The Other Human in The Loop – A Pilot Study to Find Selection Strategies for Active Learning

Daniel Kottke*, Adrian Calma*, Denis Huseljic*, Christoph Sandrock*, George Kachergis†, Bernhard Sick*
*Intelligent Embedded Systems, University of Kassel, Germany
Email: {daniel.kottke,adrian.calma,bsick}@uni-kassel.de, {dhuseljic,christoph.sandrock}@student.uni-kassel.de
†Department of Artificial Intelligence, Radboud University, Netherlands
Email: g.kachergis@donders.ru.nl

***Abstract*—Gathering data becomes increasingly simple whereas the labeling of collected instances remains difficult. Active learning provides methods to reduce the labeling effort by intelligent selection of instances. In contrast to building mathematical models or developing heuristics to solve this task, we pursue another approach: We let humans select the instances which should be labeled. Participants are asked to learn to predict the sex of 18 abstract illustrations of bugs as either male or female. This article describes the design, goal and the execution of this study with 14 groups (71 participants). In this exploratory study we analyze humans' balance between exploration and exploitation, the participants' learning behavior, the collaboration within the group as well as the question when to stop querying. The comparison of human performance with baseline active learning algorithms provides promising results which indicate that machine active learning might benefit from incorporating human strategies. Additionally, we provide the complete data and extracted spreadsheets for download.**

I. Introduction

Active learning (AL) is a machine learning paradigm which provides algorithms for mining data when little information is available. In classification tasks, this means that labels for (almost) all instances from the data pool are missing. The selection strategy of an AL algorithm selects the most useful instances from the (candidate) pool to be labeled by one or more experts or non-experts, who are generally referred to as *oracles* [1]. These labeled instances build the training set for the classifier which is retrained subsequently. This repeated procedure is called the *active learning cycle* (see Fig. 1). Using these label acquisitions, the selection strategy aims to increase the performance of the classifier as fast as possible close to the optimal value.

Especially industrial companies show a growing interest in AL as the acquisition of unlabeled databases is relatively easy and cheap. On the contrary, annotating data with high quality labels demands a substantial financial investment. Frequently, oracles are impersonated by human experts. For example, in a quality assessment system, sensors (e.g., pressure, temperature, or photo sensors) provide features for each product, but the quality (i.e., the label) of the outcome can only be determined by a human expert using visual inspection.

In this article, we change the role of the human from the labeling to the selection expert. Hence, this other human *impersonates the selection strategy* not the oracle. Thereby,

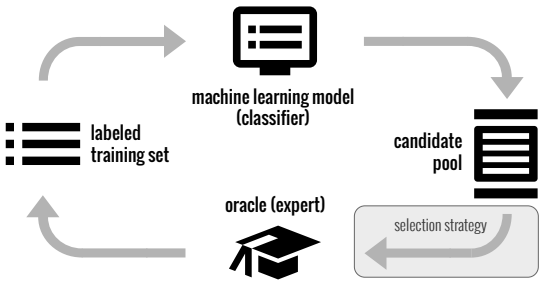

Fig. 1. The conventional active learning cycle [1], [2].

we want to *investigate the ability of humans to create and implement a strategy* for selecting the most beneficial instances. Therefore, we developed a simple game where humans are asked to separate female from male bugs based on color, dots, and eyes. The challenge is to correctly determine the sex of all 18 bugs by asking the game master to provide the true sex for of the bugs while minimizing the number of queries. The long-term goal of this study is to assess if humans' creative way of finding strategies maps to and possibly even improves the mathematical models of existing active learning algorithms.

The contributions of our work are the following:

- We developed and conducted a pilot study to investigate how humans solve an active learning task.
- We provide a detailed description of our novel experiments including the complete dataset as well as all material to reconstruct the exact setup.
- We evaluated the data and show that humans might be effective in the process of selecting the most useful instances for AL, not only for labeling data. The raw data is available as download.
- We outline one option to conduct a follow-up experiment stating possible hypotheses and challenges.

The remainder of this article is structured as follows: First, we start with a brief overview of relevant literature followed by a detailed description of our experimental setup. The process of extracting the information from the study is summarized in Sec. IV and followed by its evaluation. In Sec. VI, we describe how a future study might look like to evaluate our findings more exhaustively. Finally, we conclude our work and give a brief outlook.

II. Related Work

Active learning (AL) is a machine learning paradigm which aims to successively select information to improve the performance of a classifier [1]. This process can be visualized in the AL cycle (see Fig. 1): We start with an empty set of labeled instances and a large set of unlabeled ones. Iteratively, a selection strategy ranks the usefulness of the labeling candidates from the unlabeled set. The label of the most useful instance (or more instances in batch acquisition) is passed to the oracle to be labeled and added to the training set. This oracle usually is a human expert but can also be a simulation software or some expensive test in a lab [3]. In each iteration, the classifier is retrained based on the most recent information about the data.

