SAFETY IN THE FACE OF UNKNOWN UNKNOWNS: ALGORITHM FUSION IN DATA-DRIVEN ENGINEERING SYSTEMS

Nina Kshetry[†] *and Lav R. Varshney*[†,‡]

[†] Ensaras, Inc., Champaign, IL, USA
[‡] University of Illinois at Urbana-Champaign, Urbana, IL, USA
nina@ensaras.com, varshney@illinois.edu

ABSTRACT

Most current machine learning algorithms make highly confident yet incorrect classifications when faced with unexpected test samples from an unknown distribution different from training; such epistemic uncertainty (unknown unknowns) can have catastrophic safety implications. In this conceptual paper, we propose a method to leverage engineering science knowledge to control epistemic uncertainty and maintain decision safety. The basic idea is an algorithm fusion approach that combines data-driven learned models with physical system knowledge, to operate between the extremes of purely data-driven classifiers and purely engineering science rules. This facilitates the safe operation of data-driven engineering systems, such as wastewater treatment plants.

Index Terms— AI safety, algorithm fusion, epistemic uncertainty, metacognition, wastewater treatment

1. INTRODUCTION

Real-time classification in the presence of noisy data is a problem faced in many engineering systems, whether considering wastewater treatment plants, autonomous vehicles, the smart grid, advanced manufacturing, and sustainable buildings, among other systems that have explicit physical impacts. Motivations include quality control, regulatory compliance, and equipment protection while also providing the possibility of real-time optimization of plant operation to reduce energy and cost. The information and decision components in many such engineering systems are now based on supervised learning built on training data.

A typical decision that a real-time wastewater classification device would make is whether (1) water is safe for non-potable purposes, (2) water is safe for discharge, (3) water is suitable for microbial process, (4) water can be treated by non-reverse osmosis (RO) filtration, (5) water requires RO, (6) water will damage RO filters. Note that the classes are not disjoint, and so in principle there are $2^6 = 64$ possible classes of which 13 are actually possible due to logic constraints, cf. [1]. As examples of the impact misclassification may have: errors under (2) could lead to public health disasters such as a cholera epidemic due to wastewater discharge from a United Nations peacekeeper camp with deficient treatment [2]; errors under (4) could lead to much more energy-intensive treatment than required; and errors under (6) could damage costly equipment. Some of the 13 categories with serious deleterious impacts occur very rarely.

We had previously collected black water and gray water[1] samples from municipal wastewater treatment plants and elsewhere [3, 4]. As Fig. 1 shows, despite spending hundreds of person-hours over several months, our dataset of 235 distinct wastewater samples did not achieve balanced coverage for all 13 categories; rather 5 categories remained unseen. Note that the central difficulty here is not labeling of data for supervision, e.g. via crowdsourcing, as in other settings of machine learning [5], but in gathering unlabeled data in the first place. Yet, classifying rare events correctly is critically important for public health, environmental protection, and the safety of wastewater equipment.

This problem of gathering real (or even realistic) training data covering the entire high-dimensional universe of possibilities arises in numerous engineering systems and infrastructures where *safety* is very important and leads to the problem of *unknown unknowns* (epistemic uncertainty) for purely data-driven systems.

As a centuries-old engineering discipline, however, there is also traditional engineering and regulatory knowledge of wastewater classification expressed in terms of physical parameters measured using laboratory techniques, see e.g. US EPA discharge standards (NPDES Permit Writer's Manual). Although these parameters cannot be measured using real-time sensors, a rough approach is to train nonlinear regression models to map from sensor data to laboratory-based physical parameters, thereby establishing surrogacy relationships that can be used in conjunction with traditional wastewater rules.

In this largely conceptual paper, we propose an approach to controlling epistemic uncertainty due to limited and uneven training data in engineering domains using *algorithm fusion*.

[1] Black water contains human waste whereas gray water is wastewater from domestic sources other than bathrooms and toilets.

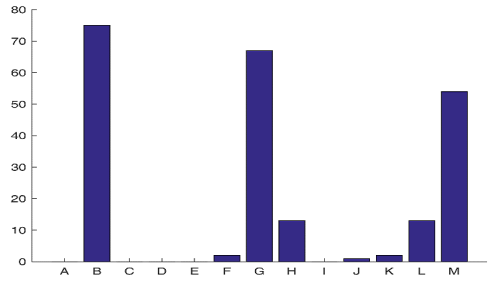

Fig. 1. Despite collecting wastewater samples from a variety of natural sources and conditions, only 8 of 13 classification categories (as expert-labeled) were covered.

