Intelligent Information Systems 9999
ISBN 666-666-666, pages 1–15

Comparing Performance of Committee Based Approaches to Active Learning

Jerzy Stefanowski and Mateusz Pachocki

Institute of Computing Science, Poznan University of Technology, Poland

Abstract

In this paper we study the use of Query by Committee strategies in active learning from unlabeled examples. In this framework we compare four algorithms for creating ensembles of classifiers: bagging, boosting, decorate and random forests. We consider the use of different measures of disagreement among components classifiers to select the most informative example to query an oracle for its label. Moreover, we introduce a new technique, based on analysing the neighborhood of examples, which is applied to create a starting training set for generating the first ensemble. The usefulness of all these approaches is experimentally evaluated. Results of our experiments confirm that accurate final classifiers could be created using a relatively small number of queries to label examples, in particular for active decorate. Simpler disagreement measures, as margins of examples or median difference are usually more effective than more advanced ones, however mainly for multi-class data. Finally, the new technique for selecting starting examples has proved to be definitely more effective than simple sampling.

Keywords: active learning, query by committee, ensembles, disagreement measures

1 Introduction

Machine learning techniques for supervised learning require a sufficiently large number of training examples. It is even claimed that the more training data a learning algorithm gets, the more accurate classifier should be induced (Davy, 2005). However, in many real-life problems only a quite limited number of labeled examples is available. Typically these examples are manually labeled by human experts - which is time consuming, costly and unrealistic in case of processing thousands of instances. It may be problematic even for more automatic framework. On the other hand, in many domains unlabeled examples are much easier to be acquired. In particular, it refers to Web mining or text classification. For instance, Blum and Mitchell (1998) considered the classification of Web pages, where hand processing of pages by humans was very difficult, while much larger amount of unlabeled pages were easy and inexpensively gathered by web crawlers. Similar examples were described for spam filtering (cf. Kiritchenko and Matwin, 2001) and categorization of text documents (cf. Liere and Tadepalli, 1997).

Therefore, several researchers postulated that for such problems learning algorithms should be able to take as much advantages of unlabeled data as possible to produce an efficient classifier without the need for large amount of labeled training data. These are motivations for using a kind of semi-supervised approaches to construct classifiers. In a recent decade we could observe a growing research interest in such approaches. The most well known proposals are *co-training* (Blum and Mitchell, 1998) or *active learning* (Cohn at al., 1994). Both of them empirically proved to work with relatively reduced number of labeled examples.

In this paper we have chosen active learning. Generally speaking, the active algorithm starts with a very small number of labeled examples and analyzes unlabeled ones. As a result it presents queries (asks for labeling an example) to the expert (or oracle). These examples are further used to improve a classifier. The key issue is to select the most valuable examples and reduce the number of queries. Since the paper (Cohn at al., 1994) various techniques have been introduced – a brief review is given in section 2. In our study, we focus attention on *Query by Committee* approaches, which have been shown to be effective in different classification tasks (cf. Melvile and Monney, 2004; Liere and Tadepalli, 1997). They are based on using a multiple classifier to select the most informative examples, which are the ones leading to a high disagreement among component classifiers. In (Abe and Mamitsuka, 1998) boosting and bagging were used in an active way to construct such classifiers. More recently Melvile and Monney (2004) introduced *ActiveDecorate* and also showed that it significantly reduced labeling costs comparing to sampling strategies.

The main aim of this paper is to experimentally compare various approaches to committee based active learning. In some sense our study is inspired by earlier promising results from (Melvile and Monney, 2004). However, our new contribution is twofold. Firstly, in our experiments we extend the number of compared approaches, in particular by taking into account random forests. Then, we also consider the use of other disagreement measures. Moreover, we decided to study the influence of choosing a starting set of labeled examples on the learning process. Thus, our other methodological contribution is an introduction of a new technique for a focused selection of this starting set.

2 Previous Research

2.1 Active Learning

According to (Cohn at al., 1994) active learning refers to the type of learning where the algorithm has some control or influence over the training data received as input. In this sense it differs from the typical *passive learning*, where the algorithm usually needs a sufficiently large amount of data provided in a static way, however it does not control the choice of examples. The main difference is that active learning algorithm can *select* some "the most informative" examples to put into the training set and does not use the complete set that may contain non-informative examples. In the literature, there exits different meanings of "informative" - usually the knowledge of the label of this examples should be the most useful to reduce the search through the hypothesis space, see more discussion in (Davy, 2005).