One of the most used techniques is uncertainty sampling [4] which aims at requesting labels for instances regarding which the classifier is most uncertain about (e.g., instances near the classifier's decision boundary). A decision theoretic approach is expected error reduction [5] where each possible label outcome of each label candidate is simulated and the generalization error of the classifier is estimated each time. Finally, the label of the instance which reduces the expected error the most is acquired. The downside of expected error reduction is the execution time. To overcome the intensive computation, probabilistic active learning [6] was introduced which is a decision-theoretic approach modeling the uncertainty of a classifier using the true posterior probability. These aforementioned approaches use mathematical models and heuristics to determine the usefulness of labeling candidates assuming that human experts provide labels. Hence, a variety of studies have investigated people's annotation performance [7], [8], [9].

In contrast to this well-known scenario, we aim to investigate whether humans might improve the selection of promising instances. This problem has recently been addressed in the field of visual analytics: Weigl et al. [10] propose MapView, a graphical data representation tool for active learning. With their approach, the authors showed that the user is able to select labeling candidates intelligently. The study by Bernard et al. [11] provides more details on the selection performance and compares it to artificial AL algorithms. Nevertheless, the focus of both studies is on visualization and representation of data rather than on humans' active selection strategies.

Research in cognitive psychology has shown that people who are allowed to self-select training instances outperform those given randomly-selected instances in a variety of learning scenarios, including simple classification problems [12], learning word-object mappings [13], and learning causal relationships [14]. Zhu et al. [15] demonstrated that people can learn and refine a two-class decision boundary in a semi-supervised setting: when given unlabeled instances in addition to labeled instances, participants learned a better decision boundary than when given only labeled instances. Using the same 3D visual stimuli and two-class learning problem as in [15], Castro et al. [16] investigated whether people learn best from labeled 1) randomly-sampled instances, 2) self-selected instances, or 3) passively-observed instances selected by an active learning algorithm. Under low levels of label noise, active learners (i.e., who selected each next instance to see labeled) converged more quickly on the decision boundary than those who passively observed randomly-selected labeled samples. However, human active learners were slower than the theoretical exponential convergence. Under high levels of label noise, people receiving machine-selected instances outperformed human active learners, perhaps because they learned to "trust the machine" [16].

More recent work has attempted to understand what selection strategies people choose to employ in such active category learning tasks and to what degree the advantage of human-selected instances can be conferred on others who passively receive the selected instances of an active learner. Markant et al. [17] investigated the difference in classification performance after people receive samples either actively selected by themselves or received from another active learner. They found that active participants, receiving the samples they chose, outperformed participants who passively received the samples chosen by those same active participants.

Markant et al. [18] examined how people choose which information to sample in an active category learning task, finding that instead of sampling instances with high uncertainty across all possibilities (i.e., label entropy), they are biased to select items that reduce uncertainty between two alternatives. That is, they favor reducing local uncertainty, for example picking instances that differ only on a single feature (i.e., margin sampling), rather than choosing instances that are expected to most reduce global uncertainty, which often vary on multiple dimensions. This preference may reflect cognitive constraints involved in evaluating which of all possible instances is most informative to sample next, even to the extent that humans may be best at considering a single feature dimension at a time. Moreover, this "control of variables" strategy is similar to that used in scientific reasoning [19]. While most active category research has involved classes distinguished by one or two features with a linear decision boundary, other research has considered how efficiently people are able to search a novel hypothesis space with several partially-overlapping features.

Kachergis et al. [20] investigated whether school-aged children could learn to search through a novel stimulus space with 16 exemplars and 11 binary features in a tablet-based game similar to 20 questions or "Guess Who?". Generally, such information search tasks offer two types of queries: (a) one can *scan* particular hypotheses (e.g., "Is it instance A?"), or (b) one can ask a *constraint-seeking* question concerning a particular feature (e.g., "Does it have feature X?"). As features might be relevant to multiple remaining hypotheses, they can often more efficiently narrow the hypothesis space, at best allowing binary search through the space. However, using feature queries optimally requires evaluating the relative informativeness of each feature, with re-evaluation after each successive query. Children chose effective but often suboptimal feature queries, and made some errors in updating the possible remaining hypotheses.

III. Pilot Study

A. Goal

The goal of our study is to find out how humans solve the active learning (AL) task. The long-term goal is to gain new insights for our active machine learning research by recognizing relevant influential factors to improve the mathematical model of a selection strategy. In this particular study, one of the main aims was to create a game which is easily understandable for anyone, in order to enable unbiased strategic thinking without knowledge in machine learning and algorithmic thinking. The experiment should be designed for both small groups and single subjects to be able to investigate collaborative problem-solving as this becomes more relevant in machine AL.

B. General Idea

The general idea is to sort pictures into two groups by finding out distinctive features of each group. Motivated by [17], we choose to let the subject learn how to discriminate bugs by their sex. To solve this task, the subject is able to ask for the correct sex (label) of a bug to identify classification hypotheses. Also, we implement an incentive to save queries.

