

Underwater Mine Classification with Imperfect Labels

David P. Williams
 NATO Undersea Research Centre
 williams@nurc.nato.int

Abstract

A new algorithm for performing classification with imperfectly labeled data 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 by drawing sets of labels as samples from a Bernoulli distribution with a specified labeling error rate. Additionally, ideas from multiple imputation are exploited to provide a principled way for determining an appropriate number of label sampling rounds to consider. The approach is demonstrated in the context of an underwater mine classification application on real synthetic aperture sonar data collected at sea, with promising results.

1. Introduction

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 set of training data. 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 data 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.

The objective of underwater mine classification tasks is to classify underwater objects as targets (*i.e.*, mines) or clutter (*e.g.*, rocks). Typically in an experimental sea trial, 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 autonomous underwater vehicle (AUV). Then, a human will manually label the objects in the sonar imagery based on the target-deployment knowledge.

However, other unknown objects may already be present at the test site. If the unknown objects are not inspected optically to verify that they are all non-mines, such objects may be labeled incorrectly. Moreover, the confusion introduced by AUV navigation errors can also contribute to flaws in the labeling process.

The classification algorithm with imperfect labels presented in this work is meant to address these issues that arise with real data collected at sea. However, the technique can be applied to diverse domains for which labeling is not trivial, such as medical diagnostics, character recognition, and document classification.

The standard approach [4] for addressing label imperfections in binary classification problems is to modify the probability of a label, y_i , of a data point \mathbf{x}_i with estimated labeling error $\epsilon_i \in [0, 0.5]$ by writing

$$p(y_i = 1 | \mathbf{x}_i, \mathbf{w}) \rightarrow \epsilon_i + (1 - 2\epsilon_i)p(y_i = 1 | \mathbf{x}_i, \mathbf{w}), \quad (1)$$

where \mathbf{w} is the classifier. When $\epsilon_i = 0$, the label is assumed to be perfect.

However, this approach never considers the case in which the label is assigned the opposite label, as it would be if the original label were indeed incorrect. Therefore, the subsequent classifier that is constructed is significantly biased toward the initial set of labels supplied with the data set. The approach we propose in this work instead explicitly allows the labels initially supplied with the data set to take on the opposite values (in binary classification problems), in effect reflecting the true uncertainty of the labeling process.

The proposed approach for dealing with imperfect labels is motivated by the insight gleaned from [7] 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.

Most other related work, such as [5] 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.

The remainder of this paper is organized in the following manner. In Sec. 2, the proposed “wisdom of crowds” approach for dealing with imperfect labels in classification problems is described. Details about the underwater mine classification task are presented in Sec. 3, with experimental results on a data set of real synthetic aperture sonar (SAS) data collected at sea shown in Sec. 4. A discussion of related work is given in Sec. 5, and concluding remarks are made in Sec. 6.

2. Classification with Imperfect Labels

The proposed approach for dealing with imperfect labels is motivated by the insight gleaned from [7] 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.

To draw an analogy to the problem of classification with imperfect labels, a set of training data labels is 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. That is, 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 only the opinion of a *single* (human) labeler. Errors can exist in this set of labels. Moreover, a different (human) labeler can produce a set of labels that differs from that of the original labeler. The stochastic, imperfect nature of the labeling process can be modeled using a Bernoulli distribution.

2.1. “The Wisdom of Crowds” Algorithm

Let $\mathbf{x}_i \in \mathbb{R}^f$ denote a vector of f features representing the i -th data point of a training set of N such points. Let $y_i \in \{0, 1\}$ denote the label, *originally supplied with the data set*, that corresponds to the i -th data point, \mathbf{x}_i . Let $\epsilon_i \in [0, 0.5]$ denote the estimated labeling error associated with the label y_i .

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

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

Thus, 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 m -th 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. Conduct M such sampling rounds 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 $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$, and denote it $\mathbf{w}_{(m)}$. Repeat the classifier learning using each of the M data sets.

