Compact Dual Ensembles for Active Learning

Amit Mandvikar[1], Huan Liu[1], and Hiroshi Motoda[2]

[1] Arizona State University, Arizona, USA.
[2] Osaka University, Osaka, Japan.
{huanliu,amitm}@asu.edu and motoda@sanken.osaka-u.ac.jp

Abstract. Generic ensemble methods can achieve excellent learning performance, but are not good candidates for active learning because of their different design purposes. We investigate how to use diversity of the member classifiers of an ensemble for efficient active learning. We empirically show, using benchmark data sets, that (1) to achieve a good (stable) ensemble, the number of classifiers needed in the ensemble varies for different data sets; (2) feature selection can be applied for classifier selection from ensembles to construct compact ensembles with high performance. Benchmark data sets and a real-world application are used to demonstrate the effectiveness of the proposed approach.

1 Introduction

Active learning is a framework in which the learner has the freedom to select which data points are added to its training set [11]. An active learner may begin with a small number of labeled instances, carefully select a few additional instances for which it requests labels, learn from the result of those requests, and then using its newly-gained knowledge, carefully choose which instances to request next. More often than not, data in forms of text (including emails), image, multi-media are unlabeled, yet many supervised learning tasks need to be performed [2, 10] in real-world applications. Active learning can significantly decrease the number of required labeled instances, thus greatly reduce expert involvement. Ensemble methods are learning algorithms that construct a set of classifiers and then classify new instances by taking a weighted or unweighted vote of their predictions. An ensemble often has smaller expected loss or error rate than any of the n individual (member) classifiers. A good ensemble is one whose members are both *accurate* and *diverse* [4]. This work explores the relationship between the two learning frameworks, attempts to take advantage of the learning performance of ensemble methods for active learning in a real-world application, and studies how to construct ensembles for effective active learning.

2 Our Approach

2.1 Ensembles and Active Learning

Active learning can be very useful where there are limited resources for labeling data, and obtaining these labels is time-consuming or difficult [11]. There exist widely used active learning methods. Some examples are: Uncertainty sampling [7] selects the instance on which the current learner has lowest certainty; Pool-based sampling [9] selects the best instances from the entire pool of unlabeled instances; and Query-by-Committee [6, 12] selects instances that have high classification variance themselves.

Constructing good ensembles of classifiers has been one of the most active areas of research in supervised learning [4]. The main discovery is that ensembles are often much more accurate than the member classifiers that make them up. A necessary and sufficient condition for an ensemble to be more accurate than any of its members is that the member classifiers are accurate and diverse. Two classifiers are diverse if they make different (or uncorrelated) errors on new data points. Many methods for constructing ensembles have been developed such as Bagging [3] and Boosting [5]. We consider Bagging in this work as it is the most straightforward way of manipulating the training data to form ensembles [4].

Disagreement or diversity of classifiers are used for different purposes for the two learning frameworks: in generic ensemble learning, diversity of classifiers is used to ensure high accuracy by voting; in active learning, disagreement of classifiers is used to identify critical instances for labeling. In order for active learning to work effectively, we need a *small* number of *highly* accurate classifiers so that they seldom disagree with each other. Since ensemble methods have shown their *robustness* in producing *highly accurate* classifiers, we have investigated the use of class-specific ensembles (dual ensembles), and shown their effectiveness in our previous work [8]. Next, we empirically investigate whether it is necessary to find compact dual ensembles and then we present a method to find them while maintaining good performance.

2.2 Observations from Experiments on Benchmark Data Sets

Ensemble's goodness can be measured by accuracy and diversity. Let $\hat{Y}(x) = \hat{y}_1(x), ...\hat{y}_n(x)$ be the set of the predictions made by member classifiers $C_1, ..., C_n$ of ensemble E on instance $\langle x, y \rangle$ where x is input, and y is the true class. The **ensemble prediction** of a uniform voting ensemble for input x under loss function l is, $\hat{y}(x) = argmin_{y \in Y} E_{c \in C}[l(\hat{y}_c(x), y]$. The **loss** of an ensemble on instance $\langle x, y \rangle$ under loss function l is given by $L(\langle x, y \rangle) = l(\hat{y}(x), y)$. The **diversity** of an ensemble on input x under loss function l is given by $\overline{D} = E_{c \in C}[l(\hat{y}_c(x), \hat{y}(x))]$. The error rate for a data set with N instances can be calculated as $e = \frac{1}{N} \sum_1^N L_i$, where L_i is the loss for instance x_i. **Accuracy** of ensemble E is $1 - e$. **Diversity** is the expected loss incurred by the predictions of the member classifiers relative to the ensemble prediction. Commonly used loss functions include square loss, absolute loss, and zero-one loss. We use zero-one loss in this work.

