Adaptive Heterogeneous Ensemble Learning Using the Context of Test Instances

Anuj Karpatne and Vipin Kumar
Department of Computer Science, University of Minnesota
Email: anuj@cs.umn.edu and kumar@cs.umn.edu

***Abstract*—We consider binary classification problems where each of the two classes shows a multi-modal distribution in the feature space, and the classification has to be performed over different test scenarios, where every test scenario only involves a subset of the positive and negative modes in the data. In such conditions, there may exist certain pairs of positive and negative modes, termed as pairs of confusing modes, which may not appear together in the same test scenario but can be highly overlapping in the feature space. Determining the class labels at such pairs of confusing modes is challenging as the labeling decisions depend not only on the feature values but also on the context of the test scenario. To overcome this challenge, we present the Adaptive Heterogeneous Ensemble Learning (AHEL) algorithm, which constructs an ensemble of classifiers in accordance with the multi-modality within the classes, and further assigns adaptive weights to classifiers based on their relevance in the context of a test scenario. We demonstrate the effectiveness of our approach in comparison with baseline approaches on a synthetic dataset and a real-world application involving global water monitoring.**

I. Introduction and Motivation

A number of binary classification problems commonly experience heterogeneity within the two classes, which is characterized by the presence of multiple modes of each of the two classes in the feature space. For example, in order to classify locations on the Earth as water or land (binary classes) using remote sensing data (explanatory features), there is a need to account for the variety of water categories (e.g. shallow water, water near swamps, etc.) and land categories (e.g. forests, shrublands, sandy soil, etc.) that exist at a global scale, resulting in a multi-modal distribution of both water and land classes. Figure 1 shows a schematic illustration of a classification problem involving multiple modes of the positive and negative classes. In such situations, different pairs of positive and negative modes can show varying degrees of overlap in the feature space. This is represented in Figure 1 as edges with varying thickness, where the thickness of an edge reflects the degree of overlap between the pair of modes. Learning a single classifier that discriminates between all varieties of positive and negative modes is then challenging, especially in the presence of highly overlapping pairs of modes. We denote this phenomena as class confusion and the pair of modes participating in a class confusion as confusing modes in the remainder of the paper.

We consider binary classification problems where the classification has to be performed over different test scenarios, and every test scenario involves only a subset of all the positive and negative modes in the data. As an illustrative example, in the context of classifying locations on the Earth as water or land, a test scenario would comprise of instances observed in the vicinity of the same water body and at the same time-step. In such a setting, different pairs of positive and negative modes may emerge or disappear in different test scenarios, and even though some modes may be participating in class confusion, the subset of modes appearing in a given test scenario can be considered to be locally separable among each other. This shows a promise in using information about the context of a test scenario for overcoming class confusion.

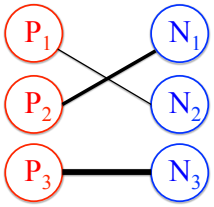


Figure 1: A schematic illustration of multi-modality within the classes, where each class comprises of three modes. Thickness of an edge shows the degree of overlap between the pair of modes.

To illustrate the importance of using the local context of a test scenario in the learning of a classifier, consider the toy dataset shown in Figure 2. This dataset comprises of instances belonging to two classes where each class comprises of two distinct modes, shown as colored circles in Figure 2. It can be observed that modes P_1 and N_1 are easily separable in the feature space, whereas modes P_2 and N_2 show class confusion. Assuming that we have access to a training dataset with adequate representation from every mode in the data, let us consider learning pair-wise classifiers, $C_{i,j}$, to distinguish between every pair of positive and negative modes, P_i and N_j. This would result in an ensemble of classifiers which can then be applied on any unlabeled instance in a test scenario to estimate its class label. Now let us consider a test scenario involving instances from P_1 and N_1, denoted by $S_{1,1}$. Since P_1 and N_1 are easily separable in the feature space and both P_1 and N_1 do not participate in any class confusion, test instances in $S_{1,1}$ would be correctly labeled even by a single classifier that discriminates between all positive and negative modes.

