Making machine learning useable by revealing internal states update – a transparent approach

Jianlong Zhou*, M. Asif Khawaja, Zhidong Li,
Jinjun Sun, Yang Wang and Fang Chen

National ICT Australia (NICTA),
Level 5, 13 Garden Street, Eveleigh, NSW 2015, Australia
Email: Jianlong.Zhou@nicta.com.au
Email: asifkhawaja@yahoo.com
Email: Zhidong.Li@nicta.com.au
Email: jsunster@gmail.com
Email: Yang.Wang@nicta.com.au
Email: Fang.Chen@nicta.com.au
*Corresponding author

Abstract: Machine learning (ML) techniques are often found difficult to apply effectively in practice because of their complexities. Therefore, making ML useable is emerging as one of active research fields recently. Furthermore, an ML algorithm is still a 'black-box'. This 'black-box' approach makes it difficult for users to understand complicated ML models. As a result, the user is uncertain about the usefulness of ML results and this affects the effectiveness of ML methods. This paper focuses on making a 'black-box' ML process transparent by presenting real-time internal status update of the ML process to users explicitly. A user study was performed to investigate the impact of revealing internal status update to users on the easiness of understanding data analysis process, meaningfulness of real-time status update, and convincingness of ML results. The study showed that revealing of the internal states of ML process can help improve easiness of understanding the data analysis process, make real-time status update more meaningful, and make ML results more convincing.

Keywords: machine learning; black-box; transparent; internal status update.

Reference to this paper should be made as follows: Zhou, J., Khawaja, M.A., Li, Z., Sun, J., Wang, Y. and Chen, F. (2016) 'Making machine learning useable by revealing internal states update – a transparent approach', *Int. J. Computational Science and Engineering*, Vol. 13, No. 4, pp.378–389.

Biographical notes: Jianlong Zhou is a Researcher of Machine Learning Research Group in NICTA. He received his PhD in Computer Science from the University of Sydney, Australia. His research interests include transparent machine learning, human-computer interaction, cognitive computing, volume visualisation, spatial augmented reality and related applications.

M. Asif Khawaja received his PhD in Computer Science from the University of New South Wales (UNSW), Australia. He was working as a Researcher in human-computer interaction at NICTA in Sydney. He specialises in human-computer interaction, cognitive load assessment, trust assessment, behavioural analysis and software engineering. Currently, he is working in the software industry as a Technical Architect.

Zhidong Li received his PhD in Computer Science from the University of New South Wales, Australia in 2014. He is currently a Research Engineer in NICTA. His research interests include machine learning, computer vision, video analysis and pattern classification.

Jinjun Sun is currently a Data Scientist in Fairfax Media, Australia. He was a Research Engineer of Machine Learning Research Group in NICTA before this role. He received his PhD in Physics from Macquarie University, Australia. His research interests include systematic architectures targeting at big data challenges.

Yang Wang is a Senior Researcher of Machine Learning research group in NICTA. He received his PhD in Computer Science from the National University of Singapore in 2004. His research interests include machine learning and information fusion techniques and their applications to intelligent infrastructure, cognitive and emotive computing.

Fang Chen is a Senior Principal Researcher of Machine Learning Research Group in NICTA. She holds a PhD in Signal and Information Processing, an MSc and BSc in Telecommunications and Electronic Systems respectively, and an MBA. Her research interests are behaviour analytics, machine learning, and pattern recognition in human and system performance prediction and evaluation. She has done extensive work on human-machine interaction and cognitive load modelling. She pioneered theoretical framework of measuring cognitive load through multimodal human behaviour, and provided much of empirical evidence on using human behaviour signals, and physiological responses to measure and monitor cognitive load.

This paper is a revised and expanded version of a paper entitled 'Transparent machine learning – revealing internal states of machine learning' presented at the International Conference on Intelligent User Interfaces 2013 Workshop on Interactive Machine Learning, Santa Monica, CA, USA, 19 March 2013.

1 Introduction

Much of machine learning (ML) research is inspired by significant problems from various fields such as biology, finance, medicine, and society (Wagstaff, 2012). Various ML algorithms offer a large number of useful ways to approach those problems that otherwise require cumbersome manual solution. Wagstaff (2012) presented a three-stage model of an ML research programme. The stage 1 and stage 2 are the preparation stages for an ML research programme and the development of solutions for a problem, respectively. The stage 3 aims to improve the impact of ML algorithms on real-world applications. The current research in ML is highly biased towards the stage 2. Most ML researches are directed towards the invention of new algorithms for learning. It is the stage 3 that directly affects the effectiveness of real-world ML applications.