The purpose of these experiments is to observe how diversity and error rate change as ensemble size increases. We use benchmark data sets from the UCI repository [1] in these experiments. We use Weka [13] implementation of Bagging [3] as the ensemble generation method and J4.8 (without pruning) as the base learning algorithm. For each data set, we run Bagging with increasing ensemble sizes from 5 to 151 and record each ensemble's error rate e and diversity

D. We run 10-fold cross validation and calculate the average values, $\overline{e}$ and $\overline{D}$. We observed that as the ensemble sizes increase, diversity values increase and approach to the maximum, and error rates decrease and become stable. The results show that smaller ensembles (with 30-70 classifiers) can achieve accuracy and diversity values similar to those of larger ensembles. We will now show a procedure for selecting compact dual ensembles from larger ensembles.

2.3 Selecting Compact Dual Ensembles via Feature Selection

The experiments with the benchmark data sets show that there exist smaller ensembles that can have similar accuracy and diversity as that of large ensembles. We need to select classifiers with these two criteria. We build our initial ensemble, E_{max} by setting $max = 100$ member classifiers. We now need to *efficiently find* a compact ensemble E_M (with M classifiers) that can have similar error rate and diversity of E_{max}. We use all the learned classifiers (C_k) to generate predictions for instances $\langle x_i, y_i \rangle : \hat{y}_i^k = C_k(x_i)$. The resulting dataset consists of instances of the form $((\hat{y}_i^1, ..., \hat{y}_i^K), y_i)$. After this data set is constructed, the problem of selecting member classifiers becomes one of feature selection. Here features actually represent member classifiers, therefore we also need to consider this special nature for the feature selection algorithm.

```
DualE: selecting compact dual ensembles

input:    $Tr$: Training data, $FSet$: All classifiers in $E_{max}$, $N$: $max$;
output:   $E_1$: Optimal ensemble for class=1, $E_0$: Optimal ensemble for class=0;

01  Generate $N$ classifiers from $T_r$ with Bagging;
02  $Tr_1 \leftarrow$ Instances($Tr$) with class label= 1;
03  $Tr_0 \leftarrow$ Instances($Tr$) with class label= 0;
04  Calculate diversity, $D_0$ and error rate, $e_0$ for $E_{max}$ on $Tr_1$;
05  $U \leftarrow N$; $L \leftarrow 0$; $M \leftarrow \frac{U-L}{2}$;
06  while $|U - M| > 1$
07      Pick $M$ classifiers from $FSet$ to form $E'$;
08      Calculate diversity, $D'$ and error rate, $e'$ for $E'$ on $Tr_1$;
09      if $(\frac{D_0-D'}{D_0} < 1\%)$ and $(\frac{e'-e_0}{e_0} < 1\%)$
10          $U \leftarrow M$; $M \leftarrow M$ - $\frac{M-L}{2}$;
11      else
12          $L \leftarrow M$; $M \leftarrow M + \frac{U-M}{2}$;
13  $E_1 \leftarrow E'$;
14  Repeat steps 5 to 12 for $Tr_0$; $E_0 \leftarrow E'$;
```

Fig. 1. Algorithm for Selecting Dual Ensembles

We design an algorithm **DualE** that takes $O(\log max)$ to determine M where max is the size of the starting ensemble (e.g., 100). In words, we test an ensemble E_M with size M which is between upper and lower bounds U and L (initialized as max and 0 respectively). If E_M's performance is similar to that of E_{max}, we set $U = M$ and $M = (L+M)/2$; otherwise, set $L = M$ and $M = (M+U)/2$. The details are given in Fig. 1. Ensemble performance is defined by error rate e and diversity D. The diversity values of the two ensembles are similar if $\frac{D_0 - D'}{D_0} \leq p$ where p is user defined $(0 < p < 1)$ and D_0 is the reference ensemble's (E_{max}'s) diversity. Similarly, error for the two ensembles is similar if $\frac{e' - e_0}{e_0} \leq p$ where e_0 is the reference ensemble's error rate.

3 Experiments

Two sets of experiments are conducted with **DualE**: one is on a image data set and other on a benchmark data set (breast). The purpose is to examine if the compact dual ensembles selected by **DualE** can work as well as the entire ensemble, E_{max}. When dual ensembles are used, it is possible that they disagree. These instances are called *uncertain* instances (UC). In context of active learning, the uncertain instances will be labeled by an expert. Prediction of E_{max} is by majority and there is no disagreement. So for E_{max} only the accuracy is reported.

