Squeezing the Ensemble Pruning: Faster and More Accurate Categorization for News Portals

Cagri Toraman and Fazli Can

Bilkent IR Group, Computer Engineering Department
Bilkent University, 06800, Ankara, Turkey
{ctoraman,canf}@cs.bilkent.edu.tr

Abstract. Recent studies show that ensemble pruning works as effective as traditional ensemble of classifiers (EoC). In this study, we analyze how ensemble pruning can improve text categorization efficiency in time-critical real-life applications such as news portals. The most crucial two phases of text categorization are training classifiers and assigning labels to new documents; but the latter is more important for efficiency of such applications. We conduct experiments on ensemble pruning-based news article categorization to measure its accuracy and time cost. The results show that our heuristics reduce the time cost of the second phase. Also we can make a trade-off between accuracy and time cost to improve both of them with appropriate pruning degrees.

Keywords: Ensemble pruning, news portal, text categorization.

1 Introduction

In real-life applications like news portals (e.g., http://news.google.com), news articles coming from various resources have to be categorized efficiently and effectively. Ensemble-based learning is a solution for accurate text categorization. It is more effective than single classifier-based learning when ensemble of classifiers are accurate and diverse [1]. However, they are not efficient due to computational workload. Ensemble pruning (selection) is a way to make it faster. There are several ensemble selection studies [4]. They do not consider how ensemble pruning works for text categorization in time-critical real-life applications like news portals. In a recent study [3], we demonstrate that ensemble pruning can increase accuracy of text categorization. In this study, we examine ensemble pruning to analyze how fast it works in news portals and heuristics to trade-off between efficiency and effectiveness.

2 Experimental Environment

We employ homogeneous (base classifiers are trained by the same algorithm) data partitioning methods to provide accurate and diverse ensembles: bagging

R. Baeza-Yates et al. (Eds.): ECIR 2012, LNCS 7224, pp. 508–511, 2012.

Table 1. Traditional ensemble of classifiers (EoC)'s accuracy and average train/test time (in min)

	Accuracy / Train Time / Test Time			
	C4.5	KNN	NB	SVM
Bagging	0.8645/15.27/0.35	0.8069/0.08/7.75	0.7913/0.60/1.79	0.8722/2.28/0.41
Disjunct	0.8494/1.02/0.33	0.8047/0.09/1.14	0.8351/0.13/1.83	0.8427/0.28/0.41
Fold	0.8582/14.20/0.35	0.7917/0.09/7.26	0.7787/0.57/1.74	0.8713/1.91/0.42
Random-size	0.8607/6.40/0.33	0.8143/0.06/3.85	0.8101/0.33/1.77	0.8273/1.06/0.41

(randomly selecting N documents with replacement where N is the training set size), disjunct (randomly selecting N/k documents for each k equal-size partitions where k is the ensemble size), fold partitioning (k equal-size partitions and k-1 partitions except k_i are trained for each partition k_i), and random-size sampling (similar to bagging, but with random-size samples). Four popular categorization algorithms are then applied: C4.5, KNN (k-nearest Neighbor), NB (Naïve Bayes), and SVM (Support Vector Machines). Ranked-based ensemble pruning is applied to prune ensembles. Base classifiers are ranked according to accuracy on a separate validation set and then pruned 10% to 90% by 10% increments.

All experiments are repeated 10 times and results are averaged. Documents are represented with term frequency vectors. The most frequent 100 unique words per category are chosen. Ensemble size is chosen as 10. Documents for validation set is selected randomly as 5% of the training set. We measure accuracy (the ratio of number of correctly classified documents to all classified documents) and time in minutes (min) for effectiveness and efficiency respectively. The experiments were conducted on Intel Core2 Duo processor (3.00 Ghz) and 4 GB of RAM.

For repeatability of the experiments we use the Reuters-21578, which is a well-known benchmark dataset [2]. After splitting it with ModApte [2], which keeps documents with at least one topic and removes ones with no topics, eliminating multi-class documents and choosing the 10 most frequent topics, we get 5,753 training and 2,254 test news articles.

3 Experimental Results

Table 1 lists accuracy values of traditional EoC. Average train and test times for each scenario are also given in the same table. Ensemble pruning degrees that can be used for keeping accuracy as the same as in traditional EoC, affirmed by unpaired t-test, are given in our recent study [3]. We apply these pruning degree values and measure train and test time for each scenario. Table 2 lists time difference between traditional EoC and our ensemble pruning heuristics.

In [3], we demonstrate that ensemble pruning also improves accuracy of traditional EoC (usually with NB and SVM) in some pruning degrees by applying unpaired t-test. Using those heuristics pruning degrees, Table 3 lists how many more news articles we can correctly categorize with train and test times.

