Be Informed and Be Involved: Effects of Uncertainty and Correlation on User's Confidence in Decision Making

Jianlong Zhou
National ICT Australia
Eveleigh NSW 2015, Australia
Jianlong.zhou@nicta.com.au

Constant Bridon
[1] National ICT Australia
Eveleigh NSW 2015, Australia
[2]Ecole Normale Supérieure,
Cachan, France
Constant.bridon@nicta.com.au

Fang Chen
National ICT Australia
Eveleigh NSW 2015, Australia
Fang.chen@nicta.com.au

Ahmad Khawaji
National ICT Australia
Eveleigh NSW 2015, Australia
Ahmad.khawaji@nicta.com.au

Yang Wang
National ICT Australia
Eveleigh NSW 2015, Australia
Yang.wang@nicta.com.au

CHI'15 Extended Abstracts, Apr 18-23, 2015, Seoul, Republic of Korea
ACM 978-1-4503-3146-3/15/04.
http://dx.doi.org/10.1145/2702613.2732769

Abstract
User's confidence in machine learning (ML) based decision making significantly affects acceptability of ML techniques. In this work, we investigate how uncertainty/correlation affects user's confidence in order to design effective user interface for ML-based intelligent systems. A user study was performed and we found that revealing of correlation helped users better understand uncertainty and thus increased confidence in model output. When correlation had the same trend with performance, correlation but not uncertainty helped users more confident in their decisions.

Author Keywords
Model output, uncertainty, correlation, confidence

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User interfaces – Evaluation/methodology

Introduction
With the fast growing use of intelligent systems in different fields, decision making has become an important topic in human-computer interaction (HCI) research. On the other hand, data analytics is becoming one of significant tools to help users understand data, get insights from data and make decisions with the

explosive increasing of data from various fields. Much of machine learning (ML) research is inspired by such expectations. The ML-based data analysis is widely used in decision making in various intelligent system. However, for a domain professional who may not have expertise in ML, an ML algorithm is still a "black-box". The "black-box" approach makes the user difficult to understand complicated ML models. As a result, the user is unconfident in the model output, and consequently in the decision making process.

Therefore, in an ML-based intelligent system, it is highly critical to know how the information from data and ML models presented in the user interface affect user's confidence. Such investigation could help develop more effective user interface for ML-based intelligent systems. Our previous research found that the type, number, and values of decision factors from ML output affect decision difficulties as well as decision qualities [10]. This paper is a natural extension of previous research. In this paper, we investigate what factors affect user's confidence during decision making.

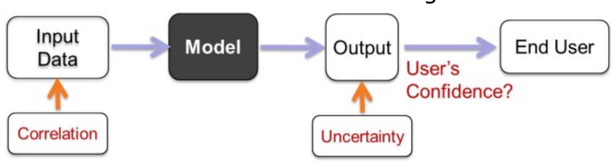

Figure 1.ML-based data analysis pipeline.

Figure 1 shows a typical ML-based data analysis pipeline. Inputs to the mathematical model (ML model) are often historical records or samples of some event. They are usually not the precise description of events. ML models are also imperfect abstractions of reality. Therefore, uncertainty is unavoidable in the prediction given by the model output. The uncertainty or reliability of model output significantly affects effectiveness of ML-based decisions. Furthermore, statistical information of input data such as correlation between variables also provides useful information to let users learn relations between variables. For example, correlation can describe how much target values are related with features in input data of the model. Discarding information such as uncertainty and correlation may lead to disaster, whereas over-conservative safety certification may result in unnecessary economic loss.

This paper aims to investigate relationships between uncertainty/correlation and user's confidence in order to design more effective user interface for ML-based intelligent systems and improve the acceptability of ML techniques. A user study was performed, and physiological signals and behaviors of participants were recorded. Analyses of subjective ratings during task time are conducted to find effects of uncertainty and correlation on user's confidence in decision making.

Related Work

Lee and Dry [6] showed that human's confidence in decision making depends on the accuracy of the advice besides the frequency of the advice. Considering that decisions are often made based on probability judgments of which users are not entirely sure, Hill [5] developed a decision rule incorporating users' confidence in probability judgments.

Decision making under uncertainty is widely investigated in decision theory[3], where uncertainty is usually considered as probabilities in utility functions. Beller et al. [1] showed that the presentation of automation uncertainty helped the automation system receive higher trust ratings and increase acceptance. However, few work is found on investigating the effect of uncertainty of model output on user's confidence.

Furthermore, statistical correlation is often used in feature selection in data analysis [4]. However, few research is done on how correlation affects user's confidence in decision making, especially decision making based on mathematic models learned from historical data. In HCI, previous work focuses on the

Task Group#	Model	Performance	Uncertainty	Correlation	
				T1-T4	T5-T6
1	A	+	+	-	+
	B	-	-	+	-
2	A	+	-	-	+
	B	-	+	+	-

Table 1. Models, performance, and uncertainty configuration for tasks in task groups. "+" means that the value of a condition for the related model is higher than that of the other model with "-" value in the same task.

