Facility Locations Utility for Uncovering Classifier Overconfidence

Karsten Maurer
Department of Statistics
Miami University
Oxford, USA
maurerkt@miamioh.edu

Walter Bennette
Information Directorate
Air Force Research Lab
Rome, USA
walter.bennette.1@us.af.mil

***Abstract*—Assessing the predictive accuracy of black box classifiers is challenging in the absence of labeled test datasets. In these scenarios we may need to rely on a human oracle to evaluate individual predictions; presenting the challenge to create query algorithms to guide the search for points that provide the most information about the classifier's predictive characteristics. Previous works have focused on developing utility models and query algorithms for discovering unknown unknowns – misclassifications with a predictive confidence above some arbitrary threshold. However, these methods tend to reward the discovery of misclassifications that occur at the rate indicated by their confidence values. These search methods may reveal nothing more than a correct assessment of predictive certainty, and as a result, we are unable to properly mitigate the risks associated with model deficiency when the model's confidence in prediction exceeds the actual model accuracy. We propose a novel problem formulation to instead search for overconfident unknown unknowns. Specifically, we propose a facility locations utility model and corresponding greedy query algorithm to search for overconfident unknown unknowns. Through robust empirical experiments we demonstrate that the greedy query algorithm with the facility locations utility model outperforms previous methods in discovering overconfident unknown unknowns.**

INTRODUCTION

Techniques such as active learning [17] and domain adaptation [14] can be used to create machine learning classifiers when large labeled datasets are not available for a particular task. For example, the black box classifiers made available through many online services (Google Cloud, Amazon Web Services, etc.) require no training data from the user and their deployment can be thought of as a kind of domain adaptation. However, with limited amounts of labeled data, users may not be able to properly evaluate a model, and are left hoping the model will be useful for their intended task. Previous works have developed human-in-the-loop methods to help evaluate classifiers in the absence of labeled data and have focused largely around algorithms that seek to uncover *unknown unknowns* (UUs): points where a classifier is highly confident in its prediction, but wrong [2].

UUs can be thought of as blind spots to a classification model, and can be caused by dataset bias during training [18], domain shift during use [19], lack of model expressibility, and other causes of poor model fit. Lakkaraju et al. [9] describe a classifier trained on a biased image dataset of cats with light fur and dogs with dark fur. When this classifier is used for inference it predicts that dogs with light fur belong to the cat class with high confidence. The light fur dogs are UUs for the classifier and reveal a deficiency of the model.

For a particular application the confidence of a prediction can be used to determine how an instance will be handled. Logically, low confidence predictions have high potential to be incorrectly classified and should have additional oversight before any action is taken. Meaning, from the viewpoint of a rational actor, UUs represent costly mistakes because minimal risk mitigation strategies will have been deployed for these high confidence predictions. However, the discovery of UUs could then allow new mitigation strategies to be formulated [11]. Additionally, as further enumerated in Bansal and Weld [3], finding UUs is valuable to understand classifier strengths and weaknesses and possibly avoid certain adversarial attacks.

While concerns with misclassifications involving high confidence predictions are well founded, there is a clear oversight in the focus of the existing literature reguarding UUs. Previous works have centered on discovering UUs defined to be misclassifications with a predictive confidence above some arbitrary threshold, τ (τ typically set to 0.65 for binary classification). However, under this definition, even for a model with perfectly calibrated confidence scores, if we sample points having a predictive confidence equal to the threshold we should expect $(1-\tau)\%$ of the points to be called an UU. Meaning, existing search algorithms for UUs do not account for the expected misclassification rates that are inherent to predictive modeling and thus ignore the fundamental purpose of confidence scores. It would be more valuable to uncover cases where misclassification is occuring at a rate higher than should be expected based on predictive confidences, thus searching for cases that reveal classifier *overconfidence*, or model miss-calibration. In this paper we build upon previous work to develop a human-in-the-loop method to identify high confidence mistakes, but adapt the methods to solve the novel problem of finding model *overconfidence*. Specifically, we develop an interactive search algorithm to uncover *overconfident unknown unknowns* based on an facility locations utility model. Identifying overconfident UUs can reveal problematic areas of the classifier and begin to hint at mitigation strategies such as model calibration [4].

