Simulation of Annotators for Active Learning: Uncertain Oracles

Adrian Calma and Bernhard Sick

Intelligent Embedded Systems
University of Kassel, Germany
{adrian.calma,bsick}@uni-kassel.de

Abstract. In real-world applications the information for previously unknown categories (labels) may come from various sources, often but not always humans. Therefore, a new problem arises: The labels are subject to uncertainty. For example, the performance of human annotators depends on many factors: e.g., expertise/experience, concentration/distraction, fatigue level, etc. Furthermore, some samples are difficult for both experts and machines to label (e.g., samples near the decision boundary). Thus, one question arises: How can one make use of annotators that can be erroneous (uncertain oracles)? A first step towards answering this question is to create experiments with humans, which involves a high time and money effort. This article addresses the following challenge: How can the expertise of erroneous human annotators be simulated? First, we discuss situations in which humans are prone to error. Second, we present methods for conducting active learning experiments with simulated uncertain oracles that possess various degrees of expertise (e.g., local/global or class/region dependent).

Keywords: Active Learning, Uncertain Oracles

1 Introduction

Consider the following problem: we have access to a large set of unlabeled images and we have the possibility to buy labels for any data point, Our first goal is to train a classifier with the highest possible accuracy. A possible approach is to label all the images and then train the classifier on the labeled data set. Now, suppose we have a limited budget, which doesn't allow us to label the all images. Our second goal is to **keep the costs to a minimum**. Thus, we need a strategy to determine which images should be labeled. A naive strategy would be to select the images at random. But, we can do better than that, if we make use of a selection strategy that selects the most *informative* images. Precisely at this point, active learning (AL) comes in, more specifically pool-based active learning (PAL). The learning cycle is presented in Figure 1: there is a **large set of unlabeled data** and our **goal** is **to train a model** (e.g., a classifier). Thus, we need to select **the most informative** data points based on a **selection strategy** and present them to an annotator (e.g., a domain expert), generally called **oracle**, for labeling. The labeled samples are added to the training set (the set with labeled data), the **classifier is updated** and, depending on the chosen **stopping criteria** (e.g., is there still money in our budget?), we continue to ask for more labels or not.

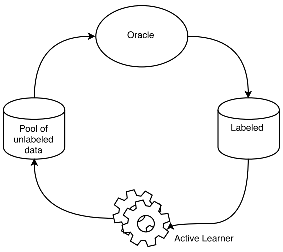

Fig. 1. Pool-based active learning cycle.

At this point we can ask ourselves: Are the labels provided by the human annotators correct? Probably not, as we can assume that humans are prone to error (Section 2). Thus, a new question arise: **How can we deal with uncertainty regarding the labels?** A first step towards answering the previous question is to develop techniques for simulating human experts prone to error. As we assume that they are unsure regarding the classification decision, we call these annotators *uncertain oracles.* Thus, this article focuses on presenting:

- cases in which the uncertain oracle misclassifies data (see Section 3) and
- techniques for simulating uncertain oracles in AL (see Section 4).

In the remainder of this article, we first present the possible causes for erroneous labels and explain what we mean by the term "uncertainty" (Section 2). Then, we present and categorize various types of expertise (Section 3). In Section 4, we introduce possible approaches for simulating error prone oracles. Then, related work will be summarized in Section 5. Finally, Section 6 concludes the article.

2 Motivation – The Problem

By now, we assumed that the answers provided by the oracles are always right. But, it is obvious that they are not always right. On the one hand, the **performance** of human annotators (human oracles) depends on multiple factors, such as: **expertise**, **experience**, level of **concentration**, level of **interest**, or level of **fatigue** [1]. On the other hand, the labels may come from simulations or test stands. Once again, it is justifiable to assume that due to imperfect simulations, sensor noise, or transmission errors, the labels are subject to uncertainty.