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

$$\begin{aligned} p(y_* = 1 | \mathbf{x}_*, \{\mathbf{w}_{(m)}\}_{m=1}^M) \\ = \frac{1}{M} \sum_{m=1}^M p(y_* = 1 | \mathbf{x}_*, \mathbf{w}_{(m)}). \end{aligned} \quad (3)$$

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

2.2. Choosing M

The proposed approach involves drawing M sets of labels (*i.e.*, sampling rounds), learning a classifier for each, and averaging the resulting predictions. The obvious question is: How many sets of labels are needed?

A similar dilemma is encountered in the work of multiple imputation [6], which replaces each missing *feature* value with a set of M samples. In the work developed for multiple imputation, 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 [6]. In our case, γ corresponds to the fraction of labels with a non-zero labeling error rate, ϵ_i . Even if $\gamma = 1$, a high efficiency can be achieved with a relatively small number of label sampling rounds, M .

Experiments that we have conducted on other application domains have provided additional insight regarding a proper choice of M . It has been observed that the number of sampling rounds needed is dependent on the classifier *stability* [8]. It is in the cases for which the classifier is unstable that accounting for label imperfections is most important. This classifier stability, generally speaking, can be seen to be a function of both the

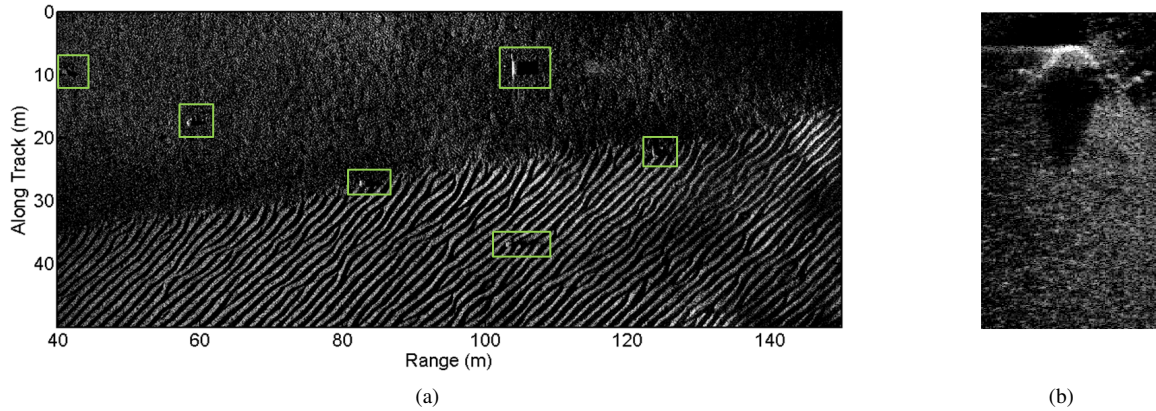


Figure 1. (a) A typical SAS image with mines indicated in green boxes. (b) A SAS image chip of a typical alarm.

total number of training data points, and the amount of class imbalance exhibited in the data set.

When a large amount of training data is available, any potentially incorrect labels will be overwhelmed by the presumably large proportion of data points with correct labels. Therefore, accounting for imperfect labels by flipping labels (according to the Bernoulli distribution) will not have a strong impact on performance because the classifier will be stable.

Thus, the number of sampling rounds M needed should be inversely proportional to the number of training data points, and directly proportional to the severity of the class imbalance. That is, accounting for labeling errors is most important when the amount of training data is low and the severity of class imbalance is high. Accordingly, in these situations the number of sampling rounds considered should be higher.

Another issue that should influence the number of sampling rounds M needed is the entropy of the labels of the data set, as dictated by the labeling error rates, $H = -\sum_{i=1}^N [\epsilon_i \log_2 \epsilon_i + (1 - \epsilon_i) \log_2 (1 - \epsilon_i)]$. A high entropy is an indication that more labels have uncertainty associated with them. In turn, it is more likely that unique sets of labels would be produced in different sampling rounds. Therefore, the number of sampling rounds M should also be proportional to the entropy of the data set labels.