C. Materials / Stimuli

We created 18 paper cards[1] with abstract illustrations of bugs (six examples are given in Fig. 2). The bugs differ in eye color, number of dots, color of dots, size of dots, and location of dots. Table I summarizes each of these features including the number of feature values and the information if this feature is relevant for classification.

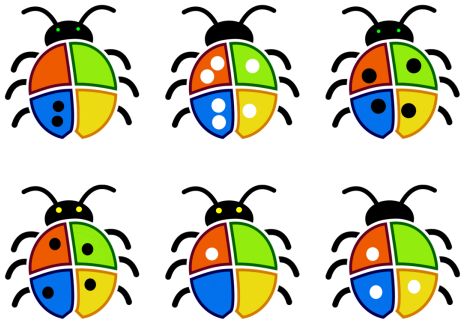

Fig. 2. Six exemplary illustrations of the bugs (upper row represents male bugs, lower row represents female bugs).

The classification rule states as follows: *Bugs with green eye color are male, bugs with yellow eye color are female. To classify bugs without eyes, the decisive feature is the number of dots: Four or more dots correspond to male bugs, three or less dots correspond to female ones.* The color, size and location of dots is irrelevant for discriminating the sex. One major point when designing the bugs was that it is clear for everyone that prior knowledge from biology is not relevant for classification. Hence, every subject was in the same situation and the success of solving the task is solely dependent on the subjects' creativity and strategic thinking.

[1]Download at www.ies.uni-kassel.de/p/bug-study#papercards

TABLE I
Characteristic features of bugs, their possible values, and whether they are relevant for correctly solving the task.

Feature	Feature Values	Relevance
Eye color	green, yellow, missing	yes
Number of dots	1–7	yes
Color of dots	black, white	no
Size of dots	small, medium, big	no
Location of dots	red, green, yellow, blue	no

D. Description of Participants

The experiment has been conducted with student groups of different size as part of a freshman game at the beginning of the first semester at the University of Kassel. There were 14 groups containing a total of 71 participants. Each group contained three to six students (more male than female) which have probably never heard of classification and active learning. The majority of them was aged between 18-20 years and did not know each other well. Each participant agreed to be filmed anonymously (bird's-eye view only capturing the playground, no faces) and that the collected data can be used for research purposes.

E. Executing the Experiment

In order to have an unbiased active learning experiment, we solely inform the participants in this study (subjects) about possible labels (male and female) and the general goal of the game but neither about selection strategies nor about classification in machine learning. During preparation of the experiment, the 18 cards have been randomly spread across the playground and equally many candies have been prepared as some sort of currency.

A schematic setup is shown in Fig. 3 with 8 bugs and 8 candies: The left and right peripheral regions of the playground have been marked with the problem classes (male, resp. female). The game master ensured that everyone understood the task prior to starting the game. Each group had 5 minutes until each bug had to be classified. During this period, each group was allowed to discuss, to rearrange the cards, and to ask for the correct sex of a bug (one after the other). The game masters, impersonated by student assistants, took over the role of the oracle telling the requested sexes of the bugs. Note that each label request had a cost of one candy. Additionally, the participants have been informed that they will loose two candies for every wrongly classified bug after the 5 minutes had passed. Hence, each group faced the trade-off between saving candies during game time and loosing more candies due to lack of information in the end.

The three participants depicted in Fig. 3 wanted to know the sex of one red and one green bug. Hence, they got the information of their sex (male and female) for the cost of two candies. With that information, the group decided to temporarily classify another bug as female whereas the other bugs remain unclassified. Next, they would probably ask for another label until they are able to classify the unlabeled bugs

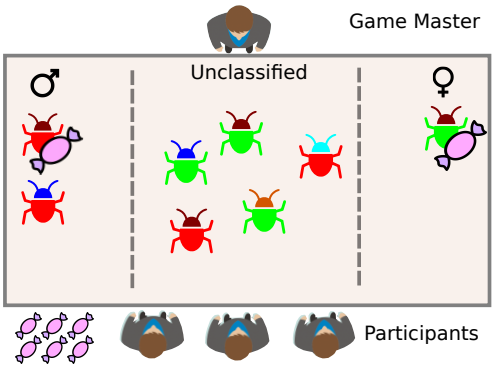

Fig. 3. Example state of our study: The three participants queried labels for two bugs (one male and one female) and payed two candies. One bug (the red one) has been classified as male based on the subjects' own inferred rule; thus they did not have to pay one candy.

with some certainty. Finally, the game master will evaluate the hypotheses and collects two candies for every wrong decision. The remaining candies serve as the reward for the group.

