Label Alteration to Improve Underwater Mine Classification

David P. Williams, *Member, IEEE*

***Abstract*—A new algorithm for performing supervised classification that intentionally alters the training labels supplied with the data set is presented. The proposed approach is motivated by the insight that the average prediction of a group of sufficiently informed people is often more accurate than the prediction of any one supposed expert. This idea that the "wisdom of crowds" can outperform a single expert is implemented in two ways. When labeling error rates can be estimated, sets of labels are drawn as samples from a Bernoulli distribution. When side information is not available, or no labeling errors are suspected, labels are intentionally altered in a structured manner. The framework is demonstrated in the context of an underwater mine classification application on synthetic aperture sonar data collected at sea, with promising results.**

***Index Terms*—Classification, ensemble methods, mine detection, underwater mines, wisdom of crowds.**

I. Introduction

A. General Motivation

In A classification problem, the label of a data point indicates the class to which it belongs. Typically, a human will be tasked to manually label the data in order to produce a training set. It is usually assumed that the labels assigned by the human are correct, with no errors. However, in many real applications, the process of compiling a training set of labeled data can be flawed with label errors. Therefore, if a different human was assigned to complete the labeling task, a different set of labels could result.

However, even without any reason to suspect imperfect labels in a particular application, there exists another motivation for working with alternative labels different from the original labels supplied with the training data set. Evidence from [1] suggests that the average prediction of a group of sufficiently informed people is often more accurate than the prediction of any one supposed expert. That is, the "wisdom of crowds" can outperform a single expert.

Taken together, these elements motivate the intentional manipulation of a set of labels when performing classification. More specifically, in this letter, we develop a label-alteration framework and demonstrate its utility on the real-world application of underwater mine classification.

Manuscript received August 6, 2010; revised August 31, 2010; accepted October 7, 2010. Date of publication November 28, 2010; date of current version April 22, 2011.

The author is with the NATO Undersea Research Centre, 19126 La Spezia (SP), Italy (e-mail: williams@nurc.nato.int).

Digital Object Identifier 10.1109/LGRS.2010.2088106

B. Motivating Application

The objective of underwater mine classification tasks is to classify objects as targets (i.e., mines) or clutter (e.g., rocks). The real-world relevance of the classification task is highlighted by the fact that even a single underwater mine can greatly disrupt commercial shipping traffic or hinder military missions.

The classification framework developed in this letter is tailored to address the following real-world scenario that commonly arises during military operations at sea. To establish a safe transit route that subsequent assets can utilize, an area must first be surveyed—typically with a sonar-equipped autonomous underwater vehicle (AUV)—and verified to be free of mines.

However, the time constraints involved in the problem make visually identifying objects—for active learning [2]—or delaying immediate classification decisions until a set of unlabeled data is available—for semisupervised learning [3]—unfeasible. Instead, a purely supervised approach must be employed.

To further complicate the task, the often-limited training data may also be potentially mislabeled. Typically, in sea (training) exercises, a set of known targets will be deployed in an area, and sonar imagery of the objects will be collected with the aid of an AUV. Then, a human will manually label the objects in the sonar imagery based on target-deployment knowledge.

However, other unknown objects may already be present at the site. If the unknown objects are not inspected optically to verify that they are all nonmines, such objects may be labeled incorrectly. Moreover, the confusion introduced by AUV navigation errors can also contribute to flaws in the labeling process. (Similar localization errors also arise in related remote-sensing tasks such as unexploded ordnance (UXO) detection [4].)

Thus, the classification task we address is when only limited labeled data—and possibly imperfectly so—is available for learning.

C. Contribution and Organization of This Letter

The main contribution of this letter is a novel supervised classification framework in which labels of data points are intentionally altered for use in the challenging regime where only limited training data are available for exploitation.

The counterintuitive classification approach is motivated by the constraints imposed by real data collected at sea, but the general framework can be applied to diverse domains in which labeling can be imperfect—such as medical diagnostics, character recognition, and document classification—as well as domains in which labels are uncorrupted.

An auxiliary contribution of this letter is the introduction of several new features for underwater mine classification that leverage environmental contextual information to provide a fuller and more informative representation of the objects.

The remainder of this letter is organized in the following manner. Section II describes the proposed classification framework motivated by the "wisdom of crowds." Details about the underwater mine classification task are presented in Section III, and experimental results on a data set of real synthetic aperture sonar (SAS) data collected at sea are shown in Section IV. A discussion of the proposed method is provided in Section V, with concluding remarks made in Section VI.