Depending on the difficulty of the labeling task, the oracles might be right in case of "easy" classification problems. The more difficult a classification task is, the likelier it is that the oracle has a higher degree of doubt (i.e., is more uncertain) about its answer. Thus, the label uncertainty depends on the difficulty of the classification task. That is, the number of steps an annotator has to perform for determining the right class, the designated time, and the risk involved by misclassification. This factors come in addition to the previously presented sources of uncertainty, such as required knowledge for problem understanding, experience regarding similar classification problems or labeling tasks, concentration, or tiredness.

What do we mean by "uncertainty"? When humans are asked to provide information about an actual situation, the confidence regarding the given answer depends on diverse factors, such as the difficulty/complexity to assess that information, previous experience, or knowledge. Certainly, there are circumstances when we cannot state our answer with absolute confidence. Thus, we tend to add additional information about the quality of our answer, i.e., to quantify and qualify our confidence [2].

On these grounds, we cannot assume that the oracles are omniscient, but we have to soften the assumption of omniscience: An oracle may be wrong. In this context, the "uncertainty" is the degree of confidence for given label. Consequently we ask ourself, how can we make use of the uncertain oracles, especially, how can we exploit the oracle's firm knowledge?

3 Human Expertise

When an expert has worked for a long period of time on a classification task, he posses more "experience". That is, he has seen and labeled more data than an oracle that just started to work on the labeling task. Therefore, such an oracle possesses *global expertise* about the classification problem. On the other hand, depending on how difficult the classification problem is or on the degree of *expertise* and experience, the oracle may bear only limited knowledge about the learning task, i.e., *local expertise*.

At this point, we assume that the expertise of an oracle (its degree of uncertainty) is time invariant.

3.1 Global Expertise

The annotators have a global expertise in the sense that their knowledge is not limited to a certain region of the input space or to a specific class. They "know" the problem in all its aspects. Still, they may possess different levels of expertise. Moreover, samples exist that are hard to label for both the learning system as well as for the oracles. For example, samples that lie near the decision boundary of a classifier are good examples for data points that might be difficult to label by the oracle and the active learner.

From a practical point of view, we may ask the oracles to provide additional information when they provide labels for samples. This is required for assessing their certainty, or rather their *uncertainty* regarding the provided answer. Such additional information may include asking for [1]:

1. a degree of confidence for one class,
2. membership probabilities for each class,
3. a difficulty estimate, or
4. a relative difficulty estimate for two data points.

In the first case, a sample is presented to the oracle, for example an image. The oracle is asked to provide a class label for the sample and to estimate his degree of confidence. Further help regarding the degree of confidence may be provided: e.g. a a graphical control element with which the oracle sets a certainty value by moving an indicator on a predefined scale (i.e., a slider). Thus, a possible answer may look like *"I select class «cat» and I rate my certainty 3 on a scale from 1 to 4, where 4 is the highest score"*.

Another possibility is to ask the oracle to provide class estimates for each of the possible categories. Given an 3-class classification problem, an answer may be *"The self estimated probability for the first class is 0%, for the second class 30%, and for the third class 70%"*.

The last two cases address cases where the oracle has to estimate how difficult it was for him to label a specific data point. Possible answer may look like *"I choose class «cat» and it was hard for me to determine it"*, if it was asked to label only one sample, or *"It was* easier *for me to label the image depicting a «cat» than the one showing a «liger»"*, if asked to label to images simultaneously.

3.2 Local Expertise

The oracles possess a *local expertise* in the sense that they do not have enough "experience", they can only recognize specific classes, they are more reliable for specific regions of the input space, or for certain features. That is, the human annotators are experts for:

1. different classes,
2. different regions of the input space, or
3. different dimensions of the input space (i.e., features, attributes).

We assume that, in some applications, the oracles have not only diverse degrees of experience and expertise, but they have various levels of proficiency for different parts of the classification problem. For example, the oracle may be more confident and adept in detecting some certain classes. The quality of the given answers and his confidence may vary over the regions of the input space or it may depend on the considered features (dimension of the input space).