F. Optimization function

As mentioned above, each group has a budget of 18 candies, which is used to query the true sex of a bug for one candy (the number of queries is denoted as q). Each misclassification err induces costs of two candies. The goal (r^*) is to save as much candies as possible as the reward r.

$$r^* = \max(r) \qquad r = 18 - q - 2 \cdot err$$

IV. Extracting Information from the Study

The whole study has been captured by video which built the basis for extracting data to analyze the results. Therefore, we created two spreadsheets[2] for every group that participated in our experiment. This includes general information like the group size and the number of active participants. Additionally, we collected information (at a rate of every second) about each query (including the bug identifier, the group's purpose and what triggered this query), actions on the playground like reordering the cards, the identification of specific features of the bugs and the current classification hypotheses for each bug (unknown, male, and female). An overview of the additional features is given in Table II. The specific features are described in greater detail in the following sections.

A. General Information about the Groups

The general information describes the basic properties of each group, i.e., the group size and the number of actively participating members of the group. As the videos have been anonymized, it was difficult to assign actions to single members. Active participation in this context is defined by contributing to the result by gestures or talking. In general, we noticed that, mostly, one or two members of a group have been active, whereby the other active members only participated partially. If a participant barely spoke, we did not count him or her as an active person.

[2]Download at www.ies.uni-kassel.de/p/bug-study#spreadsheet

TABLE II
Features of the Spreadsheet

Feature	Feature values
Time in seconds	1–300
Queried bug	1–18
Query purpose	Exploration, Exploitation, Validation
Query trigger	Eyes, dot color, dot distribution
Detected feature	Eyes, dot color, dot distribution
Bugs sorted by	Eyes, dot color, dot distribution
Hypothesis bug 1,...,18	Male, female or undecided

B. Querying the Game Master

Except for the general information, all entries have been determined w.r.t. *time* at a rate of one second. As the maximal time limit was 5 minutes, these values vary between 1 and 300. If a group finished earlier, the upcoming time values have been omitted. Each bug has a unique identifier (see backside of the card) between 1 and 18. The *queried bug*-column contains the corresponding bug at to the specific time when the sex of the bug was queried from the game master.

In addition to the bug identifier, we determined information about the group's *query purpose*. The value of a cell is either exploration, exploitation, or validation. In the sense of AL research, we define validation as an action to validate a specific classification hypothesis [1]. The term exploitation is used when a group has a broad idea how a male and female bug may look like but the queried bug has attributes of both classes [1]. The following situation constitutes an example of an exploitative query: The participants have already queried a male bug with 7 dots and a female bug with one dot. Thus, they know that the number of dots play an important role, but they do not know the correct switching point. A query is marked as exploration if the group is not able to find information about the queried bug in the data. Especially the first request, therefore is of exploratory nature as no data is available.

When studying the videos, we noticed that lots of queries have been triggered by specific features. This information has been captured in the field *query trigger*. Here, we distinguished the eye color, the dot color, and the distribution of dots. The latter summarizes the feature number of dots and location of dots, as we later noticed that an alternative decision rule arose: the location of dots. If eyes were missing and each field (red, green, yellow, and blue) contained at least one dot, the bug was male. This is equivalent to the original hypothesis: If the eyes were missing and the number of dots is greater or equal 4 the classify as *male*. Sometimes, it was unclear which feature triggered the query as it was not named. In that case, we tried to derive the value from the context, like gestures and the recent discussion.

C. Detecting Features and Resorting the Bugs

Sometimes, groups sorted the cards by some detected feature in order to improve their current classification hypothesis

or to detect characteristics which might lead to a better selection strategy. The column *bugs sorted by* contains information about this particular event each time the group mentions a particular feature to be relevant for classification, rearranges the cards by a specific feature, or speaks of its relevance. In some cases, this leads to a change in their classification hypotheses, whereas, in other cases, it is solely used to cluster the cards in order to find the best bug to query next.

As the group is faced with a full series of data processing actions, the first step was to notice the differences between the bugs. Each time, a group identifies a relevant feature, it is added to the column *detected feature*. Sometimes, a group does not name this feature explicitly. In this case, we registered it to an estimated time which is the first usage of that feature in classification or querying.

D. Classification Hypothesis

For each bug, we collected data on the group's current classification hypothesis which might be m(ale), f(emale) or unknown (in the beginning). The columns *hypothesis bug* x ($x \in \{1, \ldots, 18\}$) show the group's classification hypotheses for every point in time. First, all queried bugs are added to the corresponding cells, as the group knows the true label. Moreover, when a group member predicts the sex of a bug out loudly or sorts this bug to the left, resp. right of the table, we updated the classification hypothesis for the corresponding bug. These predictions vary with the amount of information. However, a classification hypothesis proposed by one group member is not recorded if the group rejects or ignores it.

Table III summarizes information of all 14 groups that participated in our study: group identifier, total number of participants, number of active participants, number of queries, number of correct and wrong classifications in the final status, and reward in terms of candies that are left.

V. Evaluation

The goal of this evaluation is to investigate how humans perform as selection strategy in the active learning task. As a first step, we aim to describe human behavior to see if this type of experiment is appropriate to detect human-inspired strategies which will improve active learning algorithms. We are aware that a study with only 14 groups and 71 participants might not be sufficient to show significant results. Still, as suggested by Nuzzo [21] in the article about scientific methods, we aim to perform a two-stage analysis: This pilot study serves as an exploratory study to fix the experiment design and research questions for the confirmatory experiment, which shows the importance of our pilot study. Nevertheless, the results and insights of this study should be made public to encourage interdisciplinary research in the field of active machine and human learning as performed by psychologists.

