From active towards InterActive learning: using consideration information to improve labeling correctness

Bernstein, Abraham ; Li, Jiwen

Abstract: Active learning methods have been proposed to reduce the labeling effort of human experts: based on the initially available labeled instances and information about the unlabeled data those algorithms choose only the most informative instances for labeling. They have been shown to significantly reduce the size of the required labeled dataset to generate a precise model [17]. However, active learning framework assumes "perfect" labelers, which is not true in practice (e.g., [22, 23]). In particular, an empirical study for hand-written digit recognition [5] has shown that active learning works poorly when a human labeler is used. Thus, as active learning enters the realm of practical applications, it will need to confront the practicalities and inaccuracies of human expert decision-making. Specifically, active learning approaches will have to deal with the problem that human experts are likely to make mistakes when labeling the selected instances.

DOI: https://doi.org/10.1145/1600150.1600165

From active towards itive learning: using consideration information to improve labeling correctness

Abstract

Data mining techniques have become central to many applications. Most of those applications rely on so called supervised learning algorithms, which learn from given examples in the form of data with predefined labels (e.g., classes such as spam, not spam). Labeling, however, is oftentimes expensive, as it typically requires manual work by human experts. Active learning systems reduce the human effort by choosing the most informative instances for labeling. Unfortunately, research in psychology has shown conclusively that human decisions are inaccurate, easily biased by circumstances, and far from the oracle decision making assumed in active learning research. Based on these findings we show experimentally that (human) mistakes in labeling can significantly deteriorate the performance of active learning systems. To solve this problem, we introduce consideration information - a concept from marketing - into an active learning system to bias and improve the human's labeling performance. Results (with simulated and human labelers) show that consideration information can indeed be used to exert a bias. Furthermore, we find that the choice of appropriate consideration information can be used to positively bias an expert and thereby improving the overall performance of the learning setting.

From Active Towards InterActive Learning: Using Consideration Information to Improve Labeling Correctness

Abraham Bernstein
University of Zurich - Department of Informatics
Binzmühlestrasse 14
CH-8050 Zürich, Switzerland
+41 44 635 4579
bernstein@ifi.uzh.ch

Jiwen Li
University of Zurich - Department of Informatics
Binzmühlestrasse 14
CH-8050 Zürich, Switzerland
+41 44 635 4334
li@ifi.uzh.ch

ABSTRACT

Data mining techniques have become central to many applications. Most of those applications rely on so called supervised learning algorithms, which learn from given examples in the form of data with predefined labels (e.g., classes such as spam, not spam). Labeling, however, is oftentimes expensive, as it typically requires manual work by human experts. Active learning systems reduce the human effort by choosing the most informative instances for labeling. Unfortunately, research in psychology has shown conclusively that human decisions are inaccurate, easily biased by circumstances, and far from the oracle decision making assumed in active learning research. Based on these findings we show experimentally that (human) mistakes in labeling can significantly deteriorate the performance of active learning systems. To solve this problem, we introduce consideration information – a concept from marketing – into an active learning system to bias and improve the human's labeling performance. Results (with simulated and human labelers) show that consideration information can indeed be used to exert a bias. Furthermore, we find that the choice of appropriate consideration information can be used to positively bias an expert and thereby improving the overall performance of the learning setting.

Categories and Subject Descriptors

I.2.6 [Artificial Intelligence]: Learning; I.5.1 [Pattern Recognition]: Models; H.2.8 [Database Management]: Database Applications – data mining

General Terms

Active learning, supervised learning

Keywords

Consideration information, human experts

1. INTRODUCTION

The ability to extract knowledge from data gathered and make predictions from it is absolutely central to many applications such as customer relationship management, target marketing, and fraud detection. At the center of such predictions, typically, lie prediction models that classify new entities into different classes, provide a probability distribution predicting their class membership, or predict an actual value. All those so-called supervised learning methods require labeled datasets, i.e., datasets with examples containing the class information – typically called labels – or values, of sufficiently high quality for training an estimator (model). However, in many applications, acquiring the class label of the data is much more expensive than getting the data itself (e.g., as since the labeling is done manually).

Active learning methods have been proposed to reduce the labeling effort of human experts: based on the initially available labeled instances and information about the unlabeled data those algorithms choose only the most informative instances for labeling. They have been shown to significantly reduce the size of the required labeled dataset to generate a precise model [20]. However, active learning framework assumes "perfect" labelers, which is not true in practice (e.g., [25, 26]). In particular, an empirical study for hand-written digit recognition [6] has shown that active learning works poorly when a human labeler is used. Thus, as active learning enters the realm of practical applications, it will need to confront the practicalities and inaccuracies of human expert decision-making. Specifically, *active learning approaches will have to deal with the problem that human experts are likely to make mistakes when labeling the selected instances*.

Prior supervised learning research has addressed the issues of modeling noisy human labelers in real learning systems. Common approaches include estimating the quality of labelers [9, 24, 26] or learning with uncertain labels [16, 22, 23]. However, most of these learning systems only contain mechanisms to model inaccuracies of labelers, but do not contain approaches to improve accuracies of labelers. As far as we know, only very few papers discuss possible approaches to improve the general accuracies of human labeler. One example is [21], where repeated labeling has been shown to be able to improve the labeling accuracy, thereby increasing the whole learning system's performance. However, the strategy of repeated labeling is only practical if the labeling cost is low.

In this paper we show that in a wide range of problems (even non-systematic) mistakes in labeling can significantly deteriorate the performance of active learning below the one of the simple supervised learning setting, where only the initially labeled instances are considered. Based on the psychology and marketing literature we present a solution for this problem: we argue that active learning algorithms should take the "shortcomings" of human expert labelers into account and help them to improve their rating performance by providing additional *consideration information* [11, 12, 19] for each instance to be labeled. The purpose of this consideration information is to bias the human

University of Zurich, Department of Informatics, Dynamic and Distributed Information Systems Group Working Paper.
www.ifi.uzh.ch/ddis/nc/publications.

decision maker into making fewer labeling mistakes. To this goal we extend the active learning framework with a *Consideration Information Selection Function* (CISF) to determine the consideration information. Using a simulated expert (labeler) we show that consideration information can indeed be used to exert a bias. Furthermore, we show that an appropriate CISF based on insights from the psychology can positively bias both our simulated and human experts thereby improving the overall performance of the learning setting.

