J Multimodal User Interfaces
DOI 10.1007/s12193-017-0249-8

ORIGINAL PAPER

DecisionMind: revealing human cognition states in data analytics-driven decision making with a multimodal interface

Jianlong Zhou[1] · **Fang Chen**[1]

Received: 22 November 2016 / Accepted: 14 September 2017

Abstract Despite the recognized value of machine learning (ML) techniques and high expectation of applying ML techniques within various applications, significant barriers to widespread adoption and local implementation of ML approaches still exist in the areas of trust (of ML results), comprehension (of ML processes) and related workload, as well as confidence (in decision making based on ML results) by users. This paper argues that the revealing of human cognition states with a multimodal interface during ML-based data analytics-driven decision making could provide a rich view for both ML researchers and domain experts to learn the effectiveness of ML technologies in applications. On the one hand, human cognition states could help understand to what degree users accept innovative technologies. On the other hand, through understanding human cognition states during data analytics-driven decision making, ML-based decision attributes and even ML models can be adaptively refined in order to make ML transparent. The paper also identifies examples of impact challenges and obstacles, as well as high-demand research directions in making ML transparent.

Keywords DecisionMind · Decision making · Human cognition states · Multimodal interface · Transparent machine learning

✉ Jianlong Zhou
jianlong.zhou@data61.csiro.au

Fang Chen
fang.chen@data61.csiro.au

[1] DATA61, CSIRO, Australia, 13 Garden Street, Eveleigh, NSW 2015, Australia

1 Introduction

With the rapid advancement of technologies in hardware and software, the boundary between human and technology is blurring. On the one hand, human is seamlessly involved in various interactions with technologies. On the other hand, technologies get different inputs from human and provide personalized feedback to human for various purposes such as decision making. Human factors are becoming indispensable components in technologies. Taking the recent booming data science as an example, with the rapid increase of data from different fields such as biology, finance, medicine, infrastructure, and society, users are looking to integrate their “Big Data” and advanced analytics into business operations in order to become more analytics-driven in their decision making. Much of Machine Learning (ML) research is inspired by such expectations. Various ML algorithms offer a large number of useful ways to approach those problems that otherwise require cumbersome manual solution.

Despite the recognized value of ML techniques and high expectation of applying ML techniques within various applications, users often find it difficult to effectively apply ML techniques in practice because of complicated interfaces between ML algorithms and users. To this end, most of previous work focuses on the use of visualization to depict ML process or represent ML results. Various visualization techniques can help users understand and/or interact with ML models effectively in some degree. However, they cannot provide cues on users’ cognition states such as how confident users are in ML models or decisions based on ML results. Significant barriers to widespread adoption and implementation of ML approaches still exist in the areas of trust (of ML results), comprehension (of ML processes) and related workload, as well as confidence (in recommended courses of action or decision making) by users. As a result, the User

Published online: 03 October 2017

Experience involved in real-world ML applications has been more recently identified as an area requiring research and development [1–3].

Furthermore, in most cases, to drive or improve decision making is the ultimate goal of real world ML applications [4]. Decision making has become an important topic in various areas of Human-Computer Interaction (HCI) research in recent years. And nonverbal information such as neurophysiological information is increasingly parsed and interpreted by computers to interactively construct and refine models of human's cognitive and affective states in HCI [5]. Such user's models can then be used in an adaptive fashion to enhance HCIs and make interfaces appear intelligent [6]. Therefore, the use of neurophysiological measurements to human cognition in decision making promises to provide a rich and enduring approach in building intelligent HCI systems, which adapt to users' behaviour and their decision making performance. Besides neurophysiological information, research found that human behaviour can also reflect human's mental state, such as cognitive workload and trust [7]. Imagine a computer interface that could predict and diagnose whether a decision made by a user corresponded to a high trust level and with a high confidence, by simply collecting a variety of neurophysiological information from the user. Further imagine that the interface could adaptively vary decision attributes during decision tasks to improve the decision quality, and thus resulting in the improvements of acceptance of ML models due to using these diagnoses. As a result, a decision making solution for the use of ML technologies by incorporating users' cognition states would improve both impact of ML technologies and users' motivation for the use of ML technologies.

Therefore, we strongly argue that the revealing of human cognition states with a multimodal interface during data analytics-driven decision making could provide a rich view for both ML researchers and domain experts to learn the effectiveness of ML-based intelligent systems. On the one hand, human cognition states could help understand in what degree users accept innovative technologies. On the other hand, through understanding human cognition states during data analytics-driven decision making, ML-based decision attributes and even ML models can be adaptively refined in order to make ML understandable and useable by users. The current ML-based data analytics-driven decision making systems do not take the human cognition states into consideration, which greatly affects the impact of ML technologies in real-world applications.

This paper demonstrates the link between human cognition states and ML technologies with a multimodal interface during data analytics-driven decision making. Human cognition state is integrated into the data analytics-driven decision making process. In this paper, a framework of informed decision making is proposed to demonstrate how human's behaviour and physiological signals are used to reveal human cognition states in data analytics-driven decision making. The framework aims to combine the best techniques from cognitive science, HCI, and machine learning to build powerful general-purpose tools and guidelines for data analytics-driven decision making. It helps make ML technologies more acceptable and understandable by domain users. The research is named as Transparent Machine Learning (TML) in our context. TML aims to translate ML into impacts by allowing domain users understand ML-based data-driven inferences to make trustworthy decisions confidently based on ML results, and letting ML accessible by domain users without requiring training in complex ML algorithms and mathematical concepts [1,8]. This paper focuses on the revealing of human cognition states in data analytics-driven decision making from two perspectives: physiological perspective and behavioural perspective.

