Multi-class Ensemble-Based Active Learning

Christine Körner[1] and Stefan Wrobel[1,2]

[1] Fraunhofer Institut Intelligente Analyse- und Informationssysteme, Germany
{christine.koerner, stefan.wrobel}@iais.fraunhofer.de
[2] Dept. of Computer Science III, University of Bonn, Germany

Abstract. Ensemble-based active learning has been proven to efficiently reduce the number of training instances and thus the cost of data acquisition. To determine the utility of a candidate training instance, the disagreement about its class value among the ensemble members is used. While the disagreement for binary classification is easily determined using margins, the adaption to multi-class problems is not straightforward and little studied in the literature. In this paper we consider four approaches to measure ensemble disagreement, including margins, uncertainty sampling and entropy, and evaluate them empirically on various ensemble strategies for active learning. We show that margins outperform the other disagreement measures on three of four active learning strategies. Our experiments also show that some active learning strategies are more sensitive to the choice of disagreement measure than others.

1 Introduction

Ensemble-based active learning is well-known to effectively choose training instances when resources for labeled data are limited. Its most prominent representatives are query-by-bagging and query-by-boosting [1], co-testing [2] and active-decorate [3]. All four strategies choose training instances based on the disagreement among their ensemble members. For binary-class learning problems ensemble disagreement is simply measured by the difference between positive and negative votes. However, it is not obvious how this approach can be generalized to determine ensemble disagreement in multi-class learning problems. Existing literature has consequently proposed a variety of techniques, including margins [2], uncertainty sampling [4, 5] and entropy [6, 7, 3, 8]. Surprisingly, no study exists that evaluates which of these methods is most suitable for ensemble-based active learning or whether the application of a method depends on the chosen ensemble strategy.

In this paper we compare the three disagreement measures proposed in the literature along with a "control" measure that combines different aspects of existing measurements. In a comprehensive set of experiments on 12 different learning problems, we evaluate all four disagreement measures empirically in the context of the four most prominent ensemble-based active learning strategies, namely query-by-bagging and query-by-boosting [1], co-testing [2] and active-decorate [3]. We show that margins outperform other query selection strategies

J. Fürnkranz, T. Scheffer, and M. Spiliopoulou (Eds.): ECML 2006, LNAI 4212, pp. 687–694, 2006.

on three of four active learning strategies. At the same time, we observe that for query-by-bagging and co-testing the choice of disagreement measure is essential to the success of the active learner, while query-by-boosting and active-decorate perform quite robust using different disagreement measures. The results of our experiments clearly demonstrate that from the existing disagreement measures considered in the literature, the margin-based approach should be chosen as a standard approach for multi-class ensemble-based active learning.

The paper is organized as follows. Section 2 reviews active learning strategies for ensembles. Section 3 presents the disagreement measures for the multi-class case. We evaluate the presented disagreement measures in Section 4 on 12 UCI domains and conclude the paper with future work.

2 Ensemble-Based Active Learning

In this section we review the idea of ensemble-based active learning and describe four active learning strategies which we will use in our experiments. Ensemble-based active learning originates in the query-by-committee approach by Seung et al. [9]. Query-by-committee is a form of query filtering where a stream of unlabeled instances is provided from which the algorithm chooses the most profitable for labeling [10]. The utility of a candidate instance is evaluated by an ensemble of randomly selected hypotheses from the version space, the subset of all hypotheses consistent with the training data. The stronger the committee disagrees on a class label the more valuable is the query. Assuming an infinite number of committee members and an equal number of positive and negative votes (maximal disagreement in binary classification), the knowledge about the query's true class label will halve the version pace. Query-by-committee is an iterative algorithm that adds the queried instance and its label to the training set and repeats until a desired accuracy or the quota for labeling is reached.

In the remainder of this section we review four acknowledged strategies for active learning that spring from the idea of query-by-committee but use different randomization strategies in order to create the ensembles. The presented strategies are query-by-bagging, query-by-boosting, co-testing and active-decorate.