In the remainder, we introduce the traditional active learning framework and discuss the underlying psychology theory to develop our consideration information hypotheses. Next, we introduce the consideration information active learning approach and present the experimental setup/results. We close with the discussion of the limitations and implications.

2. Theory and Hypotheses

To illustrate our approach this section first introduces the traditional active learning framework. It then continues to discuss the limitations of human experts and its implications on active learning providing the basis for the development of the hypotheses underlying this paper.

2.1 Active Learning

In traditional classification or class probability estimation (CPE) data mining approaches an induction algorithm is given an initially labeled dataset and produces a prediction model. Given that labeled instances are, usually, difficult to come by an active learning procedure uses the labeled and unlabeled data to propose a set of instances whose labels would be most fruitful in improving the overall learned model and proposes them for labeling by an expert [4]. To that end it uses an effectiveness scoring approach (such as choosing instances that minimize the local variance [20]) to compute how effective any unlabeled instance might be in improving the overall model's performance. A formal Definition of the procedure can be defined as follows:

Definition 1: Active Learning

Input: an effectiveness score (ES) calculation approach A, an initial labeled set L, an unlabeled set U, an inducer Inducer, a stopping criterion, and an integer n specifying the number of actively selected examples in each stage.

Pseudo code:

```
0. T := L.
1. WHILE stopping criterion is not met
2.   Apply Inducer to T, create Model M.
3.   FOR each element e ∈ U compute ES with A.
4.   Select UL from U using ES, where |UL| = n
5.   FOR each element e ∈ UL,
6.     have expert label e
7.   Remove UL from U and add UL (with expert's
     labels) to T.
```

Output: model M induced with Inducer from the final set T.

2.2 Active Learning with Human Experts

The limitations of human decision makers have been the subject of various research streams. In their Nobel Prize winning work Tversky and Kahneman [14] argue that people rely on a limited number of simple heuristic principles to reduce the complex tasks of assessing probabilities and predicting values to simpler judgment operations. According to their theory, people normally use two heuristics for categorization: representativeness and availability. To answer the question "What is the probability that object A belongs to class B?" humans typically rely on the *representativeness heuristics*: here probabilities are evaluated either by the degree to which A is representative of B or by the similarity between A and B. Alternatively, people assess the probability of a class by the ease with which examples can be brought to mind, which corresponds to the *availability heuristic*. Tversky and Kahnemann show experimentally that both heuristics bias people in some way and can lead to severe judgment errors. Representativeness has been demonstrated to be insensitive to prior probability of outcomes and sample size. Availability is excessively sensitive (among other things) to the retrievability of instances, imaginative instances, and illusory correlation.

Thus, as human experts are subject to the same limitations in an active learning setting they are likely to make mistakes when labeling instances. Hence, we postulate our first hypothesis:

Hypothesis 1a: The active learning performance with a human expert is generally worse than the "ideal" setting with an "oracle" decision maker.

Hypothesis 1b: With lower human labeling precision active learning is not better than learning with random sampling, and (Hypothesis 1c:) in the worst case, even worse than simple supervised learning with an initially labeled dataset.

Extending Khanemann and Tversky's work, Baron found that human thinking and decision making can be explained as the procedure of choosing between options based on evidences and personal goals [5]. His search-inference framework asserts that human thinking and decision making consists of first searching for possibilities (i.e., candidate answers to questions or resolutions to a doubt), goals (i.e., criteria to evaluate possibilities), as well as evidences (i.e., beliefs that help determine whether a possibility is likely to achieve a goal) and then choosing among the possibilities based on the evidences, which are weighed by the goals. Consequently, human decision making can be biased to a certain extent by (intentionally) making certain types of evidence available to the human decision maker (while suppressing others). This is validated in marketing [11, 12, 19], where it has been shown that consumers with a consistent set of goals can be convinced into buying a good with a carefully chosen consideration set, i.e., the set of goods under consideration as alternatives. Hauser found [12] that the consideration set concept is highly consistent with a number of central theories and results in behavioral science [2][14][17][22]. Thus, consideration sets can serve as "counter-biasing" instruments to ameliorate the typical bias found in human probability estimation.

We propose to use this finding in a practical active learning procedure: a system could bias/influence a human expert labeler by selectively showing him/her additional evidence with each instance to be labeled. We will call this additional evidence consideration information, which can have the form of labeled or unlabeled instances, the currently inferred model (or any subset thereof), or even the instances' prior distribution (or other useful statistics). Formally, we postulate our second hypothesis:

Hypothesis 2: An active learning system can bias human experts by showing them consideration information. Depending on the consideration information human experts will be biased in differ-

ent directions – either increasing or decreasing the active learning systems' overall performance.

While Hypothesis 2 states that the overall system performance can be improved by presenting the human expert with suitable consideration information, the question is which information should be chosen. We address this issue in our final hypothesis that introduces the notion of a consideration information selection function for choosing the ideal composition of the consideration information

Hypothesis 3a: There exists a consideration information selection function that improves a human labeler's precision and in turn the overall active learning system's performance.

Hypothesis 3b: The setup described in Hypothesis 3a improves the learning performance beyond that of a purely supervised learning approach (i.e., learning on the initially labeled dataset).

Note that the traditional active learning approach with a "flawless" oracle expert provides an upper bound for the possible learning performance of an active learning system. Hypothesis 1a states that this upper bound is an idealized situation and any realistic human (i.e., error-prone) labeler is likely to lower the overall system's performance. Hypothesis 1b further strengthens this statement by claiming that, depending on the human expert's mistakes in labeling, performance will fall even lower, in the worst case (Hypothesis 1c) below the performance of the supervised learning with the initially labeled dataset, which serves as a baseline performance. Hypothesis 2 now states that careful choice of provided consideration information can bias a human expert to increase (or decrease) the overall learning performance. Hypothesis 3 makes the even stronger argument that there exists a consideration information selection function, which will improve the overall system's performance not only over the worst-case (Hypothesis 3a) but over the base-line (Hypothesis 3b).