II. Classification With Altered Labels

A. Proposed Approach: "Wisdom of Crowds"

The proposed classification framework incorporating altered labels is motivated by the insight gleaned from [1] in which the average prediction of a group of sufficiently informed people is often more accurate than the prediction of any one supposed expert. That is, the "wisdom of crowds" can outperform a single expert.

A set of training data labels can be viewed as the opinion of a single (human) labeler. In turn, this set of labels—in conjunction with the training data's features—will manifest a classifier with which to make predictions on unlabeled testing data. Therefore, a direct link can be drawn between a given set of labels and a classifier, or even subsequent predictions. So one can view the (human) labeler as being ultimately responsible for the predictions that are made on unlabeled data.

In practice, one is presented with a set of labels for a data set. However, this set of labels reflects the opinion of only a *single* (human) labeler. From this initial set of labels, a collection of alternative label sets—each of which can be viewed as the opinion of a different (human) labeler—can also be generated in a specified manner.

Let $\boldsymbol{x}_i \in \mathbb{R}^d$ denote a vector of d features representing the ith data point of a training set of N such points. Let $y_i \in \{+1, -1\}$ denote the label, *originally supplied with the data set* that corresponds to the ith data point, $\boldsymbol{x}_i$. Collect this set of N training data labels as $Y = \{y_i\}_{i=1}^N$.

Suppose new sets of labels can be repeatedly generated in some prescribed manner. Denote the mth such set of N labels as $Y'_{(m)}$. Each set of labels can be viewed as the set of labels that a different human labeler assigns to the data points. Generate M such label sets so that we possess M (potentially, but not necessarily, unique) sets of N labels. Learn a classifier using the data $\{\mathbf{X}, Y'_{(m)}\}$, where $\mathbf{X} = \{\boldsymbol{x}_i\}_{i=1}^N$ and denote it $\boldsymbol{w}_{(m)}$. Repeat the classifier learning using each of the M data sets.

The proposed "wisdom of crowds" approach to classification with altered labels then averages the predictions from the M classifiers learned. More specifically, let $p(y_* = 1|\boldsymbol{x}_*, \boldsymbol{w}_{(m)})$ be the probability that a testing data point $\boldsymbol{x}_*$ belongs to class $+1$ based on classifier $\boldsymbol{w}_{(m)}$ (and in turn, the label set $Y'_{(m)}$). The final prediction is then the average of a "crowd" of M labelers

$$p\left(y_* = 1|\boldsymbol{x}_*, \{\boldsymbol{w}_{(m)}\}_{m=1}^M\right) = \frac{1}{M}\sum_{m=1}^{M} p\left(y_* = 1|\boldsymbol{x}_*, \boldsymbol{w}_{(m)}\right). \tag{1}$$

The proposed approach is general in the sense that it can be employed in conjunction with any probabilistic classification algorithm.

B. Label Alteration With Side Information

When sufficient knowledge about the problem domain and data set is possessed, this side information can be exploited to generate sets of altered labels in an intelligent manner.

The set of labels supplied with a data set reflects the opinion of only a *single* (human) labeler, so a principled process that reflects the variation possible from employing different (human) labelers is needed. The stochastic imperfect nature of the labeling process can be modeled using a Bernoulli distribution.

For the ith data point, $\boldsymbol{x}_i$, let $\epsilon_i \in [0, 0.5]$ denote the estimated labeling error associated with the label y_i *originally supplied with the data set*.

For each data point, $\boldsymbol{x}_i$, a new label (e.g., corresponding to a different human labeler) is then drawn independently from a Bernoulli distribution with parameter $1 - \epsilon_i$

$$y'_i \sim \mathcal{B}(1-\epsilon_i) = \begin{cases} y_i, & \text{with probability } 1-\epsilon_i; \\ -y_i, & \text{with probability } \epsilon_i. \end{cases} \tag{2}$$

That is, the original label y_i is flipped with probability ϵ_i.

Collect this new set of N training data labels, $\{y'_i\}_{i=1}^N$. Denote the mth such set of N labels as $Y'_{(m)}$.

To determine an appropriate number of sampling rounds, M, to consider, we exploit the work developed for multiple imputation [5], which replaces each missing *feature* value with a set of M samples. The efficiency of an estimate based on M imputations is approximately $(1 + \gamma/M)^{-1}$, where γ is the fraction of missing information for the quantity being estimated [5]. In our case, γ corresponds to the fraction of labels with a nonzero labeling error rate, ϵ_i. Even if $\gamma = 1$, high efficiency can be achieved with a relatively small number of label sampling rounds, M.