In the following manuscript we first discuss the established algorithms for discovering UUs, and demonstrate deficiencies in the utility and problem design of previous methods. We then propose our own facility locations utility model and corresponding search algorithm to maximize utility for the novel problem of finding overconfident UUs. Through robust empirical experiments we demonstrate that the greedy query algorithm with the facility locations utility model consistently results in oracle queries with superior performance in discovering overconfident UUs than previous methods. We conclude with a discussion of these results, access to the implementation and avenues for future work.

Previous Works

The search for unknown unknowns of a classification model operates with an unlabeled test set and does not require access to the original training features. This type of scenario can arise, for example, with an externally provided black box classifier. It is also assumed that an oracle can be queried to provide labels up to a certain budget and that the model can provide a realistic confidence of its prediction. Given these assumptions the search for UUs is carried out over a set of unlabeled points for which a classifier has provided predicted labels and associated confidence values.

Attenberg, Ipeirotis, and Provost [2] turned the search for classifier errors into a game to be played by humans called "Beat the Machine". Through trial and error, a utility function was derived to value high confidence mistakes (UUs). Then users "played" the game by submitting URLs to earn a monetary reward tied to the utility function.

Semi-automated methods to search for UUs have also been proposed to avoid the logistics and resources needed to crowdsource the search like "Beat the Machine" [9] [3]. Each of these methods can be distilled to the same basic components. First, a utility function is constructed to capture the value of a set of discovered UUs. Second, a strategy is developed to sample unlabeled points to maximize the designed utility, where each search strategy is driven by some estimation of a point's likelihood of being an UU. Third, all methods execute a search following the developed strategy until a labeling budget is exhausted.

Lakkaraju et al. [9] introduced the first algorithmic approach for discovering UUs with a semi-automated search directly providing unlabeled points to an oracle. Their utility function provides a unit value for each discovered UU and penalizes by the cost of labeling (for example the number of words read to evaluate a text classification). Their search strategy relies on a multi-armed bandits approach to sample from clusters of the points based on classifier confidence and a derived feature space. The bandit search is driven by tracking the average utility of a cluster, which can be viewed as an indication of the likelihood of finding an UU in that cluster.

Bansal and Weld [3] argue that the unit utility of Lakkaraju et al. [9] motivates the discovery of very similar UUs. Instead, they propose an adaptive coverage-based utility model that attempts to encourage the discovery of high confidence UUs spread throughout a feature space. They then search for UUs via a greedy algorithm to maximize utility. Like the bandit search, the greedy search relies on a clustering of a derived feature space and is driven by the observed ratio of UUs in each cluster.

Looking closer at the coverage-based utility model, it sums the prediction confidence of every test point multiplied by a similarity measure comparing it to its closest discovered UU. It has the form:

$$U(Q) = \sum_{x \in \mathbb{X}} c_x \cdot \max_{q \in S} \left\{ sim\left(x, q\right) \right\}$$

where $\mathbb{X} \subset \mathbb{R}^p$ is the set of available p-dimensional unlabeled test points, $Q \subset \mathbb{X}$ is the set of points labeled by an oracle, $S = \{x | x \in Q, y_x \neq M(x)\}$ is the set of discovered UUs for some classifier $M(x) : \mathbb{X} \rightarrow class$, c_x is the classifier's confidence in its prediction of x, and $sim(x, q)$ is a distance-based similarity metric.

Given this utility model the search for UUs is performed by greedily selecting the point q' that maximizes the expected utility increase. Meaning, q' is selected to maximize,

$$E\left[U_x\left(Q \cup q'\right)\right] = \hat{\phi}(x) \cdot c_x \cdot \max_{q \in S \cup q'} \left\{ sim\left(x, q\right) \right\},$$

where $\hat{\phi}(x) = P\left(y_x \neq M(x) | Q\right)$ is the cluster conditional probability that x is misclassified given the query set. As previously stated, this method is designed to incentivize a broader search for UUs and gives higher utility for finding misclassifications in higher confidence regions.