3. The Consideration Information Active Learning Framework

Following our hypotheses, we extended the active learning framework with a consideration information selection function (CISF) to choose the biasing consideration information. We define the CISF as follows:

Definition 2: *Consideration Information Selection Function*

Given (see Definition 1): the (currently learned) model M, the labeled training dataset T, the unlabeled dataset (or test data) U, the unlabeled instanced e, and the parameters i and m.

A consideration information selection function C is defined as

(I, CM) = C (M, T, U, e, i, m),

which returns a set of consideration instances I and a consideration model CM (part of M). Here i limits the number of biasing instances (I) and m the size of the model (CM).

Consequently, the consideration information selection function (CISF) C returns I, a set of i instances, and CM, a part of M (i.e., a partial model; e.g., a pruned sub tree of the induced decision tree) limited by m. A good CISF chooses I and M to entice the expert to correct labeling. Given this definition we can now extend the traditional active learning framework.

Definition 3: *Consideration Information Active Learning*

Input: an effectiveness score (ES) calculation approach A, an initial labeled set L, an unlabeled set U, an inducer Inducer, a stopping criterion, an integer n specifying the number of actively selected examples in each stage ,and the integers i as well as m specifying the number of instances respectively, the size of the model the expert is able to absorb.

Pseudo code:

```
0. T := L.
1. WHILE stopping criterion is not met
2.   Apply Inducer to T, create Model M.
3.   FOR each element e ∈ U compute ES with A.
4.   Select UL from U using ES, where |UL| = n
5.   FOR each element e ∈ UL,
6.     have expert label e using the
         consideration information
         C(M, T, U, e, i, m).
7.   Remove UL from U and add UL (with expert's
     labels) to T.
```

Output: model M induced with Inducer from the final set T.

The only difference between the traditional active learning setup and the consideration information active learning framework (CIAL) in Definition 3 is the consideration information selection step (highlighted grey). In this step, after the active learning system has picked a set of unlabeled instances UL to be labeled by the expert, the consideration information (I, CM) is selected for each element e of UL and is shown together with e to human expert to aid (or bias) him/her with the labeling decision.

4. Experiments

This section introduces the experimental setup and shows the first results. The experiments can be divided into two groups. The first set of experiments is specifically designed to address the hypotheses developed in section 2.2. The choice of a simulated labeler limits the generalizability of the findings. Therefore, we conduct a second experiment with human labelers to show that the human experts' labeling is consistent with the simulated one.

4.1 Hypotheses Testing: Experimental Setup

4.1.1 *Consideration Information Selection Functions*

The central element for the performance of the CIAL framework (apart from the human labeler) is the appropriateness of the CISF. We, therefore, designed a number of practical CISFs, to be used in our experiments. We limited the consideration information to labeled instances as research in marketing has shown that the use of consideration sets (i.e., sets of labeled instances) is sufficient to strongly bias people [11, 12, 19]. Consequently, we limit our current explorations to CISFs, which (using Definition 1) always have $m=0$ and, hence, return $CM = \emptyset$. In the remainder we will, thus, use the following simplified notation for the CISFs: $I = C(M, T, U, e, i)$

To validate Hypotheses 1a/b/c we define the specific function CISF1, which always returns the empty set Ø of instances. Thus:

$CISF1(M, T, U, e, 0) := \emptyset$.

When combining CISF1 with an "oracle" labeler we can get the learning curve produced by a traditional "ideal" active learning algorithm. To evaluate Hypothesis 1a, we compare this curve with the curves generated when combing CISF1 with an arbitrary human labeler. To validate Hypothesis 1b we change the CIAL's

active sampling step to random sampling (i.e., we change the active sampling function *A* to a function, which returns random effectiveness scores). Hypothesis 1b can now be validated by comparing the performance of the resulting learning curve with the ones already generated for the validation of Hypothesis 1a. We can also validate 1c by showing that some learning curves actually decrease their performance over the initial (starting) point.

We validate Hypothesis 2 with experimental evidence showing that randomly selected consideration information will bias the human expert (disorderly). Correspondingly, we design CISF2 to randomly pick *i* labeled instances as consideration information *I*.

$$CISF2(M, T, U, e, i) := \{p: p \in T \wedge p = \text{random-select}(T)\},$$

such that $i = |\{p: p \in T \wedge p = \text{random-select}(T)\}|$. Assuming that Hypothesis 2 is correct then the randomly selected labeled instances will bias the human labeler randomly and the related learning curves should distribute disorderly. Hence, the average learning curve (for a sufficient number of repetitions; we will use 100) should approximate the one using CISF1. This effect is limited by the extent of the influence of consideration information. If it has no influence, then CISF1 and CISF2 will also have similar results, as the consideration information returned by CISF2 will be ignored by the expert. In this case CISF2 is not sufficient to validate Hypothesis 2.

Studies in psychology and marketing found that similar instances provide the strongest decision making bias (see section 2). Hence, we designed CISF3 to pick the *i* most similar instances to *e* from the labeled data *T*. Thus:

$$CISF3(M, T, U, e, i) := \text{knn}(T, i),$$

where knn adopts the k-nearest neighbor algorithm [1] on *T* to find the *i* nearest neighbors. Using CISF3 we get an active learning curve, where the human labeler is biased by similar instances from data (already) labeled at the current phase of the active learning procedure. If this curve shows a significantly different performance compared to the curves generated with CISF1 then we validated Hypothesis 2.

CISF3 has one drawback. If we assume that human labelers are error-prone (even with consideration information) then the set of labeled instances *T* will contain an increasing number of incorrectly labeled instances. Choosing consideration information from *T*, thus, has an increasing chance of choosing wrongly labeled consideration information and, hence, wrongly biasing the expert. We, therefore, choose our final CISF to choose consideration information from the set initially and correctly labeled instances *L* akin to CISF3. Thus:

$$CISF4(M, L, U, e, i) := \text{knn}(L, i).$$

4.1.2 *The Simulated Decision Maker Agent*

To evaluate the CIAL framework we would need a human labeler with endless patience and expert knowledge in multiple application domains within the available datasets. Given the sheer impossibility of finding such subjects we employed a simulated decision maker based on a cognitive architecture [3][15]. Specifically, we combined a rule-based and an instance-based model (*RM* and *IM*, in Definition 4) for the simulated labeling agent. When the agent labels an instance it first checks the rule-based model. When a rule-conflict arises, our agent uses the instance-based model to do the prediction.