Despite the recognised 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. One of significant obstacles which affects an end user to use ML algorithms lies in complicated interface between ML algorithms and users, such as complex parameter settings and intermediate decisions in ML algorithms. The prevailing complex interfaces often require expert-level ML knowledge in order to fully understand. It is also difficult for ML non-users to understand complex ML models. Because of these complexities, it is very hard to see ML as a general solution for widespread applications. As a result, ML is regarded as a large bag of tricks grasped by ML experts instead of a universal tool for non-experts. It is necessary to get ML techniques out of the bag of ML experts and into various application domains from which ML could benefit.

Therefore, besides the development of ML algorithms, the research of making ML useable is emerging as one of active research fields recently. Making ML useable aims to make ML easily understandable and usable by domain professionals without requiring training in complex ML algorithms and mathematical concepts. To this end, most of previous work focuses on the application of visualisation techniques in depicting a specific ML algorithm or represent ML results (e.g., using bounding curves to represent clustering results in data space or graphs such as bar chart, pie chart to represent other ML results) (Fails and Olsen, 2003; Jakulin et al., 2005; Talbot et al., 2009; Witten et al., 2011). These approaches may help users understand ML results to some degree. However, for a domain professional who may not have expertise in ML or programming, an ML algorithm is still a 'black-box', where the user defines parameters and input data for the 'black-box' and gets output from its execution. This 'black-box' approach has obvious drawbacks: it is difficult for the user to understand the complicated ML models, such as what is going on inside the ML models and how to accomplish the learning problem (Zhou et al., 2013). As a result, the user is uncertain about the usefulness of ML results and this affects the effectiveness of ML methods. Unfortunately, there is not much research on how to make this ML 'black-box' transparent so that ML approaches can be easily understandable and usable by domain professionals.

This paper focuses on making a 'black-box' ML process transparent by presenting real-time internal status update of the ML process to users explicitly. Meaningful internal states of ML algorithms are selected and revealed during the real-time status update. Various visualisation techniques are used to allow users interactively view how the final results are obtained in ML. As a result, the 'black-box' ML becomes 'transparent'. A case study of water pipe failure history data (simulated data) analysis for future pipe failure prediction is presented to show the effectiveness of the proposed approach. A user study was performed to investigate the impact of revealing internal status update of ML to users on the *easiness* of understanding the data analysis process, *meaningfulness* of real-time status update, and *convincingness* of the ML results.

We propose in this paper that interactive ML interfaces must not only supply users with the information on input data and output results, but also enable them to perceive internal real-time status update of an ML process. As a result, a 'black-box' ML process becomes 'transparent' to users providing better understanding of the overall process. With the help of a user study, we show that revealing of the internal states of ML process can help to improve easiness of understanding the data analysis process, make real-time status update more meaningful, and make ML results more convincing. Finally, we also attempt to formulate guidelines/standards for the user interface designers of ML-based applications.

The paper is organised as follows: Section 2 presents the related work in making ML useable, such as visualisation in ML, followed by the hypotheses posed for our work. Section 3 discusses the method and approach used for the study. Section 4 describes the experiment task design, user study, participants, material used, and data collected. Section 5 presents the results and statistical analyses performed, followed by a discussion and conclusion in Sections 6 and 7, respectively.

2 Related work

Much work has been done to make ML more useable by employing various approaches. Henelius et al. (2014) proposed an iterative algorithm to find the attributes and dependencies used by classifiers in order to get understanding of which factors are of importance in classifiers. Explanations to intermediate analysis results are also investigated to make ML process transparent (Kulesza et al., 2011).

Patel et al. (2008) evaluated the current use of ML by non-experts and identified three difficulties they encounter when using ML techniques, namely:

1 difficulty in following an iterative and exploratory process

2 difficulty in under-standing the ML models

3 difficulty in evaluating performance.

Because a human usually acts as the central role in an ML process, Fails and Olsen (2003) emphasised on the importance of human involvement in an interactive ML. Their presented interactive ML model allows users to correct classifications in a continuously iterative ML loop.

Since humans have great ability to understand patterns in the natural environment through visual perception, various visualisation techniques are widely used in previous work to make ML useable. For example, to assist ML developments, much research has been done in visualising ML algorithms, such as Naïve-Bayes (Becker et al., 2002), decision trees (Ankerst et al., 1999), SVMs (Caragea et al., 2001), and HMMs (Dai and Cheng, 2008). Visualisation techniques are also frequently used to plot data instances in the projection of feature space and visualise model prediction boundaries (Frank and Hall, 2003). Furthermore, visualisation techniques are used to present an ML process, such as clustering process. Erra et al. (2011) introduced a visual clustering approach which utilises collective behavioural model. Each data item is represented by an agent visualised with a metaphor in 3D domain. Visualisation helps users to understand and guide the clustering process. Paiva et al. (2011) presented an approach that employs the similarity tree visualisation to distinguish groups of interest within the dataset. The similarity tree is built from feature spaces or similarity relationships. It is used to visually support classifications. It has advantages for both browsing and mining for their capacity of local as well as global space reconstruction with valuable layouts.