Unfortunately, and surprisingly, the coverage-based utility search consistently achieves lower coverage-based utility than the simple strategy of sequentially querying points for which the classifier is most uncertain. This is shown in Figure 1 which displays Monte Carlo medians and 90% predictions bands of the coverage-based utility for the four test datasets made available in the supplementary files to Bansal and Weld [3]. To account for the variability of greedy search algorithms due to initial conditions, searches are performed following each strategy for 1000 random samples of the test data of size n=1000, with a budget of B=100 queries. The superior performance of the most uncertain search exposes issues with the coverage-based utility model.

An issue with the coverage-based utility model is that it rewards the discovery of UUs near points for which the classifier has high confidence, not the discovery of high confidence UUs themselves. Therefore, the utility model may reward the discovery of low confidence mistakes more than the discovery of high confidence mistakes; the stated goal of the search. This is because there is no guarantee that points for which the classifier is similarly confident are confined to the same area of the feature space. Meaning, it may be better to discover the easily found low confidence mistakes than the difficult to find high confidence mistakes. This is demonstrated by results shown in Figure 1.

Given these apparent issues, we aim to construct a utility-based query algorithm that more appropriately rewards the

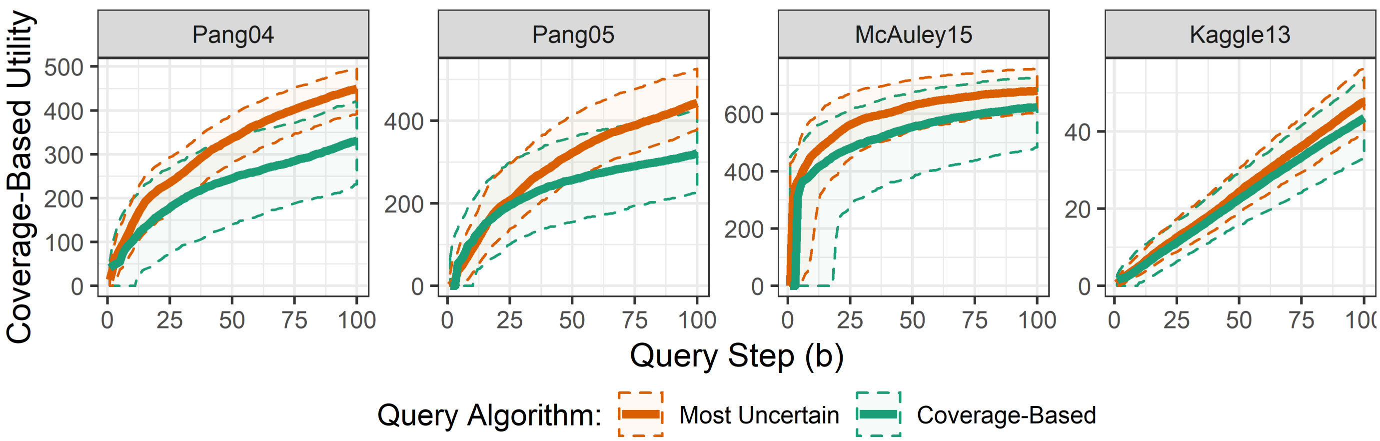

Fig. 1. Comparison of coverage-based utility outcomes achieved under most uncertain and coverage-based search algorithms. Monte Carlo medians (solid) and 90% prediction bands (dashed). In all datasets, higher coverage-based utility is achieved with most uncertain searches.

identification of UUs, and helps to identify overconfident points. Again, we believe discovering query sets where UUs exist at higher rates than expected is more valuable to a rational actor than simply finding UUs. A certain number of UUs should be expected at different confidence levels, and should be planned for. Discovering where confidence levels are incorrect can allow better mitigation strategies.