Table 2. Average train/test time difference (in min) between traditional EoC and our ensemble pruning heuristics that keeps accuracy the same

	Difference for Train Time / Test Time				
	C4.5	KNN	NB	SVM	Avg.
Bagging	1.40/0.04	-0.88/1.18	-0.21/1.60	0.12/0.24	0.10/0.76
Disjunct	0.03/0.02	-0.15/0.35	-0.23/1.07	-0.07/0.16	-0.10/0.40
Fold	*	*	-0.18/1.56	0.04/0.21	-0.07/0.88
Random-size	-0.01/0.02	-0.52/-0.02	-0.18/1.58	-0.01/0.36	-0.18/0.48
Avg.	0.47/0.02	-0.51/0.50	-0.20/1.45	0.02/0.24	-

Table 3. CLNA/avg. train time/avg. test time differences (in min) between traditional EoC and our ensemble pruning heuristics. (CLNA: Correctly Labeled News Articles).

	Difference for # of CLNA / Train Time / Test Time	
	NB	SVM
Bagging	13.0/-0.20/0.70	*
Disjunct	9.6/-0.22/0.71	3.1/-0.04/0.13
Fold	8.3/-0.20/1.01	*
Random-size	6.9/-0.19/0.48	87.9/0.05/0.15
Avg.	9.4/-0.20/0.72	45.5/-/0.14

* Our heuristics supported with unpaired t-test imply no pruning degree improves accuracy of this scenario, thus we do not conduct its experiment.

4 Discussion and Conclusion

Discussion. Consider a news portal that receives 5,000 news articles coming from RSS feeders. In the following examples, bagging with NB is considered and they are projections based on the above experimental results of 2,254 articles.

When traditional EoC is used, it takes 3.97 min to assign category labels to new documents. If our heuristics used in Table 2 are applied (i.e accuracy is maintained), the training time is then reduced to 0.42 min. This is a huge difference for on-line applications like news portals that take customer loyalty into account and present news articles to users as quickly as possible. Our pruning heuristics do not always reduce the training time; but a news portal trains its model within long periods of time while using this model for assigning labels frequently.

Another aspect is the ensemble size. Using a larger ensemble size improves accuracy in most cases as our recent study suggests. We demonstrate the same for the ensemble of 50 base classifiers (the associated tables are not given due to limited space). The time for assigning labels in an ensemble size of 50 is 18.87 min (8.51 for 2,254 articles), which is approximately 5 times more than the time for the same process in ensemble size of 10. Time reduction depends on the ensemble pruning degree. Assuming our heuristics for ensemble size of 10 is

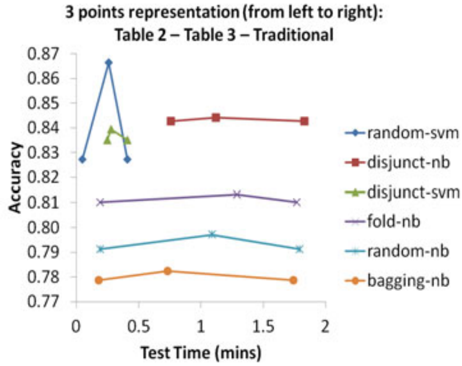

Fig. 1. Accuracy vs. train time: Trade-off between effectiveness and efficiency

appropriate for size of 50, we reduce the time for assigning labels from 18.87 to 1.99 min (0.90 for 2,254 articles). Thus, large-sized ensembles can be exploited efficiently with our pruning heuristics.

Ensemble pruning can also improve accuracy of traditional EoC. Our heuristics used in Table 3 provide approximately an additional 29 of 5,000 news articles correctly labeled. It also reduces the time for assigning labels from 3.97 to 2.41 min, yet it is larger than 0.42 min (when accuracy is maintained using Table 2). This means we can make a trade-off between time efficiency and accuracy with appropriate pruning degrees. Figure 1 illustrates the trade-off idea by considering NB and SVM cases given in Table 3. There are three points for each scenario. From left to right they represent accuracy and test time values of Table 2 (where accuracy is maintained as in traditional EoC, thus accuracy of this point is represented as the same as traditional), Table 3 (where time is sacrificed for the sake of accuracy), and Table 1 (where traditional EoC is applied).

Conclusion. In this work we analyze ensemble pruning for text categorization in time-critical real-life applications like news portals. The experimental results demonstrate that our pruning heuristics improve efficiency in most cases and even we can make a trade-off between accuracy and time.

Acknowledgments. This study is supported by TÜBİTAK grant no. 111E030.

References

1. Dietterich, T.G.: Ensemble Methods in Machine Learning. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 1–15. Springer, Heidelberg (2000)
2. Lewis, D.D., Ringuette, M.: A comparison of two learning algorithms for text categorization. In: SDAIR, pp. 81–93 (1994)
3. Toraman, C., Can, F.: Ensemble Pruning for Text Categorization Based on Data Partitioning. In: Salem, M.V.M., Shaalan, K., Oroumchian, F., Shakery, A., Khelalfa, H. (eds.) AIRS 2011. LNCS, vol. 7097, pp. 352–361. Springer, Heidelberg (2011)
4. Tsoumakas, G., Partalas, I., Vlahavas, I.: A taxonomy and short review of ensemble selection. In: ECAI, pp. 41–46 (2008)