However, if we consider a test scenario $S_{1,2}$ involving instances from P_1 and N_2, we would notice that even though P_1 and N_2 are easily separable in the feature space, the presence of class confusion between P_2 and N_2 would hamper the classification performance at N_2, since instances belonging to N_2 can be easily misclassified to be belonging to P_2. To overcome this challenge, consider the following simplistic approach: let us assign a relevance score to every pair-wise classifier, $C_{i,j}$, in accordance with its likelihood of being used in the context of a test scenario. In particular, classifiers that

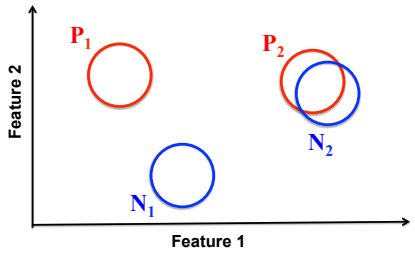


Figure 2: A toy dataset showing multi-modality within the classes, where P_2 and N_2 show class confusion.

	Test Scenario	
Classifier	$S_{1,1}$	$S_{1,2}$
$C_{1,1}$	"Relevant"	"Not Relevant"
$C_{1,2}$	"Not Relevant"	"Relevant"
$C_{2,1}$	"Not Relevant"	"Not Relevant"
$C_{2,2}$	"Not Relevant"	"Relevant"

Table I: Table summarizing whether a particular classifier, $C_{i,j}$ is relevant for a particular test scenario or not.

discriminate between modes having a higher likelihood of being observed given the distribution of instances in a test scenario would receive higher relevance scores. Using this approach, we can assign a relevance score to every pair-wise classifier for both test scenarios, $S_{1,1}$ and $S_{1,2}$, and consider it to be either "Relevant" or "Not Relevant", as summarized in Table I. For $S_{1,1}$, the only relevant classifier would then be $C_{1,1}$, which would correctly label all test instances in $S_{1,1}$. However, for $S_{1,2}$, both $C_{1,2}$ and $C_{2,2}$ would be considered as relevant, as the test instances in $S_{1,2}$ would show high likelihood for all the three modes, P_1, P_2, and N_2. However, $C_{2,2}$ would show poor cross-validation accuracy on the training set, since it discriminates between a pair of confusing modes, P_2 and N_2. $C_{2,2}$ could thus be discarded from the set of relevant classifiers, resulting in the only relevant classifier for $S_{1,2}$ to be $C_{1,2}$. $C_{1,2}$ would then be able to correctly label all test instances in $S_{1,2}$, and thus avoid class confusion in this particular situation. Note that the ability of the above simplistic scheme in overcoming class confusion arises from the fact that the distribution of test instances belonging to a test scenario contains reasonable information about its local context. We use this property as a guiding principle for motivating our proposed approach.

We propose the Adaptive Heterogeneous Ensemble Learning (AHEL) algorithm that takes into account the context of test instances belonging to a test scenario for overcoming class confusion in certain scenarios. We demonstrate the effectiveness of our approach in comparison with baseline approaches on a synthetic dataset and a real-world application involving global water monitoring. The remainder of the paper is organized as follows: Section II provides a brief overview of related work. Section III presents the proposed approach. Section IV presents experimental results. Section V includes concluding remarks and discusses directions for future work.

II. Related Work

The presence of multi-modality within the classes and its impact on classification performance has been previously discussed in [1], where the concept of modes was introduced as "small disjuncts". The impact of overlapping modes on the performance of a classifier has also been empirically analyzed in [2]. Furthermore, an ensemble learning approach for binary classification was recently presented in [3], that made use of the heterogeneity within the classes for constructing ensembles, instead of using random partitions of the input data. It was shown that such an ensemble learning method is able to capture the heterogeneity within the classes and thus result in improved classification performance. However, none of these approaches are suitable for handling the phenomena of class confusion by making use of the local context of a test scenario.

Existing approaches that make use of the context of test instances for adapting its labeling decisions involve local learning algorithms, e.g. the k-Nearest Neighbor (KNN) algorithm [4] and other concept-based local learning algorithms [5], [6]. These algorithms make use of training instances only in the local neighborhood of an individual test instance for estimating its class label. However, none of these approaches are designed to account for multi-modality within the classes and to incorporate information about a group of instances belonging to a test scenario as opposed to using the locality of an individual test instance. The use of unlabeled instances as a guide in the learning process has also been explored by semi-supervised learning [7] and transductive learning [8] approaches. The primary objective of such approaches is to address the paucity of labeled data by making use of the structure in the test instances, e.g. using clustering approaches [9]. This is different from our problem since our primary objective is to use the unlabeled instances for inferring the classification context of a test scenario involving confusing modes, even in the presence of sufficient training data. Another body of research that considers adapting the learning of a classifier in the context of a test scenario involves techniques for handling concept drift [10], [11], [12], and transfer learning approaches [13]. However none of these approaches have explored the presence of multi-modal distribution within each of the two classes, and are thus not directly relevant for our problem.