An active learning algorithm starts with an initial, very limited set of labeled examples (quite often containing very few ones) producing a first classifier. In each iteration it somehow analyses the pool of remaining unlabeled examples and presents a *query* to the *oracle* to ask for labeling single (or a few) example. Then this new labeled example is added to the current training set and the classifier is again induced using the increased data. This phase is repeated until the given stopping criterion is met.

The key issue is that the query about the selected examples should be very informative and should allow the learning algorithm to improve learning process with a small training set. As it is discussed in the literature, see e.g. (Davy, 2005), this should significantly reduce the number of labeled examples needed in order to get a sufficiently accurate classifier. The oracle is typically a human being a domain expert. Usually it is assumed that it returns the correct classification label for the particular example. Commonly experts may be reluctant to answer too many questions, so the number of possible labelings he is willing to do (i.e. number of iterations in a loop) is a candidate for the stopping criterion. Another option is to observe changes of the accuracy of the final classifier obtained in each iteration – which may be more suitable for a more automatic framework or experiments.

This general procedure was realized in different ways depending mainly of the techniques used to select the most informative examples (i.e. queries). In the first paper Cohn at al. (1994) developed a *selective sampling* method, which draw examples from the so called uncertainty region of target concept. Conceptually similar is a method called *uncertainty sampling* which selects those examples whose class membership is the most uncertain (Lewis and Catlett, 1994). The experimental evaluations shown that they may reduce the number of labeled examples.

2.2 Query by Committee

Another group of active learning methods is *Query by Committee* (QBC). According to (Sueng et al., 2004) it is a strategy that uses many copies of "hypotheses" (coming from randomized learning algorithm) to select an unlabeled example at which their classification predictions are maximally spread. Then, it also uses a final ensemble of hypotheses to classify new objects. In the first proposal the component learning algorithm was Gibbs algorithm. Although it was well-theoretically analysed, practically it was computationally intractable (Abe and Mamitsuka, 1998; Davy, 2005). Therefore, other approaches include using popular methods for learning multiple classifiers (*ensembles*). In particular, quite effective in reducing labeling costs were two proposals of Abe and Mamitsuka (1998) called *Query by Bagging* and *Query by Boosting*. They were constructed around the general scheme presented below:

Input: Learning algorithm – A; Set of labeled training examples – L; Set of unlabeled training examples – U; Number of active learning iterations – k; Number of selected examples – m (default 1)

Repeat k times

1. Generate a committee of classifiers, $C^* = EnsembleMethod(A, L)$
2. $\forall x_j \in U$ compute $Information\_Value(C^*, x_j)$, based on the current committee

3. Select a subset S of m examples that are the most informative
4. Obtain label for examples in S from oracle
5. Remove examples in S from U and add to L

Return *EnsembleMethod*(A, L)

A method used to construct a committee of classifiers makes a difference between these approaches. The first approach is based on bagging (Breiman, 1996), which works on bootstrap sampling several times the training set. On the hand, Query by Boosting uses the more adaptive approach which sequentially constructs an ensemble by changing distribution of weights assigned to the training examples. In both versions the unlabeled example was selected to a query by a *measures of disagreement* in the committee about its predicted label. Abe and Mamitsuka (1998) proposed to use the *margin* of the example – it is defined as the difference between the number of votes in the committee for the most predicted class and that for the second predicted class. Examples with the smallest margins are considered as the most informative (uncertain). Empirical studies showed that these QBC active learning approaches were able to reduce several times number of labeled comparing to random selections.

Recently Melvile and Monney (2004) introduced another QBC approach, called *Active Decorate*. Following motivations for increasing diversity of component classifiers they put into the active learning loop the specific meta-learning algorithm *Decorate*. Due to the size of this paper we skip its formal description and can only say that it uses additional artificially generated training examples to construct more diversified component classifiers in the ensemble. A set of these classifiers is constructed iteratively as in adaptive approaches. In each iteration the new classifier is added to the ensemble and is generated on the original training data augmented by a number of artificial examples generated by a specific model which assigns class labels to differ them from the current predictions of the ensemble.

Matwin et al. (2008) introduced *ALASoft* approach, where results of active learning are further processed to generate a more comprehensive model in a form of decision tree.