In average, 7.5 out of 18 bugs (3.86 male and 3.64 female) have been queried and 1.5 bugs have been wrongly classified. The best groups (ID 8, 11) received a reward of 11 candies by classifying all bugs correctly with only 7 queries. The record of unofficially tested data science professionals is a reward of 13 candies with only 5 requests and no misclassification. Note that the best result would be to classify all bugs correctly by chance without any request (probability of 0.0004%). The average time is 288.36 seconds which shows that almost every group used the full time of 5 minutes.

The following sections describe our findings and analyze the results of the pilot study in terms of human learning performance, machine learning performance, and stopping criteria.

A. First Exploration, Then Exploitation / Validation

One of the major challenges of AL is balancing exploration and exploitation [1]. The typical assumption is that one has to start with exploration due to lack of information and later refine the decision boundary by exploitation. Fig. 4 shows the relative frequency of the groups' query purpose (exploration, exploitation, validation) w.r.t. the query iteration. Note that the number of queries decreases after 5 iterations as some groups stopped querying. Obviously, the first query was completely explorative as no information was available. Over time, we see an increase of exploitation and validation. Normally, validation is not explicitly mentioned in AL algorithms. The results show that the participants of this study tend to build up

TABLE III
General information about groups and their performance.

ID	Group Size	Active Participants	No. of Queries	Correct	Wrong	Reward (Candies)
1	4	3	7	17	1	9
2	6	6	8	18	0	10
3	6	5	8	16	2	6
4	5	3	8	14	4	2
5	4	4	6	15	3	6
6	5	4	6	15	3	6
7	4	3	5	15	3	7
8	6	4	7	18	0	11
9	6	5	8	15	3	4
10	6	4	8	17	1	8
11	6	3	7	18	0	11
12	6	3	9	17	1	7
13	3	3	9	18	0	9
14	4	3	9	18	0	9

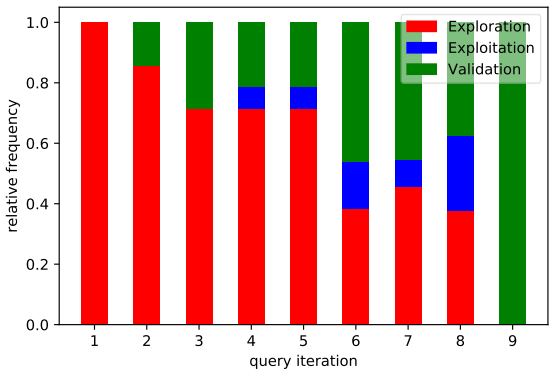

Fig. 4. Relative frequency of the query purpose (exploration, exploitation, validation) w.r.t. the query iteration.

hypotheses to be validated. Only a few queries had the purpose of exploitation.

We conclude that humans either build up hypotheses even if there is only few evidence, or data (the bugs) with more categorical features than numerical ones is not ideal for exploitative queries. Nevertheless, these results confirm the general understanding of a good selection strategy.

B. Learning Performance

Results of AL algorithms are often visualized by learning curves. The idea is to plot the performance of the iteratively trained classifier over time (resp. the number of queries). A well performing method thereby achieves high performance in a short time. Hence, we have to optimize two objectives here: 1) the final performance, and 2) the speed of improvement.

In our experiment, we did not force the participants to classify each unlabeled bug after each query. Therefore, we have an interval for the performance which considers the undecided bugs as potentially wrong to get the lower boundary and potentially correct for the upper boundary of performance, depicted as gray area in Fig. 5 for two exemplary plots of groups 1 and 3. The red star at the end marks the final decision and the time point when the group decided to stop. The vertical dashed lines emphasize the time points of queries with the color indicating the query purpose (exploration, exploitation, or validation). In order to evaluate the benefit of a single query for machine learning, we exported a machine readable dataset

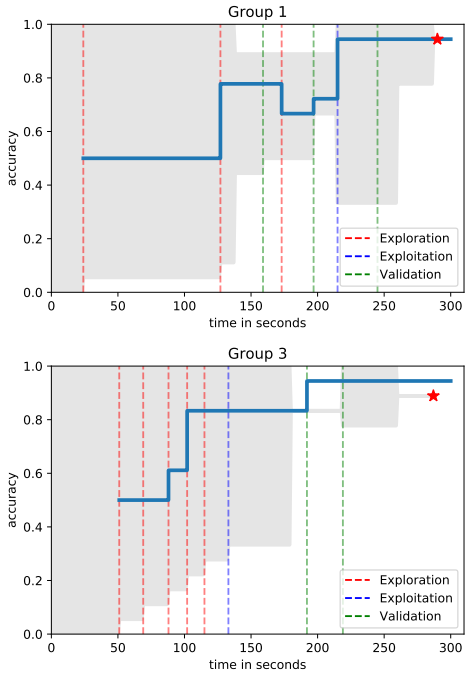

Fig. 5. Learning curves of group 1 and 3. The gray area shows the accuracy interval of the group. The accuracy of an artificial decision tree classifier trained on the queried data is shown in blue. The vertical lines show the time points and the purpose (color) of a query.