METHODOLOGY

We propose an alternative utility model based on facility location optimization methods [6]. In the facility locations problem a utility can be constructed that uses a greedy algorithm to minimize the cost, or maximize the reward, of building a series of new facilities in a supply chain, while also minimizing distances between clients to the nearest facility [7] [1]. In the UU query setting, we can draw an analog to the selection of a point to query to the establishment of a facility at that location in the feature space; evaluating the reward for selecting the point, and the distance it stands from the surrounding unobserved points. We propose a facility locations utility function as:

$$W(Q) = \sum_{q \in S} r\left(c_q\right) - \frac{1}{n} \sum_{x \in \mathbb{X}} \min_{q \in S} \left(d\left(x, q\right)\right)$$

where $r\left(c_q\right) = \log(1/(1-c_q))$ is the *reward* function for finding an UU with confidence c_q, and $d(x, q)$ is the Euclidean distance between points x and q. We use the greedy algorithm that at each iteration selects q' with the maximum expected utility, as defined in Algorithm 1.

At each iterative step in Algorithm 1, we need to select the point that will maximize the expected gain in facility location utility, given probability estimates for point misclassification, $\hat{\phi}(q'|Q) = \hat{P}(y_{q'} \neq M(q')|Q)$. To find the expected gain in utility for each point, we evaluate the utility under the possibilities that a point is either misclassified or correctly classified. These possible utility outcomes are then averaged with weights equal to the estimated probability of each outcome. Thus the optimization step requires the solution of the following:

$$\underset{q' \notin Q}{argmax} E[W(Q \cup q')] =$$

$$\underset{q' \notin Q}{argmax} \left[\begin{array}{l} \hat{\phi}(q') \cdot \left[\sum\limits_{q \in S \cup q'} r\left(c_q\right) - \frac{1}{n} \sum\limits_{x \in \mathbb{X}} \min\limits_{q \in S \cup q'} \left(d\left(x, q\right)\right) \right] + \\ \left(1 - \hat{\phi}(q')\right) \cdot \left[\sum\limits_{q \in S} r\left(c_q\right) - \frac{1}{n} \sum\limits_{x \in \mathbb{X}} \min\limits_{q \in S} \left(d\left(x, q\right)\right) \right] \end{array} \right]$$

Note that $\left[\sum_{q \in S} r\left(c_q\right) - \frac{1}{n} \sum_{x \in \mathbb{X}} \min_{q \in S} \left(d\left(x, q\right)\right)\right]$ is constant for all considered points, but cannot be removed from the argmax solution because it is multiplied by an estimated probability that is unique to each point.

Algorithm 1 Greedy Facility Location Search

Input: Test set $\mathbb{X}$, prior $\hat{\phi}\left(x|Q = \emptyset\right)$, budget B
$Q = \{\}$ {inputs that have been queried}
$y_Q = \{\}$ {oracle defined labels}
For: $b = 1, 2, ..., B$ **do:**
$q' = \text{argmax}_{q' \notin Q}\, E\left[W\left(Q \cup q'\right)\right]$
$y_{q'} = OracleQuery(q')$
$Q \leftarrow Q \cup q'$
$y_Q \leftarrow y_Q \cup y_{q'}$
$S \leftarrow \{x | x \in Q$ and $y_x \neq M(x)\}$
$b \leftarrow b + 1$
Return: Q**,** S **and** y_Q

In addition to a change in the utility structure from previous methods, we propose the use of model-based estimates for $\phi(x) = P\left(y_x \neq M(x)|Q\right)$; as an alternative to the $\hat{\phi}(x)$ based on tracking the rate of UUs in clusters found through a multi-stage clustering procedure as was done in Lakkaraju et al. [9] and Bansal and Weld. Without loss of generality, we demonstrate the use of logistic regression classifier probabilities, fitted such that:

$$\hat{\phi}(x) = logistic(c_x\hat{\beta}_0 + \sum_{j=1}^{p} x_j\hat{\beta}_j) = \frac{e^{c_x\hat{\beta}_0 + \sum_j x_j\hat{\beta}_j}}{1 + e^{c_x\hat{\beta}_0 + \sum_j x_j\hat{\beta}_j}}$$

Given that fitting the logistic regression model requires at least one misclassified and one correctly classified point, we initialize the process using $\hat{\phi}(x) = (1 - c_x)$ until both outcomes have been observed by the oracle.