C. Label Alteration Without Side Information

In the absence of sufficient side information regarding suitable ways to stochastically alter the set of originally-supplied labels, a more restrictive algorithm can be employed to generate the label sets.

Define $\overline{y}_i \in \{+1, -1\}$ to denote the opposite label of $\boldsymbol{x}_i$, so $\overline{y}_i = -y_i$.

Define $\overline{Y}_i$ to be the set Y except with $\overline{y}_i$ in place of y_i: $\overline{Y}_i = \{y_1, \ldots, y_{i-1}, \overline{y}_i, y_{i+1}, \ldots, y_N\}$. Denote the mth such set of N labels as $Y'_{(m)}$.

In this approach, each data point's label is altered for one case, so the number of label sets is determined automatically: $M = N$.

This variant can also be viewed as the appropriate version for when it is expected that no labeling errors exist in the original set of labels (i.e., $\epsilon_i = 0 \quad \forall i$).

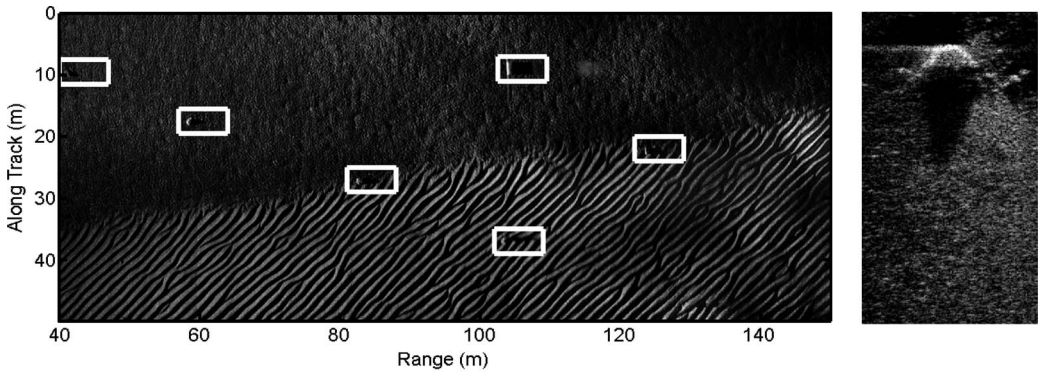

Fig. 1. (a) Typical SAS image with mines indicated in white boxes. (b) SAS image chip of a typical false alarm.

TABLE I
DATA SET DETAILS AFTER THE DETECTION PHASE

	RIGA		LIEPĀJA	
MISSION	CLUTTER	MINES	CLUTTER	MINES
1	488	36	332	36
2	337	40	180	38
3	298	37	370	39
4	226	36	10	24
5	169	21	213	36
6	105	25	493	37

III. UNDERWATER MINE CLASSIFICATION

A. Data Set

In April–May 2008, the NATO Undersea Research Centre (NURC) conducted the Colossus II sea trial in the Baltic Sea off the coast of Latvia. During this trial, high-resolution sonar data were collected by the MUSCLE AUV. This AUV is equipped with a 300-kHz sonar with a 60-kHz bandwidth that can achieve along-track and across-track image resolutions of approximately 3 and 2.5 cm, respectively. The sonar data were subsequently processed into SAS imagery. A typical SAS image, which hints at the uncertainty of the manual-labeling process, is shown in Fig. 1(a).

A set of targets was deployed at each of two sites, one in Rīga Bay and one off the coast of Liepāja. At each site, six separate AUV missions were conducted. The data from Rīga are comprised of a total of 1022 SAS images covering a total area of approximately 5.6 km^2. The data from Liepāja are comprised of a total of 578 SAS images covering a total area of approximately 3.2 km^2.

For this letter, a cascaded detection algorithm composed of two stages was applied to the SAS imagery to generate a set of alarms (i.e., mine-like objects). The algorithm searches for generic highlight-shadow patterns characteristic of mines; a typical (false) alarm detected by the algorithm is shown in Fig. 1(b). Space constraints prevent more details of the detection algorithm from being given here.

All of the alarms were then manually ground-truthed visually by inspection (while also exploiting the target-deployment location information). The characteristics of the two data sets at the end of the detection phase are summarized in Table I. These data are then passed on to the feature extraction stage.