Query-by-bagging and query-by-boosting [1] rely on sampling strategies that randomize the training data before a deterministic learning algorithm (typically C4.5) builds one classifier from each subsample. As the name implies, query-by-bagging utilizes Bagging [11] as sampling strategy, drawing each subsample with replacement. Once the ensemble is formed, the committee votes on the (binary) class values of all unlabeled instances and randomly selects a query from all instances that split the committee most evenly.

Query-by-boosting proceeds similar to query-by-bagging but uses AdaBoost [12] to create differing training sets. AdaBoost is in itself an iterative algorithm that, starting from the original sample distribution, builds a classifier but adapts the distribution to emphasize misclassified instances before the next training set is drawn. Again, an instance with the smallest margin between the number of positive and negative votes is chosen as query, but the votes are weighted according to the training error of each committee member.

Co-testing [2] is an active learning strategy that is inspired by the multi-view approach called co-training [13,14]. It utilizes two redundant views of the training data to create an ensemble and selects, in its naive approach, a query among all unlabeled instances where the two classifiers disagree. Although co-testing relies on independent views, it has been shown to perform well using random splits in domains without redundant attributes [2, 15].

Active-decorate [3] is a recent approach to ensemble-based active learning and uses artificially enhanced training sets. The underlying principle, Decorate [16], increases the size of the ensemble iteratively, starting with one classifier based on the original training set. Afterwards, it constructs new training instances assuming independent attribute distributions and labels them inversely proportional to the prediction of the current ensemble. A new classifier build from the original and artificial training data is added to the ensemble if it reduces the training error of the ensemble. Active-decorate uses margins to measure ensemble disagreement but generalizes the idea to multi-class problems.

3 Disagreement Measures for Multi-class Ensembles

For binary classification ensemble disagreement can be easily determined. It is large if the number of positive and negative votes of the ensemble are evenly split. It is small if one class prevails. However, the generalization to multi-class problems is not that straightforward. Assume that an ensemble of ten classifiers votes on two instances with four possible class values. The vote distributions for instance one and two are $d_1 = (3, 3, 2, 2)$ and $d_2 = (5, 5, 0, 0)$ respectively. Which instance should the active learner recommend? Obviously, the distribution for instance one is very homogeneous. Yet, the contradiction for instance two is fiercer as it targets class one and two.

This section describes four techniques to measure ensemble disagreement in multi-class problems, which we evaluate in Section 4. The first three techniques, margins, uncertainty sampling and entropy, are commonly used in literature. The fourth, specific disagreement, is a “control” measure which we developed to contrast existing approaches. For all measures we assume that an ensemble returns a probability distribution of the class value for each unlabeled instance obtained either by majority vote or by averaging the class distributions of the committee members. Fig. 1 visualizes all four disagreement measures for a three-class problem. Each picture illustrates the disagreement for all possible class distributions $d = (p_1, p_2, p_3)$. The x- and y-axis contain the class probabilities p_1 and p_2 respectively while p_3 is indirectly depicted by isolines with a gradient of -1. Dark colors indicate preferred sections for query selection.

Margin-based disagreement: Following the generalization of binary margins as proposed by Melville and Mooney [3], the margin in multi-class problems is calculated as difference between the first and second highest class probability. A strategy that chooses an instance with minimum margin thus evaluates the competitiveness of the most likely class label. Nevertheless, it does not consider any information about the remaining class probabilities or the level of probability on which a margin occurs.

Uncertainty sampling-based disagreement: Uncertainty sampling [4, 5] provides a second way to generalize the binary margin approach and can be applied to any classifier that provides a class label along with an estimate about the confidence in its prediction. Uncertainty sampling simply queries an instance of which the predicted class value possesses a minimum probability among all candidate instances. Thus, uncertainty sampling accounts for the level of probability. It indirectly prefers candidates with a balanced class distribution but again does not benefit from information about the remaining class probabilities.

Entropy-based disagreement: Entropy is a well-known measure in information theory to determine the disorder of a system. In ensemble-based active learning various forms, ranging from ordinary entropy [6] to Kullback-Leibler divergence [7] and Jensen-Shannon divergence [8], have been applied to train probabilistic classifiers. In our experiments we focus on ordinary entropy defined as $E = -\sum_{i=1}^{k} p_i log_2 p_i$ for a k-class problem. Again, entropy generalizes disagreement as defined for binary classification.

