Improving Multiclass Text Classification with the Support Vector Machine

Jason D. M. Rennie & Ryan Rifkin

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: There are billions of text documents available in electronic form. More and more are becoming available every day. The Web itself contains over a billion documents. Millions of people send e-mail every day. Academic publications and journals are becoming available in electronic form. These collections and many others represent a massive amount of information that is easily accessible. However, seeking value in this huge collection requires organization. Many web sites offer a hierarchically-organized view of the Web. E-mail clients offer a systems for filtering e-mail. Numerous academic communities have a Web site that allows searching on papers and shows an organization of papers. However, organizing documents by hand or creating rules for filtering is painstaking and labor-intensive. This can be greatly aided by automated classifier systems. The accuracy of such systems determines their usefulness. We propose to use the Support Vector Machine (SVM) in conjunction with Error-Correcting Output Codes (ECOC) to improve the state-of-the-art in text classification.

Motivation & Previous Work: In 1998, Joachims published results on a set of binary text classification experiments using the SVM [4]. The SVM yielded lower error than many other classification techniques. Yang followed later with experiments of her own on the same data set [5]. She used improved versions of Naive Bayes (NB) and kNN but still found that the SVM performed at least as well as all other classifiers she tried. She also found that the linear SVM performed as well as polynomial and RBF versions. Both papers used the SVM for binary text classification, leaving the multiclass problem (assigning a single label to each example) open for future research.

Berger and Ghani individually chose to attack the multiclass text classification problem using error-correcting output codes (ECOC) [2] [3]. They both chose to use Naive Bayes as the binary classifier. ECOC combines the outputs of many individual binary classifiers in an additive fashion to produce a single multiclass output. Ghani found great error reduction on Industry Sector, a data set with 105 classes. In parallel, Allwein *et. al.* wrote an article on using ECOC with loss functions for multiclass classification in non-text domains [1]. They presented a unifying framework for multiclass classification, encouraging the use of loss functions, especially when the classifier is optimized for a particular one. They also tested five different code matrices and found that the one-vs-all matrix yielded the highest error when used with the SVM.

Approach: We bridge these bodies of work by applying ECOC to text classification with the linear SVM as the binary learner. The SVM is well known to be a powerful binary classifier. In particular, it can generalize well when there are many fewer training examples in one class than the other. ECOC provides a framework for making use of many binary classifiers to produce a single multiclass output. The ECOC matrix defines how the binary classifiers are to be trained and the loss incurred for each binary output. The multiclass output is the class which incurs the least loss.

ECOC has been applied to text classification with Naive Bayes as the binary learner with great success. The SVM provides an opportunity for improvement since the SVM is known to perform binary classification better than Naive Bayes. We perform experiments on the Industry Sector and 20 Newsgroups data sets and achieve lower errors than ECOC with Naive Bayes. We achieve the lowest known error on both data sets. Unlike Allwein *et. al.*, we find that the one-vs-all matrix is very competetive with other matrices when the SVM is the binary learner. This dispells the commonly-held belief that high row and column separation is necessary for good ECOC classification. We attribute the success of the SVM to its low binary error and its ability to generalize well when the distribution of training examples is uneven (as is the case for Industry Sector).

20 Newsgroups	800		250		100		30	
	SVM	NB	SVM	NB	SVM	NB	SVM	NB
OVA	0.131	0.146	0.167	0.199	0.214	0.277	0.311	0.445
Dense 15	0.142	0.176	0.193	0.222	0.251	0.282	0.366	0.431
BCH 15	0.145	0.169	0.196	0.225	0.262	0.311	0.415	0.520
Dense 31	0.135	0.168	0.180	0.214	0.233	0.276	0.348	0.428
BCH 31	0.131	0.153	0.173	0.198	0.224	0.259	0.333	0.438
Dense 63	0.129	0.154	0.171	0.198	0.222	0.256	0.326	0.407
BCH 63	0.125	0.145	0.164	0.188	0.213	0.245	0.312	0.390

Industry Sector	52		20		10		3	
	SVM	NB	SVM	NB	SVM	NB	SVM	NB
OVA	0.072	0.357	0.176	0.568	0.341	0.725	0.650	0.828
Dense 15	0.119	0.191	0.283	0.363	0.461	0.542	0.738	0.805
BCH 15	0.106	0.182	0.261	0.352	0.438	0.518	0.717	0.771
Dense 31	0.083	0.185	0.216	0.315	0.394	0.478	0.701	0.737
BCH 31	0.076	0.140	0.198	0.292	0.371	0.462	0.676	0.743
Dense 63	0.072	0.135	0.189	0.279	0.363	0.453	0.674	0.744
BCH 63	0.067	0.128	0.176	0.272	0.343	0.443	0.653	0.734

Table 4: Above are results of multiclass classification experiments on the 20 Newsgroups (top) and Industry Sector data sets. The top row of each table indicates the number of documents/class used for training. The second row indicates the binary classifier. The far left column indicates the multiclass technique. Entries in the table are classification error. The SVM outperforms Naive Bayes (NB) across both data sets, all matrix types and all training set sizes.

Impact: The impact of our work is great. It shows that the use of the SVM with ECOC can greatly improve multiclass text classification. We have shown that it improves performance over ECOC with Naive Bayes by up to 48% and achieves the lowest known error on two data sets. Also, the error we achieve on Industry Sector is 85% lower than regular Naive Bayes (which achieves 0.434 error on Industry Sector). Naive Bayes is an algorithm which is commonly used in practice. These results warrant a re-tooling of text classifiers at large, particularly because the SVM can be made efficient. Training the SVM is significantly slower than Naive Bayes (SVM is $O(n^2)$; NB is $O(n)$), but we realize speed-ups by caching kernel products across columns of the matrix. Since we use the linear SVM, once the SVM is trained, classifying a document is no less efficient than Naive Bayes. ECOC with the linear SVM is practical and can be used in large classifier systems where efficiency is paramount.

Future Work: Matrix design is an important aspect of ECOC classification. The lowest error matrix varies by problem. We find that the BCH matrix performs best most of the time, but OVA performs as well or slightly better when there are few training examples. Allwein *et. al.* found that OVA performed worst of the matrices they tried and found that no single matrix performed best. There may be ways to identify the optimal matrix based on properties of the data set. The success of ECOC as a classification method brings renewed light to additive models and voting methods, classifiers which use a weighted sum of (possibly) non-linear classifier outputs to determine the overall classification output. Error may be further reduced by applying known techniques from this field, such as boosting.

Research Support: Jason acknowledges support from Nippon Telegraph and Telephone Corporation (MIT2000-08). Rif and Jason both acknowledge support from an NSF-ITR Grant (NSF#0085836).

References:

[1] Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

[2] Adam Berger. Error-correcting output coding for text classification. In *Proceedings of IJCAI-99 Workshop on Machine Learning for Information Filtering*, Stockholm, Sweeden, 1999.

[3] Rayid Ghani. Using error-correcting codes for text classification. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.

[4] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning (ECML)*, 1998.

[5] Yiming Yang and Xin Liu. A re-examination of text categorization methods. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99)*, 1999.