3 Our Approach to Query by Committee Based Active Learning

3.1 Algorithms for Learning Committees

In our framework for using QBC we use the generic schema presented as the pseudocode in the previous section. Following related works (Abe and Mamitsuka, 1998; Melvile and Monney, 2004) we decided to choose the most effective approaches for generating committees of classifiers: *query by bagging*, *query by boosting* and *active decorate*. C4.5 algorithm for inducing decision trees was always applied to generate component classifiers in all of these approaches (also due to previous promising experimental results)

Besides these approaches we decided to check the usefullnes of *random forests* as an approach to generate a committee in active learning. Generally speaking

random forests, introduced by Breiman (2001), is a modification of bagging applied with decision trees, where for each node of the induced tree, a subset of attributes is randomly selected. Then, the best test in the node is calculated with a particular evaluation measure over the subset of attributes. This combination of bootstrap sampling of examples with random selection of attributes should increase diversity of component tree classifiers. So, it is somehow similar to motivations of introducing the decorate but it should be less time consuming.

3.2 Disagreement Measures

Another key issue in the active learning framework concerns selection of the most informative unlabeled examples to be queries to the oracle. In first papers, like (Cohn at al., 1994), committees consisted of two classifiers only, so examples which were classified differently by them were assumed to be uncertain. This idea was also used in the first attempts of using larger committees; e.g. in (Liere and Tadepalli, 1997) just two component classifiers were randomly selected and treated in a similar way. Then, in query by bagging or boosting Abe and Mamitsuka (1998) proposed to use a *disagreement measure* of predictions of all component classifiers. They chose a *margin* of classified examples, which was defined as a difference between the number of votes in the committee for the the most often predicted class label and the number of votes for the second predicted label. Examples with the smallest value of the margin are treated as the most uncertain for the committee and therefore the most informative for active learning. Let us remark that similar definition of the margin was also considered by Breiman (2001) in the context of diversification of trees in random forests.

In our framework we decided to choose the generalized version of margins, which takes probability distributions of class predictions instead of votes (following inspirations from (Melvile and Monney, 2004)). Let us notice that, in many implementations of ensembles, the base classifier can produce class memberships for a given example (not only the single class label like in the standard way of aggregating predictions) [1].

Let $P_{C_i,y}(x)$ denotes the probability of assigning example x to class y by a base classifier C_i. Then, the probability of assigning x to class y by the complete committee C^* is defined as:

$$P_y(x) = \frac{\sum_{C_i \in C^*} P_{C_i,y}(x)}{size(C^*)} \tag{1}$$

The *generalized margin* for x is defined as the difference between the highest and the second highest predicted probabilities.

Having distributions of class probabilities for base classifiers and the final ensembles it is also possible to built other measures of disagreement for decisions of classifiers inside the committee. For instance, we can define a distance between ensemble distributions and the base classifiers. In our experiments we decided to

[1] Such a distribution of membership probabilities can be extracted from the source code of WEKA, where we implement our framework for QBC active learning

use a simple *Euclidean* distance

$$Euclidean = \sum_{i=1}^{C_{size}} \sqrt{(P_i(x) - C(x)^*)^2} \tag{2}$$

where $P_i(x)$ denotes the class probability distribution given by the i-th classifier for the example x and $C(x)^*$ corresponds to the ensemble distribution.

In the similar way defined another *median* measure, where we compare a median value from the class probability distribution of component base classifiers with a median for the ensemble. The last measure comparing probability distribution is *Jensen-Shannon divergence*, defined as

$$JS(P_1, P_2, \ldots, P_n) = H(\sum_{i=1}^{n} w_i P_i) - \sum_{i=1}^{n} w_i H(P_i) \tag{3}$$

where P_i abbreviates $P_i(x)$, w is a vote weight of classifier C_i in the ensemble of size n, and $H(P)$ is the Shannon entropy of distribution for K classes, i.e. $H(P) = -\sum_{i=1}^{K} p_j \log p_j$.

3.3 Constructing the Starting Set

Another new methodological contribution concerns the choice of the starting training set used to generate the first ensemble before starting the active learning loop. In general its source may be different. It could be naturally present in the domain problem or be somehow selected by the expert. On the other hand, in typical experiments described in literature, the expert is not available and researches are using benchmark sets of examples and simulate a decision of oracle by knowledge about class label of these examples. Usually the starting training set is constructed by a random selection of examples from the pool of labelled training data – in many studies its size is very limited; even it may contain a single example / or single examples per class.

However, algorithms like bagging or boosting may require a good sample of examples to create an ensemble, which could make reasonable decisions. This is way we decided to study in our experiments another techniques of creating this starting training set. As in the experiment the set of labelled examples is available, we propose to select the first set by a kind of a focused approach, which prefers choosing among more certain examples. We adopt an edited k-nearest neighbor method, which was introduced by (Stefanowski and Wilk, 2007) to improve classification of imbalanced minority classes. The current modification chooses, so called *safe certain* examples. A learning example is safe if it is correctly re-classied by its all *k nearest neighbors* (i.e. all its k neighbors have the same class label as it). We first scan all available examples using this principle, and then randomly select the required number of safe examples to the starting training set.

