Interpreting Intelligibility under Uncertain Data Imputation

Brian Y. Lim, Danding Wang
National University of Singapore
Singapore
brianlim@comp.nus.edu.sg
wangdanding@u.nus.edu

Tze Ping Loh, Kee Yuan Ngiam
National University Hospital
Singapore
{tze_ping_loh, kee_yuan_ngiam}@nuhs.edu.sg

ABSTRACT

Many methods have been proposed to make machine learning more interpretable, but these have mainly been evaluated with simple use cases and well-curated datasets. In contrast, real-world data presents issues that can compromise the proper interpretation of explanations by end users. In this work, we investigate the impact of missing data and imputation on how users would understand, and use explanation features and propose two approaches to provide explanation interfaces for explaining feature attribution with uncertainty due to missing data imputation. This work aims to improve the understanding and trust of intelligible healthcare analytics in clinical end users to help drive the adoption of AI.

Author Keywords
Intelligibility, Explanations, Imputation, Interfaces, Visualization, User Study.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

INTRODUCTION

Intelligibility has been proposed as a capability to enable systems to explain their inner state, reasoning mechanisms and priorities to help users understand and trust them [1, 12]. A recent review has identified that research on explainable systems typically focuses on explanation generation algorithms on systems with well-curated data or based on theory, and explanation interfaces with simple models and small datasets [1]. While some empirical user studies have shown explanations to be effective in well-behaved, albeit simple and synthetic use cases (e.g., [3, 12]), real data and systems face issues and challenges to make data processing and data mining messy. In particular, datasets often have missing data and *imputation* is typically used to estimate the true value of the missing data.

Several methods can be used to impute data, such as substituting with zeros, substituting with mean values of the missing variable, carrying forward (or backward) a nearby observed value, or model-based imputation (e.g., with hidden Markov [17]). For example, if a patient has never been tested for blood calcium (CA), we may assume that his level would be in the healthy normal range and impute the patient's reading as the mean of other patients with normal CA levels. On the other hand, if a patient had a recent high CA level, we may apply carry forward imputation to estimate his current level to be the same.

Data imputation raises a potential problem of how to interpret explanations that depend on data features input into the model. How would a user trust the importance of a feature's value in influencing an inference outcome if the value was not measured, but estimated from imputation? We hypothesize that users will have lower trust in cases of high data imputation and that some visualization methods may help to alleviate this problem.

In this position paper, we discuss the importance of considering how data pre-processing to handle practical issues of real-world data affects the usefulness and interpretation of explanations about machine learning models. We will focus on the use case of disease risk prediction using the structured data of electronic medical records (EMR). EMRs typically contain a lot of missing data, not necessarily due to errors in data collection, but because of the wide variety of tests that patients can take and that patients only take few necessary tests occasionally [17]. For example, a non-diabetic person may not need to measure his HbA1c as frequently as a diabetic, and HbA1c only needs to be measured once every three months.

Specifically, we seek to answer the following research questions:

RQ1. What information will clinicians need to interpret how a clinical decision support system with disease risk prediction makes its decision and how will this change given their awareness that some data was imputed?

RQ2. How can a suitable explanation be generated and presented to clinicians to alleviate the loss of trust in explanations due to imputation?

RQ3. How will the imputation-aware explanation model and interface be interpreted by clinicians and how will it affect their decision making?

APPROACHES: TWO EXPLANATION INTERFACES FOR IMPUTED DATA

While there are several techniques to generate explanations, such as explanations by identifying similar instances [8] or by rule associations [11], we will focus on explanations by additive feature attribution or influence scores (e.g., LIME

ExSS '18, March 11, Tokyo, Japan.