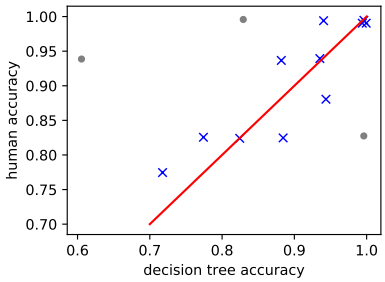

Fig. 6. Accuracy plot showing the correlation between human classification and the decision tree classifier. The red line indicates perfect matches.

as a CSV file[3]. Then, we trained an artificial decision tree classifier (with gini impurity as splitting criteria) to get predictions for every time point. The training samples consist of the queries made during the experiment by the corresponding group. This performance is plotted in blue in Fig. 5. Learning curves for all groups are available for download[4].

Fig. 6 depicts the correlation between human classification and the decision tree classifier to show that the results of both are comparable. For five groups, the accuracy of both classifiers have been completely identical (points on red line). The result of additional six groups solely varies in one misclassification. The three gray dots have been marked as outliers as these results are based on guesses not a rule (e.g., one group was not able to formulate a reasonable classification hypothesis although the queried bugs were great for classification).

C. Comparison with Baseline Algorithms

To evaluate if it is convenient to investigate the selection strategies of our participants in greater detail, we compared them to standard artificial AL baselines, namely a random strategy and a uncertainty-based strategy. We are aware of the fact that there are more advantageous methods but due to clarity of visualization and due to the fact that our participants are completely untrained, we decided to only use these standard baseline methods. Fig. 7 shows the learning curves of these baseline algorithms and learning curves from the best group, the best five groups and all groups using the previously described decision tree classifier. Random and uncertainty sampling were executed 100 times to cope with random effects. All learning curves are reported by its mean and standard deviation except for the best group (only one line). Note that uncertainty sampling used random sampling until both classes have been found. Our application is a transductive AL setting, i.e., we aim to only classify the instances from our candidate pool and do *not* aim to generalize for unseen data. Hence, the training and test set are similar. As the number of queried labels varies across the groups, we decided to only evaluate first 6 queries.

[3]Download at www.ies.uni-kassel.de/p/bug-study#csv-data
[4]Download at www.ies.uni-kassel.de/p/bug-study#learningcurves

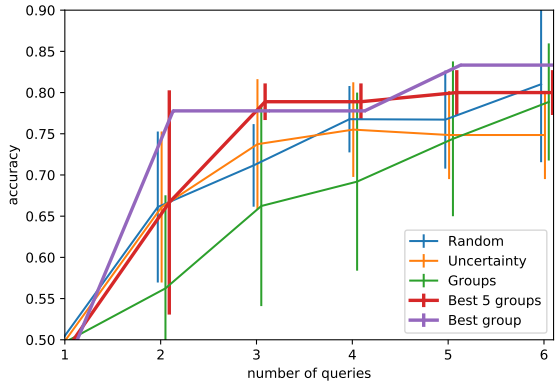

Fig. 7. Learning curves for random sampling and uncertainty sampling, as well as the results from all 14 groups, the best 5 groups and the best group. Each curve shows the mean accuracy and standard deviation.

The plot shows that the mean of all groups achieves only mediocre results but the best group, resp. the mean of the best 5 groups is advantageous compared to the baseline algorithms. The ranking of groups was based on the mean over all queries. Nevertheless, we have to be careful interpreting these results: Having tested 14 groups which additionally have no experience in performing this task might underestimate the true performance of human domain experts. Following the AL evaluation gold standard approach in [2], it is beneficial to use the same classifier for all selection strategies. This again underestimates the human performance as they did not optimize the decision tree but their own classification hypothesis. Furthermore, the selection of the best groups (resp. its ranking) should be determined in a separate experiment or task similarly to evaluating hyperparameter optimization techniques.

D. Collaboration is Beneficial for Active Learning

To evaluate the impact of collaboration during the experiment, we visualized the reward in terms of remaining candies w.r.t. the number of active participants in Fig. 8. The reward is given as box plots showing the minimum, the first quartile, the median, the third quartile, and the maximum. Note that the number of groups vary across the X-axis as we only have one group with 6 active participants.

The left plot shows the reward of the original experiment using the groups' original classification hypotheses whereas the right one shows the reward using an artificial decision tree classifier. Interestingly, a bigger group does not improve human hypothesis making, whereas the decision tree seems to benefit from collaborative selection. Our interpretation is that collaboration improves the selection of bugs as this helps in finding bugs which are controversial and therefore diverse. In contrast, it is difficult to merge these different, possibly vague human hypotheses into one general classification rule which leads to poor classification results.