Of course this method is reasonable in the experiments where expert's decisions are simulated on the set of labeled examples. For active learning with real unlabeled examples it is not directly applicable. However, first we want to use it in controlled experimental framework to study the influence of the construction

of the starting set as it was not studied in the literature. As the second point, we think that the counterpart of the k-nn method for unlabeled examples may be the use of an appropriate cluster analysis algorithm and to select among examples representing rather the inside parts of the cluster than borders or being outliers. A possible solution could be an adaptation of the density based algorithm DBSCANN that uses also an idea of local neighborhood to detect dense region and identify core points, borderline or outliers (Ester at al., 1996). Due to the time limit we were unable to make an appropriate adaptation of such an algorithm for the current paper.

4 Experimental Evaluation

The first aim of our experiments was to compare the performance of four following different approaches to Query by Committee Based Active Learning: *query by bagging*, *query by boosting*, *active decorate*, *query by random forests*. In all approaches, component classifiers were decision trees learned by J4.8 algorithm (with standard parameters). All implementations were based on the Weka toolkit[2]. The maximal size of committees for all approaches was set to 15, and for *active decorate* the number of internal iterations was equal to 50 (this choice was based on studying previous research by Melvile and Monney (2004)). The random selection of attributes in *random forests* was done as proposed in Breiman (2001). During experiments we also increased the number of trees in *random forests* to 50.

Table 1: Characteristic of data sets.

Data set	# objects	# attributes	# classes
Breast Cancer Wiscon.	683	9	2
Credit German	1000	20	2
Diabetes	768	8	2
Ionosphere	351	34	2
Soybean	683	35	15
Wine	178	13	13

We evaluated their performance on 6 data sets listed in Table 1. They come from the UCI repository [3]. Some of them are known to be hard to learn by standard algorithms. We chose them as they were previously often used in previous papers on using query by committees.

The classification accuracy was the main evaluation criterion and it was estimated by 10 fold stratified cross validation repeated 5 times. For each data set we generated a learning curve expressing the accuracy as a function of the number of available training examples. More precisely, in each experiment (fold of cross-validation) the training part was treated as a pool of unlabeled examples and in

[2]see www.cs.waikato.ac.nz/ml/weka

[3]UCI Machine Learning Repository. University of California at Irvine; see www.ics.uci.edu/ mlearn/MLRepositoru.html

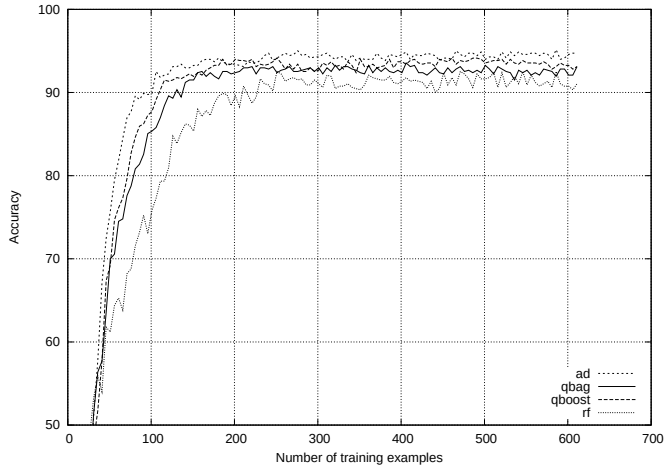

FIGURE 1: Comparing different active learners on *Soybean* – Random forests with 15 component trees.

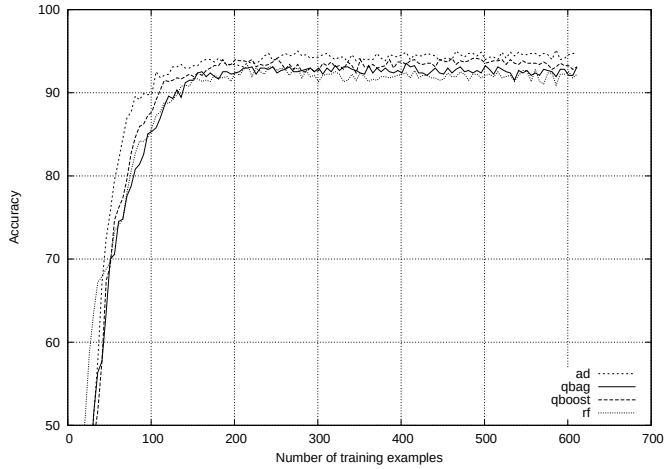

FIGURE 2: Comparing different active learners on *Soybean* – Random forests with 50 component trees.