Task	Uncertainty	Correlation
1	Yes	No
2	Yes	Yes
3	No	No
4	No	Yes
5	Yes	Yes
6	No	Yes

Table 2. Tasks performed by participants in each group**.** "Yes" means that the uncertainty or correlation is available in that task, "No" means that the uncertainty or correlation is not available in that task.

investigation of uncertainty of user's goals during interaction [8], but not uncertainty of output from data analysis models.

In summary, despite the close relations of uncertainty and correlation with decision making, few work is done to investigate how uncertainty and correlation affect user's confidence in model output, and thus affect the acceptance of ML approaches in practical applications.

Hypotheses

The following hypotheses are posed in our study:

- H1: Revealing of correlation between features and target value would give the user a better understanding of uncertainty of model output;
- H2: Uncertainty is the most important factor for user's decision and confidence in ML-based decision making;
- H3: When correlation is provided and uncertainty is not, users are less confident in model output and in their decisions.

Experiment Setup

Case Study and Experiment Data

This research used water pipe failure prediction as a case study. Water supply networks constitute one of the most crucial and valuable urban assets [7]. Utility companies use the outcomes from the failure prediction model to make renewal plan for pipes based on risk levels of failure, and thus also make reasonable budget plan for the pipe maintenance.

Water pipe failure prediction uses pipe failure historical data to predict future failure rate [7]. The data usually contains the failure records of water pipes in a given region. In this paper, different failure prediction models were set up based on one feature of pipes, e.g. material, or laid year. The uncertainty of model output was given for each model. The correlation between one feature of pipes (e.g. material) and the failure rate based on the historical data was also got from the data. Simulated data were used during the experiment.

Task Design

Each participant was told that he/she was supposed to be a manager of a water company. The water company plans to repair X% (the exact number was displayed on each diagram) pipe failures in the next financial year. He/she was asked to make a budget plan based on water pipe failure prediction models learned from the historical pipe failure records. The budget plan needs to meet following requirements:

1) Check as short length of pipes as possible (low cost).
2) The uncertainty interval of the budget should be as small as possible (high accuracy).

Participants were required to do budget plan in each task by reporting following information:

1) The length of pipes to be detected (an expected average value with the uncertainty interval to show whether the prediction is "safe" under the given scenario);

2) The model used for the decision.

We divided tasks into two groups (Task Groups, TG) based on configurations of task conditions as shown in Table 1. In TG1, uncertainty bands (range) of two models in each task were overlapping (e.g. third in Figure 2), while in TG2 they were not overlapping (e.g. bottom in Figure 2). Information for tasks includes model, performance of model, uncertainty, and correlation. The tasks were designed to cover every possible combination between performance (high, low), correlation (not given, high low), and uncertainty in model output (not given, high, low). The resulting tasks are defined in Table 2, where "T1,…" refers to different tasks as shown in Table 2. In each group, six tasks were set up with controlled conditions as shown in Table 1. Figure 2 shows examples of the presentation of correlation, performance curve of model output, and

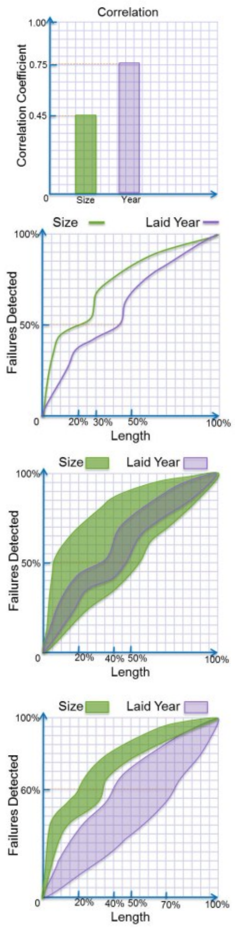

Figure 2. Presentation of different information: correlation (top), performance curve of model output (second), performance of model output with non-overlapped uncertainty (third) and overlapped uncertainty (bottom).

uncertainty used in the tasks. In each task, two models was provided and different information as shown in Table 1for both models was presented to participants based on task configurations. Participants were required to do a budget plan based on given information. There were 27 tasks performed all together by each participant: two six-task groups for two rounds, and three training tasks.

At the beginning of each decision making task, a blank screen was displayed for 6 seconds in order to allow the participant have a rest and "reset" his/her cognitive load state[9]. Then the participants started a task and diagrams with various conditions were displayed. Participants were told that they were competing against other people to reach the best budget plan in a given time period (1.5 minutes/task) in order to push participants to make their efforts for tasks.

