Class-Specific Ensembles for Active Learning in Digital Imagery

Amit Mandvikar and Huan Liu
Department of Computer Science & Engineering
Arizona State University, Tempe, AZ 85281
{amitm, hliu}@asu.edu

Abstract

In many real-world tasks of image classification, limited amounts of labeled data are available to train automatic classifiers. Consequently, extensive human expert involvement is required for instance labeling. Detecting *Egeria densa* in digital imagery is one such real-world classification task. It presents an additional challenge due to subtle spectral changes in Egeria, which makes it difficult to find a single accurate classifier. A novel solution is proposed to employ an ensemble of classifiers for *each class* (class-specific ensembles), combined with an active learning scheme. The class-specific ensembles are implicitly diverse. Diversity is required to increase the overall accuracy when combining predictions. The combined predictions of the ensembles can be used to reduce the *uncertainty* in detecting Egeria. Iterative active learning is then suggested to adapt the ensembles to the new images, unseen to the active learner. A novel solution to build compact ensembles is also presented, which are needed to expedite the re-training of the active learner. The combined results are accurate and compact ensembles, which require significantly less expert involvement for image region classification.

1 Introduction

Multimedia content is rapidly becoming a major target for data mining research. This paper is concerned with image mining - discovering patterns and knowledge from images for the purpose of classifying images or for similarity matching between images. The specific problem we address is image region classification. *Egeria densa* is an exotic submerged aquatic weed causing navigation and reservoir-pumping problems in the Sacramento-San Joaquin Delta of Northern California. As a part of a control program to manage Egeria, classification of regions in aerial images is required. This problem can be abstracted to one of classifying massive data without class labels. Relying on human experts for class labeling is not only time-consuming and costly, but also unreliable if the experts are overburdened with minute and routine tasks. Massive manual classifica-

tion becomes impractical when images are complex with many different objects (e.g., water, land, Egeria) under varying picture-taking conditions (e.g., deep water, sun glint). The main objective of the project is to relieve experts from going through all the images and pointing out locations where Egeria exists in the image. We aim to automate the process via active learning and only ask experts to label instances that the active learner is uncertain about.

The following desiderata for an image classification system present a unique challenge to data mining research for novel solutions.

1. *Reduced expert involvement.* Classification algorithms that require less expert involvement are essential in real-world applications, as human interaction forms the most serious bottleneck for efficient processing.

2. *Fewer labeled training images.* Labeled data are necessary to train automatic classifiers in a supervised fashion. The only source for such data is manual, tedious, and expensive labeling by experts. Consequently, it is sensible to ask for only a small number of labeled images for training. The reduced number of images, however, increases the difficulty for learning.

3. *Classification performance.* An image classification system can produce *certain* and *uncertain* classifications. Uncertain classifications require the intervention of human experts. Reducing the number of uncertain classifications translates directly to the reduction of expert involvement. In addition, classifications deemed certain should also be *correct*. Standard performance measures for detection problems such as accuracy, precision, recall, and F measure can be used in evaluation of correctness.

4. *Generalization.* To generalize, a classifier must perform well with unseen images. This is a central issue in pattern recognition and learning theory. A typical approach to avoid overfitting in training
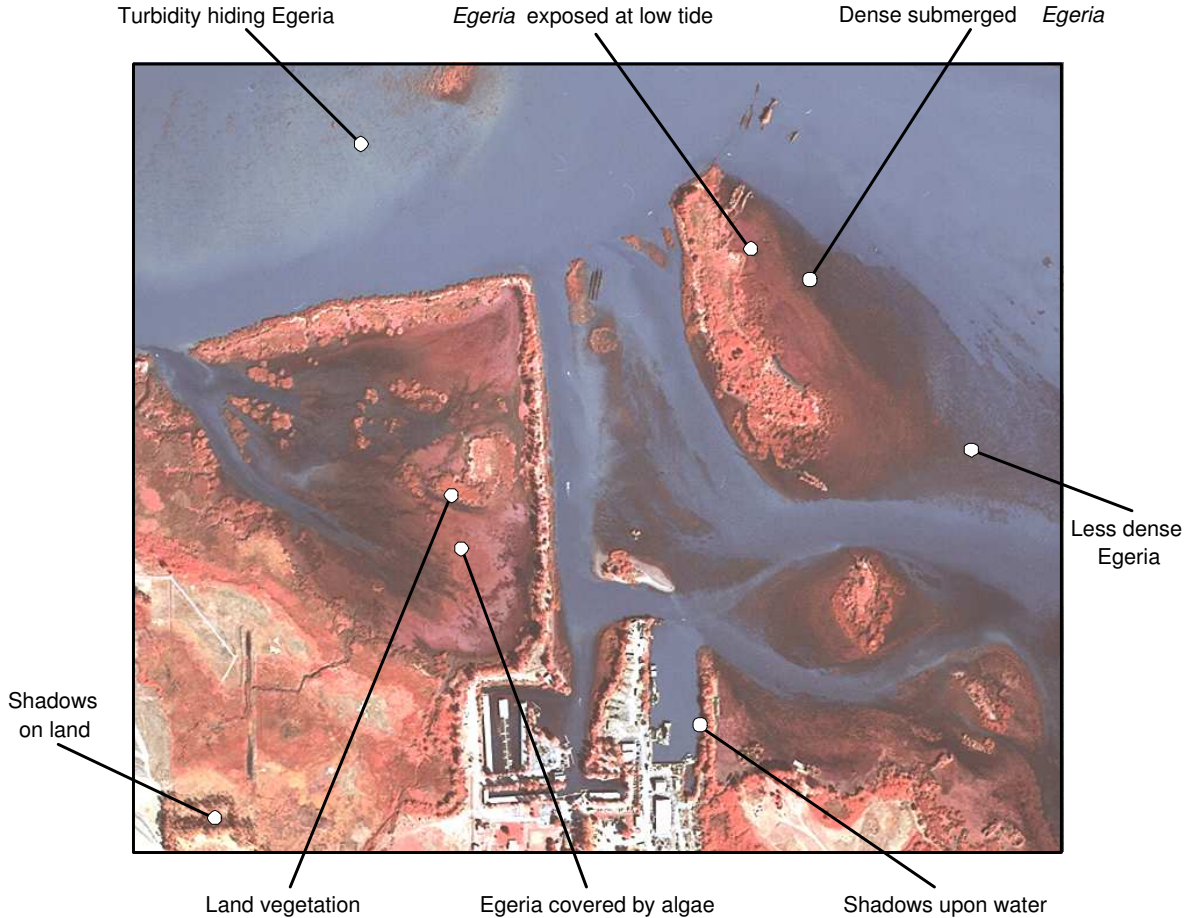
Figure 1: Scan-digitized CIR aerial photography showing spectral variations in *Egeria* and lack of spectral separation between *Egeria* and other extraneous classes.