Nomograms have also been employed for visualising trained SVMs (Jakulin et al., 2005). Using nomograms, SVMs can be effectively visualised in attribute spaces with many dimensions. In this approach, individual attributes are stacked vertically in a nomogram, packing multiple dimensions into a single one. In this way, nomograms provide a clear and comprehensive presentation of the underlying model.

EnsembleMatrix allows users to visually ensemble multiple classifiers together and provides a summary visualisation of results of these multiple classifiers (Talbot et al., 2009). It allows users to explore and build combination models through interactions.

Fiebrink et al. (Fiebrink, 2011; Fiebrink and Trueman, 2012) used human computer interaction (HCI) techniques in the design of end-user interfaces for interactive ML in real-time application domains such as music composition and performance. Their work enables domain users such as students, composers, artists to apply ML to their work effectively.

Besides ML process itself, the user interface for data input and ML output may also affect usability of ML techniques. User interface can be designed context and changed adaptively in order to improve navigation performance (Swamy et al., 2012). Virtual reality was also investigated to design user interface to increase the interest and effectiveness of users in learning advanced technology and its applications (Tarng et al., 2011).

In summary, visualisation plays significant roles in making ML useable. Most of previous work focused on using visualisation in ML algorithms and results presentations. HCI techniques are also used in ML user interfaces. However, few studies have been done on making ML transparent to users. We strongly believe that ML processes with revealing internal status update to users can improve the easiness of understanding data analysis process and make ML results more convincible. As a result, it can make ML more useable to various domain users.

2.1 Hypotheses

In this paper, the following hypotheses are posed for the transparent ML:

- Easiness: Revealing of internal states of ML processes will make ML technologies easier to understand. Users will show different ratings on easiness for tasks with/without internal state update, specifically, they will provide higher ratings on easiness for tasks with internal states update.
- Meaningfulness: Revealing of internal states of ML processes will make ML technologies more meaningful. Users will show different ratings on meaningfulness for tasks with/without internal state update, i.e., they will provide higher ratings on meaningfulness with internal states update.

- Convincingness: Users will also provide different convincingness ratings on ML results from different tasks, specifically, they will provide higher ratings on convincingness for tasks that reveal internal states update.

The visualisation that can depict changes of internal states with historical data clearer will be more helpful for users to understand ML process. Therefore, users will provide higher ratings on graphic presentations of internal state update than for numbered or textual presentations regardless of what ML method is used.

3 Method

We designed a user study to test our hypotheses but before we discuss the details of the study, this section first introduces two ML approaches whose internal status updates are revealed in the study. Then the concept of transparent ML is presented and detailed steps used to setup transparent ML are discussed, followed by the study design.

3.1 Case study

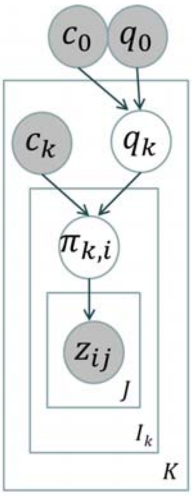
This research used water pipe failure prediction as a case study. Water supply networks constitute one of the most crucial and valuable urban assets. The combination of growing populations and ageing pipe networks requires water utilities to develop advanced risk management strategies in order to maintain their distribution systems in a financially viable way (Li et al., 2014). Especially for critical water mains (generally > 300 mm in diameter), defining based on the network location (for example, a single trunk line connecting distribution areas or under a major road) or size which infers impact potential, failure of them typically bring severe consequences due to service interruptions and negative economic and social impacts, such as flooding and traffic disruption (Li et al., 2014). If high-risk pipes can be identified before a failure occurs, it is likely that repairs can be completed with minimal service interruption, water loss and negative reputational and community impacts. Identification of an accurate predictor measure that indicates imminent failure will allow utility companies to take actions to mitigate the failure for a lower cost than repairing a full-scale failure. As the average age of the network increases, pipes fail easily with the decrease of pipe strength. It will become more important to accurately predict the risks of pipe failure and provide the right level of pipe maintenance and renewal at the right time, according to risks associated with each pipe. For example, utility companies use the outcomes from the failure prediction model to make renewal plans for pipes based on risk levels of pipes, and thus also make reasonable budget plans for the pipe maintenance.

3.2 ML approaches

Water pipe failure history data (simulated data) is used as a case study in this paper. Two ML methods are employed to predict future failure rate of water pipes: hierarchical beta process (HBP) proposed in our previous research (Li et al., 2014) and Weibull model (Ibrahim et al., 2005).

HBP is a hierarchical non-parametric model proposed to predict failure of water pipes (Li et al., 2014). In this model as shown in Figure 1, pipes are divided into K groups based on laid years and modelled as a HBP. In the top level, hyper parameters, which control across all groups of pipes by a beta distribution, are set manually according to domain experts' experience. Then, the mean failure rate (q_k) in each group can be generated from the distribution. In the middle level, the mean failure rate ($\pi_{k,i}$) of each pipe asset is generated through another beta distribution with q_k as parameter. In the bottom level, the actual failures $z_{i,j}$ are generated from a Bernoulli process year by year using $\pi_{k,i}$.