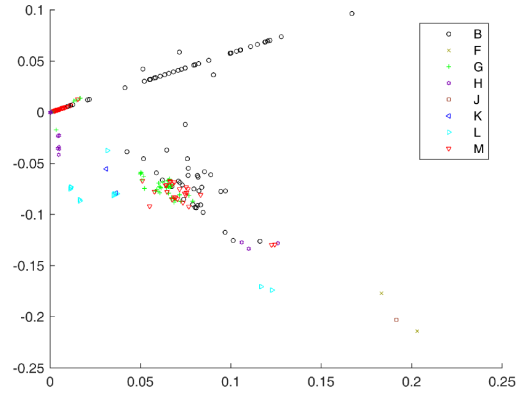

Fig. 2. Two-dimensional principal components embedding of the 235 expert-labeled wastewater training samples from Fig. 1, as projected from an eight-dimensional sensor space. Typical machine learning classifiers may be highly confident but incorrect in the white spaces.

The basic idea is to combine learned models with traditional engineering knowledge; this allows generalization not just to *known unknowns* as in typical supervised learning but also generalization to *unknown unknowns*. The learned model is given a kind of *metacognition* through a one-class classifier to know when it does not know, and should default back to rough traditional engineering knowledge and maintain safety. The question of optimally aggregating the two different kinds of models becomes a statistical signal processing problem.

Since ideas of epistemic uncertainty and AI safety may be less familiar, Sec. 2 first introduces background and also some related work. Next, Sec. 3 gives our novel algorithm fusion approach to combine data-driven and physical knowledge. Finally, Sec. 4 concludes with several suggestions for future work; indeed this short paper is largely conceptual and several novel statistical signal processing questions remain.

2. EPISTEMIC UNCERTAINTY AND SAFETY

Despite the potential benefits of real-time machine learning for engineering systems, a central difficulty is model uncertainty from incomplete knowledge [6, 7] due to limited and uneven training data. Note that test samples may differ from training in both expected and unexpected manners, and gathering all unexpected data is not viable. The difficulty is exacerbated by the fact that contrary to the intuition that unfamiliarity should lead to lack of confidence, most current machine learning algorithms (including deep learning) make highly confident yet incorrect classifications when faced with unexpected test samples from an unknown distribution different from training [8, 9]. They lack metacognition. Fig. 2 shows real-time sensor data readings from [3,4]; one can note that there is much white space even in a low-dimensional embedding.

The reason this happens is quite straightforward: the classification function is only loosely controlled by data for areas of the feature space that are unobserved in training, so the learning algorithm may extrapolate wildly without incurring much loss. Note that multiple instances of learning algorithms, e.g. having different initialization or different subsamples of training data as in ensemble methods, may be aggregated to control the individual wild behavior [8].

Differently distributed samples may specifically pose a safety risk in cyberphysical systems like wastewater treatment plants [10]: "It may be that the distribution the samples actually come from cannot be known, precluding the use of covariate shift and domain adaptation techniques. This is one form of epistemic uncertainty that is quite relevant to safety because training on a dataset from a different distribution can cause much harm." We may indeed have *unknown unknowns*, precluding domain adaptation of various kinds [11]. Methods such as meta-recognition [12] and reject options [13, 14] decline to classify when the system is likely to fail on known unknowns, but are not concerned with quality of confidence estimates and are not analyzed with respect to their ability to reject unknown unknown samples.

Though there may be epistemic uncertainty when building data-driven models, there may also be much traditional engineering knowledge (drawing on both engineering experience and on foundations from physical/biological sciences [15], so-called engineering science [16, 17]). Here we try to leverage such engineering science knowledge to control epistemic uncertainty so as to have safe classifications even for unknown unknowns. Physics-guided neural networks aim to overcome the deficiencies of typical machine learning models that do not generalize well beyond the available labeled data and may not even respect known engineering science laws [18]. As in that work, our goal is to operate between extremes of purely data-driven classifiers and purely engineering science rules; rather than a Lagrangian formulation with loss functions corresponding to data fidelity and scientific consistency as in that work [18], we develop a novel algorithm fusion approach.

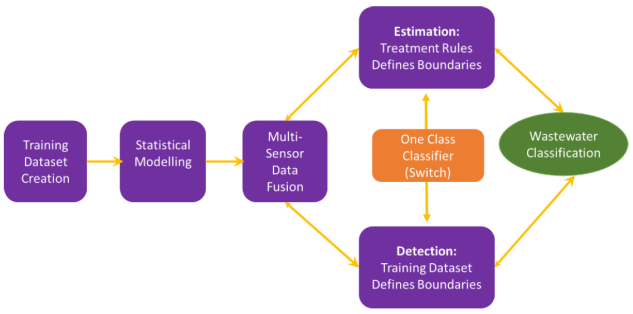

Fig. 3. Schematic diagram of basic approach to incorporating both machine learning models and engineering science rules.