It is not required to change the way the active learner queries new labels. The query approaches described in Section 3.1 can be adopted for this case too.

3.3 Disparate Features

Up to this point we assumed that the oracle and the active learner are considering the same features for solving the classification problem. But, this is not always the case. For example, complex processes happen in our brains when we examine an image. It is hard to say which "features" we consider when trying to recognize or evaluate the content of that specific image. Still the active learner "views" the same image, but it may consider additional features such as histograms or apply filters (e.g. anisotropic diffusion [3] or median [4] filters) or transformations (e.g. Fourier [5] or Hough [6] transform) on the image. Obviously, we can provide these additional information to the oracles, but the active learner might not have access to all features that were "extracted" by the oracles.

Once again, the answers expected from oracles can be implied from Section 3.1. But, you may ask yourself why we do not ask the oracle for additional information regarding the features that it considers for its decision. As we focus on classification tasks, we do not consider it in this work, but it is definitely an interesting research topic, commonly referred to as *active feature selection* [7].

4 Simulate Error Prone Annotators

A first step towards exploiting the knowledge of an uncertain oracle would be to analyze how the current AL paradigms perform in combination with multiple oracles. But, such experiments are costly both in terms of money and time. If we are able to successfully simulate uncertain oracles, then we can better investigate the performance of the selection strategies and of the classifiers without generating additional costs in this research phase. Moreover, based on the gathered knowledge from the investigation of current active learning techniques in a dedicated collaborative interactive learning (D-CIL) context, we can develop new ones, that take the uncertainty into consideration. That brings us to the following questions: how can we simulate error prone annotators (uncertain oracles)?

In the following, we will describe different approaches for simulating the uncertain oracles.

4.1 Omniscient Oracle

For the sake of completeness, we shortly describe how an omniscient oracle can be simulated and what we understand under *experience* in this context. Simulating this type of oracle is straight forward: It returns the true labels of the samples. That is, the labels are not manipulated in any way.

How can we simulate the *experience*? We define the experience as the number of samples the uncertain oracle has already seen and labeled. Thus, when we consider the complete data set for training a classifier (i.e., supervised learning) we can simulate an uncertain oracle with maximal global experience. Global, in the sense that the expertise is not limited to a region of the input space or to a specific class.

4.2 Uncertain Oracle with Global Expertise

At first, we concentrate on how to simulate uncertain oracles with global expertise and the same degree of experience. We assume that the labels near the decision boundary of the classifier are hard to classify for both the human expert (human oracle) as well as for the classifier. Thus, we can simulate an uncertain oracle by randomly altering (changing) the classes of the samples lying near the decision boundary. A legit question may arise: What is the “right” decision boundary? We do not know, but we can estimate it. As one of the goals of active learningis to be as good as a learner trained in a supervised way, we can train a classifier in a supervised way (i.e., overall data set). The decision boundary resulted from this classifier trained can be used to determine the samples for which the labels are altered.

The next challenge is to simulate oracles that have different levels of experience. For example, the oracle may have just started labeling samples for this type of problem. Thus, they have only a labeled few samples and, of course, their experience is based on a small number of data. One possible way to simulate its “experience” is to reduce the number of samples on which the classifier is trained. As the classifier is used as a model of the experience, by reducing the number of samples we increase the level of uncertainty. By doing so, we simulate an oracle that has little experience. Depending on the reduction factor, uncertain oracles with different levels of experience can be simulated. Moreover, if we can split the data in such a way, that the training set of the classifier used to simulate the uncertain oracle is larger than the pool of unlabeled data. Thus, the data from which the uncertain oracles gathered their experience is larger than the data from which the active learner can select samples for labeling, resulting in a simulated oracle with a higher degree of expertise.