Clearly, the agent simulates the primary characteristics of the human judgment theory [14] as it emulates both representative heuristics (with *IM*) and availability heuristics (with *RM*). It also inherits the gestalt of cognitive architectures. We initialize the agent's experience (i.e., *RM* and *IM*) with two randomly sampled subsets of the dataset. We call the sizes of these samples *esr* for experience size rules and *esi* for experience size instances.

Definition 4: The Simulated Labeling Agent

Given: a rule model RM and instance model IM

Input: an instance e to be labeled, consideration information I.

Pseudo code:

```
1. apply RM to e
2. If (matched rules produce impasse)
3.   add I to IM
4.   label e with IM
5.   remove I from IB
4. else
5.   label e with matched rules
```

Output: a label for e

4.1.3 *Implementation: Learning Algorithms, Datasets, and Parameters Used*

We implemented the entire learning framework based on Weka [27]. We chose the BOOTSTRAP-LV active sample selection method [20] as ES calculation method *A*. As *Inducer* we chose an unpruned J48 decision tree (Weka's C4.5 [18]). Obviously, any other active learning methods and inducer could be used.

For our agent design, we chose the PART rule learner [10] to generate the rule model *RM* and the knn [1] for the agent's instance model *IM*. Again, any rule or instance-based learner could be used to initialize the CIAL-agent. We also had to choose the sizes *esr* and *esi* (experience size rules and experience size instances). Given that the knowledge stored in the rule-based model is more condensed than in the instance-based model and assuming that human memory is limited we chose *esr* to be larger than *esi*. To ensure the robustness of our results we ran sensitivity analyses for all simulations. Specifically, we used 10%-90% of the benchmark dataset to create the agent's rule-based model and 50-450 instances to create agent's instance-based model. Effectively, we simulated different levels of "experience": little experience as *esr*=10% & *esi*=50; lots of experience as *esr*=90% & *esi*=450. Finally, considering the limitation of the human memory, we pruned *RM* to at most 20 rules.

All of the datasets we used were from the UCI machine learning repository [7]. Of the large number of datasets available from the repository we selected 8 datasets with which the BOOTSTRAP-LV method exhibited superior performance than random sampling [19]. We varied the learning phases between 20 and 30 steps depending on the size of each dataset. At each stage, the same numbers of instances were added to the current model *M*. Furthermore, we averaged the results over 100 partitions of the datasets into an initial labeled dataset *L*, an unlabeled dataset *U*, and a test dataset. To ensure comparability the same partitions were used by all CISFs.

The final parameter was the number of instances *i* to be chosen by the CSIF. Based on research in marketing [11], which finds that sizes in the range of 3-5 instances will give the maximum influence on the consumer's purchasing decision, we chose to run all

our simulations with $i = 3$. Correspondingly, we set the agent's instance-based model as a 3NN model.

4.2 Hypotheses Testing: Design and Results

4.2.1 *Confirming Hypothesis 1*

To validate Hypothesis 1 we compared the performance of various simulated labelers with the CISF1 (i.e., no consideration information) with the one of an "oracle" labeler, which represents the traditional setup. We varied the previous experience of the simulated labeler from "novice" (esr=10%, esi=50: "Agent10%-50") to "expert" (esr=90%, esi = 450: "Agent90%-450") in equidistant steps of 10%, respectively, 50 instances. Figure 1 plots a selection (to ensure readability) of the learning curves in terms of BMAE (best-estimate mean absolute error [20]) for each of the 8 datasets (omitting some of the agent experience settings to improve readability). As the graphs clearly show the overall learning performance with the simulated learner is much worse than with the "oracle" labeler confirming Hypothesis 1a. In 7 of the 8 datasets the "novice" Agent10%-50 even goes up signifying a learning performance that is even worse than simple supervised learning with the initially labeled dataset. For the other dataset (optidigits) Agent10%-50 shows little better performance than simple supervised learning with the initially labeled dataset confirming Hypothesis 1c. As the labeling agent's experience increases so does the system's overall active learning performance. In 3 of the 8 datasets (Sick-euthyroid, Hypothesis, Kr-vs-kp), however, even the agent with "lots of experience" ("Agent 90%-450") performs worse than the simple supervised case. These results clearly confirm Hypotheses 1a and 1c.

To confirm Hypothesis 1b we compared the performance of a novice (esr=20%, esi=100: Agent20%-100) as well as expert (esr=80%, esi=400: Agent80%-400) simulated labeler using CISF1 (i.e., no consideration information) with the one of an "oracle" labeler, which represents the traditional setup. For each labeler we tested BOOTSTRAP-LV and a random selection as the effectiveness scoring function A. Figure 2 plots the results for the car dataset (the others performed similarly but are omitted due to space constraints). As the graph clearly shows the simulated labelers' performance when using random sampling ("Agent-Random" in Figure 2) or BOOSTRAP-LV (i.e., "Agent-Active") is practically indistinguishable. Only the oracle labeler seems to profit from the active sampling method confirming Hypothesis 1b. A reason for this observation might be that the labeling errors cancel out the advantages of the active learning algorithms capability to adaptively select "useful" instances.

Summarizing, our experiment confirmed Hypothesis 1. Specifically, we showed that active learning looses its advantage over basic supervised learning (or randomly sampled additional labels) in environments with error-prone labelers. *Thus, the key for using active learning in practical environments is to improve the accuracy of the labelers.*

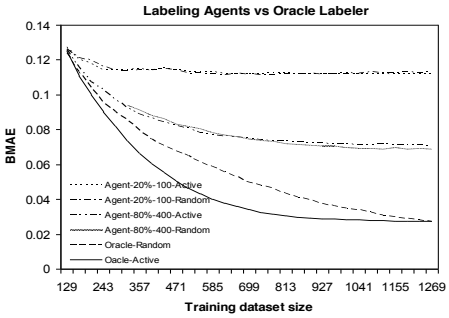


Figure 2: Comparing the Learning Performance of Active Sampling and Random Sampling with Typical Labeling Agents on the Car Dataset.