There are a few characteristics to note in the design of the facility locations utility model.

1) In the utility function, reward is only accumulated by finding UUs in the query set. This avoids the issue of placing value on points in the test set for simply having high confidence and being near a discovered UU.
2) The utility function encourages the discovery of well spread UUs by having a penalization term equal to the average minimum distance between all test points and their closest observed UU. This places value on having strong coverage of the test data by the query set, especially early in the query sequence.
3) The reward function, $r(c_x) = \log(1/(1 - c_x))$, is designed to impact the utility when more UUs than *expected* are discovered. See further discussion below.

Viewed as a geometric distribution problem with a probability $\phi(x)$ of discovering a UU, we expect to need $1/\phi(x)$ queried points like point x before discovering the first UU [5]. For heuristic insight into the reward behavior construction, if we assume that $\phi(x) = (1 - c_x)$, then our reward is a log-scaled count of the number of randomly selected points we would expect to query in order to find the UUs in our query set. The optimization step will provide the high facility-location utility for overconfident points because the reward function increases for selecting misclassifications that had high confidence, c_x, and low achieved accuracy, $\phi(x)$. Note that unlike the UU definition, this construction does not require the arbitrary definition of a confidence threshold, τ, beyond which we search for misclassifications. The reward component of the facility locations utility encourages the search procedure to select points where the model is most overconfident. We define *overconfidence* for an instance as the difference between the confidence value given by the classifier and the rates of correct classification that the model achieves for all instances at that confidence value.

Results

We empirically evaluate our facility location utility model by applying Algorithm 1 to the four datasets used in [9] and [3]: Pang04, Pang05, McAuley15 and Kaggle13; representing three text classification tasks and an image classification task, respectively. For each dataset we fit a classifier, $M(x)$, to a biased training set, then generate predicted classes and confidence values for all observations in the test set. We search for UUs in the test set belonging to a critical class using a feature space derived through singular value decomposition. The datasets and classifiers were chosen to maintain consistency with the data used to evaluate both of the previous methods, unless otherwise noted. Each dataset was obtained from the repository accompanying the work of Bansal and Weld [3].

The classifiers for Pang04, Pang05, and McAuley15 use logistic regression with unigram features. The derived feature space used for the UUs search is created with singular value decomposition on unigram features from only the test set. The classifier for the Kaggle13 dataset is a CNN (eight convolutional layers and two linear layers), and the derived feature space is created with singular value decomposition on raw pixel values.

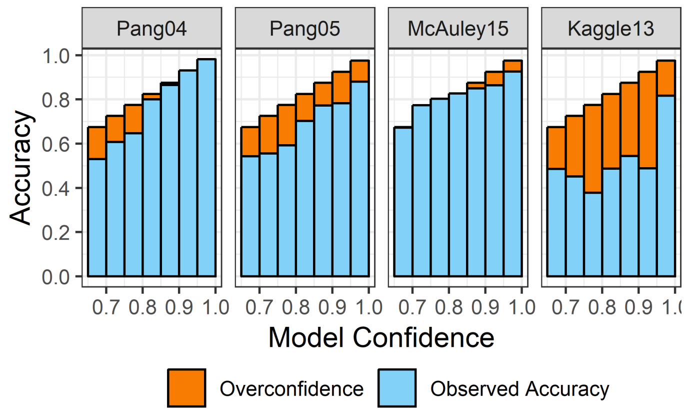

Fig. 2. Observed classifier overconfidence profiles of each experimental dataset.

Figure 2 displays the overconfidence of the models in each test dataset for observations different ranges of model confidence. The plot is created by displaying the difference between the expected and achieved accuracy for all instance within binned confidence ranges. We see that the models from the Pang04 and Pang05 datasets are most overconfident for points with relatively low confidence values. This may provide some insight as to why the simple sequential search of most-uncertain points outperformed the coverage-based utility search, as seen in Figure 1. We would expect that a simple sequential search of the most-uncertain points in Pang04 and Pang05 to also provide high facility locations utility. The predictions for McCauley15 and Kaggle13 are most overconfident for points in the higher confidence range, thus most-uncertain search should provide low facility locations utility. We see that these four datasets represent different profiles of overconfidence, thus present good variety for evaluating characteristics of the facility locations utility model.