3. Underwater Mine Classification

3.1. 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. A set of targets were 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 to collect high-resolution SAS data. The data from Rīga was comprised of a total of 1022 SAS images covering a total area of approximately 5.6 km². The data from Liepāja was comprised of a total of 578 SAS images covering a total area of approximately 3.2 km².

In this work, a detection algorithm that correlates a template, consisting of a generic highlight-shadow pattern characteristic of mines, with the scene images is used to produce a set of alarms that must subsequently be classified. A typical SAS image scene is shown in Fig. 1(a), while a typical alarm is shown in Fig. 1(b).

All of the alarms were then manually ground-truthed visually by inspection (while also exploiting the target deployment location information). After the detection stage, the average numbers of targets and clutter for a given mission in Rīga were 32.5 and 270.5, respectively. The average numbers of targets and clutter for a given mission in Liepāja were 35 and 266.3, respectively.

3.2. Feature Extraction

For each alarm, $f = 11$ features were then extracted. Five of the features were meant to establish the general shape and size of a given object, while the remaining six features were intended to capture 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 training tar-

gets. 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.

Two of the shape and size features are based on the correlation of the contact with highlight-shadow patterns characteristic of mines. The three other shape and size features are the object’s illuminated surface area, area of shadow cast, and peak echo strength.

The use of contextual information has largely been overlooked in terms of features for classification. In this work, 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. Additional features for a given contact are the proximity to the nearest contact, and the likelihood of the contact location being within sand ripples [9].

3.3. 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).

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

$$\epsilon(d) = \begin{cases} \alpha (1 + \exp\{-d\alpha + \gamma\})^{-1}, & \text{if } y = 1 \\ \alpha - \beta (1 - \exp\{-d/\gamma\}), & \text{if } y = 0, \end{cases} \quad (4)$$

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

In the figure, the cyan 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.

4. Experimental Results

Three different classification methods are applied to each of the two data sets. In all three approaches, a logistic regression algorithm is employed as the classifier. The differences among the methods are in the manner that labeling error is handled.

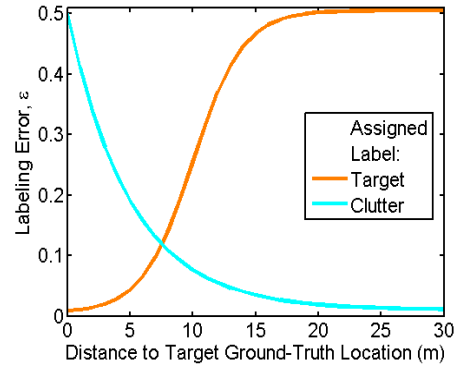


Figure 2. Definition for the labeling error.

The first approach (denoted “ $\epsilon = 0$ ”) assumes no labeling error exists, so it represents a standard logistic regression algorithm. The second approach (denoted “ $\epsilon \neq 0$ ”) accounts for labeling imperfections in the *standard* way [4], as in (1). The third approach is the proposed “wisdom of crowds” method described in Sec. 2, which involves drawing sets of labels from a Bernoulli distribution according to the estimated labeling error rate. For this proposed method, $M = 100$ sampling rounds were used.

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 is used as training data once, with the data from the remaining five missions treated as testing data.

Performance on the two underwater mine classification data sets is shown in terms of the average receiver operating characteristic (ROC) curves in Fig. 3(a) and 3(b). The corresponding area under the ROC curve (AUC), which provides a scalar summary measure of performance, is shown in Table 1. As can be seen from the figures and the table, the proposed approach performs better than the competing approaches.

Table 1. AUC (mean \pm one standard deviation from the six trials) for each of the two test sites.