In summary, the overall objectives of this study include:

- Demonstrate the link between human cognition states and ML research with physiological and behavioural signals in data analytics-driven decision making scenarios;
- Propose a framework of informed decision making to reveal human cognition states in data analytics-driven decision making towards addressing the link between human and ML technologies;
- Identify relevant challenges to human cognition states revealing and identify key research directions in human cognition states revealing in data analytics-driven decision making.

2 Physiological signals and human cognition

Extensive research has found the physiological and behavioural correlations to human cognitive states. This section reviews physiological indicators for human cognition such as cognitive load, trust, as well as confidence in decision making.

2.1 Physiological indicators and cognitive load

Moll et al. [9] reviewed evidence on brain regions identified during functional imaging of cognition activities irrespective of task constraints. It was demonstrated that the investigation of mechanisms of cognition–emotion interaction and of the neural bases is critical for understanding of the human cognition. van Gog et al. [10] used an interdisciplinary approach combining evolutionary biological theory and neuroscience within a cognitive load theory framework [11] to explain human's behaviour during observational learning. Human neurophysiological signals are also used to measure cognitive load, e.g. heart rate and heart rate variability, brain activity (e.g. changes in oxygenation and blood volume, electroencephalography (EEG)), galvanic skin response (GSR), and eyes [12].

Despite the wealthy investigation of cognitive load with the use of neurophysiological signals, little work is done on how ML results or ML models affect cognitive load during data analytics-driven decision making. Such investigation would help understand inherent cognitive load factors caused by ML in order to design cognitively effective ML models and intelligent systems.

2.2 Physiological indicators and trust

Trust is a critical social process that helps people to cooperate with others or systems and is included almost in all human–human and human–machine interactions. Krueger et al. [13] investigated neural responses for trust activities with hyperfunctional magnetic resonance imaging. The results showed that the brain paracingulate cortex (dmPFC) is critically involved in building a trust relationship by inferring another person's intentions to predict subsequent behaviour. Aimone et al. [14] investigated the neural signature of trust. The results showed that the anterior insula modulates trusting decisions that involve the possibility of betrayal. Hahn et al. [15] showed that a person's initial level of trust is determined by brain electrical activity acquired with EEG.

All these works motivate us to investigate human cognition during data analytics-driven decision making in order to understand how ML results and ML models affect human cognition.

2.3 Physiological indicators and decision making

Heekeren et al. [16] reviewed findings from human neuroimaging studies in conjunction with data analysis methods that can directly link decision making and signals in the human brain. Smith et al. [17] used fMRI to investigate the neural substrates of moral cognition in health resource allocation decision making. White et al. [18] investigated the neurophysiological correlates of confidence and uncertainty by means of fMRI. Much work has also been done on using physiological responses such as pupil dilation and skin conductance to understand human's decision making process. For example, a recent investigation [19] shows that the pupil dilation increases over the course of the decision making. Pupil dilation and GSR are also used to index confidence and decision quality in decision making [3].

Because of the neuralphysiological correlation to decision making, it is possible that certain choices in decision making can be predicted/manipulated by monitoring/manipulating specific neurons [20]. These observations motivate us to investigate the effect of ML results and ML models on users' confidence by examining users' cognition states. Such study would help improve decision efficiency in data analytics-driven decision making.

3 Human behaviour

This section highlights the connection between human behaviour and human cognition states in order to initiate the indication of human cognition changes in data analytics-driven decision making.

3.1 Behaviour and cognitive load

In cognitive load research, response-based behavioural features are defined as those that can be extracted from any user activity that is predominantly related to deliberate/voluntary task completion, for example, eye-gaze tracking, mouse clicking, digital pen input, gesture input or any other kind of interactive input used to issue system commands. For instance, Gütl et al. [21] used eye tracking to observe subjects' learning activities in real-time by monitoring their eye movements for adaptive learning purposes. Others have used mouse clicking and keyboard key-pressing behavior to make inferences about their emotional state and adapt the system's response accordingly [22]. It was also found that dialogue behaviour can be used to index cognitive load and the higher level features, such as linguistic and grammatical features, may also be extracted from user's spoken language for patterns that may be indicative of cognitive load [23].

3.2 Behaviour and trust

There are different types of behaviours that people can show when they interact with others to complete a particular task, for instance, they can cooperate or they can compete. The behaviours of cooperative and competitive individuals have a significant impact on the degree of trust which exists with their partner [7]. In addition, features of mouse movement behaviour such as movement distance, slope, and movement count also show different patterns under different trust conditions during a task [24]. Research also suggested that eye movements such as duration, sequence, and frequency of fixations can be used as indicators of trust [25].

3.3 Behaviour and decision making

Much work has been done on using behavioural information such as eye movement to understand human's decision making process [26]. Fiedler and Glockner [27] utilized eye-tracking to analyze dynamics of decision making in risk conditions. It shows that attention to an outcome of a gamble increases with its probability and its value and that attention shifts toward the subsequently favored gamble after two thirds of the decision process, indicating a gaze-cascade effect.

Since human behaviour can reflect human's cognition states, it is useful to investigate how ML results and ML models affect human behaviour and so forth human cognition states in order to make ML more acceptable by users in data analytics-driven decision making.