To confirm Hypotheses 2 and 3 we confronted the range of possible agents (i.e., simulated labelers with different levels of experience) with the four consideration information selection functions CISF1, CISF2, CISF3, and CISF4. To ensure the succinctness of our presentation we limit the discussion to a "novice" labeler with little experience (Agent-20%-100) and an expert labeler with a high level of experience (Agent-80%-400).

4.2.2 *Performance of a Simulated Novice Labeler*

Figure 3 plots the learning curves obtained on each of 8 datasets with simulated novice labeler Agent-20%-100, where the different consideration information selection functions are used. A first observation shows that the performance when using CISF1 and CISF2 is comparable for all datasets, which indicates that the average influence of randomly selected instances on Agent-20%-100 (a lower level knowledge agent) is negligible. On the assumption that any instance will bias simulated agents this result suggests that consideration information in the form of randomly selected instances biases the simulated agent indiscriminately at every labeling step. The repeated random influences by the CISF seem to cancel each other out leading to a performance similar to the setup with no consideration information (i.e., CISF1).

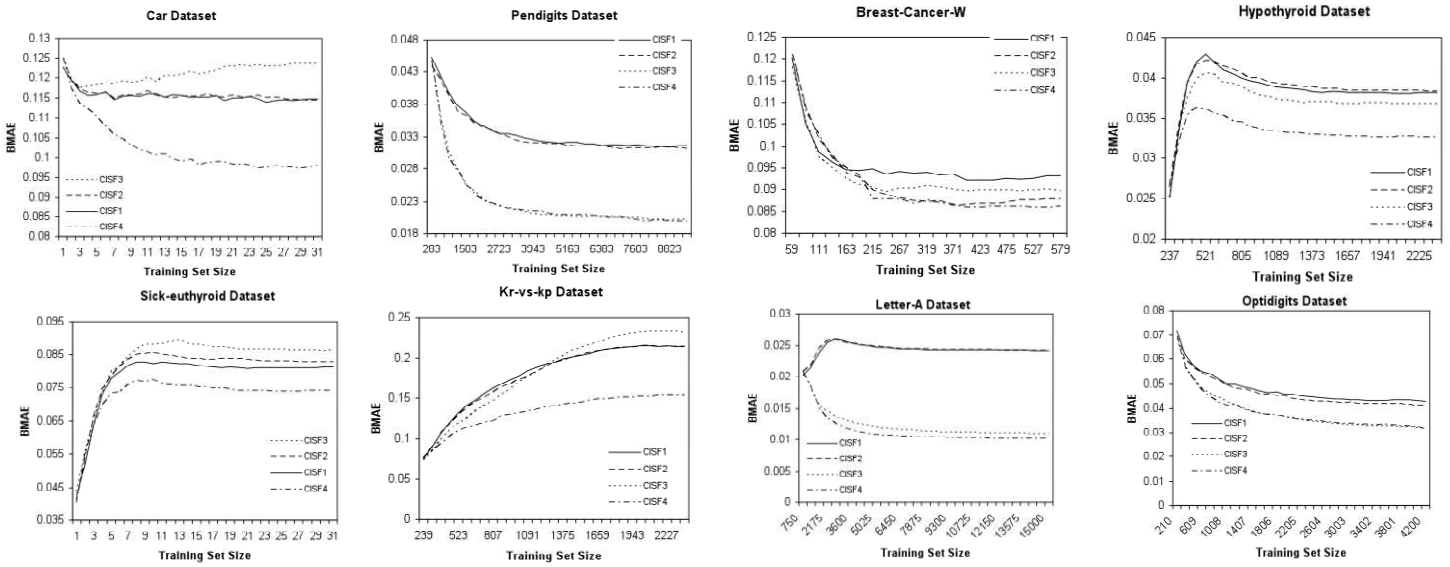


Figure 3: Influence of Consideration Information on Novice (20%-100) Labeling Agent's Performance

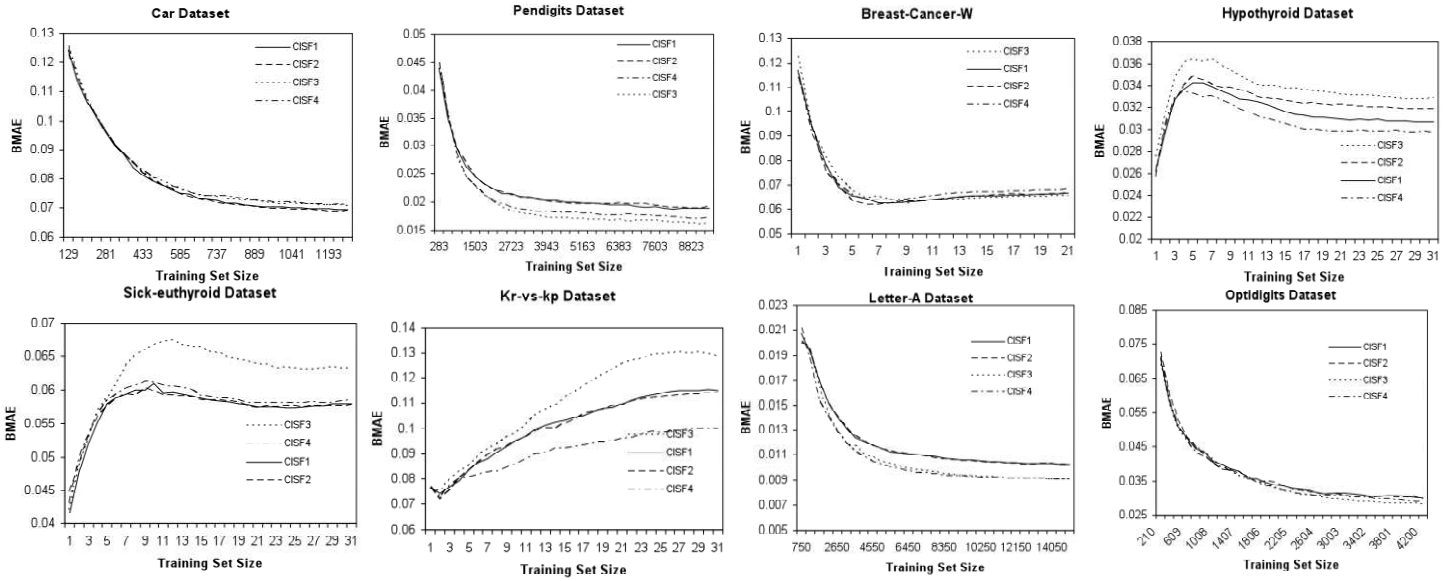


Figure 4: Influence of Consideration Information on the Expert (80%-400) Labeling Agent's Performance