is to *regularize* the structure of the classifier. In neural networks terminology, regularization causes the network to have smaller weights and forces the network response to be smooth, and hence the network is less likely to overfit [32]. Another [14, 7] is to combine the outputs of an ensemble of several, perhaps weak classifiers.

The contributions of this paper are a novel concept of *class-specific ensembles*, and learning algorithms for search of *compact ensembles* and for *iterative active learning*.

Classifying *Egeria* presents a challenging problem due to a number of variable and unfavorable conditions [13]. We observed that no single classifier can capture the subtle changes in the spectral distribution of Egeria and non-Egeria objects. We also observed that different types of classifiers are better suited to detecting different objects such as Egeria, land, and water. Individual ensembles members can learn different aspects

of the training data, so as to strengthen the accuracy of the overall ensemble [21]. Previous researchers [10] have also documented that such *diverse* ensembles are much more accurate (*stronger*) than the individual classifiers that form them. Figure 1 illustrates the spectral variations in *Egeria* that may occur even within a short distance. The figure also exhibits some problems caused by lack of spectral separation between *Egeria* and other extraneous classes. These reasons have encouraged the use of ensembles for the Egeria densa detection task. Also due to scarcity of training images, the training data typically represents only a portion of the testing images, thereby reducing classification performance for such images. We propose an active learning approach which is able to adapt the classifiers to unseen images from the domain. We introduce a novel concept of class-specific ensembles and explain why it should outperform the single ensemble approach. Our experiments show that this approach significantly reduces the number of un-

certain image regions and is better than using a single ensemble for the task of Egeria detection. An iterative active learning algorithm is presented to further reduce expert involvement in classification of new unseen images. With limited interaction with experts, our active learning scheme adapts the ensembles to new images. The iterative active learning scheme uses the expert-provided labels to re-train the ensembles for better performance. Such an scheme necessitates an compact ensemble, which can be *quickly* trained (at each iteration) interacting with experts. We present a method of combining individual classifiers to form a *compact* ensemble.

The remainder of the paper is organized as follows. Section 2 presents related work. Our approach for class-specific ensembles motivated and described in section 3. Sections 4 and 5 documents the algorithms related to active learning in detail. Section 6 provides empirical evaluation details. We conclude in section 7.

2 Related Work

Although regularization and structural risk minimization [18] can be effective in ensuring the generalization capability of a single classifier, recent research suggests that generalization can be *guaranteed* by using ensemble methods in a particular way [7]. The key point is to have individual classifiers that are *uncorrelated* or *negatively correlated* with each other [21, 8]. *Bagging* trains classifiers of the ensemble using different subsets of the training data [6]. *Boosting* gives different weights to different samples of the training data for each classifier of the ensemble [14]. *Random forests* use a randomly chosen subset of original features at each decision node of a classifier [7]. It is also possible, although less effective, to restrict each classifier to a particular feature set [11]. Researchers also focus on optimal combination strategies to obtain the final result from an ensemble. Common combination strategies are maximum posterior probability, majority, normalized product of the posteriors (maximum belief), consensus, median, and means of the outputs.