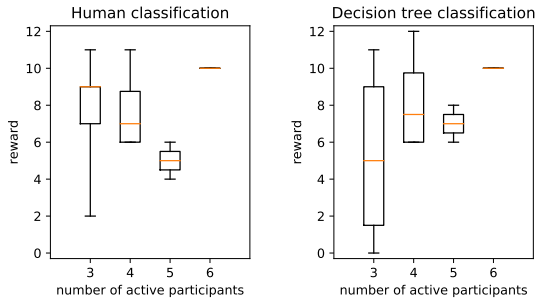

Fig. 8. Box plots showing the reward with human classification and the artificial decision tree w.r.t. the number of active participants.

E. Stopping Criterion

All in all, five groups finished prior to the final time limit of 5 minutes. In this section, we aim to investigate what might have triggered their decision to stop querying as this is a big open issue in AL research. For all five groups, the purpose of the last query has been consistently used to validate an existing hypothesis. In the second last step, only one group chose to do an exploratory query. The other four groups queried to validate or to explore. Hence, our interpretation is that groups who successfully validated their hypothesis once (or even twice) have been confident enough to stop the query process.

VI. Possible Setup for a Comprehensive Study

This exploratory study showed some advantages of collaborative human selection in an AL task and provided insights in human query behavior. As a result, we aim to extend this research and to conduct a study which is 1) larger and 2) in a more controlled environment. As having both aspects in one study would increase the effort excessively, we decided to have two separate experiments.

To conduct the first experiment, we aim to implement a highly versatile, mobile-ready web application which builds on the idea of that study. Hence, users play a game trying to find the correct hypothesis of differently difficult classification tasks on multiple datasets. To gather lots of data, we focus to have an easy and understandable user interface and we will use this tool in lectures and workshops. To evaluate learning performance and finding superior strategies, we use clustering techniques to find similar strategies. A high score board will increase the competition and live analysis will provide an interesting feature for tutorials on AL. We aware of the fact that participants might get help, might play without having understood the game, etc. Hence, we use the variety of data to describe random behavior and still show significance. To assign games played at different days to one player, we aim to use cookies or user identifiers.

Additionally, we plan to have a second comprehensive study in a fully controlled environment. Hence, subjects will be reviewed and asked to provide general information about themselves using a questionnaire. Then, we have two different datasets with three learning tasks of different difficulty. One of

these datasets will be used as a control experiment to be able to proof hypothesis across experiments. In order to evaluate the benefit of a single query for the participants, we ask them to predict the classes of each object at any point in time. The experiment is conducted with a high-resolution, multi-touch display to capture accurate, high-resolution temporal data which can be evaluated without human inspection.

Both studies will be executed in cooperation with experienced psychologists. Additionally, we plan to have some subjects use the web application in a controlled environment to hopefully find mappings from one study to the other. Thereby, we can enrich data of the big study with information from the controlled experiments.

VII. Conclusion

The aim of our research is to evaluate human active learning performance. This first, explorative study of our two-stage analysis generally showed that machine active learning research will probably benefit from creative human strategy building. In this article, we described the full experimental design including material for download such that this study can be easily replicated and validated. The evaluation of the extracted data showed that humans start with explorative queries which are later replaced by queries for validation or exploitation. This is similar to findings in artificial AL research [22]. Additionally, we provide learning curves for each group and compared their results to artificial AL baselines. The preliminary findings on collaborative aspects and the problem of when to stop active learning suggests hypotheses for further research, described in the previous section.

This study is reproducible research and therefore enables researchers to use our data for their own studies and hopefully stimulate research in this area. For example, one could investigate more complex methods to assess the usefulness of a human query. This study also contributes to research in the field of dedicated and opportunistic collaborative active learning [3], [23], visual analytics [11], and active category learning [24]. Especially, interdisciplinary research between cognitive psychology and artificial intelligence seems to be interestingly promising.

Acknowledgment

The authors thank the students of the University of Kassel for participating in this study. The research presented in this paper is conducted within the project "CIL", funded by the University of Kassel (funding program for further profiling of the university 2017-2022: "Zukunft 2017-Standard"). The financial support is gratefully acknowledged.

References

[1] B. Settles, "Active learning literature survey," University of Wisconsin-Madison, Computer Sciences 1648, 2009.

[2] D. Kottke, A. Calma, D. Huseljic, G. Krempl, and B. Sick, "Challenges of reliable, realistic and comparable active learning evaluation," in *Interactive Adaptive Learning Workshop @ ECMLPKDD 2017*, ser. CEUR Workshop Proc. 1924, 2017, pp. 2–14.