Figure 1 The diagram of HBP model (see online version for colours)



In Weibull model, failure rate of an individual pipe is assumed to grow with age. Within a class of pipes (grouped according to pipe properties such as material, laid date, etc.), the accumulated number of failures is proportional to τ^β, where τ is the age of a pipe, β is a positive parameter. Larger value of β indicates that failure rate increase faster with age. For each pipe i, the expected number of failure at year t becomes

$$\mu_i = E\left[N_{i,t}\right] = \alpha l_i \left[(\tau_i + 1)^\beta - \tau_i^\beta\right], \tag{1}$$

where $N_{i,t}$ is the number of failure for the pipe at year t, $\tau_i = t - t_{i,0}$, $t_{i,0}$ is laid date of the pipe, l_i is pipe length. α is a positive parameter influenced by various factors such as pipe material, pipe size, soil type, water pressure, etc. The parameters α and β could be learned from the training data. Given the parameters α and β, the expected failure rate (number of pipe failures per year per unit length) becomes $\alpha[(\tau + 1)^\beta - \tau^\beta]$ for pipe age τ.

3.3 Transparent ML

Previous work in transparent ML mainly focuses on the why-oriented explanation to intermediate results during ML process in order to let users understand the analysis process

(Kulesza et al., 2011). Such approaches still require learning knowledge in some degree and lack an overall image of learning process. Our study tries to make ML transparent by using the approach of dynamic visual feedback of internal states. In most cases, even if automatic ML algorithms give satisfactory results, visual feedback of ML processes is very important to offer insight into the reasons for learning failure or success. Visual feedback is also critical to support interfering with further learning processes, as well as intently correcting the learning parameters. Therefore, interactive ML interfaces must not only supply users with the information on input data and output results, but also enable them to perceive internal real-time status update of ML processes with visual feedback. Based on this consideration, a concept of transparent machine learning (TML) is proposed in this paper. In this concept, meaningful internal states of ML processes are selected and presented to users visually and meaningfully. Various visualisations are used to provide feedback and updates for internal states of ML processes. As a result, an ML process becomes a 'transparent-box'.

The TML includes following steps:

- select internal state variables that are dynamically changed and meaningful to users
- present the changing of internal state variables visually and meaningfully to users
- interact with the ML process based on explicit feedback from revealing of internal real-time status update.

TML presents selected internal states dynamically to users meaningfully (e.g., money saved, time preserved) based on domain knowledge. It provides a feedback loop that aids users learn what is going on and how to accomplish with the given learning problem. Users also have freedom to interact with the ML based on the feedback in order to improve models. TML provides a means for users to assess the model's behaviours against a variety of subjective criteria based on domain knowledge and examples. As a result, the users' understanding and trust of the system could be improved and it benefits the accuracy of learning systems as well.

4 Experiment setup

This section sets up an experiment to study the effectiveness of TML in making ML useable. The objectives of this experiment are to investigate whether revealing the internal states of a ML process makes ML techniques more usable and meaningful, whether users get more easily convinced by ML results from theoretically sound ML approaches than from relatively simple ML approaches, and what effective approaches for revealing internal states of ML processes are.

4.1 Experiment data

The water pipe failure history data contains the failure records of water pipes in a given region. It also includes various attributes of water pipes, such as identification number, laid date, length, material, diameter size, location, protective coating, surrounding soil type, etc. The objective of the analysis was to predict future failure rate of water pipes based on the history data of water pipes. Two ML methods were used in this study, which are HBP and Weibull as presented in Section 3. In HBP, the mean failure rate (q_k) in each pipe group and the mean failure rate ($\pi_{k,i}$) of each pipe were revealed to show the internal state updates in TML. In Weibull, the shape parameter of β was used to show the internal state updates in TML.

4.2 Visualisation methods for revealing internal state update

In this study, three visualisation methods of bar-chart, full number text and partial number text are used to present changes of internal states of ML processes. These three methods are commonly used by data analysts and programmes. These visualisation methods are easily to understand and easily accepted by users.