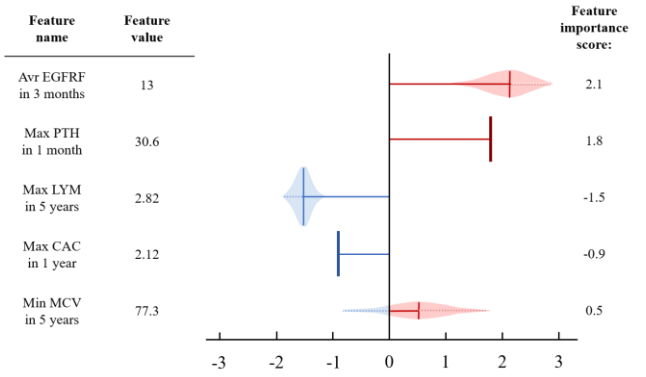

Figure 1. Mockup of feature attribution explanations with uncertainty visualizations. Each horizontal bar chart represents the influence score due to the feature of the row. The vertical line in the bar indicates the influence score calculated by current methods (e.g., LIME [15]), and the shaded region indicates the uncertainty calculated by error propagation due to missing values.

[15], QII [4], GA2M [3]). This explanation style has been popular for generating explanations for healthcare analytics (e.g., Bussone et al. [3], GA2M [3], Prospector [10]).

We propose two approaches to improve user trust in explanations given the increased uncertainty of imputations – based on expressing the uncertainty or hiding uncertain and, hence, confusing information.

Visualizing Uncertainty Distribution of Feature Attribution Scores due to Imputation

Visualizing uncertainty is a well-studied approach in HCI and information visualization to communicate errors and uncertainty to end users [6, 7, 8]. This has been shown to improve user trust and decision making, but may also lead to information overload or compromise trust [8, 13]. We will extend the typical presentation of explanations where each feature, x_i, has an influence score, $f(x_i) = f_i$. With uncertainty due to imputation, the influence score will become $f(x_i + \Delta x_i) = f_i + \Delta f_i$, where Δx_i is the error distribution of feature x_i and Δf_i is the propagated (calculated) distribution in influence score due to the error. The distribution can be calculated by assuming a Gaussian distribution or performing a Monte Carlo simulation on propagated scores based on estimated error. Drawing from various taxonomies evaluated for usability [14], we will present the uncertainty in explanations as a distribution of influence scores in the form of violin plots (see Figure 1). We choose violin plots for their ability to express more detail in a probability distribution than box plots, while also being compact.

De-emphasizing imputed features via Uncertainty Regularization

We exploit the tendency for clinicians to suppress or ignore uncertain data [16]. Therefore, this approach seeks to hide features that have high uncertainty due to imputation.

Feature Regularization is commonly used to simplify and generalize models in machine learning and to reduce overfitting, but we will leverage regularization to penalize features with higher uncertainty. Features with high uncertainty will have reduced influence scores or be hidden. Therefore, the explanation will show adjusted influence scores where some influences are reduced (e.g., horizontal bars shiftwed towards zero), or some features are not shown (influence bars hidden).

For simplicity, we leverage LIME [15] to generate explanations and use the simple linear regression with regularization as the locally approximate explainer model. Training this explainer model involves minimizing the following loss function (simplified for brevity):

$$\xi(x) = arg\min_{g} \mathcal{L}(f, g) + \Omega(g) \quad (1)$$

where, as defined in [15], $\mathcal{L}(f, g)$ is the local fidelity of the explainer model, g, with respect to the model to be explained, f and x is the data instance being explained. $\Omega(g)$ is the measure of complexity (converse of interpretability). We use Lasso regression as is common for simple linear regression with sparsity regularization, so $\Omega(g) = \lambda_1 \|\beta\|_1$. We extend this term to include a penalty for the uncertainty due to imputation, such that

$$\Omega(g) = \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_{\boldsymbol{E}}^2 \quad (2)$$

where β is the explainer model parameters (coefficients in the sparse linear model in our case), λ_1 and λ_2 are hyperparameters to tune the complexity of the explanation, and $\boldsymbol{E}$ is a diagonal matrix where the j^{th} element equals to the uncertainty ε_{0j}^2, and $\|\beta\|_{\boldsymbol{E}} = (\beta^T \boldsymbol{E} \beta)^{1/2}$. Here both sparsity and uncertainty are penalized to increase interpretability. By tuning the two hyperparameter λ_1 and λ_2, we could change the complexity of the explanation with respect to number of features shown and how much to hide or de-emphasize uncertain features.

