7}$$

Note, that the right-hand side (implicitly) depends on L through the weight vectors $w_1^{(\mathrm{svm})}, \ldots, w_d^{(\mathrm{svm})}$.

4 Experiments

The Reuters-21578 newswire benchmark dataset is the currently most widely used test collection for text categorization research.[3] Our experiments are based on the standard ModApte split which divides the dataset into 7.769 training and 3.019 test documents. Each document is associated with a subset of the 90 categories present in the dataset. In compliance with related research, the documents were represented using stemmed word frequency vectors with a TFIDF weighting scheme and elimination of common words resulting in roughly 10.000 features. For computational reasons, we restricted our experimental setup to the 10 most frequent categories in the Reuters dataset. Moreover, we used linear kernels with the default choice of $C = 10$ (and quadratic-loss) as they were demonstrated to provide an excellent basis for accurate classifiers on this dataset [6]. For normalizing the data to unit modules, we employed the aforementioned kernel modification.

[3] The Reuters-21578 newswire benchmark dataset is publicly available at `http://www.daviddlewis.com/resources/testcollections/reuters21578/`.

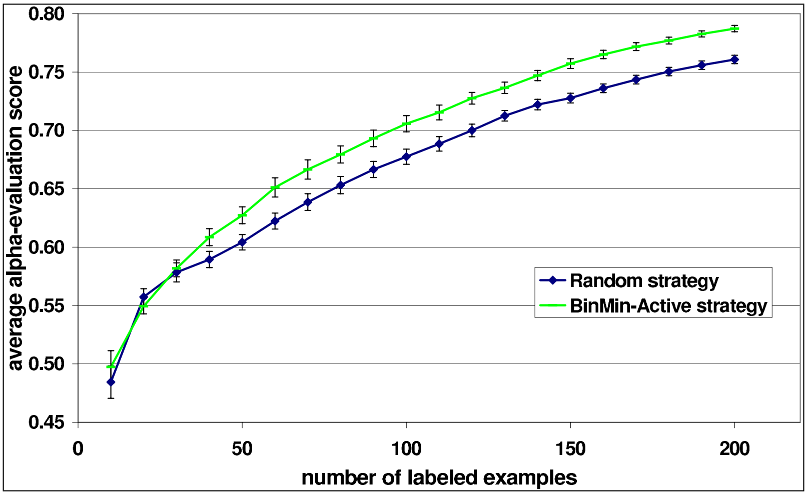

Fig. 1. Experimental learning curves for the random and active selection strategies on the Reuters-21578 text categorization benchmark dataset. This figure shows average α-evaluation scores ($\alpha = 1$) and corresponding standard errors of the mean for different numbers of labeled examples.

An initial subsample of 10 multi-label examples was randomly drawn from the training set and submitted to the active learning algorithm. Then, the target objects of the remaining examples were masked out prior to selection and the active learning strategy sequentially selected 190 examples. The accuracy of the multi-label classifiers trained on the currently labeled sets of examples was evaluated every 10 iterations. As the evaluation measure, we used the *α-evaluation score* proposed by [2]: Denote by $Y, Y' \in \mathcal{Y}$ sets of labels. Then the score$^{(\alpha)}$ is defined as

$$\text{score}^{(\alpha)}(Y, Y') \overset{\text{def}}{=} \left(\frac{|Y \cap Y'|}{|Y \cup Y'|} \right)^{\alpha} . \tag{8}$$

This similarity measure has varying properties depending on the parameter α: For $\alpha = \infty$, score$^{(\alpha)}$ evaluates to 1 only in the case of identical sets Y and Y', whereas for $\alpha = 0$, it evaluates to 1 except for the case of completely disjoint sets. We considered the intermediate choice of $\alpha = 1$, which provides a finer scale. Based on this underlying measure, the accuracy of a multi-label classifier $h : \mathcal{X} \to \mathcal{Y}$ was evaluated on the test set T:

$$\text{accuracy}_T(h) \overset{\text{def}}{=} \frac{1}{|T|} \sum_{(x,Y) \in T} \text{score}^{(\alpha)}(h(x), Y). \tag{9}$$

To compensate for effects based on the random choice of the initially labeled set, we repeated the above-described procedure 30 times and averaged the results over all runs. In addition to the proposed active selection strategy, we employed random selection of new training examples as a baseline strategy.