4 Informed decision making

As reviewed in the previous sections, human trust, cognitive load as well as decision making closely correlate to physiological and behavioural signals. However, little research is found to investigate the effects of ML results or ML models on human cognition states in ML-based data analytics-driven decision making. Human cognition states such as trust, cognitive load, and confidence play indispensable roles for effective data analytics-driven decision making. This section proposes *DecisionMind* as a framework of informed decision making to incorporate human cognition states into the data analytics-driven decision making scenario.

4.1 DecisionMind

Figure 1 shows the loop of data analytics-driven decision making with the consideration of human cognition states in the loop. As shown in Fig. 1, when an ML-based intelligent system is used for decision making, a user usually has a mental model on decisions firstly. The user then makes decisions based on different cues including his mental model. At the same time, human cognition during decision making is evaluated and is used as feedback in order to refine the decision making.

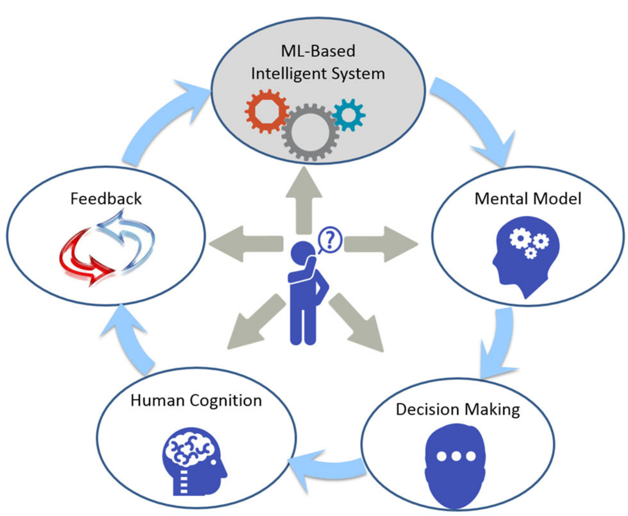

Fig. 1 The loop of data analytics-driven decision making with human factors

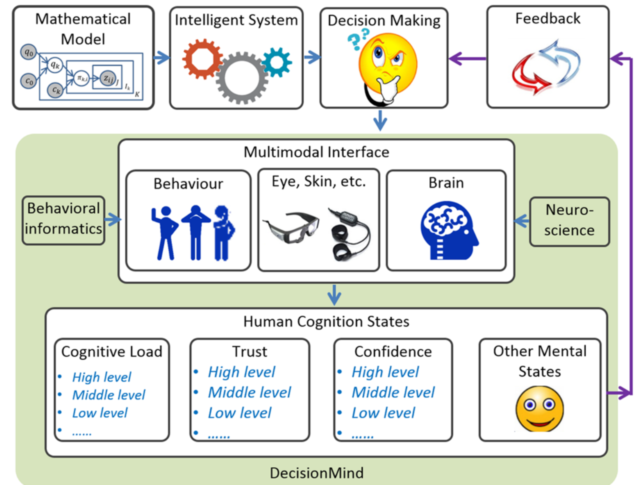

Fig. 2 Framework of informed decision making—DecisionMind

Based on this decision loop, we present a framework of informed decision making—DecisionMind (see Fig. 2). In this framework, when a user makes decisions with an ML model-based intelligent system, signals related to human cognition states are recorded at the same time with different modalities. Human cognition states during decision making are then derived from the recorded signals. Mental model during decision making is also analysed to find how it matches the final decisions. If the user's cognition (e.g. cognitive load, trust, confidence) is not in an acceptable state and the user is not satisfied with the decision quality, a feedback is sent back to the decision system to refine decision attributes and even ML models and a new decision process is started until the user satisfied with the decision performance with appropriate cognition states. During this informed decision making process, user's cognition is tracked and revealed explicitly to help the user refine decisions. DecisionMind therefore evaluates human cognition and allows human cognition quantitatively visible during data analytics-driven decision making. Imagine that a user perceives his cognition states during decision making, and further imagine that the decision attributes and even ML models could adaptively refined based on the estimated cognition states.

5 Case study

This section presents a case study of user confidence in data analytics-driven decision making to demonstrate the effectiveness of the proposed framework. An experiment is set up and physiological signals of GSR and Blood Volume Pulse (BVP) are collected to analyze human cognition states during data analytics-driven decision making.

5.1 Experiment

Water pipe failure prediction was used as a case study for this research [28]. Water supply networks constitute one of the most crucial and valuable urban assets. Identifying an accurate predictive measure for imminent failure of water pipes would allow utility companies to prioritize preventive repairs that would cost significantly less than full-scale failures. Thus, utility companies use outcomes from failure prediction models, to make renewal plans based on risk levels of pipes and also reasonable budget plans for pipe maintenance.

Water pipe failure prediction uses historical pipe failure data to predict future failure rate [28]. The data contain failure records of water pipes, and various attributes of water pipes, such as laid year, length, diameter size, surrounding soil type, etc. In this study, predictive models are simulated and they are based on different pipe features (e.g. size or laid year) with the reference of Hierarchical Beta Process (HBP) used in water pipe failure prediction [28]. The model performance curve was presented to let the participants evaluate different models. The model performance is the functional relationship between proportion of the network inspected and the proportion of pipe failures detected. Figure 3 shows the performances of two models. For example, in Fig. 3a, the model based on the feature "Size" has better performance than the one based on the feature "Laid Year", because the formal one detects more failures than the latter for a given pipe length. The uncertainty of model performance is displayed as shaded area (see Fig. 3b and c) over the thin prediction line without uncertainty for each model (Fig. 3a). Here "uncertainty" refers to an interval within which the true value of a measured quantity would lie with a given probability. This experiment is set up to determine what criteria of choice are in favor of a model, and what parameters such as uncertainty conditions influence the user confidence during the decision process.