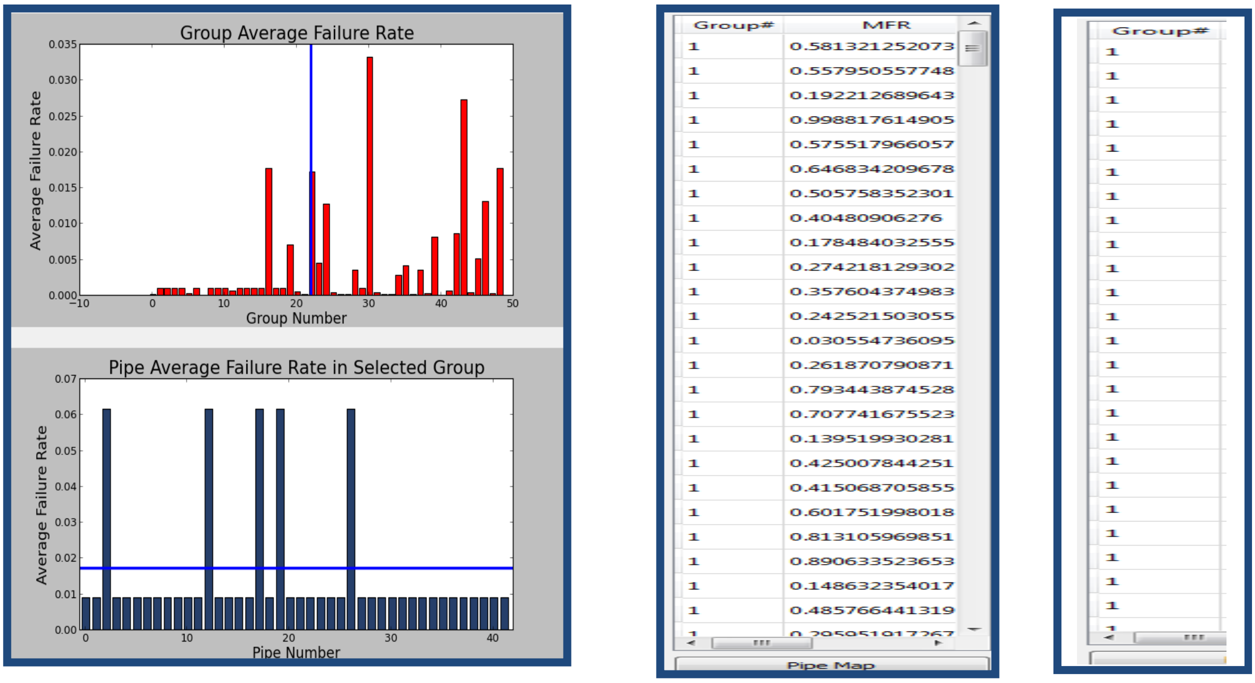
- Bar-chart [Figure 2(a)] presents changes of internal states with graphical visualisations dynamically. It allows users easily perceive changes of internal states, but lacks details of each change.
- Full number text [Figure 2(b)] presents changes of internal states with text numbers. The presented information includes group numbers and average failure rate of pipes. This method provides details of changes of internal states on each step. However, because of too many details, it makes users difficult to grasp the overall patterns of changes.
- Partial number text visualisation method [Figure 2(c)] only presents group numbers and the average failure rate of the selected group to users. This method allows users easily to perceive changes of internal states, but lacks details of changes such as detailed values and patterns of changes.

These visualisations were designed to be interactive so the user can also select a specific data point or slice to view more details about that chunk.

4.3 Task design

As mentioned, two ML methods and three visualisation approaches were used for setting up tasks in the experiment. Therefore, there were six tasks all together performed by each participant. The task configurations are shown in Table 1.

Figure 2 Different presentation methods for real-time status update (see online version for colours)