METHOD	RIGA	LIEPĀJA
$\epsilon = 0$	0.9389 \pm 0.0237	0.9775 \pm 0.0165
$\epsilon \neq 0$	0.9653 \pm 0.0108	0.9823 \pm 0.0263
PROPOSED	0.9726 \pm 0.0052	0.9932 \pm 0.0033

For the proposed approach, 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

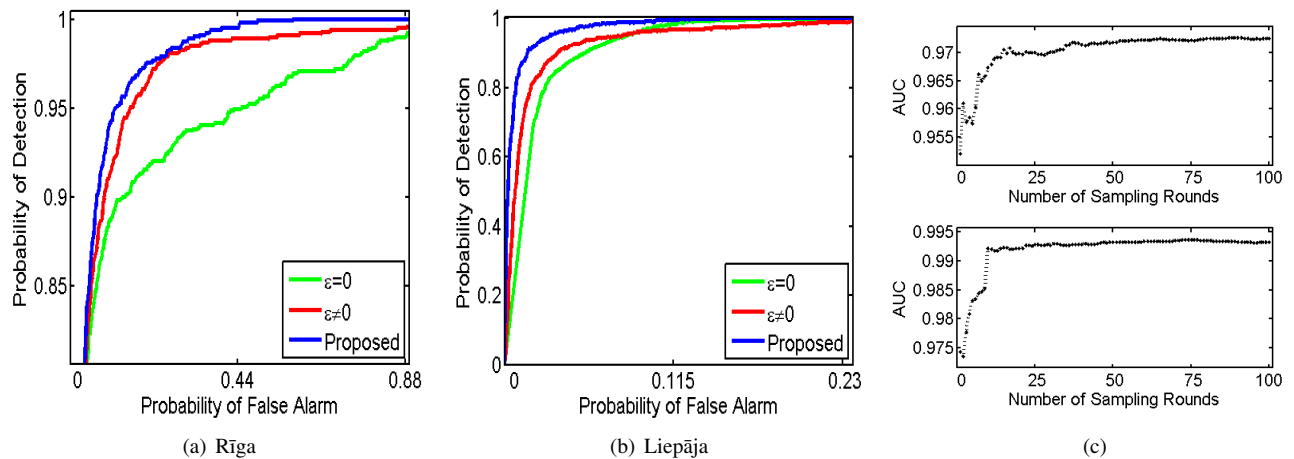


Figure 3. (a-b) Average ROC curves and (c) evolution of the AUC for the proposed method (top: Rīga; bottom: Liepāja).

the performance stabilized quickly, after approximately only $M = 20$ sampling rounds.

5. Discussion

There are several strands of related work that are similar to the proposed “wisdom of crowds” approach.

Multiple imputation [6] addresses the case of missing *feature values* by drawing samples from an assumed distribution. By imputing the missing values multiple times, multiple replicas of the data set are created. Therefore, the proposed “wisdom of crowds” approach can be viewed as a form of multiple imputation, except for labels rather than features.

Several other methods are also based on the idea of combining the results of multiple classifiers to improve performance. Among these are the mixture of experts [3], and ensemble methods like bagging [1] and boosting [2]. However, none of the aforementioned methods consider creating multiple classifiers by purposely altering the labels of data points, as is done in the proposed method.

6. Conclusion

A new, elegantly simple algorithm for performing classification with potentially imperfectly labeled data was presented. The approach is general in that it can be used in conjunction with any classification method. The proposed technique, inspired by the idea of the “wisdom of crowds,” was demonstrated on two underwater mine classification data sets of real SAS data collected at sea.

Arguments for determining an appropriate number of label sampling rounds to consider were also provided.

Interestingly, 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 result suggests that providing a measure of confidence for each manually supplied label can be beneficial when working with real data.

References

- [1] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.
- [2] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121:256–285, 1995.
- [3] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.
- [4] M. Opper and O. Winther. Mean field methods for classification with Gaussian processes. In *Advances in NIPS II*, pages 309–315, 1998.
- [5] 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 *Proceedings of ICML*, pages 889–896, 2009.
- [6] D. Rubin. *Multiple imputation for nonresponse in surveys*. John Wiley and Sons, 1987.
- [7] J. Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday, 2004.
- [8] C. Tomasi. Past performance and future results. *Nature*, 408:378, 2004.
- [9] D. Williams and E. Coiras. On sand ripple detection in synthetic aperture sonar imagery. In *Proceedings of ICASSP*, pages 1074–1077, 2010.