Investigating the learning curves further we note that CISF3, which chooses the instances most similar to the one to be labeled, performs significantly different to the case where no consideration information is provided (CISF1). Clearly, CISF3 supplies a set of instances that strongly bias the novice labler. We can, thus, conclude that similar instances can indeed be used to bias a labeler. Combining these observations with the result of using CISF2 we can confirm our Hypothesis 2 for all 8 datasets and the novice labeling simulation. Furthermore, the experiment also shows that the influence exerted by CISF3 is unreliable as the direction of influence differs among. For example, the car dataset learning curve clearly shows that CISF3 initially seems to provide the correctly labeled instances, but then, probably due to dominantly choosing wrongly labeled instances as consideration information by CISF3, deteriorates to ultimately end up close to the initial (i.e., supervised learning) performance. For the pendigits dataset, however, where CISF3 dominantly chose correctly labeled instances as the consideration information, the learning performance is significantly better than when using CISF1. The data also shows that the probability of choosing a wrongly labeled instance is small at the onset and only rises as the active learning system progresses through its stages. In the car, Sick-euthyroid, and Kr-vs-kp datasets the performance of CISF3 initially starts out to be at least as good as the one of CISF1 but eventually, presumably due to the choice of wrongly labeled instances, deteriorates to levels, which are worse than the one of CISF1.

Addressing this issue CISF4 picks its consideration information instances only from the initially labeled dataset, which are presumed to be correct. As the graphs show CISF4 consistently seems to bias novice labeler correctly, resulting in an overall better performance than with the other consideration information selection functions. Note that this superior performance is

achieved even though the pool of instances from which the consideration information is chosen is much smaller than with CISF3. The experiments for the novice labeler seem to reconfirm our Hypothesis 3a for our datasets and a novice labeler. CISF4 is not optimal. In particular, we find that CISF4 only confirms Hypothesis 3b for 4 of the 7 datasets, which is insufficient as a validation. Nevertheless, CISF4 seems to approach a desirable consideration information selection function, as it confirms Hypothesis 3b for some of the datasets.

4.2.3 *Performance of a Simulated Expert Labeler*

The learning results for the simulated expert labeler Agent-80%-400 show similar tendencies as the results of the novice labeler (see Figure 4), also confirming our hypotheses. However, the results also show some significant differences. First, the learning curves of CISF1 are considerably better for the simulated expert labeler (Figure 4) than the novice agent (Figure 3) reconfirming our assumption on the influence of the labeler's experience on the overall learning performance. Second, the influence of using CISF3 or CISF4 is not as obvious as in Figure 3. Extreme examples are the Car and Optidigits datasets, where CISF3 or CISF4 have a strong influence on the simulated novice labeler's (and the overall algorithm's) performance. With the simulated expert labeler, however, the influence is practically impossible to discern. This result supports people's everyday experience that experts are more difficult to bias than novices, as their experience provides them with a larger pool of examples to ground their decisions. Furthermore, as clearly shown in the Hypothyroid, Sick-euthyroid, and Kr-vs-kp datasets, we find that the positive influence provided by CISF4 over CISF1is only marginal while the negative influence of CISF3 is quite clear. We believe that this is due to the limited pool of information (i.e., the set of initially labeled instances), which CISF4 can explore. Hence, as the experience of the agent increases the contribution of the information in the initially labeled data decreases.

4.2.4 *Summary of the Experiments*

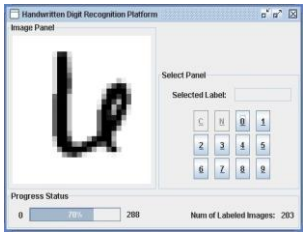
From the above simulation results with both the simulated novice and expert labeler we find that similar instances can indeed be used as consideration information to bias the labeler. We also demonstrated that randomly selected instances bias the agent aimlessly. Most importantly, we showed that using correctly labeled similar instances (e.g., from the initial dataset) does bias (simulated) labelers in the correct direction. The influence of the bias is, however, more dominant with novice labelers than with experienced labelers. Consequently, we can confirm Hypothesis 3a for the 8 datasets and our labeling simulation. Unfortunately, we cannot confirm Hypothesis 3b, as the performance for three out of the eight datasets is actually worse than the purely supervised case. We believe that this is due to deficiencies in our agent design (see also discussion below) but have to leave this question open to further investigation.

4.3 Experiments Showing the Congruence of the Human and Simulated Labeler

The biggest limitation of the experiments so far is the simulated labeler. The human decision making process is arguably too complex to be conclusively simulated by a computer-based agent limiting the generalizability of our findings. Specifically, our labeling agent design incorporated many design decisions. As an example, consider the choice of instances to populate the agent's experience (i.e., instance-based model): we chose a random set of instances, where humans adaptively chose the most "useful" instances (supposedly, according to some evolutionary criteria).