III. Proposed Approach

Notations: Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_1^n$ denote the training dataset with n labeled instances, where $\mathbf{x}_i \in \mathbb{R}^d$ is a d-dimensional feature vector and $y_i \in \{-1, +1\}$ is its binary response label. Let us assume that this training dataset comprises of n_+ positively labeled instances, denoted by $\mathcal{X}_+ = \{\mathbf{x}_i\}_1^{n_+}$, and n_- negatively labeled instances, denoted by $\mathcal{X}_- = \{\mathbf{x}_i\}_1^{n_-}$. Given this training dataset, our objective is to estimate the binary response, $y \in \{-1, 1\}$, for every test instance, $\mathbf{x}$, belonging to a test scenario, $\mathcal{X}_S = \{\mathbf{x}_i\}_1^s$.

We present the Adaptive Heterogeneous Ensemble Learning (AHEL) algorithm that comprises of the following steps:

A. Learning the Multi-modality in Training Data:

We assume that our training dataset, $\mathcal{D}$, contains a variety of instances from all possible positive and negative modes in the data, but explicit information about the multi-modal structure of the two classes is not known and needs to be inferred. To achieve this, we consider clustering the training instances belonging to each of the two classes separately, similar to the approach used in [3]. This results in the decomposition of the

positive class, $\mathcal{X}_+$, into m_+ clusters or modes and the negative class, $\mathcal{X}_-$, into m_- clusters or modes, respectively. The choice of the clustering algorithm and the number of clusters, m_+ and m_-, used for representing the multi-modality within the classes depends on the characteristics of the data. For every cluster label c, let $\mathcal{X}_c$ denote the set of training instances with cluster label c, where c can either be one of the positive cluster labels, P_1 to P_{m_+}, or the negative cluster labels, N_1 to N_{m_-}.

We further consider every cluster label c to have an associated conditional probability distribution, $\mathcal{P}(\mathbf{x}|c)$, for every instance $\mathbf{x} \in \mathbb{R}^d$. This can either be available as a by-product of the clustering algorithm or can be inferred from the distribution of instances in $\mathcal{X}_c$. As an example, we consider $\mathcal{P}(x|c)$ to follow a normal distribution in the feature space with the sample mean, $\bar{\mathbf{x}_c}$, as its center and with unit variance, whenever $\mathcal{P}(x|c)$ is not explicitly available during the clustering process. However, it should be noted that the choice of the probability distribution used for representing $\mathcal{P}(x|c)$ depends on the target application and can be acquired via domain knowledge.

B. Constructing an Ensemble of Classifiers:

We construct an ensemble of classifiers to discriminate between every pair of positive and negative cluster labels in $\mathcal{D}$, similar in essence to the Bipartite One-vs-One (BOVO) ensemble construction strategy proposed in [3]. This ensures adequate representation of every mode in the ensemble construction process, along with maintaining sufficient diversity among the classifiers. This can be contrasted with traditional ensemble learning approaches for binary classification, e.g. bagging, boosting, and random forests, which make use of random partitions of the training data as opposed to using a stratified sampling of the training instances in accordance with the multi-modal structure of the two classes.

For every pair of positive and negative cluster labels, (P_i, N_j), we learn a classifier, f_l, to discriminate between $\mathcal{X}_{P_i}$ and $\mathcal{X}_{N_j}$, using an appropriate choice of the base classifier. This results in the learning of an ensemble of classifiers, $\{f_1, \ldots, f_{m^\star}\}$, where $m^\star = m_+ \times m_-$. We further compute the cross-validation accuracy of every classifier, f_l, using 5-fold cross-validation on $\mathcal{X}_{P_i}$ and $\mathcal{X}_{N_j}$, and use it as a measure of the accuracy of f_l, denoted by $Acc(f_l)$.

C. Assigning Adaptive Weights to Classifiers:

For every classifier, f_l, we assign it a weight, $w(f_l, \mathcal{X}_S)$, representing its importance of being used for classification in the context of a test scenario, $\mathcal{X}_S$. In particular, we want to assign higher weights to classifiers that discriminate between pairs of modes that have a higher likelihood of being observed, given the distribution of instances in a test scenario, $\mathcal{X}_S$. Such a weighting scheme is achieved as follows.