Three groups of tasks with no uncertainty presentation (Control task), overlapping uncertainty (OLUT task) and non-overlapping uncertainty (Non-OLUT task) were conducted by participants during the experiment. The order of tasks were randomized. The nature of task was on-screen budget estimation with expected variation to be noted as upper and lower limits.

26 participants were recruited with the range of ages from twenties to forties and an average age of 30 years. Of all participants, 9 were females. Educational qualifications were largely postgraduate (13 PhD, 6 Masters, 4 Bachelors, 3 other).

GSR and BVP (Blood Volume Pulse) devices from ProComp Infiniti of Thought Technology Ltd were used to collect skin conductance responses and BVP signals of subjects respectively. BVP measures the blood volume in the skin capillary bed in the finger with photoplethysmography (PPG)

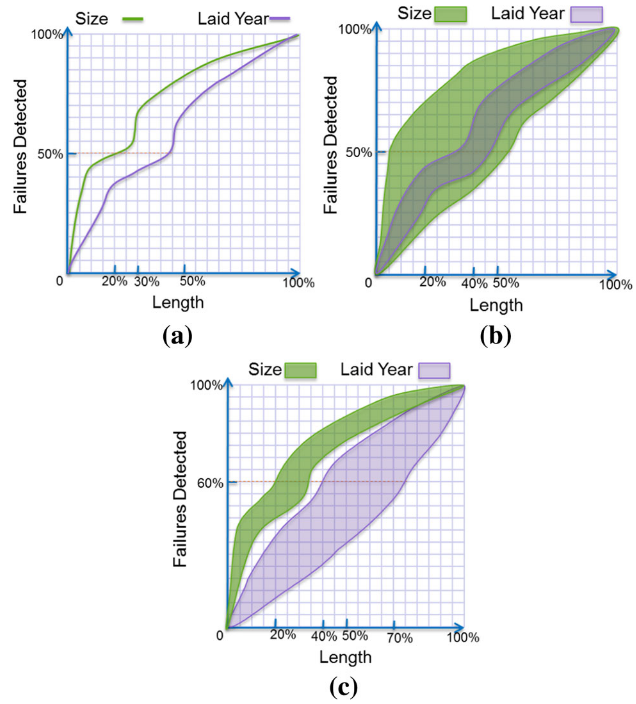

Fig. 3 Performance of predictive models: **a** without uncertainty, **b** with overlapping uncertainty, **c** with non-overlapping uncertainty

in BVP sensors [29], reflects the emotional state of humans. BVP is often used as an indicator of affective processes and emotional arousal. GSR and BVP sensors were attached to subjects' left hand fingers. Besides, after each decision making task, participants were asked to rate the confidence level of the decisions they made using a 9-point Likert scale (1: least confident, and 9: most confident). All participants were right-handed. Different tasks were presented on a 21-in. Dell monitor with a screen resolution of 1024 by 768 pixels.

This section analyzes two modalities of GSR and BVP for user confidence in data analytics-driven decision making.

5.2 Subjective ratings of user confidence

Figure 4 shows average subjective ratings of participants' confidence in decision making tasks under different conditions. A Friedman test showed that there was a statistically significant difference among the three tasks in confidence levels, $\chi^2(2) = 13.481$, $p < .001$. The post-hoc Wilcoxon tests with a Bonferroni correction applied resulting in a new significance level set at $p < .017$ ($0.05/3 = 0.017$ because we have three conditions/tasks) was then applied to find pairwise differences between tasks.

It was found that users were significantly more confident in Non-OLUT task than in OLUT task ($Z = 79.0$, $p < .001$). The result suggests that when uncertainty was presented to users, non-overlapping uncertainty made users more confident in decision making than overlapping uncertainty as we

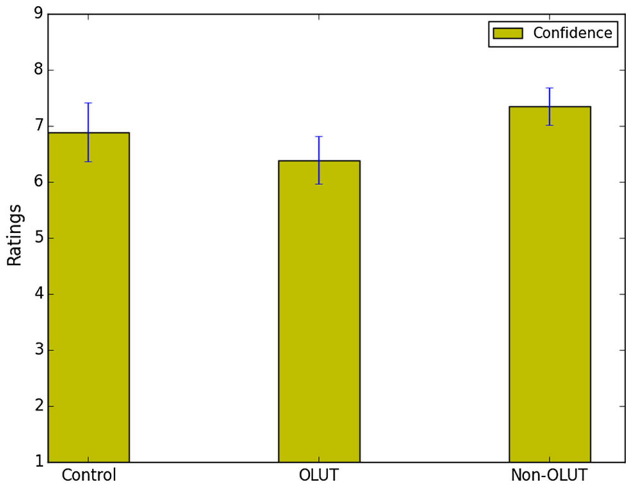

Fig. 4 Average subjective ratings of participants' confidence in decision making in uncertainty-based tasks

expected. However, we did not find significant differences between Control task and OLUT task or between Control task and Non-OLUT task as we expected.

5.3 GSR for user confidence

In this section, GSR responses from subjects are analyzed. Figure 5 shows an example of GSR signals of a participant in one task session. Various features are firstly extracted from GSR signals. GSR features are then used to classify confidence levels in order to show the potential of using GSR in indexing user confidence in decision making. The GSR data analysis is divided into following steps: (1) data calibration, (2) signal smoothing, (3) extrema detection, (4) feature encoding, (5) feature significance test, (6) confidence level classification.