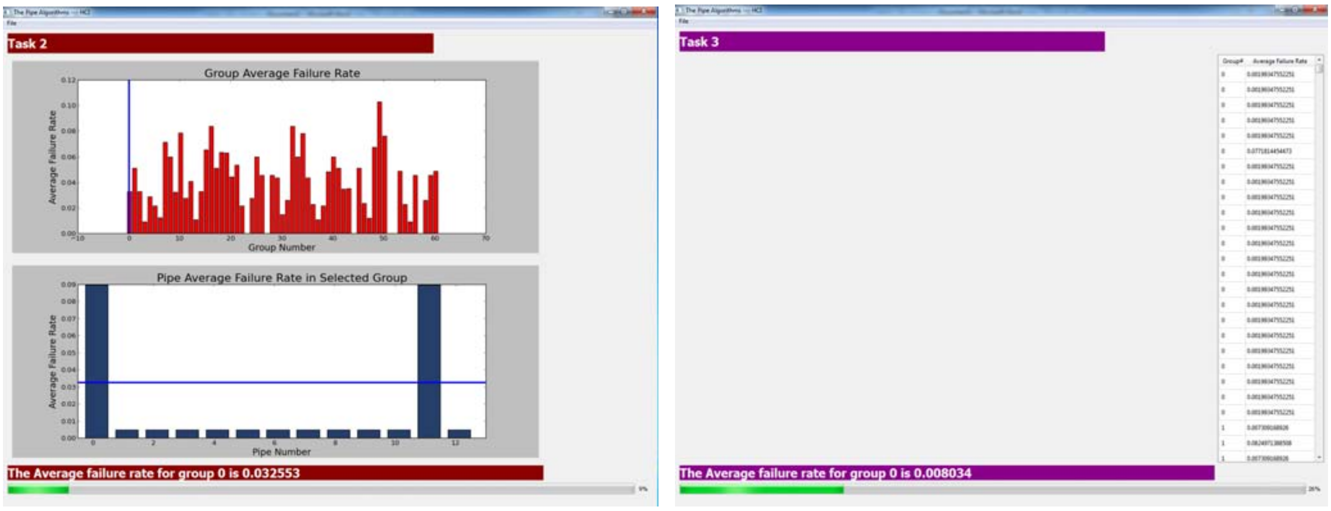


Group#	MFR
1	0.581321252073
1	0.557950557748
1	0.192212689643
1	0.998817614905
1	0.575517966057
1	0.646834209678
1	0.505758352301
1	0.40480906276
1	0.178484032555
1	0.274218129302
1	0.357604374983
1	0.242521503055
1	0.030554736095
1	0.261870790871
1	0.793443874528
1	0.707741675523
1	0.139519930281
1	0.425007844251
1	0.415068705855
1	0.601751998018
1	0.813105969851
1	0.890633523653
1	0.148632354017
1	0.485766441319
1	0.295951917267

Group#
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1

(a) (b) (c)

Figure 3 Screenshots of two tasks performed in the study (see online version for colours)



The Average failure rate for group 0 is 0.032553

Group#	Average Failure Rate
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
0	[illegible]
1	[illegible]
1	[illegible]
1	[illegible]

The Average failure rate for group 0 is 0.008034

(a) (b)

Table 1 Tasks performed by participants

Task #	*ML method*	*Presentation method*
1	HBP	Partial number text
2	HBP	Graph (bar-chart)
3	HBP	Full number text
4	Weibull	Partial number text
5	Weibull	Graph (bar-chart)
6	Weibull	Full number text

Participants were required to do water pipe failure rate prediction based on historical data using two ML methods with three conditions (revealing internal states with partial number text, bar-chart, or full number text). All participants were provided training on the concept used in the experiment and how to interact with the system before performing the six tasks. During the training, each participant was also introduced about the water pipe failure history data and data analysis goals of the study. Then an example task was practised by each participant to experience the data analysis process and get familiar with the user interface used in the experiment. Figure 3 presents screenshots of the custom user interaction system showing two of the tasks used in the study. The six tasks were then performed separately and questionnaires were also answered after each task. Finally, a questionnaire on the overall experiment was answered to complete the experiment. During the experiment, names of ML methods were not known to participants; they were presented with only the task numbers as in Table 1. The order of tasks was also randomised to avoid bias and/or any possible learning effects.

4.4 Participants

A total of 30 participants were involved in the study. Ages of participants range from early 20s to 40s. Participants were rewarded movie voucher or chocolate bars after the experiment for the compensation of their time. Participants were from three groups with different background (ten participants from each group respectively):

1. researchers who are not doing ML or data mining research (i.e., non-ML researchers)
2. researchers who are doing ML or data mining research (ML researchers)
3. administrative staff.

Each group of participants has their own specialities, for example, ML researchers were good at using ML algorithms while administrative staffs were good at analysing information in different forms. Various groups of participants were included in order to analyse differences of responses with different knowledge background.

4.5 Data collection

In this study, participants were required answer questionnaires after every task to rate the TML using a seven-point Likert-scale (1 = totally disagree, and 7 = totally agree) on three aspects of each task:

- easiness of understanding the analysis process
- meaningfulness of real-time status update
- convincingness on the analysis results.

Participants were also required to compare which presentation method gives more easily understandable and meaningful information on changes of internal states. They also ranked which task makes the participant feel most confident about the results of the water pipe prediction based on the overall experiment.

5 Results and analyses

We performed Friedman test with post-hoc analysis using Wilcoxon signed-rank tests to analyse the mean differences in participant responses. The Friedman test is the non-parametric alternative to the one-way ANOVA with repeated measures and is used to test for differences between groups when the dependent variable being measured is ordinal, e.g., a seven-point Likert-scale from strongly disagree through to strongly agree (Field, 2009). Because our experiment used non-parametric ordinal ratings, we used Friedman test and post-hoc Wilcoxon signed-rank tests to analyse experiment results.

5.1 Analyses of overall ratings

Firstly, we analysed the overall ratings from 30 participants in all groups. Figure 3 shows the overall comparison of ratings on meaningfulness, convincingness, and easiness of understanding of the real-time status update during ML processes for the six tasks presented in Table 1. Differences in easiness, meaningfulness, and convincingness were analysed respectively with Friedman test and post-hoc pair-wise tests (Wilcoxon signed-rank tests).