For every test instance $\mathbf{x}$ belonging to $\mathcal{X}_S$, we compute its probability of being generated from a mode c as $\mathcal{P}(\mathbf{x}|c)$. We can then assign a relevance score to every mode c, denoted by $\mathcal{R}(c, \mathcal{X}_S)$, which indicates its likelihood of being observed given the distribution of instances in $\mathcal{X}_S$, defined as:

$$\mathcal{R}(c, \mathcal{X}_S) = \sum_{\mathbf{x} \in \mathcal{X}_S} \mathcal{P}(\mathbf{x}|c) \tag{1}$$

For a classifier, f_l, that discriminates between P_i and N_j, the relevance of using f_l in the context of $\mathcal{X}_S$, denoted by $\mathcal{R}(f_l, \mathcal{X}_S)$, depends on the relevance of observing modes P_i and N_j in $\mathcal{X}_S$, and can be estimated as:

$$\mathcal{R}(f_l, \mathcal{X}_S) = \mathcal{R}(P_i, \mathcal{X}_S) \times \mathcal{R}(N_j, \mathcal{X}_S) \tag{2}$$

$\mathcal{R}(f_l, \mathcal{X}_S)$ ensures that classifiers receive high weights only if both the modes involved in learning f_l have a high likelihood of being observed in $\mathcal{X}_S$. Each classifier f_l is further assigned a score, $\alpha(f_l)$, denoting its ability to differentiate between its pair of participating modes. $\alpha(f_l)$ can be computed as:

$$\alpha(f_l) = \begin{cases} Acc(f_l), & \text{if } Acc(f_l) > 0.6. \\ 0, & \text{otherwise.} \end{cases}$$

The weight of a classifier f_l in the context of test scenario $\mathcal{X}_S$ is then estimated as:

$$w(f_l, \mathcal{X}_S) = \alpha(f_l) \times \mathcal{R}(f_l, \mathcal{X}_S) \tag{3}$$

To illustrate the usefulness of $w(f_l, \mathcal{X}_S)$ in choosing the appropriate set of classifiers, especially in the presence of class confusion, consider a test scenario $\mathcal{X}_S$ that involves instances from P_c and N_{nc}, such that P_c shows class confusion with some other mode N_c not present in $\mathcal{X}_S$. In such a situation, P_c, N_c, and N_{nc} would receive the highest relevance scores in the context of $\mathcal{X}_S$. By taking the products of the relevance scores, the two classifiers that would receive the highest relevance scores would then be the ones that separate (P_c and N_c) and (P_c and N_{nc}). On the other hand, none of the pair-wise classifiers separating P_c, N_c, and N_{nc} from some other mode, O, will have a high relevance score, due to the low relevance score of O. The classifier separating (P_c and N_c) will eventually receive a low weight owing to its poor cross-validation accuracy and will be discarded. Thus, the classifier separating (P_c and N_{nc}) will be appropriately selected with the highest weight, resulting in adequate classification performance even in the presence of class confusion.

Note that our proposed weighting scheme inherently assumes that every test scenario involves a subset of positive and negative modes that are separable among each other but may show class confusion with other modes observed globally that are not present in the current test scenario. It is also assumed that a test scenario involving a confusing mode has instances from both the classes, thus requiring the use of a classifier in the first place. Furthermore, the ability of the above weighting scheme in avoiding class confusion hinges on the presence of atleast a single non-confusing mode in the test scenario, which can dominate the assignment of relevance scores to classifiers.

D. Combining Ensemble Responses:

We apply the ensemble of classifiers on a test instance, $\mathbf{x} \in \mathcal{X}_S$, to obtain a vector of ensemble responses, $f(\mathbf{x}) = [f_1(\mathbf{x}), ..., f_{m^\star}(\mathbf{x})]$. For each ensemble response, $f_l(\mathbf{x})$, we compute its loss w.r.t. a cluster label, c, as follows:

$$\text{Loss}(c, f_l) = \begin{cases} L(+f_l), & \text{if } c = P_i. \\ L(-f_l), & \text{if } c = N_j. \\ 0, & \text{otherwise.} \end{cases}$$

where, P_i and N_j are the positive and negative cluster labels used for learning f_l, and $L(z)$ is an appropriate loss function, e.g. the hinge loss function, $L(z) = \max\{1-z, 0\}$, commonly used with support vector machines (SVMs) as base classifiers. The combined loss of all ensemble responses w.r.t a cluster label c is then defined as:

$$\text{Loss}(c, f(\mathbf{x})) = \sum_{l=1}^{m^\star} w(f_l, \mathcal{X}_S)\text{Loss}(c, f_l) \qquad (4)$$