Another possibility to simulate uncertain oracles with different levels of experience is to alternate the parameter values of the classifiers. For example, we can simulate the expertise of an uncertain oracle with a classifier trained with default parameters. For a better expertise, we can imply heuristics (e.g., grid search) to find suitable parameters for the classifiers.

Furthermore, the expertise can be simulated by different types of classifiers. We can use generative or discriminative classifier for simulating the expertise of an expert.

Last but not least, we can add noise to the feature values. Of course, this is not always possible, as it depends on the type of feature (i.e., nominal, continuous, ordinal, etc.) and on the values range. By doing so, we can simulate uncertain oracles that have an experience built on similar samples.

In a nutshell, we can simulate oracles with global and various degrees of expertise by

- modifying (altering) the classes of the samples lying near the decision boundary,
- training different classifier types for various uncertain oracles, and
- training a classifier
  - on training sets of different size (more or less samples than in the pool of unlabeled data),
  - using different parametrization strategies and parameter sets, or
  - adding noise to the feature values (if possible and if it makes sense).

Additionally, any combination of the previous simulation can be implied. For example, if we want to simulate an oracle with little global expertise based on similar samples, we can reduce the training set of the classifier and add noise to the feature values.

4.3 Uncertain Oracle with Local Expertise

The expertise of an oracle can be restricted to a certain class or to a specific region of the input space. Thus, to simulate a better expertise with respect to one or more classes of our choice, we can change the labels of the samples belonging to the classes for which we would like to simulate a little (or no) expertise. It is also possible to exclude the samples belonging to one class, which translates to "the uncertain oracle has no expertise regarding this specific class". One possible approach is to train a generative classifier on these data. The resulting classifier estimates the processes that are supposed to generate the data, i. e. one process generates samples belonging to one class. That is, a process generates samples belonging to only one class. Therefore, we can artificially change the labels of the estimated processes, which results in an erroneousness classification of samples that were assumed to be generated by that process.

The expertise of the uncertain oracles may be restricted to a specific region of the input space. Depending on the feature values, the labeling quality can suffer. For example, an uncertain oracle is more accurate regarding samples that lie in regions of the input space, which have been previously seen or learned by the oracle. We propose two ways to simulate the local expertise: (1) by using various classifier types and (2) by deliberately altering the class affiliations of the samples lying in those regions.

By using different classifier types, the regions of the input space are modeled in different ways and, thus, the result of the classification may vary.

By modifying the classes of the samples lying in specific regions of the input space, the result of the classifier is modified. That is, for samples lying in these regions, the expertise of the uncertain oracle is diminished.

The difference between class based experience and region based experience is showed in Figure 2. Here, we have a region of the input space where two classes strongly overlap, *green* ∘'s and *blue* +. If we assume that a human expert has firm knowledge about class *green* ∘, then he will probably label the samples that belong to the green class correctly and the others not (higher error rate for *blue* + and *red* △). On the other had, assuming that the oracle correctly labels samples in a given region of the input space leads us to the conclusion that it labels correctly all the samples in the specified region. For example, the uncertain oracle has a region based expertise for samples having feature values $\in [-1.5, 1.5]$, will lead to correct class affiliation for samples lying in this region. In this concrete case, samples lying in the square defined by $(-1.5, -1.5)$ and $(1.5, 1.5)$ and belonging to either class are labeled correctly.

An overview of the introduced simulation methods is presented in Figure 3. The core of the simulation techniques is the assumption regarding which features are considered. The described simulation methods can be applied for both cases:

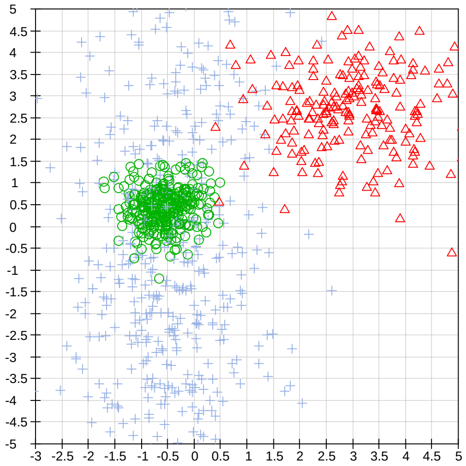

Fig. 2. Samples belonging to three classes (*green* ∘'s, *blue* +'s, and *red* △)'s depicted in the input space, whereby the processes generating samples belonging to *green* ∘'s and *blue* +'s strongly overlap.

when the uncertain oracle considers the same features as the active learner and when not.