To reconfirm that an automated procedure could bias humans to correct labeling we ran a second experiment with human experts. Specifically, we designed a handwritten digit recognition experiment where 7 human subjects were asked to recognize images of handwritten digits. We choose the MNIST digit handwritten image dataset, which is a typical digit recognition benchmark in the machine learning and pattern recognition community with 70'000 images of handwritten digits with their correct labels (i.e., their correct number). Since we couldn't confront our subjects with 70'000 images we decided to reduce the overall set of examples. To that end we preprocessed all the image files and reduced the number of image features from 728 to 50 using a principal component analysis (pca) [13]. We then trained a support vector machine (svm) [8] and only considered pictures where the support vector machine failed to recognize for our test. Using this procedure we identified 288 images that were very difficult to classify for the support vector machine.

a)



b)

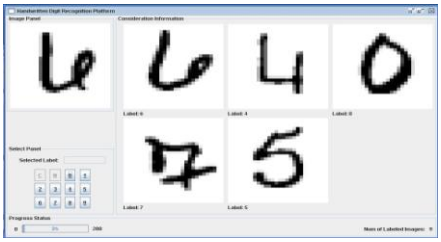


Figure 5: Human Labeler Experiment User Interface

7 randomly selected human subjects were then asked to classify the 288 digits using the user interface shown in Figure 5a) without any consideration information (i.e., according to CISF0). We then slightly adapted CISF4 as this task wasn't a binary classification task, such as all the experiments with the automated labeler as follows: rather than choosing the k (in the experiments 5) nearest neighbors from the whole dataset we chose the nearest neighbor from each of the 10 classes (numbers 0, .., 9) and then discard the 5 most dissimilar ones. As a measure of similarity we used the Euclidian distance on 10 features of the images determined using principal component analysis from the 728 original ones. 5 different randomly selected subjects classified the 288 digits using the interface shown in Figure 5b). Table 1 shows the error count and error rate for the subjects determining the labels of the images without consideration information (x1, …, x7) and with the consideration information (y1, …, y5). A t-test adjusted for small numbers of samples confirms that relative error values for subjects with consideration information are drawn from a sample with a

significantly smaller mean than the others with p= 0.0015[1] confirming that we have successfully biased our subjects into correct labeling.

Table 2: Error Rate and Error Count with and without without consideration information

no consideration information	Subject	x_1	x_2	x_3	x_4	x_5	x_6	x_7
	Error Count	58	89	67	58	80	85	84
	Error Rate	0.2001	0.3090	0.2326	0.2014	0.2778	0.2951	0.2917

with consideration information	Subject	y_1	y_2	y_3	y_4	y_5
	Error Count	50	56	46	49	57
	Error Rate	0.1736	0.1944	0.1597	0.1701	0.1979

5. Discussion and Limitations

The experiments showed conclusively that active-learning research should take the error-proneness of human labelers into account. Not surprising, we showed that an error-prone labeler is likely to have a worse performance than an oracle labeler. Furthermore, in most datasets we even found that an error-prone labeler performs worse than the simple supervised learning setting with the initially labeled data casting doubt on the usefulness of active learning in practical settings with highly error-prone labelers. We also showed that by varying the degree of precision of a labeler the usefulness of active learning varies with the accuracy of the labeler. Interchangeably employing a random or active effectiveness score calculation approach we also found that the active learning approach did not perform better than a random one when confronted with an error-prone labeler. Consequently, given the findings in psychology about the inaccuracies of human decision-making and assuming a high cost of labeling it seems that helping *labelers to make less labeling mistakes is the key for the practical use of active learning in settings with human labelers*. Note that these findings are independent of our chosen labeling agent simulation approach, as it only relies on the error-proneness of the labeling agent but not on any specific feature of their decision making. In fact, early in our experimentation we found similar results when using labeling agents, which introduced random error (compared to an oracle labeler) at different error-rates. The findings are, however, limited by the number of datasets employed.

Introducing the notion of consideration information to influence a labeler the experiments also showed that an active-learning procedure can indeed influence a simulated labeler – in some cases even in the desired direction. We found that *a consideration information active learning procedure with an appropriate consideration information selection function outperformed the traditional active learning procedure when an error-prone labeler was involved*. Addressing the limitation of the simulated labeler, we ran a second experiment to show the congruence of human labelers with the simulation. Specifically, we showed in a digit recognition task that an automated consideration information selection function is capable of choosing instances that can indeed significantly bias human labelers into better labeling. Therefore, it is feasible to assume that the results with the simulated labeler will generalize to the human labeler setting – an experiment we still intend to undertake.

[1] Since it is difficult to ascertain normality with such a small number of examples we also ran the Man-Whitney test which was also highly significant with p=0.0013

One underlying assumption of our work, and, hence, a limitation of the findings is that we assume the correctness of the initial labeled dataset. While this assumption underlies most active learning studies it may not be true in practical applications. Issues of data-quality might introduce errors into the data. Furthermore, even the initially labeled instances might have been labeled by a human labeler, who again is likely to introduce mistakes. As we have seen in the experimental evaluation when comparing CISF3 with CISF4, basing the consideration information on wrongly labeled initial information can substantially mislead the (human) labeler and in turn deteriorate the overall learning performance. As a consequence we have to answer the following questions: What is the influence of incorrect consideration information on the labeler? How much should labelers trust consideration information? Perhaps every labeled instance should be associated with its lineage and a trustworthiness ratio. We believe, however, that this problem goes beyond the scope of this investigation and should be subject of future work.

Our second experiment illustrates the usefulness of consideration information. Not only did the subjects perform better in the task they often expressed that the consideration information provided them with the context that was needed to accomplish the task appropriately. Consider, for example, the digit shown in Figure 5a. Taken by itself it can be seen as sloppily written number “4”. The context of the consideration information does, however, suggest that it is more likely to be a number “6”. Hence, as one subject expressed at the end of the test, the consideration information did provide some automatically inferred context reminding them that a “4” can be written differently elsewhere.

The second experiment has two shortcomings, however. First, the number of subjects is very small. While we took this into account in the statistical analysis the small number aggravates the standard experimental problem of whether the subject pool adequately represents typical experts in labeling tasks. While this hampers the generalizability of our finding we believe that evidence from psychology and marketing strongly supports our finding. Second, the task chosen might be seen as inadequate. We agree that when people talk of experts they typically envision heavily trained professionals such as physicians or engineers and not the general population with their exceptional capability to recognize handwritten information. The task, though, exhibits exactly the features we were searching for: a situation where people are usually adequate and machines sometimes fail miserably. But we agree that it would be more satisfactory to choose a task one associates with experts – an endeavor we intend to undertake in the future.