In active learning, Freund et al. [15] analyzed and suggested the selective sampling with the Query-by-Committee algorithm (QBC: introduced by [31]) that uses a committee of perceptrons to sample from a training data set to reduce prediction error. Abe and Mamitsuka [1] proposed two variants of the QBC algorithm, query-by-bagging and query-by-boosting. Both of them did better than QBC, C4.5, and boosting with C4.5.

Recent research [19, 16, 29, 30] concentrates on algorithms to process data automatically and algorithms that involve much less human expert involvement. A variant of an active learning algorithm has been suggested [19] that learns from specific unlabeled instances

via uncertainty sampling. The goal is to reduce the number of queries to human experts. Hakkani-Tur et al. [16] suggested a similar approach in the domain of automatic speech recognition (ASR). The difference between their approaches is in their sampling methods that select the most informative examples for learning.

In related research work, Muslea et al. [23] used selective sampling instead of uncertainty sampling to filter out the most informative unlabeled instances. The authors used two disjoint sets of features *(views)* to learn separate classifiers and then proceeded to label the most informative unlabeled instance for which the two classifiers disagree, add it to training data and relearn. They suggest that choosing the contention instances for which both classifiers are most confident provides maximal improvement. The authors continued their research [24] and experimented to prove that their algorithm, Co-Testing + Co-EM *(Co-EMT)* outperforms the algorithms EM [26], Co-Training [5] and Co-EM [25] on artificial as well as real-world domains.

Other researchers [29, 30] mentioned that most of the previous work on active learning focused on improving accuracy rather than reducing expert involvement. Instead they concentrated on using class probability estimates to get the class probability rankings, which enabled effective sampling from unlabeled instances. The authors proved that their sampling technique is better (in terms of training data set size) than uncertainty sampling or bootstrapping.

3 Class-Specific Ensembles

We first discuss the need for class-specific ensembles for a binary class prediction task, and then elaborate on how to learn these class-specific ensembles. Since we restrict our discussion only for a 2-class problem, for easier readability we will refer to the binary class-specific ensembles as *dual ensembles.*
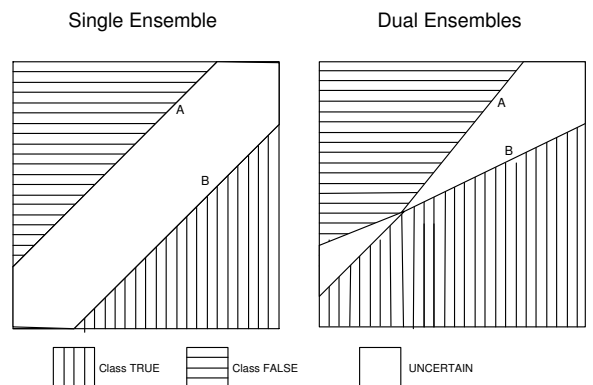

Figure 2: An illustrative example for two types of ensembles

The examples in Figure 2 illustrate the difference between a single ensemble and dual ensembles. The figure shows the 2-D map of the possible outputs of the two types of ensembles, a single ensemble (on the left) and a dual ensemble (on the right). An ensemble contains a fixed number of classifiers, each of which learns the separation between the classes, *True* and *False*. The lines A and B separate the regions of the output map into certain and uncertain predictions. A single ensemble can produce three outputs based on consensus (the degree of agreement among the member classifiers): *True, False, Uncertain*, as shown in the left of Figure 2. The region between lines A and B is uncertain as the ensemble cannot reach a definite consensus; the posterior probability for the instances in this region is neither close to 1 nor to 0. The *True* and *False* parts do not overlap because in such an ensemble learning, the focus is on one class *True* and the other class *False* is determined by default. In a more general setting, the class distributions for *True* and *False* are not exactly reversed, i.e. similar image regions could have different class labels, while dissimilar image regions could have the same class label (this can in seen in Figure 1). In such a domain the class distributions vary remarkably from image to image. It can be observed that most learners would not be highly certain about their decisions on *some* instances for such images. These instances, depicted by the regions between the lines A and B are *don't know*s or uncertain. We will refer to these uncertain regions or instances as *UC* in the rest of the paper. For a single ensemble such uncertain decisions are observed when there is no obvious consensus among all the classifiers within the ensemble. On the other hand, a dual ensemble, consists of two separate ensembles, one for each class. For class = *True*, an ensemble can predict if the class is *True* or *Not True*. The other ensemble predicts if the class is *False* or *Not False*. When the two ensembles do not agree with their classifications, the prediction is deemed as *Uncertain*. Such a scenario is depicted in the right of Figure 2.

We will now try to interpret why dual ensembles should work better than single ensembles. As mentioned earlier, diversity is an essential attribute for an ensemble to achieve high generalization [3]. For the best performance, the errors of the individual member of an ensemble should exhibit as low correlation as possible [33]. This *error diversity* could be incorporated *explicitly* or *implicitly* [8]. Explicit methods measure diversity in some manner and directly incorporate this knowledge into the construction or combination of the classifiers, as in Adaboost [14]. Implicit methods utilize purely stochastic perturbations to encourage diver-

sity, as in Bagging [6]. Considering a committee of two ensembles (the dual ensembles), we need to implicitly encourage the diversity between the two ensembles to obtain a highly accurate committee. For this we should force the dual ensembles to learn different aspects of the final distribution that we are trying to model. This can be achieved by specifically tuning each ensemble in the committee for one class, wherein each ensemble tries to predict its class with high accuracy.