4.4 Motivating Example: Generative Classifier based Simulation

One possible way to simulate the expertise of an uncertain oracle is by means of a generative classifier, e.g. a classifier based on mixture models, which is based on a probabilistic mixture modeling approach. That is, for a given D-dimensional input sample $\mathbf{x}'$ we can compute the posterior distribution $p(c|\mathbf{x}')$, i.e., the probabilities for class c membership given the input $\mathbf{x}'$. To minimize the risk of classification errors we then select the class with the highest posterior probability (cf. the principle of *winner-takes-all*). Thus, the "uncertainty" can be computed as $1 - p(c'|j)$, where $c' = \text{argmax}_c\, p(c|j)$. In case of other classifier types (e.g., Support Vector Machines), Platt scaling [8] can be used to transform the outputs into probability distributions.

5 Related Work

In [9], the authors simulate oracles with different types of accuracies: 10% of samples are incorrect, 20% unknown, and 70% uncertain knowledge. k-means clustering is implied in [10] to generate the concepts and to assign the oracles to different clusters, in order to simulate the experience (in this article called "knowledge sets"). Clustering is also used in [11], where some clusters represent regions for which the oracles give unsure as feedback. Virtual oracles for binary classification, with different labeling qualities, controlled by two parameters that

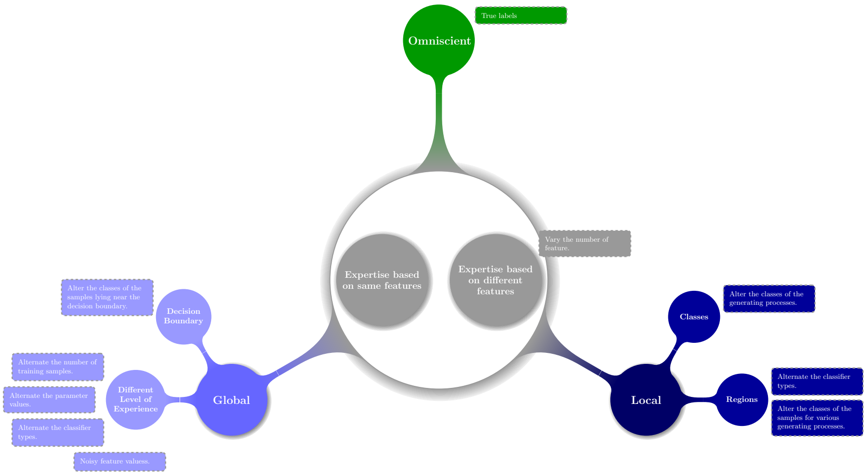
Omniscient
True labels
Expertise based on same features
Expertise based on different features
Vary the number of feature.
Global
Decision Boundary
Alter the classes of the samples lying near the decision boundary.
Different Level of Experience
Alternate the number of training samples.
Alternate the parameter values.
Alternate the classifier types.
Noisy feature valuess.
Local
Classes
Alter the classes of the generating processes.
Regions
Alternate the classifier types.
Alter the classes of the samples for various generating processes.

Fig. 3. Types of expertise and possible simulation practices.

represent the label accuracy regarding the two classes are presented in [12]. In [13], a uniform distribution is implied to simulate various behavior of the oracles. Randomly flipping labels with a specific probability [14] and ranges for the noise rate [15] are also applied to simulate uncertain oracles. A Gauss distribution [16] has also been use to simulate the expertise of oracles. But also multiple oracles have been simulated, where their label quality does not vary [17].

