Active Learning of SVM and Decision Tree Classifiers for Text Categorization

Peter Bednár

Dept. of Cybernetics and Artificial Intelligence, Technical University of Kosice,
Letna 9, 042 00 Kosice, Slovakia, Peter.Bednar@tuke.sk

Abstract: Many real applications require large training dataset for supervised learning. In this paper we will present one method of active learning, which allows reducing the number of training examples required for effective learning. Presented algorithm is based on the simple heuristic that selects examples according to the confidence of the classifier prediction for the given example. This heuristic doesn't require validation set and can be used effectively to select small set of labeled examples.

Keywords: active learning, text categorization

1 Introduction

Effective application of the supervised learning methods in real domains requires the large training dataset. However in some cases, the training examples are inaccessible or it is too expansive to obtain labeled dataset.

Even when the large dataset can be obtained relatively cheap, it can be too expansive to classify training examples to the predefined classes. The example of such a domain is the text categorization task. The goal of the text categorization is to assign text documents into the predefined classes (categories) according to its text content. In the case where the user plays role of the expert who classify training examples (for examples in the system for e-mail filtering), using of automatic methods based on the supervised learning becomes very limited.

This paper describes some methods of the active learning [1][2][3], where the learning algorithm selects next training examples according the to confidence of the model prediction. In the next chapter, base selection algorithm is described. Following chapters describes two algorithms that we have used for experiments - SVM and decision trees. At the end, we will present the results of some experiments and conclusion.

2 Active Learning

The primary motivation for active learning comes from the time or expense of obtaining labeled training dataset. In this paper, we will describe the method of active learning based on the selection of the training examples. In these settings, the learned is presented with a large corpus of unlabeled examples, and is given the option of labeling some subset of them. Since we can assign the "cost" to the labeling of each example, the goal of active learning is to choose a small subset of unlabeled examples that maximizes the classification accuracy.

The heuristic of active selection of examples is based on the "confidence" of the prediction. This heuristic doesn't require validation set and can be used effectively to select small set of labeled examples. For the binary classification in the domain *X* we assume that the classification is based on the sign of the decision function *f*: *X* → [-1,1]. The confidence of the prediction than can be formulated as the absolute value of function *f*. The algorithms which we will apply to this problem are the Support Vector Machine and the decision tree classifier.

2.1 Support Vector Machine

Given the examples in the domain *X*, a linear support vector machine [4] is defined in terms of the hyperplane

$$w.x + b = 0 \tag{1}$$

corresponding to the decision function

$$f(x) = sign(w.x + b) \tag{2}$$

for $\boldsymbol{w} \in \Re^N$ and $b \in \Re$. The support vector machine attempts to find, among all possible hyperplanes in *N*-dimensional document space, the hyperplane (***w*** and *b* parameters) that separates the positive and negative training examples with widest margin. This can be formulated as a quadratic optimalization problem

$$\max_{w,b}\left\{\min_{x_i}\left\{\| x - x_i \| : x \in \Re^N, w.x + b = 0\right\}\right\} \tag{3}$$

When the data are not separable by hyperplane, a soft margin classifier is used which requires that the misclassification cost *C* is assigned to each misclassified training examples. The problem 3 is usually optimized by introducing the Lagrange multipliers α_i, and recasting the problem in the terms of its dual form.

minimize $$L_D = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j \tag{4}$$

subject to $0 \le \alpha_i \le C, \sum_i \alpha_i y_i = 0$

2.2 Decision Trees

A decision tree classifier [5] is a tree in which internal nodes are labeled by attributes (words occurrences in the case of text categorization), branches departing from them are labeled by tests the weight that attribute has in the test document, and leafs are labeled by categories. Decision tree categorizes a test document by recursively testing the weights that the attributed labeling the internal nodes have in document vector, until a leaf reached.

The most common approach to inducing a decision tree is to partition the labeled examples recursively until a stopping criterion is met. The partition is defined by selecting the test which divide all examples to the disjoint subsets assigned to the test branches, passing each example to the corresponding branch, and treating each block of the partition as a subproblem, for which a subtree is build recursively. A common stopping criterion for a subset of examples is that they all have the same class.

Since the misclassification probability for the given example can be estimated according to the misclassification probability computed for the leaf that covers this example, we have directly computed confidence of the prediction as the *Laplace* estimation of the leaf's misclassification error:

$$conf_i = \frac{p + 0.5}{p + n + 1} \tag{5}$$

where p (n) is the number of positive (negative) training examples assigned to the leaf.

3 Experiments

We have used Reuters-21578 dataset for our experiments. This dataset contains 21578 articles from the newspapers of the Reuters agency. Only documents that were assigned to at least one of 90 categories were used for the experiments. The examples were divided into the training set of 7769 documents and test set of 3019 (i.e. ModApte split commonly used in experiments).

Text was divided according to the occurrences of the whitespaces and transformed to the lowercase. The diacritic prefixes and suffixes were removed from the words. The stemming (transformation to the base form or morphological root of the word) was not performed. We have removed all words without letters (stop words - i.e. functional words etc. were not removed).

The main goal of the experiments was to find how effectively can active selection of examples improves classification accuracy contrary to the random selection. At the beginning, the training set was initialized randomly with one positive and one negative example. Than we have added stepwise into the training set 1 (2-100), 10 (100-200), 50 (200-1000) and 200 (1000-7769) examples from the former ModApte training set. We have used the whole ModApte testing set for the testing. The experiments were repeated 5 times with different random initialization.

