Supervised Discovery of Unknown Unknowns through Test Sample Mining (Student Abstract)

Zheng Wang,[1] **Bruno Abrahao,**[1*] **Ece Kamar**[2]
[1]NYU Shanghai, China
[2]Microsoft Research, USA
{zheng.wang, abrahao}@nyu.edu, eckamar@microsoft.com

Abstract

Given a fixed hypothesis space, defined to model class structure in a particular domain of application, unknown unknowns (u.u.s) are data examples that form classes in the feature space whose structure is not represented in a trained model. Accordingly, this leads to incorrect class prediction with high confidence, which represents one of the major sources of blind spots in machine learning. Our method seeks to reduce the structural mismatch between the training model and that of the target space in a supervised way. We illuminate further structure through cross-validation on a modified training model, set up to mine and trap u.u.s in a marginal training class, created from examples of a random sample of the test set. Contrary to previous approaches, our method simplifies the solution, as it does not rely on budgeted queries to an Oracle whose outcomes inform adjustments to training. In addition, our empirically results exhibit consistent performance improvements over baselines, on both synthetic and real-world data sets.

Introduction

Blind spots in classification tasks represent a major source of algorithmic bias in machine learning, leading to errors, unfairness, and other problems. Among the several forms of blind spots, *unknown unknowns* (u.u.s) are commonly referred as data points that belong to classes that are not well represented in a training model. As a result, classifiers often give incorrect class membership predictions to this examples with high confidence (Attenberg, Ipeirotis, and Provost 2015; Lakkaraju et al. 2017). Different from the traditional prediction errors, u.u.s arise due to **the systematic structural mismatch between the trained model and the target space**, i.e., given a fixed hypothesis space, we denote examples of a given class that are not well represented in the training data unknown unknowns. Recent work has addressed the algorithmic challenges to tackle this form of "data blindness", some of which rely on budgeted queries to an oracle, whose outcomes guide adjustments in the training model (Lakkaraju et al. 2017).

*Supported by a National Natural Science Foundation of China (NSFC) grant #61850410536.

The conceptual idea of a "class" is subject to the interpretation of the observer. Moreover, different hypothesis spaces will conceptually separate the feature space into different classes. For instance, adding more features may impact the ability to recognize further patterns, i.e., *feature blindness*. In this work, we employ a working definition of an u.u. class by making an assumption about their geometric structure. That is, given a fixed hypothesis space, the u.u. classes form separable clusters in the feature space, which make them distinguishable from known structures in the target space. This definition is without loss of generality, as it allows for any abstraction of conceptual blindness to examples.

Under the preceding assumption, a key component of our approach consists of a Random Test Sampling and Cross-Validation method, which illuminates further structure in the target space through cross-validation on a modified training model, which is set up to mine and trap u.u.s in a marginal training class. Our method has proved its effectiveness in identifying u.u.s on both synthetic and real-world data sets, outperforming traditional approaches while bearing the simplicity of its algorithmic design. Particularly, we assume no presence of an oracle or human expert, which makes the method feasible and scalable to real-world applications.

Framework Formulation

For convenience, we set our scope on multi-class classification. Let M be the classifier. Let $X = \{X_1, X_2, \cdots, X_m\}$ be the training set, where X_i denotes a specific class of data with a unique label. Let Y be the test set (as a representative of the target space), for which we assume the potential existence of one or more u.u. classes. To detect the u.u.s, we design a **Random Test Sampling and Cross-Validation (RTSCV)** framework, which consists of the following steps:

1. Random sampling phase: We randomly select a small portion of the test data, denoted X_s, and add it to the existing X as a new dummy (sample) class X_s. In this way, we obtain a new training set $\widetilde{X} = \{X_1, X_2, \cdots, X_m, X_s\}$.
2. We perform an $(m+1)$-fold cross-validation on $\widetilde{X}$. We then extract all the samples in $\widetilde{X}$ assigned to the sample class X_s and denote it X_u, which we expect to have

retained, mostly, u.u. examples in X_s.

3. Last, we add X_u to the initial training set X and obtain the final training set $\overline{X} = \{X_1, X_2, \cdots, X_m, X_u\}$. Then, we train an $(m+1)$-class classifier M with this training set and test the model on the test set Y.
4. To discover further u.u. classes, we repeat the process until class X_u after cross-validation ends up empty.

Theoretical Intuition

The key to our approach lies on the successful disentanglement of the u.u. class from the original training classes during cross-validation, which exploits the potential assumption that cross-validation is not only able to identify (Brodley and Friedl 1999), but also to *relabel* the mislabeled data.