Participants and Apparatus

26 participants were recruited from three groups, with the range of ages from twenties to forties: 9 ML researchers, 8 non-ML researchers, and 9 administrative staff. Of all participants, 9 were females.

GSR and BVP devices from ProComp Infiniti of Thought Technology Ltd were used to respectively collect skin conductance responses and blood volume pulse of subjects.. Different diagrams of model performance, uncertainty and correlation were presented on a 21-inch Dell monitor with a screen resolution of 1024 by 768 pixels.

Data Collection

After each decision making task, participants were asked to rate the confidence level in the model output and thus the budget plan they made using a 9-point Likert scale (1=least confident, and 9=most confident). Besides, GSR, BVP signals, and mouse movement were collected during task time.

Preliminary Analyses

Statistical analyses with Wilcoxon test on subjective ratings were performed to test our hypotheses because we are comparing correlated data. Decision criteria used by participants and decision results were also analyzed in our hypotheses testing.

H1This hypothesis was directly tested based on the subjective ratings in Task 2 and Task 5. Users seemed to better understand uncertainty when correlation and uncertainty had the opposite trend, i.e. when the correlation of a feature was high and the associated model uncertainty was low and vice versa. Even for the worst rating (Task 2 in TG2), correlation was rated useful by users (M = 5.8, SD = 1.41). However, the ratings seemed to be significantly different between Task 2 and Task 5 in a TG (TG1: Z=56.0, p<.001, and TG2: Z= 104.0, p<.001). It emphasizes the fact that a given user did not handle this question the same way for both Tasks. Then, a better explanation of the link between correlation and uncertainty could lead to better confidence in the model output. In conclusion, users found correlation was statistically helpful for every possible scenario as we hypothesized.

H2In order to test hypothesis H2, the Task 1 of each task group was taken as the ground truth and was pair-wise compared to Task 2 and Task 5 respectively.

In TG1, for Task 1, i.e. with uncertainty being provided only, "low uncertainty" was used as the dominant decision reason by 53% of participants and the mean confidence rating was M = 6.38, SD = 1.52. However, for both Task 2 and Task 5, the dominant decision reason was "high correlation", used by 55% and 58% of all participants respectively, with an equivalent or even better confidence ratings (Task 2: M = 6.94 and SD = 1.45 and Z= 137.5, p=.003, Task 5: M = 6.35and SD = 1.68 and Z= 376, p=.841). It means that users switched their decisions between Task 1 and Task 5 but

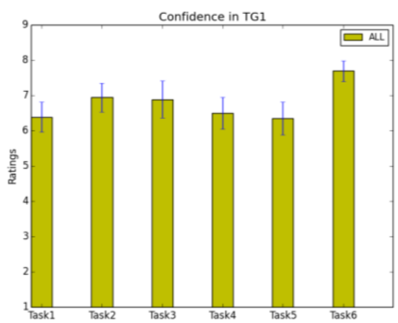

Figure 3.Average subjective ratings of participants' confidence in model output in TG1.

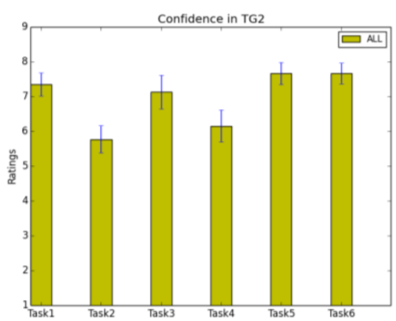

Figure 4. Average subjective ratings of participants' confidence in model output in TG2.

had a strongly similar rating. Between Task 1 and Task 2, decision was the same but confidence was improved, and between Task 1 and Task 5, confidence was the same but decision was different. Therefore, correlation positively affected both decision and confidence, overcoming uncertainty. In TG2, for Task 1, 2 and 5, "high performance" was used as the dominant criteria for decision making, by 62%, 51% and 48% of participants respectively. Users' confidence was significantly higher in Task 1 (M = 7.34, SD = 1.20) than in Task 2 (M = 5.78, SD = 1.41) (Z=44.5, p<.001). User's confidence was also significantly higher in Task 5 (M = 7.67, SD = 1.13) than in Task 1 (Z= 116.5, p=.023). It shows that the decision was strongly affected by performance, and confidence was mainly affected by the trends of correlation and performance. The only positive impact occurred with the case when correlation and performance shared the same trend. And in every scenario, uncertainty was irrelevant to users' confidence.

In summary, uncertainty was not the most dominant criteria of decision making especially when correlation was provided. Performance was also significant, and the most impacting factor was the couple correlation - performance, contrary to what we hypothesized.

H3 In this testing, Task 3 of each task group was taken as the ground truth, and was compared pair-wise to Task 4 and Task 6 respectively.