The first graph (Figure 1) shows the dependency of the classification accuracy on the number of training examples for SVM classifier (error bars show standard deviations).

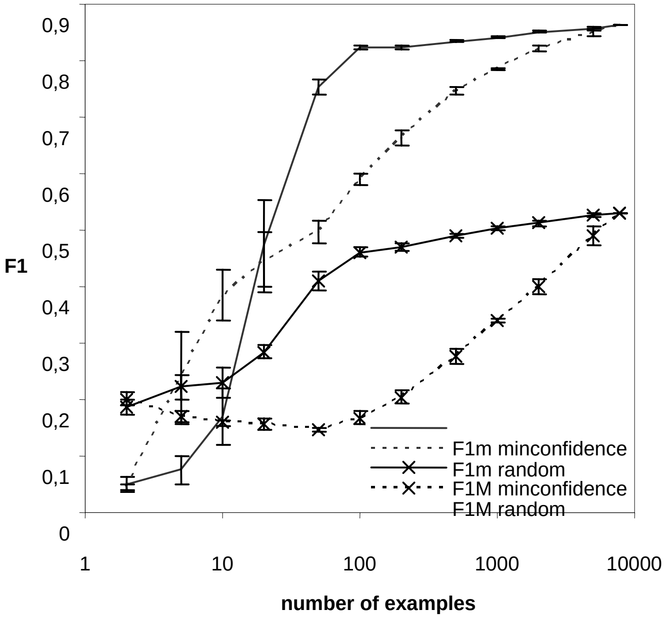

Figure 1

Dependency of classification accuracy on the number of training examples for SVM classifier

As the accuracy measure we have used F_1 measure adopted from information retrieval which is defined as

$$F_1 = \frac{2a}{2a+b+c} \tag{6}$$

where *a b* and *c* are corresponding entries from two-way contingency table (i.e. *a* is the number of correctly classified positive examples, *b* is the number of false positive, etc.). There are two ways of computing the F_1 average for all categories: macro-average where value is computed for each category and these are averaged to final value or micro-averaged where we first obtain global values for *a*, *b* and *c*.

According to the results, micro and macro averaged F_1 measure of active selection is higher than the random selection almost for every size of the training set. For example for macro-average with active selection we have obtained accuracy 46% with only 100 training examples (this is only 1.2% of all examples). We needed approximately 3400 randomly selected examples to achieve the same accuracy. The results for micro-average are similar.

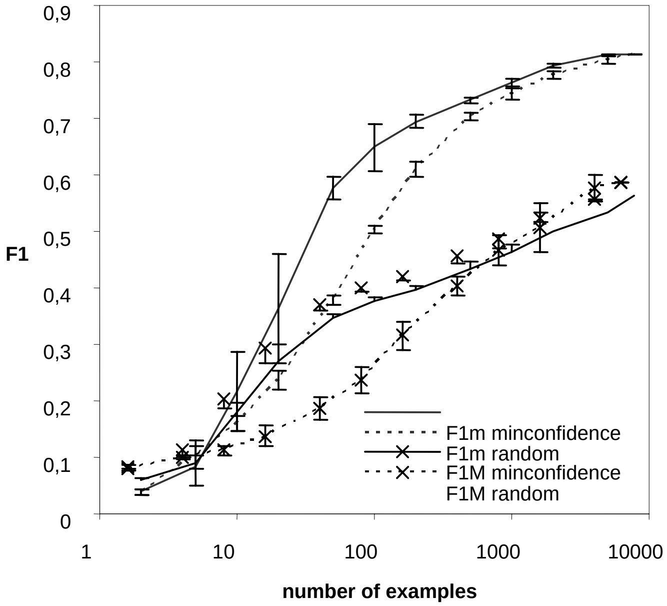

Figure 2

Dependency of classification accuracy on the number of training examples for decision tree classifier

The Figure 2 shows experiment with the same settings for decision tree classifier. The difference between active and random selection is not so noticeable as for the

SVM classifier, but again with active selection we can obtain higher accuracy with the lower number of training examples.

Conclusions

In this paper we have presented one method of active learning, which allows reducing the number of training examples required for effective learning. This method is based on the selection of examples for which the classiifer has lowest confidence of its class prediction. We have tested this method on two algorithms - support vector machine and decision tree classifier. In both cases, active selection of examples obtains better classification accuracy with lower number of training examples than non active random selection of training data. In the future, we will extend these results with some new strategies for active selection based for example on the agreement of various hypotheses in ensemble classifiers.

Acknowledgement

The work presented in this paper was supported by the Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic within the project ”Document classification and annotation for the Semantic web” No. 1/1060/04 and by the German-Slovak research project DAAD No. 8/2004 “Text Mining for Metadata Extraction and Semantic Retrieval”.

References

[1] Cohn, D., Ghahramani, Z., Jordan, M.: Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4, 129-145, (1996)

[2] Tong, S., Koller, D.: Support Vector Machine active learning with Application to Text Categorization, *Machine Learning Research*, Vol. 2, 45-66, (2001)

[3] Schohn, G. Cohn, D.: Less Is More: Active Learning with Support Vector Machines, *Proc. 17th International Conference Machine Learning*, 839-846, 2000

[4] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer Verlag

[5] Apté, C., Damerau, F., Weiss, S.: Automated learning of decition rules for text categorization. ACM Transactions on Information Systems, 12(3):233-251, (1994)