B. Feature Extraction

For each alarm, $d = 11$ features are extracted. Five of the features are meant to establish the general shape and size of a given object, while the remaining six features are intended to capture the contextual information from the scene.

The motivation for the former set is that the features should be invariant to *specific* target types. In practice, there is a real possibility of encountering a target type that was not among the (controlled) set of deployed targets. For example, many old mines deployed during the World Wars are still in the ocean today. Therefore, in order to be able to correctly classify such targets, the features that are used to describe a given alarm should capture the inherent characteristics belonging to the entire class of targets. That is, the features should not be intimately tied to specific target types.

In this letter, two features are based on the correlation of the contact with highlight-shadow patterns characteristic of mines. Three other features are the object's illuminated surface area, area of shadow cast, and peak echo strength. Variants of such shape and size features are in common use [6].

The use of contextual information, however, has largely been overlooked in terms of features for underwater mine classification. In this letter, we exploit contextual information in the form of contact density and seabed type. More specifically, four features are based on the density of contacts within circles of different radii from the contact. (Among other things, these features are used to help establish the presence of boulder fields.) Additional features for a given contact are the proximity to the nearest contact and the likelihood of the contact location being within sand ripples [7]. These features capture operational intelligence from experience with the problem domain.

C. Labeling Error

The historical navigation accuracy of the AUV was used to define the labeling-error equation appropriately (i.e., to accurately reflect the challenges encountered with location uncertainty and contact association in real data collected at sea). Once equation parameters that resulted in reasonable curves were found, the parameters were fixed (i.e., no tuning was done to affect the experimental results).

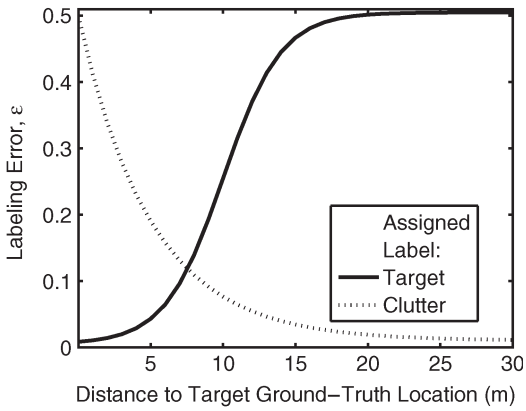

Fig. 2. Definition for the labeling error.

The labeling error for a given object as a function of distance (in meters), d, between it and the nearest recorded target ground-truth location was defined to be

$$\epsilon(d) = \begin{cases} \alpha\left(1+\exp\{-d\alpha+\gamma\}\right)^{-1}, & \text{if } y=+1 \\ \alpha-\beta\left(1-\exp\{-d/\gamma\}\right), & \text{if } y=-1 \end{cases} \quad (3)$$

where $\alpha = 0.5$, $\beta = 0.49$, $\gamma = 5$, and y are the labels manually assigned to the object. This equation is plotted in Fig. 2.

In the figure, the dotted curve (i.e., for objects manually labeled as clutter) is asymptotic to $\epsilon = 0.01$ to reflect the fact that there is always a chance that an object that was already present at the site is actually a target.

IV. Experimental Results

To evaluate the proposed framework, three different methods were considered, all of which use a simple logistic-regression model as the classifier. The differences among the methods pertain to the labels that are used to learn each classifier.

The first method uses the training data with the originally-supplied set of labels, which provides a baseline on performance to which the other methods can be compared. The other two methods are the variants of the proposed label-alteration framework described in Sections II-B and C.

For the variant that involves drawing sets of labels from a Bernoulli distribution using (3) (i.e., "with side information"), the final performance shown in the ensuing results corresponds to that achieved using $M = 100$ sampling rounds.

To assess classification performance of the methods, six-fold cross validation—using the natural data divisions by mission—is used at each of the two sites. For a given site, data from one of the missions are used as training data once, with the data from the remaining five missions treated as testing data.

Performance on the two data sets is shown in terms of the average receiver operating characteristic (ROC) curves in Fig. 3(a) and (b).

The area under an ROC curve (AUC), given by the Wilcoxon statistic [8], is a popular summary measure of performance for binary classification problems (higher AUC values indicate better performance). The classification results in terms of the AUC are presented in Table 2. Entries in bold indicate that the result is statistically significant and better than the baseline approach at a 95% confidence interval, according to a paired t-test and to the Wilcoxon signed-ranks test [9].