5.3.1 GSR features

This subsection shows the steps to extract and encode GSR features in this study. The 6-s GSR values before the task start time are used to calibrate GSR during the task time in order to compensate the differences between tasks of a subject. A Hann window function [30] is then convoluted to GSR signals to remove noises. The smoothed signal is also normalized using Z-Normalization to omit subjective differences between various signals before the feature extraction.

Both statistical features and extrema-based features [31] are extracted and analysed. These features include (see Fig. 5):

- Mean of GSR (summation of GSR values over task time divided by task time) $\mu_{\mathcal{G}}$;
- Variance of GSR $\sigma_{\mathcal{G}}$;

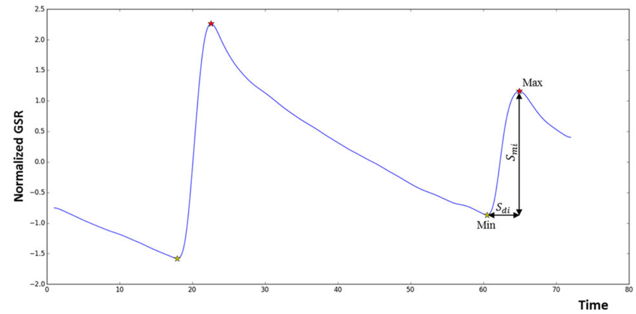

Fig. 5 Extremas and extrema features of GSR

- Number of responses S_f, which is the number of peaks in a GSR signal;
- Sum of duration $S_d = \sum S_{di}$;
- Sum of magnitude $S_m = \sum S_{mi}$;
- Sum of estimated area $S_a = \sum S_{ai}$.

S_f, S_d, S_m, and S_a are features of the GSR orienting response [31]. The definition of magnitude S_{mi} and duration S_{di} are defined as shown in Fig. 5. The area of response is estimated by $S_{ai} = \frac{1}{2} S_{mi} S_{di}$.

5.3.2 GSR feature significance test

In this subsection, one-way ANOVA tests with post-hoc analysis using t-tests were performed to evaluate confidence discrimination of features among different tasks.

An ANOVA test found that features of S_d ($F(2, 39) = 3.817$, $p = .029$), S_m ($F(2, 39) = 3.539$, $p = .036$), and S_a ($F(2, 39) = 4.52$, $p = .016$) showed statistically significant differences among three tasks (two uncertainty-based tasks plus control task). Post-hoc analysis with t-tests were then conducted with a Bonferroni correction (significance level set at $p < .017$ as discussed in the previous section) for all pairwise differences of significant features. There were no significant differences found between tasks. Furthermore, we used a readjusted significance alpha level of 0.025 (0.05/2 by considering actual two conditions of with/without uncertainty revealing) to see if we can find any other pair-wise differences that we expected. Using this new alpha level, the results showed that OLUT task had significantly higher S_d than both Control task ($t = 2.353$, $p = .024$) and Non-OLUT task ($t = 2.396$, $p = .022$).

The results suggest that overlapping uncertainty made features such as S_d values increased significantly. Therefore, less confidence level tasks made GSR feature S_d values significantly higher. These findings suggest that GSR features can be used to index user confidence in decision making tasks.

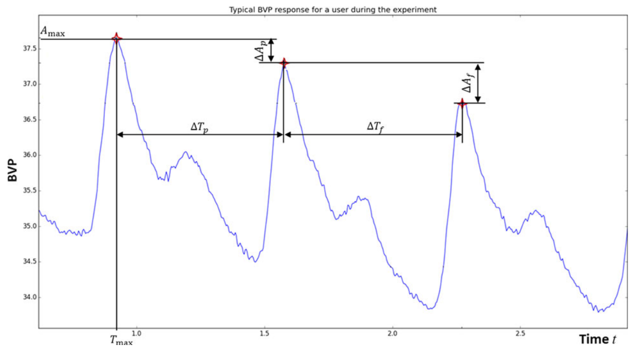

Fig. 6 An example of the BVP signal and the features of the signal

5.4 BVP for user confidence

5.4.1 Analysis of BVP signals

A typical BVP signal collected during the experiment is given in Fig. 6. Three beat intervals are displayed in this figure: the intermediary maxima are referred to dicrotic notches. As a periodical signal, BVP is associated to three frequency bands: Very Low Frequency (VLF) (0.00–0.04 Hz), Low Frequency (LF) (0.05–0.15 Hz), and High Frequency (HF) (0.16–0.40 Hz). The time length of BVP signal in this study, i.e. the task time length, is too short (< 5 min) to consider VLF activity. On the other hand, LF band reflects sympathetic activity and HF band is related to parasympathetic activity [32]. Kristal-Boneh et al. [33] showed that when experiencing mental stress, the sympathetic activity of the heart increases whereas the parasympathetic activity decreases. Therefore the ratio of LF/HF in power can be used as stress indicator. It is the ratio of the signal power in LF band over the signal power in HF band.

The feature extraction process of BVP signals is divided into the following steps: (1) signal normalization, (2) signal smoothing, (3) extrema detection, and (4) feature encoding.

We observed that BVP was highly subjective, and it differs from person to person. Therefore, BVP signals are normalized using Z-Normalization to compensate subjective differences between various signals before the extrema detection:

$$S_N = \frac{S - \mu}{\sigma} \tag{1}$$

where μ and σ are mean and variance of the BVP signal respectively, S and S_N are the original and normalized BVP signals respectively.