The image mining problem, that we study here, is to classify Egeria Densa in aerial images. To automate Egeria classification, we ask experts to label images, but want to minimize the task. Active learning is employed to reduce this expert involvement. The idea is to let experts label some instances, learn from these labeled instances, and then apply the active learner to unseen images. We have 17 images with 5329 instances, represented by 13 attributes of color, texture and edge. One image is used for training and the rest for testing. We first train an initial ensemble E_{max} with $max = 100$ on the training image, then obtain accuracy of E_{max} for the 17 testing images. Dual E_s are the ensembles selected by using **DualE**. Table 3 clearly shows that accuracy of E_s is similar to that of E_{max}. The number of uncertain regions is also relatively small. This demonstrates the effectiveness of using the dual ensembles, E_s to reduce expert involvement for labeling.

Table 1. E_s vs. E_{max} for Image data

Image	Dual E_s		E_{max}	
	Acc%	#UC	Acc%	Acc Gain%
1	81.91	1	81.90	-0.0122
2	90.00	0	90.00	0.0000
3	78.28	38	79.28	1.2775
4	87.09	34	86.47	-0.7119
5	79.41	0	79.73	0.4029
6	84.51	88	84.77	0.3076
7	85.00	3	85.41	0.4823
8	85.95	18	86.6	0.7562
9	71.46	0	72.32	1.2035
10	91.08	2	90.8	-0.3074
11	89.15	31	88.82	-0.3702
12	75.91	0	76.02	0.1449
13	66.84	0	67.38	0.8079
14	73.06	49	73.73	0.9170
15	83.1	1	83.24	0.1684
16	76.57	14	76.82	0.3265
17	87.67	31	88.42	0.8555
Average	81.58	18.24	81.86	0.3676

For the breast dataset, we design a new 3-fold cross validation scheme, which uses 1-fold for training, the remaining 2 folds for testing. This is repeated for all the 3 folds of the training data. The results for the breast data set are shown in Table 2. We also randomly select member classifiers to form random dual ensembles, Dual E_r. We do so 10 times and report the average accuracy and number of uncertain instances. In Table 2, Dual E_s are the selected ensembles (using **DualE**), and E_{max} is the initial ensemble. Accuracy gains for Dual E_r and E_{max} (and UC Incr for Dual E_r) against E_s are reported. Comparing dual

Table 2. Comparing E_s with E_r and E_{max} for Breast data

	Dual E_s		Dual E_r				E_{max}	
	Acc%	#UC	Acc%	#UC	Acc Gain%	UC Incr%	Acc%	Acc Gain%
Fold 1	95.9227	3	94.0773	13.6	-1.9238	353.33	96.1373	0.2237
Fold 2	97.2103	5	94.4206	15.2	-2.8698	204.00	96.9957	-0.2208
Fold 3	94.8498	12	93.5193	8.7	-1.4027	-27.50	94.4206	-0.4525
Average	95.9943	6.67	94.0057	12.5	-2.0655	176.61	95.8512	-0.1498

E_s and dual E_r, dual E_r exhibit lower accuracy and more uncertain instances. Comparing dual E_s and E_{max}, we observe no significant change in accuracy. This is consistent with what we want (maintain both accuracy and diversity).

4 Conclusions

In this work, we point out that (1) generic ensemble methods are not suitable for active learning (2) dual ensembles are very good for active learning if we can build compact dual ensembles. Our empirical study suggests that there exist such compact ensembles. We propose **DualE** that can find compact ensembles with good performance via feature selection. Experiments on a benchmark and an image data set exhibit the effectiveness of dual ensembles for active learning.

References

1. C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998. http://www.ics.uci.edu/∼mlearn/MLRepository.html.
2. A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT: Proc. of Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*, 1998.
3. L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.
4. T.G. Dietterich. Ensemble methods in machine learning. In *Proc. of Intl. Workshop on Multiple Classifier Systems*, pp. 1–15. Springer-Verlag, 2000.
5. Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In *Proc. of Intl. Conf. on Machine Learning*, pp. 148–156. Morgan Kaufmann, 1996.
6. Y. Freund, H. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28:133–168, 1997.
7. D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *Proc. of ACM-SIGR Conference on Research and Development in Information Retrieval*, pp. 3 – 12, 1994.
8. A. Mandvikar and H Liu. Class-Specific Ensembles for Active Learning in Digital Imagery. In *Proc. of SIAM Intl. Conf. on Data Mining*, in press, 2004.
9. A. McCallum and K. Nigam. Employing EM in pool-based active learning for text classification. In *Proc. of Intl. Conf. on Machine Learning*, pp. 350–358, 1998.
10. K. Nigam, A. K. Mccallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents usingEM. *Machine Learning*, 39:103–134, 2000.
11. N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proc. of Intl. Conf. On Machine Learning*, 2001.
12. H.S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proc. of Workshop on Computational Learning Theory*, pp. 287–294, 1992. ACM Press.
13. I.H. Witten and E. Frank. *Data Mining - Practical Machine Learning Tools and Techniques with JAVA Implementations.* Morgan Kaufmann Publishers, 2000.