each iteration of active learning the given approach processed them and selected the most informative example to add to the current training set. We did not stop active learning until all examples were added, so curves illustrate performance of all algorithms finishing with the same number of examples. Let us remark that we present graphically results averaged over several runs of cross validation and curves are not additionally smoothed, so one can observe slight fluctuations of plots. In all figures, plots for compared approaches are described using the fol-

lowing abbreviations: **qbag** – *query by bagging*, **qboost** –*query by boosting*, **ad** – *active decorate*, **rf** –*query by random forests*.

All compared approaches begin with a starting training set of one labeled example (randomly chosen), then a *single unlabled example* was selected in each iteration of active learning. In these experiments the *margin* of the example was used as the disagreement measure. Firstly we noticed that *random forests* nearly always needed more labelled examples than other approaches and its figure was quite often the lowest , i.e. dominated by others, see the position **rf** at Figure 1. We hypothesized that keeping the same number of component classifiers as bagging or boosting may be too restrictive for this kind of ensemble which due to randomization of trees may need more components, (cf. Breiman, 2001). Following it, we stepwise increased this number, observing that for 50 trees the classification results of *random forests* became comparable to other approaches – compare Figures 1 and 2. Thus, the rest of experiments was run with this parameter.

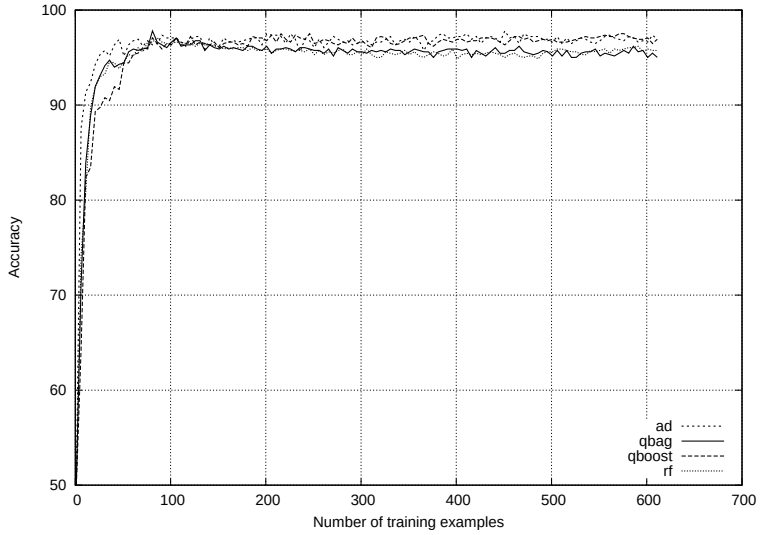

FIGURE 3: Comparing different active learners on *Breast* data.

Due to the page limits we could present only the most representative learning curves – see e.g. Fig. 2 or 3. One can notice that the highest classification accuracy is achieved by using quite a small number of additional training examples. For other data sets we obtained shapes of plots being similar to one of them although they "stabilized" at different levels of accuracy, e.g. plot for *Wine* is quite similar to *Soybean*. Nearly for all data set, active decorate approach led to the fastest increase of the accuracy, i.e. its plot was the most bent to the left upper corner of the figure. Other approaches generally worked comparably - depending on the data set one of them was slightly more effective. For data sets like *Soybean, Wine, Ionosphere* the differences between all compared approaches was the most visible – see Fig. 2, while for other data sets their plots were closer each other.

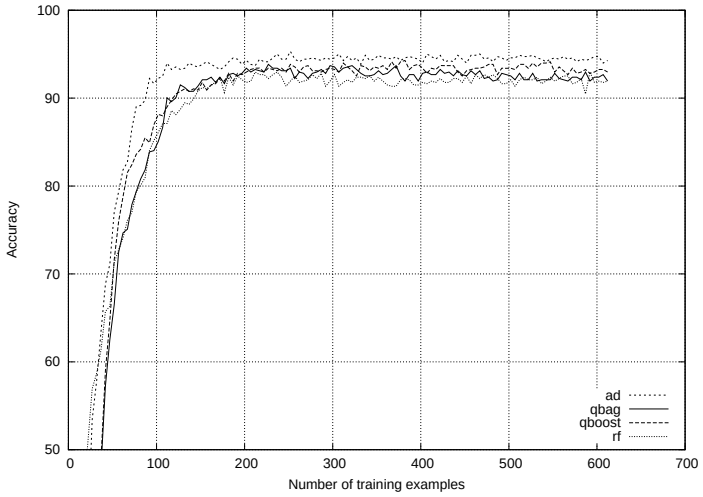

FIGURE 4: The effect of creating the starting training set in active learners with edited k-nn on *Soybean*.