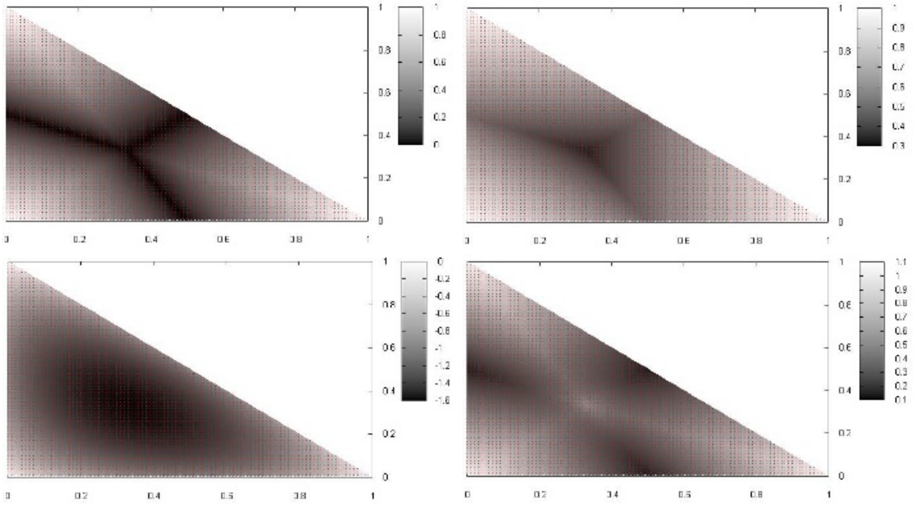

Fig. 1. Visualization of disagreement measures for a three-class problem; top left: margin-based; top right: uncertainty sampling-based; bottom left: entropy-based; bottom right: specific disagreement

Specific disagreement ("control"): The above approaches select queries either by degree of competition between the first two predominant strategies (dark lines, Fig. 1) or according to homogeneity of distribution (dark centers, Fig. 1). Yet, Muslea [15] pointed out that disagreement between two differing predictions also increases with the level of confidence. We therefore designed a "control" measure, specific disagreement, which combines different aspects of the above measures to indicate disagreement on a narrow subset of class values. Our measure combines margin-based disagreement ($margin$) with the maximal class probability (max), normalized with the total number of class values ($|c|$): $specific\,disagreement = margin + 0.5\,\frac{1}{(|c|\cdot max)^3}$.

4 Experiments

We evaluated the disagreement measures introduced in Section 3 on 12 data sets from the UCI repository [17] as given in Table 1[1]. We applied all measures to the ensemble-based active learning strategies query-by-bagging, query-by-boosting, co-testing and active-decorate. With the exception of co-testing, each ensemble consisted of 20 committee members. The ensembles used C4.5 as base learner and were configured according to their default parameters in the WEKA toolkit [18]. We initiated the active learners with 50 randomly drawn training instances and evaluated the experiments based on 2x10-fold cross validation. During each iteration we added 1 query to the training set and proceeded until all available data was used or a maximum of 250 queries were issued. In order to ascertain the effect of active learning, each ensemble strategy was additionally evaluated on a random sequence of training instances.

Table 1. Characteristics of UCI data sets

data set	attributes			inst	accuracy		data set	attributes			inst	accuracy
	num	sym	class		C4.5			num	sym	class		C4.5
abalone	7	1	3	4177	60.25		optdigits	64	0	10	5620	90.69
bupa	6	0	2	345	68.70		pima	8	0	2	768	73.83
car	0	6	4	1728	92.65		segment.	19	0	7	2310	97.23
ecoli	7	0	8	336	85.07		vehicle	18	0	4	846	71.95
glass	9	0	7	214	65.88		wdbc	30	0	2	569	94.01
letter	16	0	26	20000	88.06		yeast	8	0	10	1484	56.10

We apply two techniques to compare the performance of disagreement measures. The first performs a pairwise comparison of disagreement measures. The second ranks the disagreement measures according to their number of queries that are necessary to reach some target error rate. Both techniques are conducted independently for each active learning strategy.