We will show that using dual ensembles provides better classification and also fewer uncertain classifications, when each ensemble is specifically tuned for detecting one class. The subsequent problem is to identify the relevant classifiers tuned (via learning) towards specific classes to form dual ensembles.

4 In Search of Compact Class-Specific Ensembles

We may tend to use as many classifiers in an ensemble as possible because

- Each classification algorithm may have a different view of the training image and capture varied aspects of the image, as different classification algorithms have different biases and assumptions.

- No single classifier can completely cover the entire domain. In other words, some algorithms may succeed in capturing some latent information about the domain, while others may fail.

However, problems can result from using too many classification algorithms, as follows.

- Using more classification algorithms can result in longer overall training time, especially if we use iterative active learning where we have to re-train each learner at every iteration.

- Some classification algorithms may be prone to overfitting in the image domain. If these algorithms are included in the ensemble, there may be a high risk of allowing the ensemble to overfit the training image(s).

The above analysis suggests the necessity of searching for a relevant compact set of classifiers to form each class-specific ensemble. Exhaustive search for the best combination is impractical because the search space is exponential in the total number of classification algorithms for consideration. Thus we need an efficient methodology to find the optimally compact combination of classifiers for the dual ensembles. Since we already have diverse ensembles, we need to discuss other suitable performance measures for defining an optimally

compact ensemble. This should be followed by a proper learning algorithm that searches for compact ensembles while optimizing the measures.

4.1 Performance Measures

Precision, Recall, and Accuracy are the common criteria used for performance comparison. These measures are defined in terms of the instances that are relevant and the instances that are correctly classified (or retrieved). The true positives (TP) and true negatives (TN) are the correctly classified instances. A false positive (FP) is when the outcome is incorrectly predicted as YES when it is in fact NO. A false negative (FN) is when the outcome is incorrectly classified as NO when in fact it is YES. Precision, recall, and accuracy are defined in terms of TP, TN, FN, and FP.

- $Precision = TP/(TP + FP)$: the fraction of the classified information which is relevant.

- $Recall = TP/(TP + FN)$: the fraction of the classified relevant information versus all relevant information.

- $Accuracy = (TP+TN)/(TP+FP+TN+FN)$: the overall success rate of the classifier.

Accuracy takes into account the true negatives (TN) in its numerator. If a particular image has a large number of class "negative" that are classified correctly, then the resultant accuracy may be misleadingly high, overshadowing the others (TP, FN, FP). Particularly, in our application, we are mainly concerned with detecting Egeria (true positives). It has also been noted in [27] that accuracy may not provide a good measure for classification. Since both precision and recall have only TP in their numerator, they are suitable for performance measuring. In addition, we consider *reduction in uncertain regions* (UC) as a third measure.

High precision or high recall alone is not a good performance measure as each describes only one aspect of classification. Together they provide a good measure: for example, the F measure (F) [22, 28].

- $F = 2 * P * R/(P + R)$: the harmonic mean of precision and recall.

If both, precision and recall are 1 then F is 1, which means all and only positive instances are classified as positive. When either of precision or recall is 0 then F measure is 0. Hence, F measure is a good measure for both generality and accuracy.

4.2 Our Approach

Among many learning algorithms for classification, clustering, and association rules, we observe that association rule algorithms [2] can search the attribute space to find the best combination of attribute-values associated with a class. $A \Rightarrow B$ is an association rule that satisfies the minimum support and minimum confidence. The support for a rule is the joint probability $P(A, B)$ and the confidence is the conditional probability $P(B|A)$, where A and B are itemsets of attribute values (e.g., $a_1 = v_1, a_2 = v_2, b_1 = c_1, b_2 = c_2$). In our case, B is a class value ($b = c$), and A is a combination of attribute values. Thus the confidence of a rule gives us the measure of accuracy of the rule, while the support gives us the measure of generality of the rule. Association rules with high support and confidence are those both general and accurate. There are efficient algorithms to learn association rules from data [17, 2]. Since precision and recall are parallel to confidence and support, we employ association rule algorithms to search for compact dual ensembles. This approach is different from feature selection [4, 9, 20], where the attribute space is searched to find the best combination of attributes rather than attribute-values.

In order to search for compact ensembles, we need a data set that links the predictions of all classifiers to the label of each image region. This new data set can be obtained by applying all the classification algorithms to the training image so that each classifier is a feature (i.e., column) and its value is the prediction of the classifier. For each image region (one instance in the new data set), there are predictions of all the classifiers and also the class label 'Egeria' or not, given by experts. We are concerned only with those association rules that have the class label 'Egeria' or 'non-Egeria' as the consequent. We will restrict our search to such rules and obtain rules with the maximum number of features (classifiers) in the precedent without a significant loss in support or confidence. The best rule for each class label indicates the best combination of classifiers for the ensemble. Thus the ensembles obtained from this procedure are optimal in terms of both support and confidence. Detailed algorithm for this procedure is discussed next.