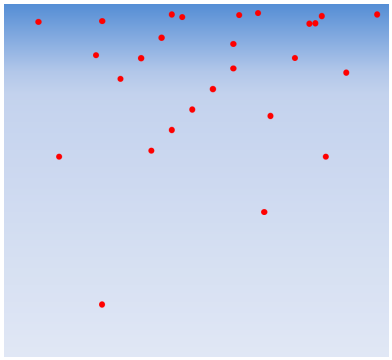

Fig. 4. Weighting (blue gradient) allocates more weight (darker) to the learned model where there is more training data in the sensor space (red dots), than in areas where there is little training data.

3. ALGORITHM FUSION TO COMBINE DATA-DRIVEN AND PHYSICAL KNOWLEDGE

To combine data-driven and physics-driven models for safety, we consider hard/soft switching, and also propose a statistical learning framework for learning with epistemic uncertainty.

3.1. Hard Switching

Let us first consider a very basic approach to incorporating both a learned model and traditional engineering science rules (operating on physical parameters estimated from sensor readings), as shown in Fig. 3. As inspired by outlier detection (and especially universal information theory approaches to the problem [19]), we use available data to train a one-class classifier [20]; outlier detection and one-class classification determine whether a test sample comes from the same distribution as the training data. In operation, when sensor data is within the support of the learned model (which is trained on the same (labeled) data as the one-class classifier), i.e. when the one-class classifier declares it is within range, then the learned model is used; when the sensor data is outside of range, a combination of estimation algorithms and traditional wastewater classification rules are used.

As detailed in [3, 4] using appropriately weighted misclassification cost metrics, we previously demonstrated that this basic approach is effective for the wastewater application, in using an off-the-shelf SVM-based one-class classifier [20] to perform hard switching, random forest classifier as the learned model, and LMSE estimator for physical parameters from sensor data [21] together with wastewater theory-derived rules for non-potable reuse from the US Army (TB MED 577) and for discharge from US EPA standards (NPDES Permit Writer's Manual) as the physical knowledge-based model. This effectiveness is both in improving classification performance within the support of the learned model and in preventing wild behavior outside the support, compared to either model acting alone.

Note that this setting where the learned model is better than the traditional model within support (improves accuracy), whereas the traditional model is better outside the support (controls wild behavior), has strong similarities to the so-called algorithmic noise tolerance (ANT) approach to circuit design [22]. In ANT, there is a main block that is either right on in terms of negligible estimation error or has error that is very large, as well as an estimator block that always has non-negligible noise but never very large. With this analogy and the binary hypothesis test for choosing between the two models here, for insight we can directly analyze the Bayes-optimal likelihood ratio test (with appropriately-designed threshold) developed therein [22]. Note that in practice with limited data we would use a one-class classifier.

3.2. Soft Switching

Can an algorithm fusion approach softer than such hard switching be more effective?

Taking such a hybrid approach of using both models simultaneously boils down to estimating the likelihood of the learned model's effectiveness in different parts of sensor data space and then using that as a weight in an aggregation procedure [22]. This is depicted schematically in Fig. 4. In particular, we will need to not just perform one-class classification but rather to perform distribution support estimation [23, 24] to get a one-class posterior distribution on the learned model.

The basic style of result in the information-theoretic density estimation literature is that given n independent samples from an unknown discrete probability distribution $P = (p_1, p_2, \ldots, p_S)$, with unknown support size S, consider the problem of estimating a functional of the distribution of the form: $F(P) = \sum_{i=1}^{S} f(p_i)$, where $f : (0, 1] \rightarrow \mathbb{R}$ is a continuous function. Then there is a general procedure for constructing minimax rate-optimal estimators for these functionals under L_2 loss. Extensions to real-valued data is also possible [25]. Of particular importance to us, such procedures return not only the minimax estimator but also the corresponding minimax risk, which can then be used directly

for weighting in the aggregation function of algorithm fusion. Moreover, there are converse bounds for such estimation questions [23] to determine basic limits of this architecture.

3.3. PAC Learning Framework

To formalize the general performance limits of learning with epistemic uncertainty further, let us extend basics of traditional statistical learning theory [26] in characterizing sample size required for algorithms to learn a family of concepts. In particular, theoretical learning guarantees for an optimal algorithm depend on the complexity of the concept classes considered and the size of the training sample.

