On Newsgroup, it beats centroid classifier by 4 percents; on Sector-48, it beats centroid classifier by 11 percents. More encouraging, it yields better performance than SVM classifier on Sector-48. This improvement also indicates that Model-Refinement can effectively reduce the bias incurred by ECOC.

Table 1: The MicroF1 of different methods

Method Dataset	Centroid	MR +Centroid	ECOC +Centroid	ECOC + MR +Centroid	SVM
Sector-48	0.7985	0.8671	0.6422	**0.9122**	0.8948
NewsGroup	0.8371	0.8697	0.8085	**0.8788**	0.8777

Table 2: The MacroF1 of different methods

Method Dataset	Centroid	MR +Centroid	ECOC +Centroid	ECOC + MR +Centroid	SVM
Sector-48	0.8097	0.8701	0.6559	**0.9138**	0.8970
NewsGroup	0.8331	0.8661	0.7936	0.8757	0.8759

Table 3 and 4 report the classification accuracy of combining ECOC with Model-Refinement on two datasets vs. the length BCH coding. For Model-Refinement, we fix its *MaxIteration* as 8; the number of features is fixed as 10,000.

Table 3: the MicroF1 vs. the length of BCH coding

Bit Dataset	15bit	31bit	63bit
Sector-48	0.8461	0.8948	0.9105
NewsGroup	0.8463	0.8745	0.8788

Table 4: the MacroF1 vs. the length of BCH coding

Bit Dataset	15bit	31bit	63bit
Sector-48	0.8459	0.8961	0.9122
NewsGroup	0.8430	0.8714	0.8757

We can clearly observe that increasing the length of the codes increases the classification accuracy. However, the increase in accuracy is not directly proportional to the increase in the length of the code. As the codes get larger, the accuracies start leveling off as we can observe from the two tables.

5. Conclusion Remarks

In this work, we examine the use of ECOC for improving centroid text classifier. The implementation framework is to decompose multi-class problems into multiple binary problems and then learn the individual binary classification problems by centroid classifier. Meanwhile, Model-Refinement is employed to reduce the bias incurred by ECOC.

In order to investigate the effectiveness and robustness of proposed method, we conduct an extensive experiment on two commonly used corpora, i.e., Industry Sector and Newsgroup. The experimental results indicate that the combination of ECOC with Model-Refinement makes a considerable performance improvement over traditional centroid classifier, and even performs comparably with SVM classifier.

References

Berger, A. *Error-correcting output coding for text classification*. In Proceedings of IJCAI, 1999.

Chai, K., Chieu, H. and Ng, H. *Bayesian online classifiers for text classification and filtering*. SIGIR. 2002, 97-104

Ghani, R. *Using error-correcting codes for text classification*. ICML. 2000

Ghani, R. *Combining labeled and unlabeled data for multiclass text categorization*. ICML. 2002

Han, E. and Karypis, G. *Centroid-Based Document Classification Analysis & Experimental Result*. PKDD. 2000.

Liu, Y., Yang, Y. and Carbonell, J. *Boosting to Correct Inductive Bias in Text Classification*. CIKM. 2002, 348-355

Rennie, J. and Rifkin, R. *Improving multiclass text classification with the support vector machine*. In MIT. AI Memo AIM-2001-026, 2001.

Sebastiani, F. *Machine learning in automated text categorization*. ACM Computing Surveys, 2002,34(1): 1-47.

Tan, S., Cheng, X., Ghanem, M., Wang, B. and Xu, H. *A novel refinement approach for text categorization*. CIKM. 2005, 469-476

Yang, Y. and Pedersen, J. *A Comparative Study on Feature Selection in Text Categorization*. ICML. 1997, 412-420.