As with the evaluation of the coverage-based utility, we run the facility locations queries on 1000 random samples of size n=1000 from each of the datasets, using a budget B=100. In the following subsections we evaluate the utility outcomes of the facility locations queries in comparison to the most-uncertain search method, and compare the ability of several algorithms to discover overconfident points.

Facility Location Utility Outcomes

To evaluate the queries generated by the facility locations utility model we collect query results from running Algo-

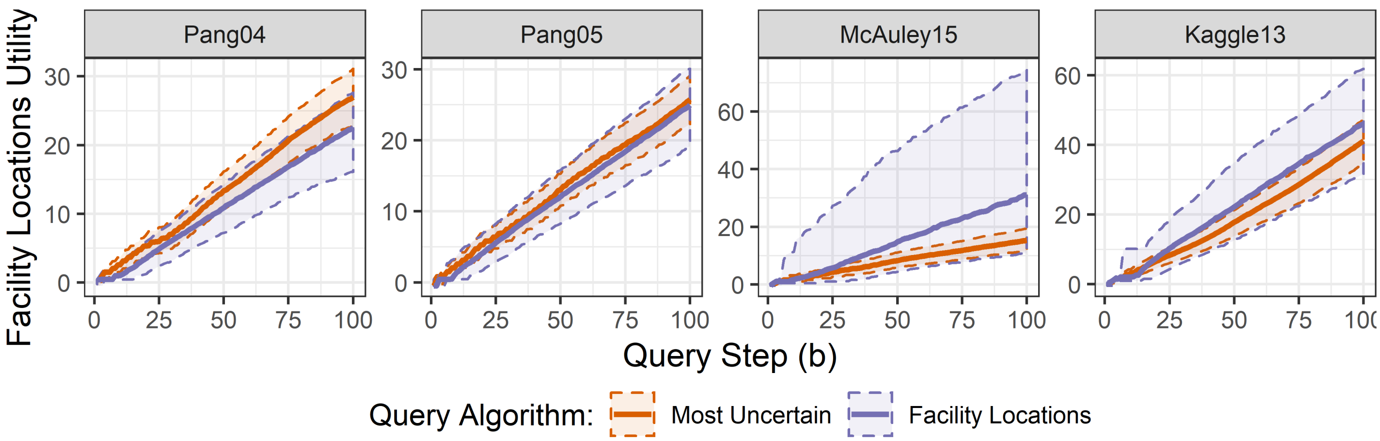

Fig. 3. Comparison of facility locations utility outcomes achieved under most uncertain and coverage-based search algorithms. Monte Carlo medians (solid) and 90% prediction bands (dashed). Higher facility locations utility is typically achieved with facility locations search in all cases.

rithm 1 on repeated random samples from the test sets, thus allowing Monte Carlo estimates to be used for utility characteristics.

Figure 3 displays the Monte Carlo medians and 90% predictions bands for the facility locations utility gains under our algorithms and under the most-uncertain selection method — sequentially searches with points with progressively higher confidence values, starting just above τ=0.65. This done to provide a parallel with the comparison completed in Figure 1; with the notable difference that each case is done with respect to the utility function over which each was optimized. It would be fundamentally unfair to evaluate facility-location searches with coverage-based utility or to evaluate coverage-based searches with facility-location utility. Thus we first compare their preformance against common baseline search method using their respective utilities. We see that the most-uncertain search provides the stronger utility for the Pang04 and Pang05 cases. It may seem that this reflects poorly on our algorithm, but in truth, we know from Figure 2 that this has occured because the model is more overconfident in the points just beyond the τ=0.65 threshold. For McAuley15 where overconfidence is most severe near τ=0.9, we see inconsistent, but higher utility outcomes from the facility locations search, and consistently low utility outcomes from the most-uncertain search. In the last case of Kaggle13, where the overconfidence profile is multi-modal, the facility locations search provided less consistent, but typically stronger utility outcomes than most-uncertain searches. Thus in all scenarios, the facility locations utility model is properly placing value on the pursuit of the most overconfident points, as per its design.