6. Conclusions and Implications

Research in Active Learning will need to confront the practicalities and inaccuracies in human expert decision-making. The psychology literature has a long tradition of investigating those questions. When researching the active learning literature for this study, however, we only found one study, where human labelers answered an algorithm’s questions on a hand-written digit recognition task [6]. The experiment showed that the human labeler’s decisions are likely to be error-prone and, therefore, detrimental to the overall performance of the investigated query-learning system.

Using insights from psychology, we propose to use a more interactive learning procedure to mitigate the inaccuracies of human labelers with consideration information to positively bias the hu-

man labeler. Our experiments show first that helping error-prone labelers to make fewer mistakes is the key for the practical use of active learning as the procedure may otherwise worsen the overall prediction quality below the baseline supervised case. Second, we showed that a consideration information active learning procedure with an appropriate consideration information selection function outperformed the traditional active learning procedure when an error-prone labeler was involved. In particular, we found that both our simulated labelers as well as human subjects were positively biased by (correctly) labeled instances most similar to the one they had to label.

As we noted, the limitations of our evaluation warrants further experiments with human labelers in a real task. Nonetheless, we believe that using consideration information to bias labelers is a fruitful avenue towards mitigating the effects of error-prone labelers – a goal that will increase the practicality of many data mining approaches.

7. ACKNOWLEDGMENTS

We would like to thank Rebecca Hamilton and Haym Hirsh for some input on our initial ideas. We would also like to thank the people at the UCI repository for providing the datasets and Maytal Saar-Tschechansky for providing BOOSTRAP-LV.

8. REFERENCES

[1] Aha, D., Kibler, D., and Albert, M. "Instance-based learning algorithms," Machine Learning (6:1) 1991, pp 37-66.

[2] Alba, W.J., and Hutchinson, J.W. "Effects of Context and Post-Category on Recall of Competing Brands," Journal of Consumer Research (13:3) 1987, pp 411-454.

[3] Anderson, J.R. The Architecture of Cognition Harvard University Press, Cambridge, MA, 1983.

[4] Angluin, D. "Queries and concept learning.," Machine Learning (2) 1988, pp 319-342.

[5] Baron, J. Thinking AND Deciding Cambridge University Press, 1998.

[6] Baum, E.B., and Lang, K. "Query learning can work poorly when a human oracle is used," International Joint Conference in Neural Networks (IJCNN'92), Beijing, China, 1992.

[7] Blake, C., and Merz, C.J. "UCI repository of machine learning databases," Department of Information and Computer Science, 1998 [http://www.ics.uci.edu/~mlean/MLRepository.html]. Irvine, CA: University of California, 1998.

[8] Boser, B., Guyon, I., and Vapnik, V. “A training algorithm for optimal margin classifiers.” Proceedings of the Fifth Annual Workshop on Computational Learning, 1992.

[9] Dawid, A. P., and Skene, A. M. “Maximum likelihood estimation of observer error-rates using the EM algorithm.” Applied Statistics 28, 1 (Sept. 1979), pp 20-28.

[10] Frank, E., and Witten, I.H. "Generating Accurate Rule Sets Without Global Optimization," The Fifteenth International Conference in Machine Learning (ICML), Morgan Kaufmann Publishers, Madison, WI, 1998, pp. 144-151.

[11] Hamilton, R.W. "Why Do People Suggest What They Do Not Want? Using Context Effects to Influence Others' Choices," Journal of Consumer Research (29) 2003, pp 492-506.

[12] Hauser, J.R., and Wernerfelt, B. "An evaluation Cost Model of Consideration Sets," Journal of Consumer Research (16:4) 1990, pp 393-408.

[13] Jolliffe, I.T. Principal Component Analysis Springer; 2nd edition, 2002.

[14] Kahneman, D., Slovic, P., and Tversky, A. Judgment under uncertainty: Heuristics and biases Cambridge University Press, 1982.

[15] Laird, J.E., Newell, A., and Rosenbloom, P.S. "SOAR: An architecture for general intelligence," Artificial Intelligence (33:1) 1987, pp 1-64.

[16] Lugosi, G. “Learning with an unreliable teacher.” Pattern Recognition 25, 1 (Jan. 1992), pp 79-87.

[17] Miller, A.G. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing In formation," The Psychological Review (63) 1956, pp 81-97.

[18] Quinlan, J.R. C4.5: Programs for machine learning Morgan Kaufman, San Mateo, 1993.

[19] Roberts, J.H., and James, L.L. "Consideration: Review of Research and Prospects for Future Insights," Jounral of Marketing research (34:3) 1997, pp 406-410.

[20] Saar-Tsechansky, M., and Provost, F. "Active Sampling for Class Probability Estimation and Ranking," Machine Learning (54:2) 2004, pp 153-178.

[21] Sheng, S., Provost, F., and Ipeirotis, P. “Get another label? Improving data quality and data mining using multiple, noisy labelers.” Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 2008, pp 614-622.

[22] Silverman, B. W. “Some asymptotic properties of the probabilistic teacher.” IEEE Transactions on Information Theory 26, 2 (Mar 1980), pp 246-249.

[23] Smyth, P. “Learning with probabilistic supervision.” In Computational Learning Theory and Natural Learning Systems. Vol. III: Selecting Good Models, T. Petsche, Ed. MIT Press, Apr. 1995.

[24] Smyth, P. “Bounds on the mean classification error rate of multiple experts.” Pattern Recognition Letters 17, 12 (May 1996)

[25] Smyth, P., Burl, M.C, Fayyad, U. M., and Perona, P. “Knowledge discovery in large image databases: Dealing with uncertainties in ground truth.” In Knowledge Discovery in Databases: Papers from the 1994 AAAI Workshop (KDD-94) 1994, pp 109-120.

[26] Smyth, P., Burl, M.C, Fayyad, U. M., and Perona, P. “Inferring ground truth from subjective labeling of Venus images.” In NIPS 1994, pp 1085-1092.

[27] Witten, I., and Frank, E. Data Mining: Practical machine learning tools with Java implementations Morgan Kaufmann, San Francisco, 2000.

[28] Wright, P. "Consumer Choice Strategies: Simplifying vs. Optimizing," Journal of Marketing Research (12:2) 1975, pp 60-67.