We choose $\hat{c}$ as the cluster label which provides the minimum loss, $\hat{c} = \arg\min_c \text{Loss}(c, f(\mathbf{x}))$. The test instance $\mathbf{x}$ is then classified as positive if $\hat{c}$ is a positive cluster label, otherwise it is classified as negative.

IV. Experimental Results

We compared the performance of AHEL with the baseline approach of learning a single non-linear classifier, termed as the GLOBAL approach. We also compared our results with the Bipartite One-vs-One (BOVO) ensemble learning approach that was presented in [3], which is able to handle heterogeneity within the classes but is unable to adapt its learning using the local context of a test scenario. In order to compare our performance with local learning algorithms, we considered the k-nearest neighbor (KNN) algorithm with $k = 5$ as a baseline approach. Furthermore, in order to emphasize the importance of using the distribution of an entire group of instances belonging to a test scenario as opposed to an individual test instance, we considered a variant of our algorithm that uses instance-specific information for assigning weights to ensemble classifiers, termed as the Instance-specific Heterogeneous Ensemble Learning (IHEL) algorithm. Specifically, IHEL considers the relevance of using a classifier f_l on a test instance $\mathbf{x}$ as $\mathcal{R}(f_l, \mathbf{x}) = \max(\mathcal{P}(\mathbf{x}|P_i), \mathcal{P}(\mathbf{x}|N_j))$, where f_l discriminates between P_i and N_j. IHEL thus follows the same formulation as AHEL, except for the fact that it uses $\mathcal{R}(f_l, \mathbf{x})$ in place of $\mathcal{R}(f_l, \mathcal{X}_S)$.

We used support vector machines (SVMs) with radial basis function (RBF) kernel as the base classifier for the GLOBAL approach and all ensemble learning methods used in this paper. The optimal hyper-parameters of SVM were chosen using 5-fold cross-validation on the training set in every experiment. The number of positive and negative clusters were kept equal in all experiments ($m_+ = m_- = m$). The classification error rate was used as the evaluation metric for comparing the performance of classification algorithms in every experiment.

A. Results on Synthetic Dataset:

We considered the synthetic dataset shown in Figure 3, which comprises of 10 positive and 10 negative modes, where every mode is generated using a bi-variate Gaussian distribution. Note that some pairs of modes in this dataset are easily separable (e.g. P_7 and N_7), while others show a high degree of class confusion (e.g. P_1 and N_1). These synthetic modes are representative of the variety of positive and negative modes that are experienced in real-world classification problems. We randomly sampled 200 instances each from every positive and negative mode for constructing the global training dataset. To simulate a variety of test scenarios, we randomly sampled 1000 instances each from every pair of positive and negative modes,

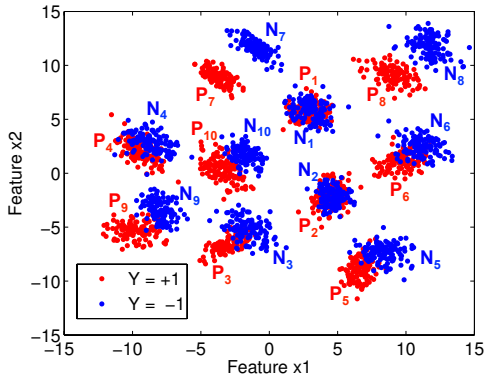


Figure 3: Synthetic dataset with 10 positive modes: P_1 to P_{10}, and 10 negative modes: N_1 to N_{10}, with varying degrees of class confusion among pairs of modes.

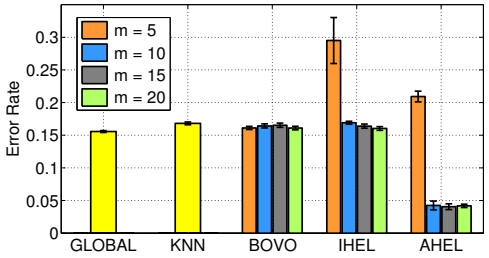


Figure 4: Comparing classification performance on synthetic dataset.