As can be seen from the figures and the table, the proposed approach achieved better performance than the case in which the labels were assumed to be perfect, which, to the best of our knowledge, they are. This improvement in performance can be directly attributed to the use of a "crowd" of classifiers and to the manner in which the training data labels was manipulated, since the same training data (features) and classification method (i.e., logistic regression) were used for all approaches.

For the proposed approach with side information, the evolution of the AUC as a function of the number of sampling rounds, M, used, is shown in Fig. 3(c). This figure shows that the performance stabilizes quickly, after approximately only $M = 20$ sampling rounds.

V. Discussion

The idea of averaging predictions from multiple classifiers, which falls under the umbrella term "ensemble methods" [10], is admittedly simple and not new. Instead, the novelty of the proposed algorithm lies in the way the collection of classifiers is generated: by altering labels in a prescribed manner and being able to do so even in the absence of side information.

In many ensemble methods—such as bagging [11] or mixture of experts [12]—multiple classifiers are learned as a result of dividing the training data into subsets or partitioning the feature space. When the initial set of training data available is already small, such an approach is inappropriate. In contrast, the proposed approach—like boosting algorithms [13]—always uses all of the available data to learn each classifier, a key distinction.

The proposed approach also deals with potentially imperfect labels. Most related work in this vein, such as [14] and references therein, instead addresses the different scenario in which *multiple* labelers provide conflicting labels for a data set, rather than the case in which a *single* set of potentially imperfect labels is provided. Other related work typically assumes that there is a single common labeling error for all data points of a given class (e.g., [15]). In contrast, the proposed approach that samples from a Bernoulli distribution permits more nuanced labeling knowledge to be exploited.

Active learning [2] and semisupervised learning [3] are well established methods for improving performance beyond a supervised approach, but they require additional information to do so. The proposed framework can instead boost performance when no unlabeled data is available and label acquisition is impossible. The generality of the framework, however, also means that the proposed method could be incorporated into approaches that employ semisupervised classifiers (instead of the supervised logistic regression used in this letter) or active learning when unlabeled data *are* available.

At the crux of the proposed framework is the generation of multiple sets of labels, and in turn, a collection of classifiers. Although it is likely that some of the individual classifiers degrade performance if examined alone, in aggregate, the wisdom of crowds prevails, leading to significant classification improvement.

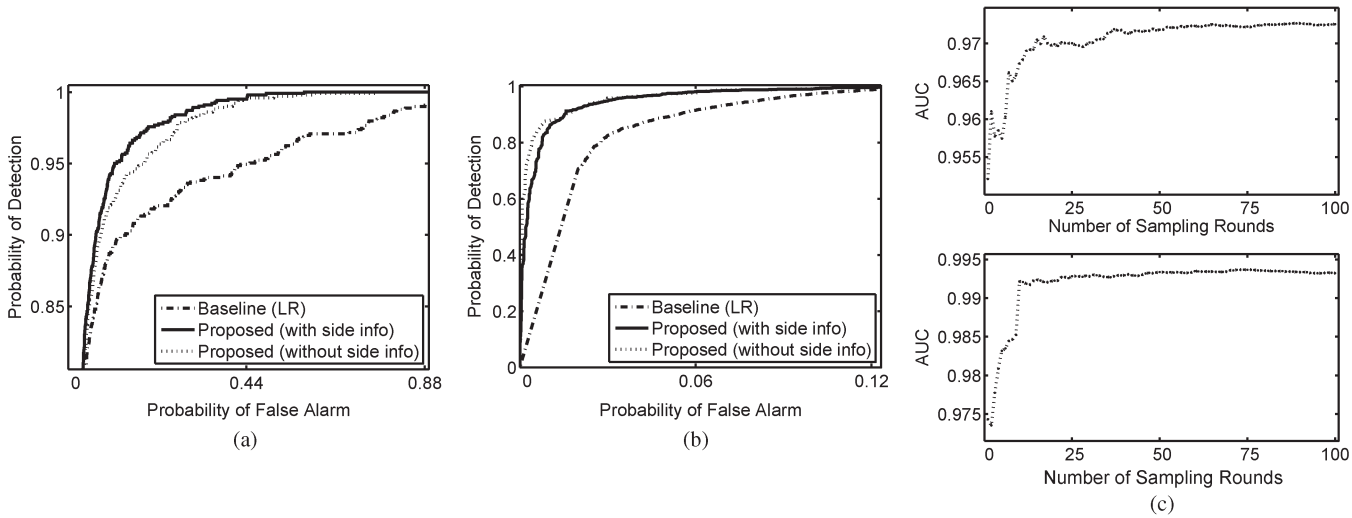

Fig. 3. (a)–(b) Average ROC curves and (c) evolution of the AUC for the proposed method (top: Rīga; bottom: Liepâja).