As depicted in Figure 1, active learning significantly outperforms random selection starting at about 40 selection steps (at least at the 0.05 significance level). This pattern is not only typical for active learning in multi-label classification but also for other categories of learning problems where active learning becomes more effective once the labeled data is sufficient to train an adequate intermediate model.

5 Related Work

In the field of active learning, there are two principle categories of approaches: So-called *query learning* [1] refers to a learning model where the learning algorithm is given the ability to request true class labels corresponding to examples generated from the entire input domain. In contrast to this, in *selective sampling* the learner is restricted to request labels associated with examples from a finite set of examples (*pool-based model*) or the learning algorithm has to decide whether to request the corresponding true labels for sequentially presented single examples (*stream-based model*). Research in the field of pool-based active learning with kernel machines has mainly focused on binary classification. Beyond this category, multiclass classification [12] and label-ranking [3] are among those categories of learning problems which were demonstrated to benefit substantially from the active learning framework in terms of the number of labeled examples necessary to attain a certain level of accuracy.

6 Conclusion

We introduced a novel generalization of pool-based active learning to the category of multi-label classification problems which is based on the common one-versus-all binary decomposition scheme. From a theoretical perspective, a generalized view of the version space model provides an appealing motivation of our approach. An experimental study on the well-known Reuters-21578 text categorization benchmark dataset demonstrates the efficiency of our approach in terms of the number of labeled examples necessary to attain a certain level of accuracy. Moreover, as it is reasonable to assume that acquiring target objects in multi-label classification learning is more expensive than for less complex domain like binary classification, the benefits of active learning in this context become even more obvious and suggest that it is a promising approach in reducing the cost of learning.

References

1. ANGLUIN, D. (1988). Queries and concept learning. *Journal of Machine Learning*, 2:319–342.

2. BOUTELL, M.R., LUO, J., SHEN, X., and BROWN, C.M. (2004). Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771.
3. BRINKER, K. (2004). Active learning of label ranking functions. In Greiner, R. and Schuurmans, D., editors, *Proceedings of the Twenty-First International Conference on Machine Learning (ICML 2004)*, pages 129–136.
4. CORTES, C., and VAPNIK, V. (1995). Support vector networks. *Journal of Machine Learning*, 20:273 – 297.
5. GRÜNBAUM, B. (1960). Partitions of mass-distributions and convex bodies by hyperplanes. *Pacific J. Math.*, 10:1257–1261.
6. JOACHIMS, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In Nédellec, C. and Rouveirol, C., editors, *Proceedings of the European Conference on Machine Learning (ECML 1998)*, pages 137–142, Berlin. Springer.
7. LEWIS, D.D., and GALE, W.A. (1994). A sequential algorithm for training text classifiers. In Croft, W. B. and van Rijsbergen, C. J., editors, *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, pages 3–12, Dublin, IE. Springer Verlag, Heidelberg, DE.
8. McCALLUM, A.K., and NIGAM, K. (1998). Employing EM in pool-based active learning for text classification. In: Shavlik, J.W., editor, *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998)*, pages 350–358, Madison, US. Morgan Kaufmann Publishers, San Francisco, US.
9. MITCHELL, T.M. (1982). Generalization as search. *Journal of Artificial Intelligence*, 18:203–226.
10. ROY, N., and McCALLUM, A. (2001). Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 441–448. Morgan Kaufmann, San Francisco, CA.
11. SHAWE-TAYLOR, J., and CRISTIANINI, N. (1999). Further results on the margin distribution. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (COLT 1999)*, pages 278–285. ACM Press.
12. TONG, S., and KOLLER, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66.
13. VAPNIK, V. (1998). *Statistical Learning Theory.* John Wiley, N.Y.