In the next experiments we studied the use of new technique for selecting the starting training set. Referring to literature on active learning and co-training we decided to select 2% of the data to the starting set for all approaches. In Fig. 4 we show results for *Soybean* data. One can notice that using this k-nn techniques all compared approaches needed less additional examples to be labeled to achieve the level of the high accuracy comparing to starting from the random selected examples. Such an observation also held for other data sets. To study more precisely the change of the number of examples selected by active learning to be labeled we performed an additional analysis. As a "target accuracy" we assumed the classification accuracy which was obtained by a learning approach in the stable part of the plot (as we checked, it was comparable to the accuracy obtained by the passive version of the multiple classifier for the high number of training examples). Then, analysing consecutive iterations of active learning we determined the minimum number of learning examples required by the approach to achieve this target accuracy. To extend comparison we also evaluated the passive version of approaches (e.g. bagging used without active learning phase) – which were evaluated in simple on line processing of randomly ordered examples until they achieved an accuracy similar to the target one (i.e. with the threshold 1%). The numbers of necessary examples are summarized in Table 2. In parentheses, we put information on the ratio of using the total data sets. All approaches are described as: first line denotes a passive version of the committee, the second line corresponds to its active version with a random choice of the starting set and the third is its modification with k-nn choice. Bold fonts indicate the best results, i.e. the highest reduction of queries - number of examples to be labeled. For random forests we presented results for 50 trees (denoted as RF with 50) and additionally some results for using 15 trees (the last two lines in the table).

TABLE 2: Reduction of the number of training examples to achieve the target accuracy - for Random forests results are presented for 50 trees and for 15.

Approach	wine	ionosphere	breast	soybean	diabetes	credit-g
Decorate	37 (0.23)	56 (0.18)	35 (0.06)	347 (0.57)	172 (0.25)	174 (0.19)
AD	27 (0.17)	36 (0.11)	37 (0.06)	125 (0.20)	410 (0.59)	267 (0.30)
AD k-nn	**22 (0.14)**	**32 (0.10)**	**30 (0.05)**	**105 (0.17)**	**48 (0.07)**	**161 (0.18)**
Bagging	122 (0.76)	206 (0.65)	388 (0.63)	373 (0.61)	126 (0.18)	297 (0.33)
QBag	52 (0.33)	**46 (0.15)**	53 (0.09)	144 (0.23)	**98 (0.14)**	**215 (0.24)**
QBag k-nn	**40 (0.25)**	48 (0.15)	**53 (0.09)**	**126 (0.21)**	118 (0.17)	293 (0.33)
Boosting	111 (0.69)	162 (0.51)	60 (0.10)	269 (0.44)	201 (0.29)	251 (0.28)
QBoost	**75 (0.47)**	**102 (0.32)**	62 (0.10)	149 (0.24)	199 (0.29)	203 (0.23)
Qboost k-nn	103 (0.64)	130 (0.41)	**54 (0.09)**	**145 (0.24)**	**38 (0.05)**	**170 (0.19)**
RF (50)	158 (0.99)	167 (0.53)	105 (0.17)	176 (0.25)	418 (0.68)	479 (0.53)
ARF (50)	**56 (0.35)**	71 (0.23)	56 (0.09)	76 (0.11)	**151 (0.25)**	272 (0.30)
ARF k-nn	128 (0.80)	**67 (0.21)**	**48 (0.08)**	**76 (0.11)**	152 (0.25)	**205 (0.22)**
RF (15)	115 (0.72)	251 (0.80)	105 (0.17)	411 (0.67)	176 (0.25)	377 (0.42)
ARF (15)	**58 (0.36)**	118 (0.37)	65 (0.11)	212 (0.35)	139 (0.20)	**170 (0.19)**
ARF k-nn	63 (0.39)	**86 (0.27)**	**48 (0.08)**	**190 (0.31)**	**111 (0.16)**	250 (0.28)

Moreover, we studied the influence of changing the number of examples selected in each iteration of active learning. Let us remind that up to now we always pose a query about single, the most valuable example in each iteration. Generally speaking extending this number for more than a few examples did not influence the highest accuracy obtained by all active learning approaches, however for data sets with larger number of classes, e.g. *Soybean*, it caused slowing down the increase of the learning curve – which means that more learning examples were necessary to obtain "stabilization point" at the curve (around twice more). On the other hand it also resulted in decreasing time cost of learning.