P_i and N_j, to construct 100 test scenarios, $S_{i,j}$. The random sampling procedure for obtaining the training and test sets was repeated 10 times.

Figure 4 compares the error rates of competing classification algorithms on the overall test set, comprising of instances from all possible 100 test scenarios. The bisecting K-means (BKM) algorithm [14] was used as the preferred clustering strategy for BOVO, IHEL, and AHEL, with varying number of clusters, m. It can be seen that both GLOBAL and BOVO have error rates close to 0.15, since they are unable to incorporate the local context of test scenarios for overcoming class confusion. Furthermore, techniques that use instance-specific context of individual test instances, namely KNN and IHEL, show no significant improvement than GLOBAL. In contrast, AHEL shows a significant reduction in the error rate for $m \geq 10$ when compared with all the baseline approaches, since it uses the overall distribution of instances belonging to a test scenario for adapting its learning.

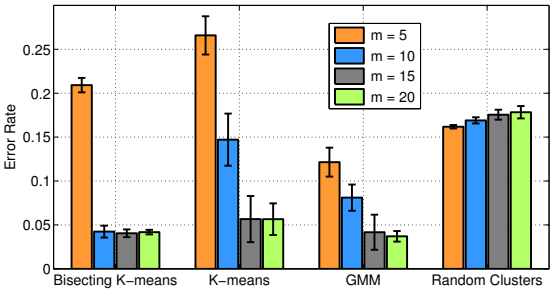


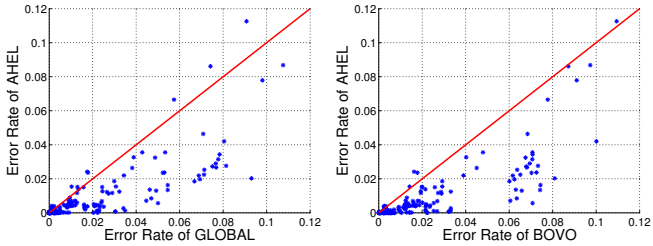
Figure 5: Varying the clustering strategy used in AHEL.

Figure 5 compares the performance of AHEL using varying clustering algorithms and number of clusters (m) used to represent the multi-modality within the classes. It can be seen that the performance of AHEL is initially poor for $m = 5$ because the clustering is unable to capture the heterogeneity within the classes, resulting in under-clustering, which degrades the performance of AHEL. However, as m is increased from 5 to 20, AHEL is able to adequately capture the heterogeneity within the classes and thus show drastic improvements in classification performance for all clustering algorithms. Note that the the performance of AHEL using Bisecting K-means is better than that of AHEL using K-means and Gaussian Mixture Model (GMM) clustering for $m \geq 10$, due to the tendency of K-means and GMM clustering to merge larger clusters and thus exhibit under-clustering. However, the performance of AHEL does not deteriorate even in the presence of over-clustering as m is increased from 10 to 20. Instead, the variance of the error rates of AHEL keeps decreasing as m is increased beyond 10, demonstrating the robustness of AHEL even with a large number of ensemble classifiers. Figure 5 also shows that the performance of AHEL is significantly better when a meaningful clustering strategy is used (e.g. BKM, K-means, and GMM), instead of using an artificial partitioning of the data into random clusters, demonstrating the utility of using information about the multi-modality within the two classes while learning classifier ensembles.

B. Global Water Monitoring Results:

We consider a real-world application of AHEL for monitoring water bodies at a global scale using remote sensing variables. Monitoring water bodies is important for effective water management and for understanding the impact of human actions and climate change on water bodies. To this end, remote sensing variables capture a variety of information about the Earth's surface that can be used for labeling every location on the Earth at a given time as water or land (binary classes). However, the presence of a rich variety of land and water categories that exist at a global scale makes it challenging to perform global water monitoring. There is an opportunity to overcome this challenge by using the local context of a test scenario, involving test instances observed in the vicinity of the same water body at the same time-step.