The normalized signal S_N is then smoothed with a Hanning window [30]. The window size, which behaves as a cut off frequency, is chosen because of the maximal admissible heart frequency in a normal situation, namely 200 beats/minute [34,35]. This filtering is efficient to remove the dicrotic notches from the original signal, because they are not part of stress reaction [29].

Extremum detection is then performed on the smoothed signal (labeled as red star in Fig. 6). The extrema contained in LF and HF frequency band width are marked with a red star in Fig. 6. Because of noise, 396 out of 624 signals were analyzed and more than 16,000 extrema data points were obtained in this study.

Two types of BVP features were extracted in this study:

- *Summary features* summary features include: (1) mean of BVP (μ_{bvp}), (2) variance of BVP (σ_{bvp}), and (3) LF/HF ratio (R_{LH}).
- *Dynamic features* dynamic features are defined with extrema points as shown in Fig. 6. They are: (4) BVP amplitude at each extrema point (A_{max}), (5) normalized time at each extrema point (T_{max}), which is the time period at the extrema point divided by the task time length, (6) delta time precedent (ΔT_p), (7) delta time following (ΔT_f), (8) delta amplitude precedent (ΔA_p), and (9) delta amplitude following (ΔA_f).

5.4.2 BVP feature significance test

In this subsection, one-way ANOVA tests with post-hoc analysis using t-tests were performed to evaluate confidence discrimination of features among different tasks.

An ANOVA test found that features of T_{max} ($F(2, 2304) = 30.009$, $p < .000$), A_{max} ($F(2, 2304) = 4.530$, $p = .011$), ΔT_p ($F(2, 2304) = 6.694$, $p = .001$), and ΔT_f ($F(2, 2304) = 7.574$, $p = .001$) showed statistically significant differences among three tasks. Post-hoc analysis with t-tests were then conducted with a Bonferroni correction (significance level set at $p < .017$ as discussed in the previous section) for all pairwise differences of significant features. The post-hoc tests showed that Control task took significantly smaller T_{max} than both Non-OLUT task ($t = 4.980$, $p < .000$) and OLUT task ($t = 7.954$, $p < .000$) respectively. It was also found that Control task took significantly larger A_{max} than both Non-OLUT task ($t = 2.721$, $p = .007$) and OLUT task ($t = 2.550$, $p = .011$) respectively. The post-hoc tests found that OLUT task had significantly larger ΔT_p than both Control task ($t = 2.897$, $p = .004$) and Non-OLUT task ($t = 2.984$, $p = .003$). It was found that OLUT task also had significantly larger ΔT_f than both Control task ($t = 2.728$, $p = .006$) and Non-OLUT task ($t = 3.472$, $p = .001$).

These results suggest that overlapping uncertainty made BVP delta time features such as ΔT_p and ΔT_f values increased significantly. However, both overlapping uncertainty and non-overlapping uncertainty made BVP max features such as T_{max} and A_{max} values decreased significantly. As a result, BVP features show significant differences among tasks with different confidence levels, e.g. less confidence level tasks made BVP features of ΔT_p and ΔT_f values significantly higher. Therefore, BVP features can be used as indicators of user confidence levels in decision making tasks.

6 Discussion

This section discusses challenges and obstacles of revealing human cognition states in data analytics-driven decision making. The applications and future research directions are derived based on these discussions.

6.1 Challenges

DecisionMind is an ambitious framework for making ML transparent in real-world applications. The research challenges involved in DecisionMind mainly include:

- *Fundamentals*
  - Fundamental theories on relations between cognitive science and data analytics.
  - Fundamental questions regarding associated factors (e.g. uncertainty) with data analytics and decision making.
- *Protocol*
  - How to build a protocol that can capture a wide range of physiological and behavioural information during data analytics-driven decision making and link the capturing to human cognition states.
  - How to implement the protocol at the client and within the data analytics-driven decision making.
- *Cognitive response*
  - How to evaluate human trust on ML models, ML results and even decisions. With the evaluation, how to improve human trust in data analytics-driven applications.
  - How ML results and ML models affect human cognitive load during decision making.
- *Decision making*
  - Decision making profile. How to define an effective decision making profile based on data analytics.
  - Decision quality. How decision quality can be evaluated in data analytics-driven decision making.
- *Affecting factors*
  - Affecting factors and human cognition states. How associated factors such as various uncertainty affect user confidence, trust, and cognitive load in data analytics-driven decision making.
  - Presentation/visualization. Effective presentation/visualization methods for data analytics results and associated factors.

6.2 Obstacles

Let us imagine a researcher in ML and HCI who is motivated to tackle problems of revealing human cognition states in data analytics-driven decision making. What obstacles to success a researcher may meet?