6 Conclusion

In this article, we addressed a challenge in the field of AL and, especially, in the field of D-CIL [1], where oracles might be wrong for various reasons. Thus, the queried labels are subject to uncertainty. The research regarding uncertain oracles is still in its infancy, so we proposed **simulation methods for uncertain oracles** in order to help the research go further. The simulation methods will help investigate the performance of the current AL techniques and understand their advantages and disadvantages. Moreover, new questions for future research arise: How can we exploit the uncertain oracles? Is it necessary to re-query labels for already labeled samples? How can we learn (model) the expertise of an uncertain oracle? How do we decide whether the uncertain oracle is erroneous or the process to be learned are nondeterministic? How do we decide whom to ask next?

References

1. Calma, A., Leimeister, J.M., Lukowicz, P., Oeste-Reiß, S., Reitmaier, T., Schmidt, A., Sick, B., Stumme, G., Zweig, K.A.: From active learning to dedicated collaborative interactive learning. In: International Conference on Architecture of Computing Systems, Nuremberg, Germany (2016) 1–8

2. Motro, A., Smets, P., eds.: Uncertainty Management in Information Systems – From Needs to Solutions. Springer US (1997)
3. Weickert, J.: Anisotropic Diffusion in Image Processing. B.G. Teubner Stuttgart (1998)
4. Zhu, Y., Huang, C.: An improved median filtering algorithm for image noise reduction. Physics Procedia **25** (2012) 609–616
5. Cochran, W., Cooley, J., Favin, D., Helms, H., Kaenel, R., Lang, W., Maling, G., Nelson, D., Rader, C., Welch, P.: What is the fast Fourier transform? Proceedings of the IEEE **55** (1967) 1664 – 1674
6. Nixon, M.S., Aguado, A.S.: Feature Extraction and Image Processing. Academic Press (2008)
7. Liua, H., Motoda, H., Yua, L.: A selective sampling approach to active feature selection. Artificial Intelligence **159** (2004) 49–74
8. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers **10** (1999) 61–74
9. Fang, M., Zhu, X.: Active learning with uncertain labeling knowledge. Pattern Recognition Letters **43** (2013) 98–108
10. Fang, M., Zhu, X., Li, B., Ding, W., Wu, X.: Self-Taught Active Learning from Crowds. In: 2012 IEEE 12th International Conference on Data Mining (ICDM), Brussels, Belgium (2012) 1–6
11. Zhong, J., Tang, K., Zhou, Z.H.: Active Learning from Crowds with Unsure Option. In: 24th International Conference on Artificial Intelligence, AAAI Press (2015) 1061–1067
12. Jing, Z., Xindong, W., S., S.V.: Active Learning With Imbalanced Multiple Noisy Labeling. IEEE Transactions on Cybernetics **45** (2015) 1081–1093
13. Kumar, A., Lease, M.: Modeling Annotator Accuracies for Supervised Learning. In: WSDM 2011 Workshop on Crowdsourcing for Search and Data Mining (CSDM 11), Hong Kong, China (2011) 19–22
14. Yan, Y., Rosales, R.: Active learning from multiple knowledge sources. In: 15th International Conference on Artificial Intelligence and Statistics (AISTATS). Volume XX., La Palma, Canary Islands (2012)
15. Du, J., Ling, C.X.: Active learning with human-like noisy oracle. In: IEEE 10th International Conference on Data Mining, Sydney, Australia (2010) 797–802
16. Zhao, L.: An Active Learning Approach for Jointly Estimating Worker Performance and Annotation Reliability with Crowdsourced Data. ArXiv (2014) 1–18
17. Shu, Z., Sheng, V.S., Li, J.: Learning from crowds with active learning and self-healing. Neural Computing and Applications (2017) 1–12