The theoretical intuition of our approach lies on the property that the samples from the test set make up a high variance class, whose boundary encompasses all other classes. Similarly to a hierarchical classification argument, examples that belong to known classes are likely placed in the right classes due to the conciseness and specificity of the representation. On the contrary, u.u.s are placed in the high-variance all-encompassing class due to dissimilarity.

As a 2-dimensional illustration, Figure 1 shows a setting where we consider three classes of samples from multivariate Gaussian distribution. Note that the application of our method have numerically isolated the u.u.s.

Our ongoing work aims to provide complete performance guarantee based on properties of the feature space, such as class separability, intra-class variance, and sample size.

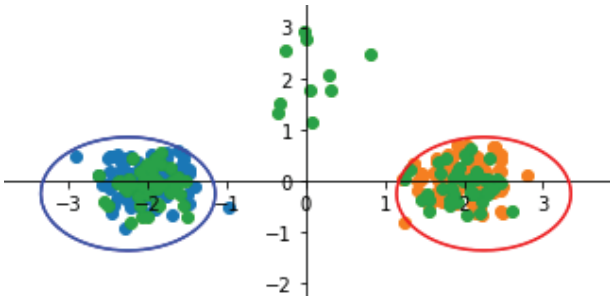

Figure 1: 2D example of a Bayesian decision boundary separating the original training classes (blue and orange) and the sample class (green). Our methods identifies the green samples outside the boundary as u.u.s.

Experimental Evaluation

We present experimental results on both synthetic and real-world data sets. For the synthetic data, we generated 10 known classes and one u.u. class. We compare our approach with a benchmark obtained by directly applying the "blind" classification algorithm. In addition, we use classification with probabilistic threshold (CPT) and clustering with side information as baselines. We show results for Support Vector Machine (SVM) and K-Nearest-Neighbors (KNN).

Figure 2 plots classification performance as the standard accuracy score, against class separability, measured by the Silhouette score. Our method achieves a consistent performance improvement over the benchmark score and other baselines under different class separability levels. In particular, when the Silhouette score is low, our method could achieve near-perfect structural reconstruction, indicating the exhaustive identification of u.u.s.

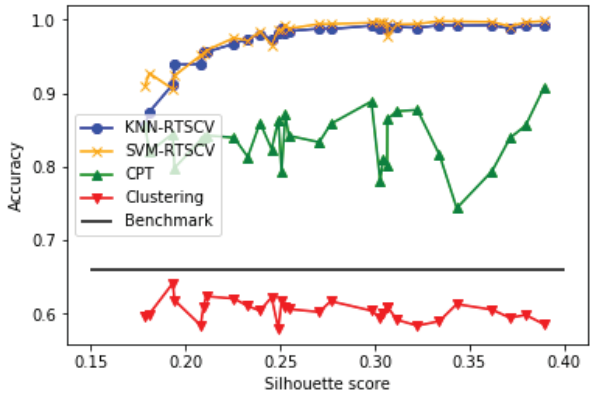

Figure 2: Comparison with benchmark and baseline methods against class separability measured by Silhouette score.

	Class 1	Class 2	Class 3
Iris	**1.0** (0.61)	**0.81** (0.52)	**0.83** (0.60)
Wine	**0.91** (0.62)	**0.85** (0.65)	**0.85** (0.71)
Knowledge	**0.88** (0.64)	**0.92** (0.66)	**0.87** (0.63)

Table 1: Performance on real-world data sets. Bold scores represent the classification accuracy after applying our method, with the scores in parenthesis denoting the benchmark accuracy. Column label denotes the index of the class omitted from the training set.

We also illustrate the approach with real-world standard classification benchmark data sets, Iris, Wine, and Knowledge Modelling[1]. For each data set, we biased the training set by removing one class of training samples while keeping the test set unchanged. We show results for KNN classification. Table 1 shows enhanced performance of the classification tasks across all three data sets after applying our method to identify unknown unknowns.

References

Attenberg, J.; Ipeirotis, P.; and Provost, F. 2015. Beat the machine: Challenging humans to find a predictive model's "unknown unknowns". *Journal of Data and Information Quality* 6(1):1:1–1:17.

Brodley, C. E., and Friedl, M. A. 1999. Identifying mislabeled training data. *Journal of Artificial Intelligence Research* 11(1):131–167.

Lakkaraju, H.; Kamar, E.; Caruana, R.; and Horvitz, E. 2017. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In *Proc. of the Thirty-First AAAI Conference on Artificial Intelligence*.

[1] http://archive.ics.uci.edu/ml