FUTURE USER EXPERIMENTS: DISEASE RISK PREDICTION USE CASE

We will investigate the impact of missing data on user trust in the explanations with an application use case in predictive healthcare analytics on electronic medical records (EMR). We will specifically focus on diagnosing hyperparathyroidism and recruit clinicians as the target user. We aim to improve their understanding, trust and decision making when using intelligible disease risk prediction. We will conduct two user studies:

Formative user study: to understand the *usability breakdowns* in interpreting explanations given the awareness that some data features are based on data imputations, and user *requirements* for intelligibility. We will present users with several inference instances (i) without explanations, (ii) with explanations, and (iii) with missing data indicated. To understand how users interpret the explanation information and make their decisions, we will have them *think aloud* as they examine several use cases and conduct structured *interviews*. While we already

have hypothesized two approaches to generating uncertainty-aware explanations, with this initial study, we aim to learn more explanation approaches which users may want to better characterize the uncertainty due to imputation and what could be shown to regain their trust.

Evaluative user study: we will implement our two explanation interfaces into diagnostic dashboard prototypes and perform a comparative evaluation with baselines of no explanation and with basic feature attribution explanations (e.g., LIME [15]). We note that the amount of uncertainty can confound the user's level of trust in the system [13]. Therefore, we will control both the system confidence level and amount of imputation in patient cases used in the experiment scenarios. These will be varied as a secondary independent variable. We will measure the accuracy of user diagnosis (correct/wrong with respect to labels from hospital discharge reports), speed of decision (from first viewing patient data to final decision), confidence in diagnosis (7-point Likert scale), trust in the system prediction (7-point Likert scale), and understanding of the patient case (coded from transcribed interviews and think aloud (e.g., see [12, 13]).

CONCLUSION

In this position paper, we have discussed the importance of considering how data pre-processing, specifically data imputation, may compromise the interpretation and trust of explainable AI. We briefly presented two approaches to address the resultant uncertainty by either visualizing the uncertainty or by hiding it. We propose two experiments to understand the impact of missing data on the requirements for explainable AI and to evaluate the efficacy of the proposed solutions.

REFERENCES

1. Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., Kankanhalli, M. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '18.*
2. Bellotti, V., & Edwards, K. (2001). Intelligibility and accountability: human considerations in context-aware systems. *Human–Computer Interaction*, *16*(2-4), 193-212.
3. Bussone, A., Stumpf, S., & O'Sullivan, D. (2015, October). The role of explanations on trust and reliance in clinical decision support systems. In *Healthcare Informatics (ICHI), 2015 International Conference on* (pp. 160-169). IEEE.
4. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015, August). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721-1730). ACM.
5. Datta, A., Sen, S., & Zick, Y. (2016, May). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Security and Privacy (SP), 2016 IEEE Symposium on* (pp. 598-617). IEEE.
6. Jung, M. F., Sirkin, D., Gür, T. M., & Steinert, M. (2015, April). Displayed uncertainty improves driving experience and behavior: The case of range anxiety in an electric car. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 2201-2210). ACM.
7. Kay, M., Morris, D., & Kientz, J. A. (2013, September). There's no such thing as gaining a pound: Reconsidering the bathroom scale user interface. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (pp. 401-410). ACM.
8. Kay, M., Kola, T., Hullman, J. R., & Munson, S. A. (2016, May). When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5092-5103). ACM.
9. Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730*.
10. Krause, J., Perer, A., & Ng, K. (2016, May). Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5686-5697). ACM.
11. Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, *9*(3), 1350-1371.
12. Lim, B. Y., Dey, A. K., & Avrahami, D. (2009, April). Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2119-2128). ACM.
13. Lim, B. Y., & Dey, A. K. (2011, September). Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th international conference on Ubiquitous computing* (pp. 415-424). ACM.
14. Pang, A. T., Wittenbrink, C. M., & Lodha, S. K. (1997). Approaches to uncertainty visualization. *The Visual Computer*, *13*(8), 370-390.
15. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on*

Knowledge Discovery and Data Mining (pp. 1135-1144). ACM.

16. Simpkin, A. L., & Schwartzstein, R. M. (2016). Tolerating uncertainty—the next medical revolution?. *New England Journal of Medicine*, *375*(18), 1713-1715.
17. Zheng, K., Gao, J., Ngiam, K. Y., Ooi, B. C., & Yip, W. L. J. (2017, August). Resolving the bias in electronic medical records. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2171-2180). ACM.