5.1.1 Easiness

Friedman test showed that there was a statistically significant difference between the six tasks for the easiness of understanding, $\chi^2(5) = 56.226$, $p = 0.000$; < 0.05. Post-hoc analysis with Wilcoxon signed-rank tests was then conducted with a Bonferroni correction applied resulting in a significance level set at $p < 0.008$ for all pair-wise differences. The tests show that task 2 (HBP + Graph) was significantly easiest to understand than any other tasks ($Z < -3.880$, $p = 0.000$), followed by task 5 (Weibull + Graph), which was significantly easier than task 1 ($Z = -2.886$, $p = 0.004$) and task 4 ($Z = -3.251$, $p = 0.001$) where both tasks 1 and 4 used only partial number output. Task 6, which used full number output without any graphical visualisation, was easier to understand than task 4 that used only partial number output ($Z = -2.691$, $p = 0.007$). There were no other significant differences found for the ratings of easiness. These results suggest that ML methods that reveal the internal states update of the process graphically and interactively are significantly easier to understand for the end users, as we hypothesised.

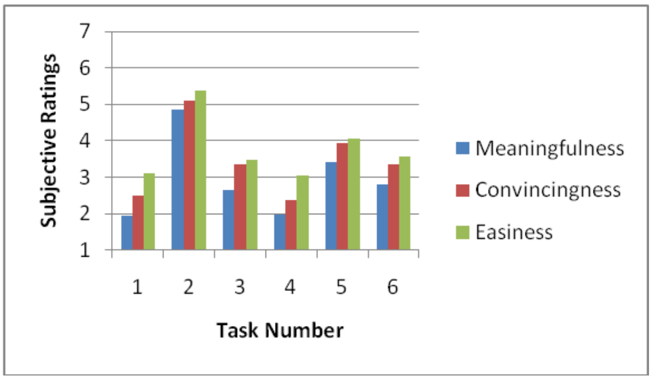
5.1.2 Meaningfulness

There was a statistically significant difference between the six tasks for the meaningfulness too, $\chi^2(5) = 89.777$, $p = 0.000$. Post-hoc Wilcoxon tests with Bonferroni correction showed that task 2 was significantly the most meaningful than any other tasks ($Z < -3.985$, $p = 0.000$), followed by task 5, which was significantly more meaningful than task 1 ($Z = -4.087$, $p = 0.004$) and task 4 ($Z = -4.064$, $p = 0.000$). Although, both tasks 2 and 5 used graphical visualisation for the results, participants found task 2 that involved HBP ML method the most meaningful. Task 3 was significantly more meaningful than task 1 ($Z = -3.063$, $p = 0.002$), and task 6 was significantly more meaningful than task 1 ($Z = -3.225$, $p = 0.001$) and task 4 ($Z = -3.093$, $p = 0.002$). Both tasks 3 and 6 used textual but full detailed number output, so the statistics suggest that even though the results are textual only (i.e., no graphical representation), still the participants find the full numbered text output more meaningful overall than partial number text. All these findings are in line without hypotheses. There were no other significant differences for ratings on meaningfulness between any other tasks.

5.1.3 Convincingness

For convincingness, there was a statistically significant difference between the six tasks, $\chi^2(5) = 78.448$, $p = 0.000$. Post-hoc tests showed that task 2 was significantly the most convincing than any other tasks ($Z < -3.304$, $p < 0.001$). Task 5 was significantly more convincing than task 1 ($Z = -3.981$, $p = 0.000$) and task 4 ($Z = -4.146$, p=0.000). Once again, both tasks 2 and 5 that use graphical visualisations are found to be more convincing than other tasks but task 2 with HBP method is the most convincing. Task 6 was significantly more convincing than task 1 ($Z = -3.225$, $p = 0.001$) and task 4 ($Z = -3.450$, $p = 0.001$). Task 3 was significantly more convincing than task 4 ($Z = -2.931$, $p = 0.003$). Similar to previous results for easiness and meaningfulness, these results suggest that in the absence of graphical visualisations, full number text output makes the ML method results more convincing than partial text output. There were no other significant differences found for ratings on convincingness between other tasks.

Figure 4 Comparison of meaningfulness, convincingness, and easiness of understanding of real-time status update in the six tasks based on the ratings from all 30 participants (see online version for colours)



5.2 Analyses of responses of individual groups

As mentioned in Section 4.4, there were three groups of participants: administrative staff, non-ML researchers, and ML researchers, each group with ten participants. Different participant groups have different knowledge background. For example, administrative staff are more experienced in detecting changes in number values and may not have knowledge in ML at all, non-ML researchers are experienced in various data representations but may not have extensive knowledge in ML, while ML researchers have more knowledge in ML, such as models used and detailed performance evaluation approaches. Therefore, we are also interested in finding whether there were any significant differences in responses from individual groups in order to understand the effects of knowledge background on responses. We performed Friedman test with post-hoc tests (Wilcoxon signed-rank tests) on group ratings for easiness, meaningfulness, and convincingness. Figures 4, 5, and 6 illustrate the comparison between groups for ratings on easiness, meaningfulness, and convincingness, respectively.