- *Neuroscience* Despite the conspicuous progress in neuroscience for understanding human's neural activities, there are still many unsolved questions on quantitative evaluation of human cognition. This is one of major obstacles in understanding human cognition in decision making. Current research in neuroscience uses different latest techniques such as imaging techniques (e.g. fMRI) to understand differences in brain or other physiological signals when conducting tasks. However, these are not as precise as expected. There are still no concrete theories and evidence for evaluating human cognition states quantitatively, and linking human cognition with decision making. These obstacles could not always be there, but more likely be understood precisely with the advancement of neurobiological and genome research with modern tools such as imaging or microscopic techniques.
- *Risk* With the advancement of computing techniques and ML algorithms, ML-based intelligent systems are becoming more stable, less prone to error. Despite the improvement, human can still feel risky when making decisions relying on machines because it raises new concerns. For instance, how trustworthy of the prediction? When an error from the system occurs, where can we track back the error sources? How risky to take actions based on decisions from the system? These concerns are especially significant in modern complex high-risk domains such as aviation, medicine, and finance. These concerns must be addressed to increase the impact of ML in the real-world.
- *Complexity* Despite the proliferation of neuroscience and ML algorithms, these fields themselves have not yet matured enough for domain users to apply those techniques to their own problems with few efforts. This is because of the complexity of these fields. Firstly, it is often difficult to understand cognitive neurobiological information for domain users, let alone use them for the evaluation of human cognition states. Furthermore, high abstract ML algorithms are "black-boxes" to domain users and the associated complex parameter settings are even nightmares for domain users. These complexities

significantly affect the impact of data analytics-driven decision making. Simplifying and maturing neurophysiological results and ML algorithms can help relieve this obstacle and permit wider uses of cognition revealing in data analytics-driven decision making.
- *Generalization* Human trust, cognitive load, or confidence may be different in conducting a same task because of users' social background such as education, age, gender, etc. and other factors. People from different domains may also show differences of their attitude in conducting a same task. Therefore, the generalization becomes one obvious obstacle if data analytics-driven decision making with cognition states revealing is conducted by different users from different domains.

6.3 Research directions

The goal of making ML transparent is to produce powerful general-purpose tools and guideline for building human-machine interfaces by revealing human cognition states in data analytics-driven decision making. Such interfaces help data analytics-based intelligent systems make high quality decisions confidently. These include guidelines for interface design, software implementations and various sensors as well as other hardware. The research directions include:

- Research into the paradigm of human decision making process, including human and social factors as well as the role of technology and information in data analytics-driven decision making.
- Research of User Experiences that engender feelings of trust, cognitive load, and confidence in data analytics-driven decision making.
- Customer risk profile and uncertainty oriented decision making support solution based on data analytics.
- Visualization approaches for data, ML process, decision making, and associated factors (e.g. uncertainty).

6.4 Applications

Making ML transparent aims to build powerful general-purpose tools and guidelines for ML-based data analytics-driven decision making in order to increase impact of ML in real-world applications. It has wide applications to serve the purposes of intelligent, smart, effective, efficient, and high quality decision making. The research outcomes can be used by any potential users of ML technologies to make ML easier and more acceptable for them to get done what they want to do confidently. The research outcomes can also help providers of ML expertise to make a more efficient market place for them to sell their capabilities.

The research outcomes could benefit impact of ML technologies from following perspectives:

- *Human side* It allows users learn their cognition states in real-time during data analytics-driven decision making. It helps users understand whether they really accept and understand ML technologies physiologically from their cognition perspectives.
- *Environment side* Users' cognition does not cheat them and allows them perceive risk factors implicitly. Therefore, the revealing of human cognition states could motivate users to use ML technologies more confidently and make users more enjoy the use of ML technologies in their decision making.
- *Technical side* It links human cognition with ML technologies explicitly. As a result, the acceptance and understanding of ML technologies by users can be evaluated with human cognition which is more meaningful both for domain users and ML experts.
- *Education side* The revealing of human cognition could help developers of ML technologies learn in what degree users understand ML technologies, which could also motivate ML developers refine ML technologies in order to make them more understandable.

7 Conclusions

Machine learning offers a large number of powerful ways to approach problems that otherwise require manual solutions. However, because of "black-box" of ML approaches and complexities, users often feel uncertain and lack confidence in decisions in ML-based intelligent systems and it is very hard to see ML as a general solution for widespread applications. This paper argued that the revealing of human cognition states with a multimodal interface during data analytics-driven decision making could help make ML technologies more acceptable by users and improve impact of ML in real-world applications. Human cognition states can be revealed through physiological and behavioural signals. The paper also identified typical examples of impact challenges and real obstacles in making ML useable and transparent. Based on challenges and obstacles, the paper demonstrated high-demand research directions and applications.

Acknowledgements Authors thank all volunteer participants for the experiment. This work was supported in part by the Asian Office of Aerospace Research & Development (AOARD) under grant No. FA2386-14-1-0022 AOARD 134131.

References

1. Zhou J, Khawaja MA, Li Z et al (2016) Making machine learning useable by revealing internal states update—a transparent approach. Int J Comput Sci Eng 13:378–389