During pairwise comparison we conduct a z-test on the (over 20 trials averaged) ensemble accuracies after a new training instance has been added. We count for the first 200 queries how often each strategy significantly outperforms the other.[2] If the difference between the individual counts exceeds a threshold

[1] For letter and optdigits we used only the first 5000 and 2500 instances respectively.

[2] Note, that the z-tests are not independent because queries are added sequentially.

of 20, we assign one point to the superior disagreement measure on the given data set. We aggregate the scores over all UCI data sets and calculate total wins, losses and ties per disagreement measure. Table 2 shows the results of pairwise comparison separately for each active learning strategy. A score of 2 in cell (2, 1) means, for example, that the disagreement measure in row 2 outperformed the disagreement measure in column 1 on 2 out of 12 UCI data sets.

The second evaluation technique estimates how efficient an active learner uses the data and is similar to measures used in [1, 3]. It compares the number of training instances necessary to reach a certain target error rate, calculated as average error of the last 50 training examples given a random sequence of queries. In contrast to [3] we record the training set size on the third occurrence the target error rate is reached. This correction proved necessary because the error rate on consecutive queries showed great variation. We ranked the results for each UCI data set and calculated average ranks as shown in Table 3.

Table 2. Pairwise comparison of disagreement measures

query-by-boosting

	mar.	unc.	ent.	spe.	ran.	total: +	-	0
mar.	0	1	1	0	7	9	6	45
unc.	2	0	1	0	6	9	3	48
ent.	2	0	0	1	7	10	4	46
spe.	2	1	1	0	7	11	1	48
ran.	0	1	1	0	0	2	27	31

query-by-bagging

	mar.	unc.	ent.	spe.	ran.	total: +	-	0
mar.	0	4	4	7	9	24	0	36
unc.	0	0	4	5	5	14	4	42
ent.	0	0	0	2	7	9	10	41
spe.	0	0	1	0	5	6	16	38
ran.	0	0	1	2	0	3	26	31

co-testing

	mar.	unc.	ent.	spe.	ran.	total: +	-	0
mar.	0	5	6	3	5	19	0	41
unc.	0	0	3	0	3	6	12	42
ent.	0	0	0	0	3	3	20	37
spe.	0	4	5	0	5	14	5	41
ran.	0	3	6	2	0	11	16	33

active-decorate

	mar.	unc.	ent.	spe.	ran.	total: +	-	0
mar.	0	2	2	6	7	17	1	42
unc.	1	0	1	6	7	15	3	42
ent.	0	0	0	5	7	12	5	43
spe.	0	0	1	0	5	6	19	35
ran.	0	1	1	2	0	4	26	30

Table 3. Comparison by number of training instances using ranks

active learner	margin	unc. samp.	entropy	specific	random
query-by-boosting	2.79	3.08	2.38	2.00	4.75
query-by-bagging	1.92	2.42	2.88	3.38	4.42
co-testing	1.50	3.17	4.00	2.88	3.46
active-decorate	2.25	2.50	2.67	3.67	3.92

Before we compare the performance of disagreement measures for each active learning strategy, we would like to direct the attention on the consistent results of both evaluation techniques. High scores in pairwise comparison correspond to first ranks and vice versa. Furthermore, insignificant differences in the total scores of pairwise comparison are reflected in small variation between ranks.

When query-by-boosting serves as active learner, all disagreement measures are distinct superior to a random sequence of queries. The differences between individual measures though is very small and shows, except for a slight advantage of specific disagreement, no distinction. The results for query-by-bagging support again the general superiority of any disagreement measure over random query selection. Yet, a clear distinction between the disagreement measures exists. Margins achieve the best results, followed by uncertainty sampling, entropy and finally specific disagreement. The performance of co-testing is closely connected to the applied disagreement measure. Again, margins dominate the other approaches. Note, that only margins and specific disagreement perform better than a random strategy. The results for active-decorate show a general superiority of all disagreement measures over a random sequence, in the case of specific disagreement the distinction is only marginal. The descending order of margin-, uncertainty sampling- and entropy-based disagreement as found in query-by-bagging and co-testing is preserved, although the distance between the methods is much smaller.