In the final part of the experiments we compared the use of four different measures of disagreement *margin*, *JS-divergence*, *median* and *euclidean* to be used for selecting examples to be labeled. We evaluated them for all active learning approaches with k-nn selection of the starting set. Results for many data sets did not show significant differences between using these measures (see. e.g. Fig 6). More visible differences were noticed for all data sets with non-binary decision class. To illustrate it we show in Fig. 5 again results for *Soybean* data. One can notice that simpler measures as margin or median were superior to JS-divergence and euclidean distance of probability distributions.

The last issue concerns the computation costs of each QBC approach. In Table 3 we summarize time of achieving by each approach the "stabilization" point of classification accuracy at the learning curve.

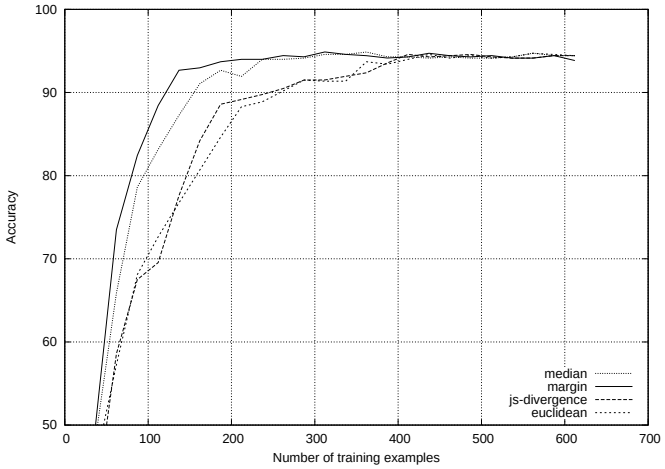

FIGURE 5: Comparing measures of disagreement in active decorate on *Soybean*.

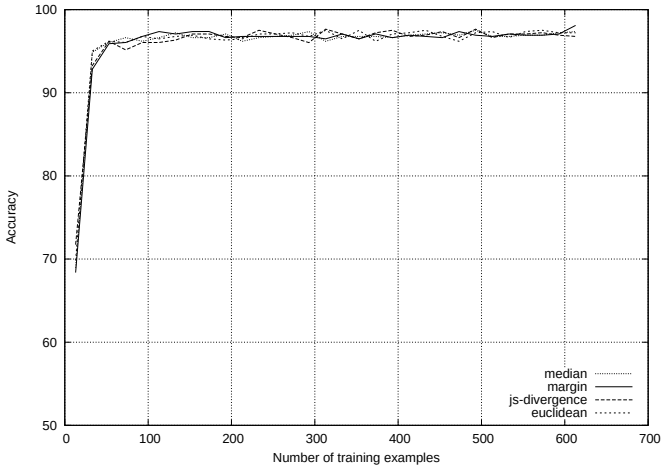

FIGURE 6: Comparing measures of disagreement in active decorate on *Breast*.

5 Conclusions

Results of our experiments with different approaches to construct active learning systems based on query by committee clearly confirmed that by using a relatively small number of examples – well selected for labeling – it is possible to generate a final classifier characterized by an accuracy comparable to passive approaches using much larger set of examples. This conclusion is consistent with earlier empirical studies on related QBC active learning (cf. Abe and Mamitsuka, 1998; Melvile and Monney, 2004). One can notice in Table 2 that the reduction of the number of

TABLE 3: Committee training time (in seconds) of different active learners.

Approach	wine	ionosphere	breast	soybean	diabetes	credit-g
Active Decorate	0.42	1.8	0.7	4.0	3.1	3.08
Query by Bagging	0.04	0.2	0.01	0.28	0.4	0.38
Query by Boosting	0.07	0.36	0.11	0.63	0.14	0.44
Random Forests	0.02	0.1	0.01	0.33	0.19	0.33

examples necessary for efficient active learning varies between 5% and 18% of the total size of training data depending on the data set and particular approach.