Efficient Discovery of Overconfidence

We compare the queries gathered by the coverage-based utility algorithm from [3], the bandit search algorithm from [9], and our facility locations utility algorithm. Given that all of the searches rely on their own utility function, it does not make sense to compare their selections on the utility values directly. Instead we compare the efficiency of the search for UUs, using a summary statistic that we call the *standardized discovery ratio* (SDR). The SDR is an adaptation of the *standardized mortality ratio* used in biostatistics to evaluate the mortality rate for a given sample of patients, which standardizes using their initial risk of death [20] [16]. In our case we use an analog that evaluates the misclassification rate, standardized by the initial model confidence. The SDR is computed as

$$|S| / \sum_{x=1}^{B} (1 - c_x)$$

thus counting the number of discovered misclassifications, divided by the number of misclassifications expected based on the confidence values of the queried points. The SDR can be interpreted as the number of times more misclassifications were found than were expected based on model confidence; making it a natural metric for evaluating overconfidence.

Figure 4 compares the Monte Carlo medians and 90% central prediction intervals for the SDR values associated with 1000 random samples of size n=1000 from each of the datasets, using each of the four query algorithms: facility locations, coverage-based, bandit, and most uncertain. The SDR intervals for Pang04 and Pang05 reveal that all four algorithms are similarly efficient at discovering overconfident UUs in situations where the overconfident points fall just beyond the defined threshold, $\tau = 0.65$. The SDR intervals for McAuley15 and Kaggle13, where overconfidence was most prevalent for points far beyond the threshold, the facility locations utility algorithm typically provides the most efficient discovery of overconfident points. For Kaggle13, the median SDR for the facility locations algorithm is 1.2 times larger than the coverage-based utility algorithm and 1.6 times larger than both the most uncertain and bandit algorithms.

Discussion & Conclusions

Previous literature has defined unknown unknowns as any highly confident predictions that result in misclassification, possibly with respect to a critical class. This definition ignores the unavoidable uncertainties of predictive modeling. It should

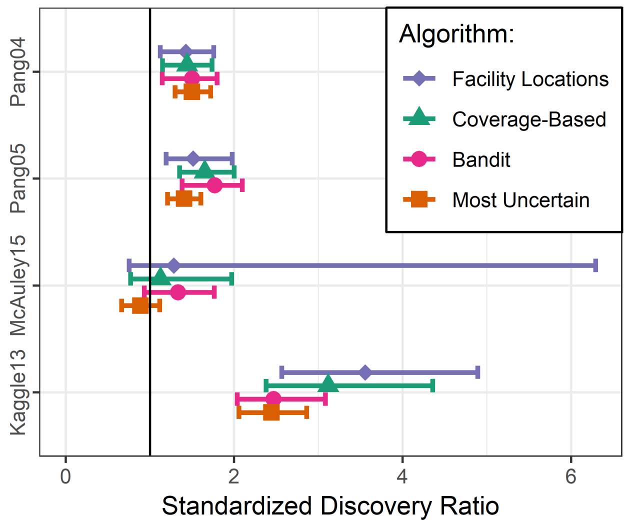

Fig. 4. Monte Carlo medians and 90% prediction intervals of the standardize discovery ratio.

be expected that classifier predictions are imperfect, this is why confidence values exist! The actions taken as a result of the predictions should take into account the inherent uncertainty. However, in the case where the claimed confidence is overstated, a rational actor cannot properly mitigate the risk posed by misclassification. Unlike the previous works that propose utility functions that seek to uncover high confidence misclassifications, the facility locations utility that we propose is designed to seek out *overconfident* misclassifications.