We used the seven reflectance bands collected by the MODerate-resolution Imaging Spectoradiometer (MODIS) instruments onboard NASA's satellites as the set of features for classification, which are available at 500m resolution for every 8 days. Ground truth information was obtained via the Shuttle Radar Topography Mission's (SRTM) Water Body Dataset (SWBD), which provides a mapping of all water bodies for a large fraction of the Earth (60^o S to 60^{o}N), but for a single date: Feb 18, 2000. We considered a diverse set of 99 lakes collected from different regions of the world for the purpose of evaluation. For each lake, we created a buffer region of 20 pixels at 500m resolution around the periphery of the water body, and used the buffer region as well as the interior of the water body to construct the evaluation dataset. After removing instances at the immediate boundaries of the water bodies and ignoring instances with missing values, this evaluation dataset comprised of $\approx$ 1.3 million data instances, where every instance had an associated binary label of water (positive) or land (negative). We randomly sampled 2000 instances each from both classes to construct the global training dataset. The remainder of the evaluation dataset was considered for testing. Since different pairs of water and land categories appear together in different regions of the world and at different times, we needed to consider test scenarios involving different pairs of water and land categories for the purpose of evaluation. To achieve this, we first clustered the water and land classes in the test set into $m = 15$ clusters each using the Bisecting K-means clustering algorithm. Every pair of water and land clusters, (W_i, L_j), was then considered as a different test scenario, $S_{i,j}$. We repeated the sampling procedure for obtaining the training and test sets 10 times.

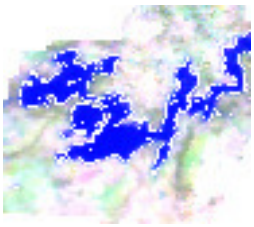


(a) Comparing AHEL and GLOBAL (b) Comparing AHEL and BOVO

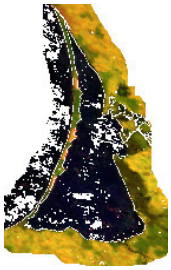
Figure 6: Scatter plots of mean error rates of Global, BOVO, and AHEL across all test scenarios.

Figure 6 presents scatter plots comparing the performance of AHEL with baseline approaches individually across all 225 test scenarios. Every point on a scatter plot compares the mean error rate of two classification algorithms on a particular test scenario, where the red line in each scatter plot shows the plot of $y = x$ for ease of comparison. It can be seen that AHEL shows drastic improvements in classification performance than GLOBAL and BOVO across a vast majority of test scenarios. In order to assess the statistical significance of the differences in the classification performance, we computed the p-value of AHEL showing lower mean error rate than GLOBAL and BOVO over all 225 test scenarios using one-tailed Wilcoxon signed rank tests, which came out to be equal to 1.74×10^{-25} and 2.02×10^{-35} respectively. This shows that the improvements in classification performance of AHEL are statistically significant.

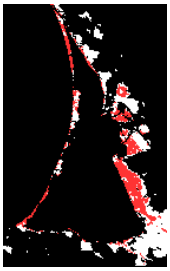
We next analyze the differences in the performance of AHEL and baseline approaches over two illustrative test scenarios, $S_{5,1}$ and $S_{10,1}$. Figure 7 compares the classification performance of GLOBAL and AHEL on the test scenario $S_{5,1}$ involving W_5 and L_1. Figure 7(a) shows the false color composite image (using the 7^{th}, 5^{th}, and 4^{th} bands, as red, green and blue colors respectively) of Gariep Dam in South Africa, which has all its water instances coming from W_5, shown in blue color. Figure 7(b) shows the false color composite image of Curonian Lagoon in Russia, which has a portion of its land from the land category L_1, indicated as red and white pixels in Figures 7(c) and 7(d). For these instances belonging to category L_1, Figures 7(c) and 7(d) show the misclassifications (errors) of GLOBAL and AHEL respectively as red pixels. It can be observed that GLOBAL is making errors over a large portion of L_1 as compared to AHEL. This is because L_1 comprises of land instances that appear very close to shallow water (see the false color in Figure 7(b)), resulting in its class confusion in the global training set. However, the false color of



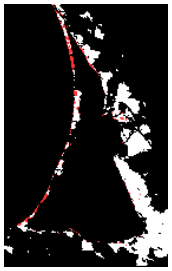
(a) False color composite image of water in Gariep Dam, South Africa



(b) False color composite image of land near Curonian Lagoon, Russia



(c) Errors of GLOBAL (shown in red) over L_1 (shown in white)



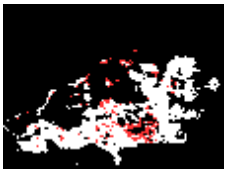
(d) Errors of AHEL (shown in red) over L_1 (shown in white)

Figure 7: Comparing GLOBAL and AHEL at $S_{5,1}$.