4.3 Algorithm for selecting compact Class-Specific ensembles

The algorithm is presented in Figure 4 and illustrated in Figure 3. It takes as input the entire set of classification algorithms E and training data Tr with class labels l_{Tr}, and produces as output the compact ensembles for class label *yes* and class label *no*. The major steps are (i) creating a new data set D (steps 1-3) by training all the classifiers E with the entire training dataset, (ii) learning association rules from D for dual classes (steps 4 and 5), and (iii) finding the best association rules for each class (steps 6-10).
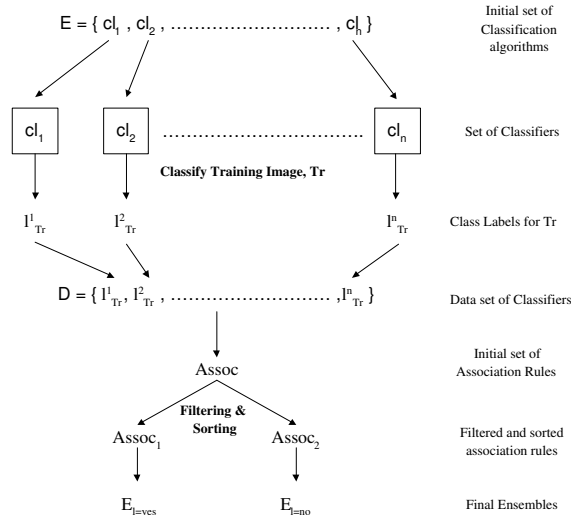
Figure 3: Illustrative example for Algorithm in Figure 4

input: Tr, E : Set of n classification algorithms
output: $E_{l=yes}, E_{l=no}$

01   Train E with Tr to obtain n classifiers, cl_1 to cl_n;
02   Obtain class labels, l_{Tr}^1 to l_{Tr}^n for Tr using cl_1 to cl_n;
03   Form a data set, $D \leftarrow \{l_{Tr}^1, l_{Tr}^2, ..., l_{Tr}^n, l_{Tr}\}$;
04   Learn association rules, $Assoc$ from D;
05    $Assoc_1 \leftarrow$ Filter $(Assoc$ / consequent is $l_{Tr} = yes)$;
06    $m_{l=yes} \leftarrow$ Max $(Assoc_1, supp * conf)$;
07    $Assoc_1 \leftarrow$ Filter $(Assoc_1/supp * conf \geq 0.9 * m_{l=yes})$;
08    $Assoc_1 \leftarrow$ Sort $(Assoc_1, length(precedent))$;
09    $E_{l=yes} \leftarrow$ Precedent(First($Assoc_1$));
10   Repeat steps 5 to 9 for $l_{Tr} = no$ to obtain $E_{l=no}$;

Figure 4: Algorithm for selecting compact Class-Specific Ensembles

Rules with support-confidence product $> 90\%$ of the maximum support-confidence for Tr are considered for selection. Each rule set is ranked according to *length* - the total number of classification algorithms in the precedent. This is because such rules have the maximum number of tightly bound classifiers in predicting the class label. The longest rule from each set is selected to obtain the ensemble for each class.

The next task is to use the dual ensembles ($E_{l=yes}$ and $E_{l=no}$) to determine certain and uncertain instances. We need to decide the maximum number of classifiers in an ensemble that should agree on a prediction to reach a decision of "certain" or not for each ensemble.

An ensemble with all classifiers being required to agree on a prediction would lead to high precision, but low recall; an ensemble with few classifiers being required to agree would lead to high recall and low precision. Thus, we need to find the maximum number of classifiers with which the ensemble gives the best estimated precision and recall, and hence the best F measure. The training image is used again for this task. $E_{l=yes}$ is certain only if all $n_{l=yes}$ classifiers agree on *YES*. The F measure (F_0) is recorded. If $(n_{l=yes} - 1)$ classifiers agree, then F_1 is checked. This process is repeated to find F_k for $(n_{l=yes} - k)$ classifiers by incrementing k until 1 classifier remains. The agreement threshold for $E_{l=yes}$ is then the maximum number of classifiers with highest F measure. The same procedure is repeated for ensemble $E_{l=no}$.

The dual ensembles $E_{l=yes}$ and $E_{l=no}$ work together to decide if an instance's prediction is certain or not following the rule of majority. In predicting an instance, if both $E_{l=yes}$ and $E_{l=no}$ are certain and agree with their predictions, the instance is considered certain and labeled with the prediction; if they are certain and disagree, the instance is considered uncertain; and if both are uncertain, the instance is uncertain.