TABLE II
AUC (MEAN ± ONE STANDARD DEVIATION FROM THE SIX TRIALS) FOR EACH OF THE TWO TEST SITES

METHOD	RIGA	LIEPĀJA
BASELINE	0.9389 ± 0.0237	0.9775 ± 0.0165
WITH SIDE INFO	$\mathbf{0.9726 \pm 0.0052}$	$\mathbf{0.9932 \pm 0.0033}$
WITHOUT SIDE INFO	$\mathbf{0.9676 \pm 0.0082}$	$\mathbf{0.9938 \pm 0.0044}$

Side information—in terms of specifying the parameter of the Bernoulli distribution of the labeling error—allows a diversity of classifiers to be generated quickly, because multiple labels in the set can be altered simultaneously. Without side information, a conservative approach that alters only a single label at a time must be taken. This choice prevents gross unjustified label changes that could lead to degraded classification performance, but it also means that more label sets must be generated to achieve sufficient classifier diversity. Moreover, it is expected that the conservative approach can effect significant performance improvement only when the training data set is relatively small and hence unstable [16].

The biggest drawback of the proposed framework is that numerous classifiers must be trained. However, this issue is mitigated by the perfectly parallelizable form of the algorithm, the relatively low cost of computational power, and the fact that the learning can be performed offline sans time constraints.

VI. CONCLUSION

A new elegantly simple algorithm for performing classification by intentionally altering labels was presented. The approach is general in that it can be used in conjunction with any probabilistic classification method. Importantly, the framework can be employed when labeling errors are suspected, as well as in the absence of information suggesting such errors.

In the future, additional experiments will be conducted on data for other remote-sensing applications, such as land mine [17] and UXO detections [4], [18], [19]. Other future research will investigate the performance of the proposed method when larger sets of training data, different classification algorithms, and other prediction-aggregation methods are employed.

REFERENCES

[1] J. Surowiecki, *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. New York: Doubleday, 2004.
[2] D. Mackay, "Information-based objective functions for active data selection," *Neural Comput.*, vol. 4, no. 4, pp. 590–604, Jul. 1992.
[3] X. Zhu, "Semi-supervised learning literature survey," Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, Tech. Rep. 1530, 2005.
[4] S. Tantum, Y. Yu, and L. Collins, "Bayesian mitigation of sensor position errors to improve unexploded ordnance detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 1, pp. 103–107, Jan. 2008.
[5] D. Rubin, *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: Wiley, 1987.
[6] G. Dobeck, J. Hyland, and L. Smedley, "Automated detection/classification of sea mines in sonar imagery," *Proc. SPIE*, vol. 3079, pp. 90–110, 1997.
[7] D. Williams and E. Coiras, "On sand ripple detection in synthetic aperture sonar imagery," in *Proc. ICASSP*, 2010, pp. 1074–1077.
[8] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982.
[9] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.
[10] T. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*. New York: Springer-Verlag, 2001, pp. 1–15.
[11] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.
[12] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, 1991.
[13] Y. Freund, "Boosting a weak learning algorithm by majority," *Inf. Comput.*, vol. 121, no. 2, pp. 256–285, Sep. 1995.
[14] V. Raykar, S. Yu, L. Zhao, A. Jerebko, C. Florin, G. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: Whom to trust when everyone lies a bit," in *Proc. ICML*, 2009, pp. 889–896.
[15] N. Lawrence and B. Schölkopf, "Estimating a kernel Fisher discriminant in the presence of label noise," in *Proc. ICML*, 2001, pp. 306–313.
[16] C. Tomasi, "Past performance and future results," *Nature*, vol. 408, p. 378, 2004.
[17] D. Williams, V. Myers, and M. Silvious, "Mine classification with imbalanced data," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 3, pp. 528–532, Jul. 2009.
[18] Q. Liu, X. Liao, and L. Carin, "Detection of unexploded ordnance via efficient semi-supervised and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 9, pp. 2558–2567, Sep. 2008.
[19] L. Beran and D. Oldenburg, "Selecting a discrimination algorithm for unexploded ordnance remediation," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 9, pp. 2547–2557, Sep. 2008.