However, we observed that the choice of component techniques used in our framework influenced the results to a different extend. Among four compared methods applied to generate a committee the best reduction of the number of examples was mainly obtained by *active dectorate*. Looking at the second line of the results for each approach presented in Table 2 one can notice that *active dectorate* was the winner for the following data sets *wine, ionosphere, breast cancer* and *soybean*. However, it was the worst in case of *credit german* and *diabetes*, where other query approaches were better. We also noticed that the performance of the new considered approach *random forests* clearly depends on the number of component trees. Setting it to the similar number as in bagging or boosting made it the worst accurate approach. Increasing the number of trees to 50 resulted in quite comparable performance as other studied QBC approaches.

On the other hand, our experiments also showed that the good accuracy performance of *active dectorate* was obtained at the cost of the highest computation time. Although one could expect it knowing the nature of additional internal steps of generating artificial examples inside this approach, the range of the difference to *query by bagging* or *boosting* was quite large. *Query by random forests* was definitely the fastest approach – which is a new observation considering previous research on QBC. To sum up, depending on the problem at hand and preferences (query reduction vs time costs), one should carefully choose the most appropriate approach to generate the ensemble.

The other new observation from our experiments is that the way how the starting set of examples is provided to the active framework has a significant influence on the performance of all compared approach (see figures 1 - 4 or Table 2). The introduced edited k-nn method nearly always reduced the number of necessary examples to labeling. In particular, it helped the most to *active decorate* which achieved the best results for all data sets. We can hypothesize that such a choice of more certainly classified examples may influence the quality of the first constructed ensembles in their decisions and for *active decorate* it may give additional chance to generate better artificial examples.

Finally, the choice of the disagreement measure for predictions of committee was not significant for the majority of our data sets. Only for two multi-class data

sets *soybean* and *wine* simpler measures like *margins* or *median distance* were more useful than more complicated ones, as *JS-divergence* (see Figure 5). Looking into some literature discussions (Melvile and Monney, 2004) we found remarks that measures based on idea of *margins* are more directly oriented to identify decision boundaries - which may be more suitable for active learning than the reduction of uncertainty with predicted class probability distribution, which is somehow offered by the more complicated form of divergence.

In future research we want to follow the observation about the influence of constructing the starting training sets for QBC active learning and we are going to study the use of an unsupervised approach to select unlabeled examples to the starting set, e.g. by means of adaptation of density-based clustering algorithms.

References

Naoki Abe and Hiroshi Mamitsuka, Query learning strategies using boosting and bagging, in *Proceedings of the Fifteenth International Conference on Machine Learning* ICML-98, 1–10.

Avrim Blum and Tom Mitchell (1998), Combining labeled and unlabeled data with co-training, in *Proceedings of the Workshop on Computational Learning Theory*, Morgan Kaufmann Publishers, 92–100.

Leo Breiman (1996), Bagging predictors, *Machine Learning*, 24(2): 123–140.

Leo Breiman (2001), Random forests, *Machine Learning*, 45(1): 5–32.

David Cohn, Les Atlas, and Richard Ladner (1994), Improving generalization with active learning. *Machine Learning*, 15(2): 201–221.

Michael Davy (2005), A review of active learning and co-training in text classification. *Dep. of Computer Science, Trinity College Dublin, Research Report*, TCD-CS-2005-64, 39 pp.

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu (1996), A density-based algorithm for discovering clusters in large spatial databases, in *Proc. of KDD'96*.

Svetlana Kiritchenko and Stan Matwin (2001), Email classification with co-training, in Proceedings of the CASCON 01 Conference, 8–19.

David D. Lewis and Jason Catlett (1994), Heterogeneous Uncertainty Sampling for Supervised Learning, in *Proceedings of the 11th ICML Conf.*, 148–156.

Ray Liere and Prasad Tadepalli (1997), Active learning with committes for text categorization, in *Proc. of 14th AAAI Conf.*, 591–596.

Prem Melville and Raymon Mooney (2004), Diverse ensembles for active learning, in *Proceedings of the 21st Int. Conference on Machine Learning*, 584–591.

Prem Melville and Raymon Mooney (2003), Constructing diverse classifier ensembles using artificial training examples, in *Proceedings of IJCAI*, 505–510.

Jerzy Stefanowski, Szymon Wilk (2007), Improving rule based classifiers induced by MODLEM by selective pre-processing of imbalanced data, in *Proc. of the RSKD Workshop at ECML/PKDD*, 54-65.

Jiang Su, Stan Matwin, Jelber Shirabad, and Jin Huang (2008), Active learning with automatic soft labeling for induction of decision trees (manuscript).

H. S. Seung, Manfred Opper and Haim Sompolinsky (1992), Query by Committee, in *Proc. 5th COLT Conf.*, 287–294.