5 Adapting via Iterative Active Learning

Clearly, the ensembles obtained from restricted training data have their limitations: when applied to some of the unseen images, they might result in a large number of uncertain instances. Instead of asking experts to resolve all these uncertain instances, we propose an iterative active learning approach that only requires experts to resolve a small number of instances. An active learner may begin with a small set of labeled training data, and predict class labels for the unlabeled testing data. The prediction can result in two sets of data: certain or uncertain. Some of the instances with uncertain predictions are presented to human experts to assign the correct class labels. The active learner is then *retrained* with the old and the newly labeled data to improve its prediction. In short, active learning is basically a supervised learning algorithm, requiring an expert to resolve its uncertain classifications. If we can have an effective active learner, we can significantly reduce the number of instances with uncertain predictions. The combination of the compact dual ensembles, is such an effective active learner as it not only reduces the uncertain classifications but is also compact so as to facilitate the iterative re-training of the classifiers in the ensembles.

The algorithm is presented in Figure 5 which takes as input Tr, a new image Ts, the number of uncertain

Figure 5: Iterative Active Learning Algorithm

instances m for experts to resolve at each iteration, and the compact dual ensembles. The suggested iterative active learner only requires a small number ($m = 25$) of additional labeled instances at each iteration to adapt the original ensembles to an new testing image. These additional instances should be reasonably less so that the expert is not overwhelmed at every iteration. The algorithm returns the adapted dual ensembles for the new image, Ts. The essence of the algorithm is to use a small amount of the expert's input to iteratively adapt the dual ensembles to a new image so expert involvement can be further reduced while maintaining the F measure. The oracle in this case is the human expert. The improvement stops when Fgain and UCgain both are insignificant ($< 5\%$ and $< 10\%$ respectively) or if UC_{new} is smaller than m.

6 Experiments and Evaluations

We performed experiments with a set of real-world image data. Each image is 300×300 pixels in TIF format (RGB). The extracted features are color, texture, and edge features. There are 13 features in total (details can be found in [12]). The template for feature extraction is 8×8 pixels. With 50% overlap between neighboring

regions, there are a total of 74×74 or 5476 regions (instances) per image. We designed 4 experiments to evaluate the following:

1. How the class-specific ensembles fare against single ensembles;

2. Whether we need to *learn* the class-specific ensembles;

3. How the class-specific ensembles fare against classification rules determined by experts; and

4. Whether the class-specific ensembles are applicable to new unseen images.

With the principal goal of reducing the burden on experts, we use only one image for training and apply the learned results to other 16 unseen testing images of different areas for detection. Among the classification algorithms available in the machine-learning package WEKA [34], we select all that can be applied to the image domain to ensure a variety of classification algorithms. There are six categories:

(a) Decision Tree based algorithms such as C4.5, Decision Stump, Id3, Alternating Decision Tree;

(b) Rule/Discretization based algorithms like Decision List Learner (PART), One Rule, PRISM, Hyper Pipes, Voting Feature Intervals;

(c) Neural Networks based algorithms such as Voted Perceptrons, Kernel Density Estimators, Logistic;

(d) Support Vector Machine based algorithms like Sequential Minimal Optimization for SVMs;

(e) Probability Estimators such as Naive Bayesian Classifier, Naive Bayesian Classifier-simple; and

(f) Instance Based algorithms such as Instance Based1, Decision Table.

We apply the algorithm in Figure 4 with the complete set of classification algorithms as input. The compact class-specific ensembles found by the algorithm are given below. The two ensembles are composed of different combinations of classifiers.

$E_{l=yes}$: C4.5, Alternating Decision Trees, Decision Trees (PART), PRISM, Hyper Pipes, Kernel Density, Logistic, Decision Tables $\Rightarrow$ 'Class = **yes**'.

$E_{l=no}$: Id3, Alternating Decision Trees, Decision Trees (PART), PRISM, Kernel Density, Instance Based1, Decision Tables $\Rightarrow$ 'Class = **no**'.

#	Compact Dual Ensembles (A)		Compact Single Ensembles (B)		Random Dual Ensembles (C)		Domain Expert's Rules (D)	
	F*	UC*	F*	UC*	F*	UC*	F*	UC*
1	0.7557	0	0.7557	0.5	0.7561	23.6	0.8509	35
2	0.7851	0	0.7851	0	0.77011	523.2	0.8040	582
3	0.6609	7	0.6611	23	0.67101	305.3	0.7401	50
4	0.7101	8	0.7103	22.5	0.7053	161.5	0.7785	18
5	0.5920	9	0.5921	14	0.5989	290.6	0.7467	86
6	0.7711	20	0.7543	72	0.7428	230.2	0.7755	95
7	0.8169	5	0.8162	18.5	0.8091	224.3	0.7980	121
8	0.4540	159	0.4415	209.5	0.5139	349.9	0.7327	253
9	0.5069	29	0.5120	13.5	0.5121	252.3	0.4586	33
10	0.4950	44	0.4923	53.5	0.5425	152	0.4627	134
11	0.4403	66	0.4197	107	0.4122	241.1	0.5644	129
12	0.6806	8	0.6811	9.5	0.6677	85.5	0.6780	63
13	0.6002	16	0.6008	22	0.5962	121.9	0.5835	58
14	0.6736	24	0.6722	41.5	0.6954	396	0.7091	99
15	0.5850	14	0.5867	3	0.6044	268.4	0.6291	245
16	0.8024	12	0.8011	22	0.8039	85.4	0.8132	41
17	0.6957	7	0.6936	21.5	0.7254	340.3	0.7011	134
	Avg UC* Insts	25.18	Avg UC*	38.44	Avg UC*	238.32	Avg UC*	128
Comparative Results			Avg UC* Incr	52.7%	Avg UC* Incr	846.6%	Avg UC* Incr	408.4%
			Avg F Gain	-0.55%	Avg F Gain	1.26%	Avg F Gain	8.60%