Through repeated random initialization in our computational experiments, we thoroughly tested the outcomes of our facility locations utility algorithm against the bandit, coverage-based, and most uncertain search algorithms. We have demonstrated the ability of our greedy algorithm, using logistic regression probability estimates for $\hat{\phi}(x)$ in the optimization step, to consistently obtain strong facility locations utility in four data scenarios with disparate overconfidence profiles. This is important because in real-world applications we would not know the overconfidence behavior a priori to our query search, so we require a versatile estimation method. We have also demonstrated that oracle queries gathered using a facility locations utility search tend to have higher standardized discovery ratios than the alternative algorithms, thus represent a more efficient use of the constrained budget for queries.

The source code and datasets needed for replicating the experimental results discussed in this paper are available online in the supplemental materials for this manuscript. Also, an on open source implementation in R [15] of the facility locations algorithm and associated functions through the `uuutils` R package can be accessed through the github repository at www.github.com/kmaurer/uuutils.

Future work related to these methods include the exploration of ways to initialize the query set to overcome the relatively slow learning that is seen in low budget cases and exploring the use of our query algorithm within a confidence recalibration strategy to allow more appropriate risk mitigation.

Acknowledgments

Support for this research was provided by the Visiting Summer Faculty Program and Extension Grants sponsored by the Griffiss Institute, in collaboration with the Air Force Research Lab.

We would like to thank Dr. Eric Heim, Dylan Elliott, and Dr. John Bailer for providing valuable feedback throughout the development of these methods.

Cleared for public release by WPAFB Public Affairs on 18 Sep 2018, case number 88ABW2018-4629. Any opinions, findings and conclusions or recommendations expressed in this work are those of the author and do not necessarily reflect the view of the United States Air Force.

References

[1] Arya, V.; Garg, N.; Khandekar, R.; Meyerson, A.; Munagala, K.; and Pandit, V. 2004. Local search heuristics for k-median and facility location problems. *SIAM Journal on computing* 33(3):544–562.

[2] Attenberg, J.; Ipeirotis, P.; and Provost, F. 2015. Beat the Machine. *Journal of Data and Information Quality* 6(1):1–17.

[3] Bansal, G., and Weld, D. S. 2018. A coverage-based utility model for identifying unknown unknowns. In *AAAI*.

[4] Bella, A.; Ferri, C.; Hernández-Orallo, J.; and Ramírez-Quintana, M. J. 2010. Calibration of machine learning models. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. IGI Global. 128–146.

[5] Casella, G., and Berger, R. L. 2002. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

[6] Farahani, R. Z., and Hekmatfar, M. 2009. *Facility location: concepts, models, algorithms and case studies*. Springer.

[7] Guha, S., and Khuller, S. 1999. Greedy strikes back: Improved facility location algorithms. *Journal of algorithms* 31(1):228–248.

[8] Kaggle. 2013. www.kaggle.com/c/dogs-vs-cats. Accessed: 2018-08-28.

[9] Lakkaraju, H.; Kamar, E.; Caruana, R.; and Horvitz, E. 2017. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In *AAAI*.

[10] McAuley, J.; Pandey, R.; and Leskovec, J. 2015. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. ACM.

[11] Nushi, B.; Kamar, E.; Horvitz, E.; and Kossmann, D. 2016. On Human Intellect and Machine Failures: Troubleshooting Integrative Machine Learning Systems.

[12] Pang, B., and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, 271. Association for Computational Linguistics.

[13] Pang, B., and Lee, L. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, 115–124. Association for Computational Linguistics.

[14] Patel, V.; Gopalan, R.; Li, R.; and Chellappa, R. 2014. Visual Domain Adaptation: An Overview of Recent Advances. 1–34.

[15] R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[16] Rosner, B. 2015. *Fundamentals of biostatistics*. Nelson Education.

[17] Settles, B. 2010. Active Learning Literature Survey. *Machine Learning* 15(2):201–221.

[18] Stock, P., and Cisse, M. 2017. Convnets and imagenet beyond accuracy: Explanations, bias detection, adversarial examples and model criticism. *arXiv preprint arXiv:1711.11443*.

[19] Sugiyama, M.; Lawrence, N. D.; Schwaighofer, A.; et al. 2017. *Dataset shift in machine learning*. The MIT Press.

[20] Taylor, P. 2013. Standardized mortality ratios. *International journal of epidemiology* 42(6):1882–1890.