In TG1, for Task 3, users' confidence was high (M = 6.88,SD = 1.88), and "high performance" was chosen as the dominant decision criteria by all participants. However, 38% of participants decided to switch their decision in Task 4 and therefore used "high correlation" as the criteria for decision making, even if performance and correlation had opposite trends in this scenario. The associated confidence for Task 4 was slightly lower (M = 6.5, SD = 1.63), and significantly different (Z=217.5, p= .023). However, when correlation and performance had the same trend (Task 6 scenario), users' confidence was improved significantly (M = 7.69, SD = 1.05) compare to Task 4(Z=105.0, p<.001). The comparison with task 3 has the same result (Z=167.5, p<.01). It shows that correlation affected decision when it had an opposite trend to performance and slightly lowered confidence ratings. But when correlation and performance shared the same trend, decision remained the same and confidence was significantly improved.

In TG2, we got similar results than in TG1; the differences between decision results and confidence ratings were quantitatively more pronounced and followed the same conclusion.

In summary, if correlation was provided to participants and uncertainty was not, users' confidence and decision were affected. When correlation and performance had opposite trend, decision statistically changed and confidence lowered, as we expected. However, when correlation and performance shared the same trend, decision remained the same and confidence significantly improved, contrary to our hypothesis.

Discussion and Ongoing Work

This study found that: 1) Correlation did help users better understand uncertainty and thus increase users' confidence as we expected, but for a given user, this confidence could be improved with a better explanation of the link between correlation and uncertainty; 2) Uncertainty was not the dominant criteria for decision making. Moreover, users changed their initial decisions when they were given additional information thus increasing their confidence, both in their resulting decision and in model output; 3)Users tended to be more confident in their decisions and in model outputs when correlation shared the same trend with performance. This could be because of the "grounding communication" referred to by psychologists [2]. Because of grounding, confidence in model output is resolved through a drive towards a mutual understanding or common ground (correlation had the same trend with performance) in the process.

Our current work focuses on analyzing physiological signals (GSR and BVP) as well as behavioral signals (mouse movement) of participants for confidence classifications during ML-based decision making: 1) We investigate what physiological features can be used to index user's confidence levels during decision making. 2) We also investigate how different information such as uncertainty and correlation affect patterns of physiological signals and thus user's confidence levels, in order to design effective user interface for ML-based intelligent systems. 3) User's behavioral signals are also analyzed to find relations between confidence and user behaviors. Our ultimate goal is to set up a framework of measurable confidence in decision making in order to dynamically update confidence levels in ML-based intelligent systems.

Summary

This work investigated the effect of uncertainty and correlation on user's confidence in model output in ML-based decision making. It was found that correlation played more significant roles than uncertainty in helping users more confident in ML-based decision making. The proposed work can be used in confidence-aware user interface design in ML-based intelligent systems. The future work focuses on analyzing physiological signals and behavioral signals of participants in indexing confidence in ML-based decision making.

Acknowledgements

This work is partly supported by AOARD under grant No. FA2386-14-1-0007 AOARD 134144 and FA2386-14-1-0022 AOARD 134131.

References

1. Beller, J., Heesen, M., and Vollrath, M. Improving the Driver–Automation Interaction An Approach Using Automation Uncertainty. *Human Factors: The Journal of the Human Factors and Ergonomics Society 55*, 6 (2013), 1130–1141.
2. Clark, H.H. and Brennan, S.E. Grounding in communication. In *Perspectives on socially shared cognition*. American Psychological Association, 1991, 127–149.
3. Damghani, K.K., Taghavifard, M.T., and Moghaddam, R.T. Decision Making Under Uncertain and Risky Situations. *Enterprise Risk Management Symposium Monograph Society of Actuaries-Schaumburg, Illinois*, (2009).
4. Guyon, I. and Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res. 3*, (2003), 1157–1182.
5. Hill, B. Confidence and decision. *Games and Economic Behavior 82*, C (2013), 675–692.
6. Lee, M.D. and Dry, M.J. Decision making and confidence given uncertain advice. *Cognitive Science 30*, 6 (2006), 1081–1095.
7. Li, Z., Zhang, B., Wang, Y., et al. Water Pipe Condition Assessment: A Hierarchical Beta Process Approach for Sparse Incident Data. *Machine Learning 95*, 1 (2014), 11–26.
8. Salovaara, A. *Proactive Reasoning in Mobile HCI: Tackling Uncertainty and Appropriate Task Support.* Helsinki Institute for Information Technology, 2004.
9. Wang, W., Li, Z., Wang, Y., and Chen, F. Indexing Cognitive Workload Based on Pupillary Response Under Luminance and Emotional Changes. *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, ACM (2013), 247–256.
10. Zhou, J., Sun, J., Chen, F., et al. Measurable Decision Making with GSR and Pupillary Analysis for Intelligent User Interface. *ACM Transactions on Computer-Human Interaction*, 21, 6 (2015), 33.