Table 1: Experimental Results: Comparing Compact Dual Ensembles with Compact Single Ensembles, Random Dual Ensembles, and Domain Expert Rules

Let F and number of uncertain instances for the kth testing image from ensemble i be F_k^i and UC_k^i, and let the corresponding values from ensemble j be F_k^j and UC_k^j. We calculate F measure gain and uncertain instance increase averaged over n testing images as follows:

$$(6.1) \quad AverageUCIncr = \frac{\sum_{k=1}^{n} UC_k^j - \sum_{k=1}^{n} UC_k^i}{\sum_{k=1}^{n} UC_k^i}$$

$$(6.2) \quad AverageFGain = \frac{\sum_{k=1}^{n} \frac{F_k^j - F_k^i}{F_k^i}}{n}$$

Table 1 summarizes experimental results in four columns (A, B, C, D). Image #1 is the training image. The last two rows show the average Fgain and average UCincrease w.r.t. results in Column A.

Experiment 1. We compare the dual ensembles ($E_{l=yes}$ and $E_{l=no}$) with the single ensembles (either $E_{l=yes}$ or $E_{l=no}$). The results are shown in Column B. The average UCincrease is almost 53% and the average Fgain is -0.55%. It is evident that in general, dual ensembles are not only more accurate, but also separate certain and uncertain instances better than single ensembles, except for 2 cases (images #9 and #15).

Experiment 2. We compare the dual ensembles to 10 randomly selected dual ensembles. We wish to check if the dual ensembles could be found by chance. Each classifier is randomly chosen from one of the categories mentioned earlier and learns from the training image. Although the average Fgain is only increased by 1.26%, the UC increases significantly by 846.6% as shown in Column C of Table 1. We conclude that it is necessary to search for the compact dual ensembles, as random dual ensembles work poorly in reducing UC.

Experiment 3. We compare the dual ensembles to results obtained by the classification rules of domain experts. The experts' rules outperform the compact dual ensembles in terms of Fgain by 8.6%, but the number of uncertain instances (UC) increases by 408.4% (in Column D of Table 1). The high Fgain and high UC for the expert classification rules is due to the fact that an expert can only directly work on the former (designing highly general and accurate rules), but not on the latter (finding low UC rules). Our system is particularly designed to compensate in this shortcoming.

Experiment 4. We explore if the dual ensembles can be further improved via iterative active learning. This function would be very useful in adapting to new unseen images. We can observe in Table 1, some of the unseen testing images have a high number of uncertain instances. It is impractical to overwhelm expert to resolve such a high number of uncertain instance. The algorithm in Figure 5 iteratively selects a small number of certain and uncertain instances (from such images)

#	Before Iterative AL		After Iterative AL		Fgain	UCincr	# runs	# queries
	F Measure	UC	F Measure	UC				
8	0.4540	159	0.5762	47	26.90%	-70.44%	3	75
9	0.5069	29	0.5069	29	0.0%	0.0%	1	25
10	0.4950	44	0.5385	11	8.77%	-75.0%	2	50
11	0.4403	66	0.5547	18	25.96%	-72.73%	3	75
	Average UC Insts	74.5	Average UC Insts	26.25	15.41%	-64.77%	2.25	56.25

Table 2: Experiment 4 Results: Before and After Iterative Active Learning

and adds them into the original training data after experts resolve these selected uncertain instances.

The results are shown in Table 2. After a few more iterations of learning, three out of the four images with $UC > 25$ achieve Fgain (15.41%) and negative UCincrease (-64.77%). This experiment results suggest that it is practical to adapt the learned dual ensembles to new unseen images to achieve high performance in terms of Fgain and reduced uncertain instances.

7 Conclusion

We present a novel approach to active learning with class-specific ensembles of various classifiers. One ensemble is trained for *each class*. These class-specific ensembles are implicitly diverse as they focus on different aspects of the concept to be modeled. The search for the class-specific ensembles is transformed to finding classifiers specifically tuned to each class. This is achieved by discovering association rules between classifiers and each class label. The learned ensembles can also be adapted to new images via iterative active learning. Extensive experiments were conducted in the real-world domain of detecting 'Egeria' in aerial images. The experiments compared the performance of the dual ensembles with single ensembles, with randomly selected dual ensembles, and with classification rules determined by domain experts. The class-specific ensembles outperformed them in terms of uncertain region reduction by 52.7%, 846.6%, and 408.4% respectively. Thus, they can significantly reduce expert involvement in instance labeling. The experimental results show that both components of our solution (class specific ensembles and active learning) together can significantly reduce expert involvement without compromising performance.