2. Zhou J, Bridon C, Chen F et al (2015) Be informed and be involved: effects of uncertainty and correlation on user's confidence in decision making. In: Proceedings of ACM SIGCHI conference on human factors in computing systems (CHI2015) works-in-progress
3. Zhou J, Sun J, Chen F et al (2015) Measurable decision making with GSR and pupillary analysis for intelligent user interface. ACM Trans Comput Hum Interact 21:33
4. Helgee EA (2010) Improving drug discovery decision making using machine learning and graph theory in QSAR modeling. Ph.D. thesis, University of Gothenburg
5. Wang W, Li Z, Wang Y, Chen F (2013) Indexing cognitive workload based on pupillary response under luminance and emotional changes. In: Proceedings of the 2013 international conference on intelligent user interfaces. ACM, New York, NY, USA, pp 247–256
6. Holzinger A (2013) Human–computer interaction and knowledge discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In: Cuzzocrea A, Kittl C, Simos DE et al (eds) Availability, reliability, and security in information systems and HCI. Springer, Berlin, pp 319–328
7. Khawaji A, Chen F, Marcus N, Zhou J (2013) Trust and cooperation in text-based computer-mediated communication. In: Proceedings of the 25th Australian computer–human interaction conference: augmentation, application, innovation, collaboration. ACM, New York, NY, USA, pp 37–40
8. Zhou J, Chen F (2015) Making machine learning useable. Int J Intell Syst Technol Appl 14:91. doi:10.1504/IJISTA.2015.074069
9. Moll J, De Oliveira-Souza R, Zahn R (2008) The neural basis of moral cognition: sentiments, concepts, and values. Ann N Y Acad Sci 1124:161–180. doi:10.1196/annals.1440.005
10. van Gog T, Paas F, Marcus N et al (2008) The mirror neuron system and observational learning: implications for the effectiveness of dynamic visualizations. Educ Psychol Rev 21:21–30. doi:10.1007/s10648-008-9094-3
11. Ayres P, Paas F (2009) Interdisciplinary perspectives inspiring a new generation of cognitive load research. Educ Psychol Rev 21:1–9. doi:10.1007/s10648-008-9090-7
12. Chen F, Zhou J, Wang Y et al (2016) Robust multimodal cognitive load measurement. Springer, Berlin
13. Krueger F, McCabe K, Moll J et al (2007) Neural correlates of trust. PNAS 104:20084–20089. doi:10.1073/pnas.0710103104
14. Aimone JA, Houser D, Weber B (2014) Neural signatures of betrayal aversion: an fMRI study of trust. Proc Biol Sci 281:20132127. doi:10.1098/rspb.2013.2127
15. Hahn T, Notebaert K, Anderl C et al (2015) How to trust a perfect stranger: predicting initial trust behavior from resting-state brain-electrical connectivity. Soc Cogn Affect Neurosci 10:809–813. doi:10.1093/scan/nsu122
16. Heekeren HR, Marrett S, Ungerleider LG (2008) The neural systems that mediate human perceptual decision making. Nat Rev Neurosci 9:467–479. doi:10.1038/nrn2374
17. Smith LJ, Anand P, Benattayallah A, Hodgson TL (2015) An fMRI investigation of moral cognition in healthcare decision making. J Neurosci Psychol Econ 8:116–133. doi:10.1037/npe0000038
18. White TP, Engen NH, Sørensen S et al (2014) Uncertainty and confidence from the triple-network perspective: voxel-based meta-analyses. Brain Cogn 85:191–200. doi:10.1016/j.bandc.2013.12.002
19. Franco-Watkins AM, Johnson JG (2011) Applying the decision moving window to risky choice: comparison of eye-tracking and mouse-tracing methods. Judgm Decis Mak 6:740–749
20. Schall JD (2001) Neural basis of deciding, choosing and acting. Nat Rev Neurosci 2:33–42
21. Guetl C, Pivec M, Trummer C et al (2005) ADELE (adaptive e-learning with eye-tracking): theoretical background, system architecture and application scenarios. Eur J Open Distance E Learn 2: 1–8
22. Arshad S, Wang Y, Chen F (2015) Interactive mouse stream as real-time indicator of user's cognitive load. In: Proceedings of the 33rd annual ACM conference extended abstracts on human factors in computing systems, pp 1025–1030
23. Khawaja MA, Chen F, Marcus N (2014) Measuring cognitive load using linguistic features: implications for usability evaluation and adaptive interaction design. Int J Hum Comput Interact 30:343–368. doi:10.1080/10447318.2013.860579
24. Khawaji A, Chen F, Zhou J, Marcus N (2014) Trust and cognitive load in the text-chat environment: the role of mouse movement. In: Proceedings of the 26th Australian computer–human interaction conference on designing futures: the future of design, pp 324–327
25. Jenkins Q, Jiang X (2010) Measuring trust and application of eye tracking in human robotic interaction. In: Proceedings of the 2010 industrial engineering research conference
26. Preuschoff K, Hart BM, Einhäuser W (2011) Pupil dilation signals surprise: evidence for noradrenaline's role in decision making. Front Neurosci 5:115. doi:10.3389/fnins.2011.00115
27. Fiedler S, Glockner A (2012) The dynamics of decision making in risky choice: an eye-tracking analysis. Front Psychol 3:1–18
28. Li Z, Zhang B, Wang Y et al (2014) Water pipe condition assessment: a hierarchical beta process approach for sparse incident data. Mach Learn 95:11–26
29. Peper E, Harvey R, Lin I-M et al (2007) Is there more to blood volume pulse than heart rate variability respiratory sinus arrhythmia, and cardiorespiratory synchrony? Biofeedback 35:54–61
30. Oppenheim AV, Schafer RW (2010) Discrete-time signal processing. Pearson, Upper Saddle River
31. Healey J, Picard R (2000) SmartCar: detecting driver stress. In: Proceedings of 15th international conference on pattern recognition 2000. pp 218–221
32. Zhai J, Barreto AB, Chin C, Li C (2005) Realization of stress detection using psychophysiological signals for improvement of human–computer interactions. Proc IEEE SoutheastCon 2005:415–420
33. Kristal-Boneh E, Raifel M, Froom P, Ribak J (1995) Heart rate variability in health and disease. Scand J Work Environ Health 21:85–95
34. Karvonen MJ, Kentala E, Mustala O (1957) The effects of training on heart rate: a longitudinal study. Ann Med Exp Biol Fenn 35:307–315
35. Tanaka H, Monahan KD, Seals DR (2001) Age-predicted maximal heart rate revisited. J Am Coll Cardiol 37:153–156