The Probably Almost Correct (PAC) framework for supervised learning helps define the class of learnable concepts in terms of the number of sample points needed to achieve an approximate solution. Here, *examples* are instances of data used for learning; *features* are the set of attributes associated to an example; *labels* are values or categories assigned to examples; and a *hypothesis set* is a set of functions mapping features to the set of labels.

Let us denote the set of all possible examples as $\mathcal{X}$ and the set of all possible labels as $\mathcal{Y}$. For illustrative purposes here, consider binary classification so $\mathcal{Y} = \{0, 1\}$. A concept $c : \mathcal{X} \to \mathcal{Y}$ is a mapping from $\mathcal{X}$ to $\mathcal{Y}$ and a *concept class* is a set of concepts we may wish to learn and is denoted $\mathcal{C}$. Assume examples are independently and identically distributed (i.i.d.) according to some fixed but unknown distribution D. Then the learner considers a fixed set of possible concepts H, called a *hypothesis set*. The learner receives a sample $S = (x_1, \ldots, x_m)$ drawn i.i.d. according to D as well as labels $(c(x_1), \ldots, c(x_m))$ which are based on a specific target concept $c \in \mathcal{C}$ to learn. The task is to use the labeled sample S to select a hypothesis $h_S \in H$ that has small *generalization error* with respect to the concept c. The traditional generalization error (in the setting of known unknowns) of a hypothesis $h \in H$ is denoted $R(h)$ and defined as follows.

Definition 1. *Given a hypothesis $h \in H$, a target concept $c \in C$, and an underlying distribution D, the* generalization error *of h is defined by $R(h) = \Pr_{x \sim D}[h(x) \neq c(x)]$.*

A concept class $\mathcal{C}$ is PAC-learnable if the hypothesis returned by a possible algorithm after observing a number of points polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$ is approximately correct (error at most ϵ) with high probability (at least $1 - \delta$). The PAC framework establishes sample complexity and generalization bounds. As an example, consider a finite hypothesis set H such that the target concept c is in H.

Theorem 1. *Let H be a finite set of functions mapping from $\mathcal{X}$ to $\mathcal{Y}$. Let $\mathcal{A}$ be an algorithm that for any target concept $c \in H$ and i.i.d. sample S returns a consistent hypothesis h_S (i.e. its empirical error is zero). Then for any $\epsilon, \delta > 0$, the inequality $\Pr_{S \sim D^m}[R(h_S) \leq \epsilon] \geq 1 - \delta$ holds if $m \geq \frac{1}{\epsilon}(\log |H| + \log \frac{1}{\delta})$. This sample complexity result is equivalent to a generalization bound: for any $\epsilon, \delta > 0$, with probability at least $1 - \delta$, $R(h_S) \leq \frac{1}{m}(\log |H| + \log \frac{1}{\delta})$.*

Although the generalization error of a hypothesis is not directly accessible to the learner since both the distribution D and the target concept c are unknown, the learner can measure the empirical error of a hypothesis on the labeled sample S. Note that the generalization error is measured with respect to D, but when we have test samples that are from a completely unknown distribution π in some class of distributions Π, we should define a notion of epistemic uncertain generalization error as follows.

Definition 2. *Given a hypothesis $h \in H$, a target concept $c \in C$, and an underlying distribution D, the* epistemic uncertain generalization error *of h is defined by $R_e(h) = \sup_{\pi \in \Pi} \Pr_{x \sim \pi}[h(x) \neq c(x)]$.*

The goal is to prove a lower bound similar to Thm. 1 under epistemic uncertainty. One approach may be to convert universal hypothesis testing and anomaly detection lower bounds [27, 28] into statistical learning lower bounds, following [29, 30]. The reason that universal information-theoretic techniques may be useful in settings of epistemic uncertainty is they too consider worst-case performance over $\pi \in \Pi$.

4. CONCLUSION

This paper described an AI safety question that arises in engineering systems design where training data may not capture rare but deleterious phenomena, yet traditional physics-based engineering knowledge exists. As we have suggested in this short conceptual paper, this basic setting introduces several interesting signal processing questions on model aggregation and statistical learning questions from a PAC perspective.

5. REFERENCES

[1] G. Hommel and G. Bernhard, "Bonferroni procedures for logically related hypotheses," *J. Stat. Plan. Inference*, vol. 82, no. 1-2, pp. 119–128, Dec. 1999.

[2] R. Piarroux, R. Barrais, B. Faucher, R. Haus, M. Piarroux, J. Gaudart, R. Magloire, and D. Raoult, "Understanding the cholera epidemic, Haiti," *Emerging Infectious Diseases*, vol. 17, no. 7, pp. 1161–1168, Jul. 2011.