[3] A. Calma, J. M. Leimeister, P. Lukowicz, S. Oeste-Reiß, T. Reitmaier, A. Schmidt, B. Sick, G. Stumme, and A. K. Zweig, "From active learning to dedicated collaborative interactive learning," in *29th Int. Conf. on Architecture of Computing Systems, Workshop Proceedings*. Nuremberg, Germany: VDI Verlag, 2016, pp. 1–8.

[4] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proc. of the 17th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*. Dublin, Ireland: Springer-Verlag New York, Inc., 1994, pp. 3–12.

[5] N. Roy and A. Mccallum, "Toward optimal active learning through monte carlo estimation of error reduction," in *Proc. of the Int. Conf. on Machine Learning*, 2001.

[6] D. Kottke, G. Krempl, D. Lang, J. Teschner, and M. Spiliopoulou, "Multi-class probabilistic active learning," in *ECAI*, ser. Frontiers in Artificial Intelligence and Applications, vol. 285. IOS Press, 2016, pp. 586–594.

[7] M. Behrisch, F. Korkmaz, L. Shao, and T. Schreck, "Feedback-driven interactive exploration of large multidimensional data supported by visual classifier," *2014 IEEE Conf. on Visual Analytics Science and Technology, VAST 2014 - Proceedings*, pp. 43–52, 2015.

[8] T. Ertl, H. Bosch, S. Koch, and F. Heimerl, "Visual Classifier Training for Text Document Retrieval," *IEEE Transactions on Visualization & Computer Graphics*, vol. 18, pp. 2839–2848, 2012.

[9] B. Höferlin, R. Netzel, M. Höferlin, D. Weiskopf, and G. Heidemann, "Inter-active learning of ad-hoc classifiers for video visual analytics," *IEEE Conf. on Visual Analytics Science and Technology 2012, VAST*, pp. 23–32, 2012.

[10] E. Weigl, A. Walch, U. Neissl, P. Meyer-Heye, T. Radauer, E. Lughofer, W. Heidl, and C. Eitzinger, "Mapview: Graphical data representation for active learning." in *Active Learning Workshop @ iKnow*, ser. CEUR Workshop Proc. 1707, 2016, pp. 3–8.

[11] J. Bernard, M. Hutter, M. Zeppelzauer, D. Fellner, and M. Sedlmair, "Comparing Visual-Interactive Labeling with Active Learning: An Experimental Study," *IEEE Transactions on Visualization and Computer Graphics*, 2017.

[12] D. Markant and T. Gureckis, "The impact of self-directed learning in a perceptual categorization task," *Journal of Experimental Psychology: General*, 2012.

[13] G. Kachergis, C. Yu, and R. M. Shiffrin, "Actively learning object names across ambiguous situations," *Topics in Cognitive Science*, 2013.

[14] D. A. Lagnado and S. Sloman, "The advantage of timely intervention," *Journal of Exp. Psych.: Learning, Memory, and Cognition*, vol. 30, pp. 856–876, 2004.

[15] X. Zhu, T. Rogers, R. Qian, and C. Kalish, "Humans Perform Semi-supervised Classification Too," *Proc. of the 22Nd National Conf. on Artificial Intelligence (AAAI'07)*, pp. 864–869, 2007.

[16] R. Castro, C. W. Kalish, R. Nowak, R. Qian, T. Rogers, and X. Zhu, "Human Active Learning," *Advances in Neural Information Processing Systems*, pp. 241–248, 2008.

[17] D. B. Markant and T. M. Gureckis, "Is it better to select or to receive? Learning via active and passive hypothesis testing," *Journal of Experimental Psychology: General*, vol. 143, no. 1, pp. 94–122, 2014.

[18] D. B. Markant, B. Settles, and T. M. Gureckis, "Self-directed learning favors local, rather than global, uncertainty," *Cognitive Science*, pp. 1–21, 2015.

[19] D. Kuhn and D. Dean, "Is developing scientific thinking all about learning to control variables?" *Psychological Science*, vol. 16, no. 11, pp. 866–870, 2005.

[20] G. Kachergis, M. Rhodes, and T. Gureckis, "Desirable difficulties during the development of active inquiry skills," *Cognition*, vol. 166, pp. 407–417, 2017.

[21] R. Nuzzo, "Scientific method: statistical errors," *Nature News*, vol. 506, no. 7487, p. 150, 2014.

[22] T. Reitmaier and B. Sick, "Let us know your decision: Pool-based active training of a generative classifier with the selection strategy 4ds," *Information Sciences*, vol. 230, pp. 106–131, 2013.

[23] G. Bahle, A. Calma, J. M. Leimeister, P. Lukowicz, S. Oeste-Reiß, T. Reitmaier, A. Schmidt, B. Sick, G. Stumme, and K. A. Zweig, "Lifelong learning and collaboration of smart technical systems in open-ended environments – opportunistic collaborative interactive learning," in *Self-Improving System Integration*, Würzburg, DE, 2016, pp. 1–10.

[24] D. B. Markant, A. Ruggeri, T. M. Gureckis, and F. Xu, "Enhanced memory as a common effect of active learning," *Mind, Brain, and Education*, 2016.