Acknowledgments

References

[1] N. Abe and H. Mamitsuka. Query learning using boosting and bagging. In *Proceeding of the 15th International Conference on Machine Learning*, pages 1–10, 1998.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. Int. Conf. Very Large Data Bases*, pages 487–499, Santiago, Chile, September 1994.

[3] Y. Baram, R. El-Yaniv, and K. Luz. Online choice of active learning algorithms. In Tom Fawcett and Nina Mishra, editors, *20th International Conference on Machine Learning (ICML'03)*, pages 19–26, Washington DC, USA, August 2003.

[4] A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271, 1997.

[5] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Workshop on Computational Learning Theory*. Morgan Kaufmann, 1998.

[6] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[7] L. Breiman. Random forests. Technical report, Statistics Department, University of California Berkeley, 2001.

[8] G. Brown and J. Wyatt. The Use of the Ambiguity Decomposition in Neural Netork Ensemble Learning Methods. In Tom Fawcett and Nina Mishra, editors, *20th International Conference on Machine Learning (ICML'03)*, pages 67–74, Washington DC, USA, August 2003.

[9] M. Dash and H. Liu. Feature selection methods for classifications. *Intelligent Data Analysis: An International Journal*, 1(3), 1997.

[10] T. Dietterich. Ensemble methods in machine learning. *Lecture Notes in Computer Science*, 1857:1–15, 2000.

[11] R. Duin and D. Tax. Experiments with classifier combining rules. In *Multiple Classifier Systems, First International Workshop, MCS 2000*, volume 1857, pages 16–29. Springer, 2000.

[12] P. Foschi, N. Kolippakkam, H. Liu, and A. Mandvikar. Feature extraction for image mining. In *International Workshop on Multimedia Information Systems (MIS 2002)*, pages 103 – 109, October 2002.

[13] P. Foschi and H. Liu. Active learning for classifying

a spectrally variable subject. In *2nd International Workshop on Pattern Recognition for Remote Sensing (PRRS 2002), Niagara Falls, Canada*, pages 115–124, August 2002.

[14] Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the 13th International Conference on Machine Learning*, pages 148–156. Morgan Kaufmann, 1996.

[15] Y. Freund, H. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[16] D. Hakkani-Tur, G. Riccardi, and A. Gorin. Active learning for automatic speech recognition. In *International Conference on Acoustics Speech and Signal Processing 2002*, 2002.

[17] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *Proceedings of Special Interest Group on Management of Data, SIGMOD-2000*, pages 1–12, 2000.

[18] S. Haykins. *Neural Networks: A comprehensive foundation (IInd Edition)*. Prentice Hall, 1999.

[19] V. Iyengar, C. Apte, and T. Zhang. Active learning using adaptive resampling. In *Proceedings of 6th ACM International Conference on Knowledge Discovery and Data Mining*, pages 92–98, 2000.

[20] R. Kohavi and G.H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[21] Y. Liu and X. Yao. Ensemble learning via negative correlation. *Neural Networks*, 12(10):1399–1404, 1999.

[22] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel. Performance measures for information extraction. In *Proceedings of DARPA Broadcast News Workshop, 1999.*, 1999.

[23] I. Muslea, S. Minton, and C. Knoblock. Selective sampling with redundant views. In *Proceedings of the National Conference on Artificial Intelligence*, pages 621–626, 2000.

[24] I. Muslea, S. Minton, and C. Knoblock. Active + semi-supervised learning = robust multi-view learning. In *Proceedings of the 19th International Conference on Machine Learning*, pages 435–442, 2002.

[25] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of Conference on Information and Knowledge Management*, pages 86–93, 2000.

[26] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Proceedings of the International Conference on Machine Learning*, 39(2/3):103–134, 2000.

[27] F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the 15th International Conference on Machine Learning*, pages 445–453, 1998.

[28] C. Van Rijsbergen. *Information Retrieval, 2nd edition*. Butterworths, 1979.

[29] M. Saar-Tsechensky and F. Provost. Active learning for class probability estimation. In *Proceedings of International Joint Conference on AI*, pages 911–920, 2001.

[30] M. Saar-Tsechensky and F. Provost. Active sampling for class probability estimation. In *Proceedings of Machine Learning*, 2002.

[31] H. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the 5th ACM workshop on Computational Learning Theory (COLT-92)*, pages 287–294, 1992.

[32] A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill-posed Problems*. W.H. Winston ed., 1977.

[33] N. Ueda and R. Nakano. Generalization error of ensemble estimators. In *Proceedings of International Conference on Neural Networks*, pages 90–95, 1996.

[34] I. Witten and E. Frank. *Data Mining: Practical Machine Learning tools and techniques with java implementations*. Morgan Kauffmann, 2000.