[3] N. Kshetry and L. R. Varshney, "Optimal wastewater management using noisy sensor fusion," in *2017 Arab-American Frontiers of Science, Engineering, and Medicine Symposium*. Rabat, Morocco: National Academies, Nov. 2017.

[4] ——, "Optimal wastewater management using advanced analytics," in *2018 Illinois Wastewater Professionals Conference*, Springfield, IL, Apr. 2018.

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," arXiv:1409.0575 [cs.CV]., Sep. 2014.

[6] J. Attenberg, P. Ipeirotis, and F. Provost, "Beat the machine: Challenging humans to find a predictive model's 'unknown unknowns'," *J. Data Inf. Qual.*, vol. 6, no. 1, Mar. 2015.

[7] T. G. Dietterich, "Steps toward robust artificial intelligence," *A. I. Mag.*, vol. 38, no. 3, pp. 3–24, Fall 2017.

[8] Z. Li and D. Hoiem, "$\mathcal{G}$-distillation: Reducing overconfident errors on novel samples," arXiv:1804.03166 [cs.CV]., Apr. 2018.

[9] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan, "Do deep generative models know what they don't know?" arXiv:1810.09136 [stat.ML]., Oct. 2018.

[10] K. R. Varshney and H. Alemzadeh, "On the safety of machine learning: Cyber-physical systems, decision sciences, and data products," *Big Data*, vol. 5, no. 3, pp. 246–255, Sep. 2017.

[11] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, Eds., *Dataset Shift in Machine Learning*. Cambridge, MA: MIT Press, 2009.

[12] W. J. Scheirer, A. Rocha, R. J. Micheals, and T. E. Boult, "Meta-recognition: The theory and practice of recognition score analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1689–1695, Aug. 2011.

[13] G. D. Forney, Jr., "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 2, pp. 206–220, Mar. 1968.

[14] Q. Li, A. Vempaty, L. R. Varshney, and P. K. Varshney, "Multi-object classification via crowdsourcing with a reject option," *IEEE Trans. Signal Process.*, vol. 65, no. 4, pp. 1068–1081, Feb. 2017.

[15] E. W. Rice, R. B. Baird, A. D. Eaton, and L. S. Clesceri, Eds., *Standard Methods for the Examination of Water and Wastewater*, 22nd ed. American Public Health Association, 2012.

[16] S. Y. Auyang, *Engineering—An Endless Frontier*. Cambridge, MA: Harvard University Press, 2004.

[17] R. R. Kline, "The paradox of "engineering science": A Cold War debate about education in the U.S." *IEEE Technol. Soc. Mag.*, vol. 19, no. 3, pp. 19–25, Fall 2000.

[18] A. Karpatne, W. Watkins, J. Read, and V. Kumar, "Physics-guided neural networks (PGNN): An application in lake temperature modeling," arXiv:1710.11431v2 [cs.LG]., Feb. 2018.

[19] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal outlier hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4066–4082, Jul. 2014.

[20] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.

[21] A. S. Willsky, G. W. Wornell, and J. H. Shapiro, *Stochastic Processes, Detection and Estimation: 6.432 Course Notes*. Massachusetts Institute of Technology, Fall 2002.

[22] E. P. Kim and N. R. Shanbhag, "Statistical analysis of algorithmic noise tolerance," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2013)*, May 2013, pp. 2731–2735.

[23] G. Valiant and P. Valiant, "Estimating the unseen: an $n/log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs," in *Proc. 43rd Annu. ACM Symp. Theory Comput. (STOC'11)*, Jun. 2011, pp. 685–694.

[24] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2835–2885, May 2015.

[25] Y. Han, J. Jiao, T. Weissman, and Y. Wu, "Optimal rates of entropy estimation over Lipschitz balls," arXiv:1711.02141 [math.ST]., Nov. 2017.

[26] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. Cambridge, MA: MIT Press, 2012.

[27] Y. Li, "Universal outlier hypothesis testing with applications to anomaly detection," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2015.

[28] Y. Bu, S. Zou, Y. Liang, and V. V. Veeravalli, "Universal outlying sequence detection for continuous observations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2016)*, Mar. 2016, pp. 4254–4258.

[29] P. Massart and É. Nédélec, "Risk bounds for statistical learning," *Ann. Stat.*, vol. 34, no. 5, pp. 2326–2366, Oct. 2006.

[30] B. Hajek and M. Raginsky, *ECE 543: Statistical Learning Theory*. University of Illinois at Urbana-Champaign, Spring 2018.