(a) False color composite image of water in Lake Tana, Ethiopia

(b) False color composite image of land near Burullus Lake, Egypt

(c) Errors of BOVO (shown in red) over L_1 (shown in white)



(d) Errors of AHEL (shown in red) over L_1 (shown in white)

Figure 8: Comparing BOVO and AHEL at $S_{10,1}$.

W_5 in Figure 7(a) can be seen to be very different from that of L_1 in Figure 7(b). Hence, in the local context of $S_{5,1}$, AHEL is able to handle the class confusion and thus show improved classification performance. The mean error rates of GLOBAL and AHEL for $S_{5,1}$ are 0.081 and 0.027 respectively. Figure 8 presents a similar analysis of the performance of BOVO and AHEL for the test scenario $S_{10,1}$. The mean error rates of BOVO and AHEL for $S_{10,1}$ are 0.07 and 0.019 respectively.

V. Conclusions and Future Work

We consider binary classification problems where both classes show a multi-modal distribution in the feature space and the classification has to be performed over different test scenarios, where every test scenario involves only a subset of all the positive and negative modes in the data. We propose the Adaptive Heterogeneous Ensemble Learning (AHEL) algorithm that constructs an ensemble of classifiers to discriminate between every pair of positive and negative modes, and uses the local context of test scenarios for adaptively weighting the ensemble of classifiers. We demonstrate the effectiveness of AHEL in comparison with baseline approaches on a synthetic dataset and a real-world application involving global water monitoring. Future extensions of our work could explore variants of our weighting scheme that can account for the imbalance among the classes, commonly experienced in real-world classification problems. Future work can also focus on studying the theoretical properties of AHEL, which can help in generalizing it to handle a broader family of class confusion scenarios in the presence of multi-modality within the classes.

VI. Acknowledgments

This research was supported in part by the NSF Awards 1029711 and 0905581, NASA Award NNX12AP37G, the UMII Felllowship, and the Doctoral Dissertation Fellowship. Access to computing facilities was provided by the Minnesota Supercomputing Institute.

References

[1] T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 40–49, 2004.

[2] R. C. Prati, G. E. Batista, and M. C. Monard, "Class imbalances versus class overlapping: an analysis of a learning system behavior," in *MICAI 2004: Advances in Artificial Intelligence*. Springer, 2004, pp. 312–321.

[3] A. Karpatne, A. Khandelwal, and V. Kumar, "Ensemble learning methods for binary classification with multi-modality within the classes," in *SDM*, 2015.

[4] T. Cover and P. Hart, "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions on*, vol. 13, no. 1, pp. 21–27, 1967.

[5] L. Bottou and V. Vapnik, "Local learning algorithms," *Neural computation*, vol. 4, no. 6, pp. 888–900, 1992.

[6] V. Vapnik and L. Bottou, "Local algorithms for pattern recognition and dependencies estimation," *Neural Computation*, vol. 5, no. 6, pp. 893–909, 1993.

[7] X. Zhu, "Semi-supervised learning literature survey," vol. 2, 2006, p. 3.

[8] T. Joachims *et al.*, "Transductive learning via spectral graph partitioning," in *ICML*, vol. 3, 2003, pp. 290–297.

[9] O. Chapelle, J. Weston, and B. Schölkopf, "Cluster kernels for semi-supervised learning," in *Advances in neural information processing systems*, 2002, pp. 585–592.

[10] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys (CSUR)*, vol. 46, no. 4, p. 44, 2014.

[11] J. Z. Kolter and M. Maloof, "Dynamic weighted majority: A new ensemble method for tracking concept drift," in *ICDM*, 2003, pp. 123–130.

[12] L. L. Minku, A. P. White, and X. Yao, "The impact of diversity on online ensemble learning in the presence of concept drift," *TKDE*, vol. 22, no. 5, pp. 730–742, 2010.

[13] S. J. Pan and Q. Yang, "A survey on transfer learning," *TKDE*, vol. 22, no. 10, pp. 1345–1359, 2010.

[14] M. Steinbach, G. Karypis, V. Kumar *et al.*, "A comparison of document clustering techniques," in *KDD workshop on text mining*, vol. 400. Boston, 2000, pp. 525–526.