To summarize the results, whenever a distinction between disagreement measures is obvious, margins receive the best results followed by uncertainty sampling and entropy. The quality of specific disagreement varies for different active learning strategies. While query-by-bagging and co-testing react very sensitive to different disagreement measures (in case of co-testing the application of uncertainty sampling and entropy even leads to worse results than a random query selection), query-by-boosting and active-decorate perform robust on all disagreement measures.

How can we explain the results? The poor performance of entropy-based disagreement may have already been anticipated from Fig. 1. It shows a broad and unspecific selection of queries. In fact, to give an example, the entropy of two distributions $d_1 = (0.5, 0.5, 0)$ and $d_2 = (0.77, 0.015, 0.015)$ is equal. Yet, d_1 is a good candidate for querying while d_2 is not. The good performance of margin-based disagreement can be explained by its focus on competitive strategies, among which it selects equally between uniform and non-uniform distributions. Neither uncertainty sampling-based nor specific disagreement, which shift the focus to a more or less uniform distribution respectively, perform as well. It implies that both, very undirected ensemble decisions as well as decisions which focus on a few highly confident choices, are essential to active learning and should not be disregarded. The robustness of query-by-boosting and active-decorate took us by surprise. We believe that as both ensemble methods interfere with the distribution of their training data, they are able to compensate the choice of less informative queries. However, a definite answer to this behavior needs further research.

5 Conclusion and Future Work

In this paper we present a detailed study which compares commonly used disagreement measures for multi-class ensemble-based active learning. We compare the measures empirically on four active learning strategies, namely query-by-boosting, query-by-bagging, co-testing and active-decorate. In a comprehensive set of experiments on 12 UCI domains we show the superiority of margin-based disagreement, which should be used as a standard approach. In addition, our evaluation shows that the sensibility to disagreement measures varies between active learning strategies. In future work we would like to improve the margin-based approach by enhancing it with further information on the class distribution. We also plan to expand our studies to include disagreement measures that base on the individual class probability distributions of the ensemble members.

References

1. Abe, N., Mamitsuka, H.: Query learning strategies using boosting and bagging. In: Proc. of ICML-98, Morgan Kaufmann (1998) 1–9
2. Muslea, I., Minton, S., Knoblock, C.A.: Selective sampling with redundant views. In: Proc. of AAAI-00, AAAI Press / The MIT Press (2000) 621–626
3. Melville, P., Mooney, R.: Diverse ensembles for active learning. In: Proc. of ICML-04, ACM (2004) 584–591
4. Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: Proc. of ICML-94, ACM (1994) 148–156
5. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: Proc. of SIGIR-94, ACM / Springer (1994) 3–12
6. Dagan, I., Engelson, S.: Committee-based sampling for training probabilistic classifiers. In: Proc. of ICML-95, Morgan Kaufmann (1995) 150–157
7. McCallum, A., Nigam, K.: Employing em and pool-based active learning for text classification. In: Proc. of ICML-98, Morgan Kaufmann (1998) 350–358
8. Melville, P., Yang, S.M., Saar-Tsechansky, M., Mooney, R.: Active learning for probability estimation using jensen-shannon divergence. In: Proc. of ECML-05, Springer (2005) 268–279
9. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proc. of COLT-92, ACM (1992) 287–294
10. Freund, Y., Seung, H.S., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Machine Learning **28**(2-3) (1997) 133–168
11. Breiman, L.: Bagging predictors. Technical report 421, University of California, Berkeley (1994)
12. Freund, Y., Schapire, R.: Experiments with a new boosting algorithm. In: Proc. of ICML-96, Morgan Kaufmann (1996) 148–156
13. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proc. of COLT-98, ACM (1998) 92–100
14. Nigam, K., Ghani, R.: Analyzing the effectiveness and applicability of co-training. In: Proc. of CIKM-00, ACM (2000) 86–93
15. Muslea, I.: Active Learning with Multiple Views. PhD thesis, University of Southern California (2002)
16. Melville, P., Mooney, R.: Constructing diverse classifier ensembles using artificial training examples. In: Proc. of IJCAI-03, Morgan Kaufmann (2003) 505–510
17. Blake, C.L., Merz, C.J.: Uci repository of of machine learning databases. (http://www.ics.uci.edu/∼mlearn/MLRepository.html)
18. Witten, I.H., Frank, E.: Data Mining - Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (2000)