Figure 5 Comparison of easiness of understanding among various groups (see online version for colours)

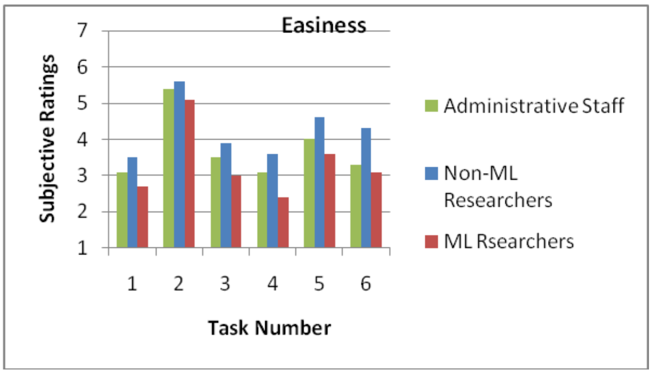


Figure 6 Comparison of meaningfulness among various groups (see online version for colours)

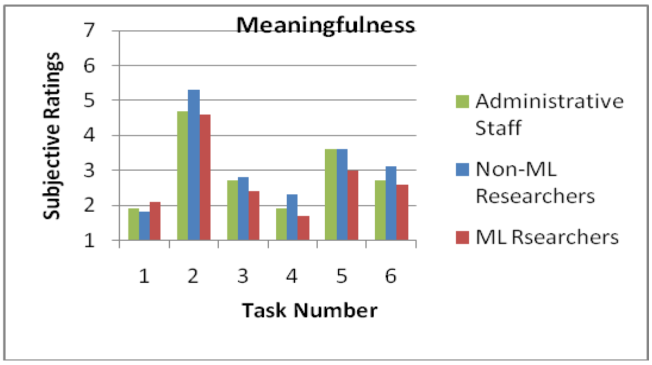
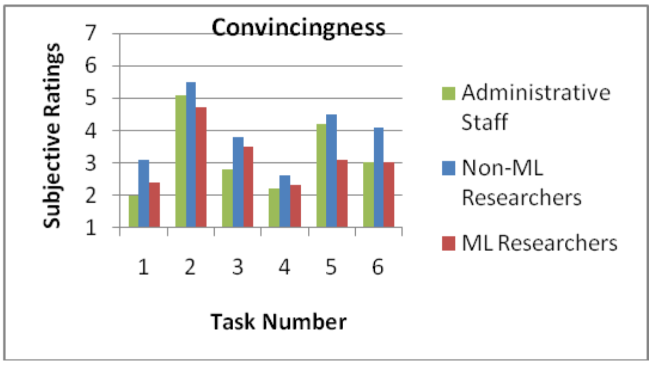


Figure 7 Comparison of convincingness among various groups (see online version for colours)



5.2.1 Easiness

- *Administrative staff*: Friedman test showed that for the participants of administrative staff, there was a statistically significant difference between the six tasks for the easiness of understanding, $\chi^2(5) = 15.768$, $p = 0.008$; < 0.05. For pair-wise differences, post-hoc analyses with Wilcoxon signed-rank tests were conducted with a Bonferroni adjusted alpha level set at 0.008, instead of 0.05). However, at this adjusted significance level of 0.008, the tests did not show any pair-wise differences for tasks. This could be due to a smaller group size (ten subjects) which may have reduced the overall statistical power (Nakagawa, 2004). Also, the Bonferroni adjustment is used to avoid any possible Type-I errors, but it is known that Bonferroni adjustment over corrects the alpha level and may cause Type-II errors and hence reduce the overall statistical power (Perneger, 1998; Rothman, 2010). Therefore, to avoid any potential Type-II errors we used a readjusted significance alpha level of 0.017 to see if we can find any other pair-wise differences that we expected. This adjusted significance alpha level of 0.017 was calculated by dividing the original alpha level of 0.05 by 3 based on the fact that for each ML method we have three conditions to test. Using this new alpha level of 0.017, the results showed that task 2 was significantly easier to understand than task 5 ($Z = -2.588, p = 0.010$). Task 2 was also found to be significantly easier to understand by the group than task 1 ($Z = -2.446, p = 0.014$) and task 4 ($Z = -2.395, p = 0.017$). These results are in line with our previous ones and suggest that ML methods revealing internal state update using graphical visualisation are easier to understand than non-graphical and textual number results and HBP ML method is easier to understand than the Weibull method regardless of what visualisation method is used. There were no significant differences between other tasks.

- *Non-ML researchers*: For non-ML researchers also, the Friedman test showed a statistically significant difference between the six tasks for the easiness of understanding, $\chi^2(5) = 16.199, p = 0.006; < 0.05$. Post-hoc Wilcoxon signed-rank tests with adjusted alpha level of 0.017 found only one pair-wise significant difference between task 1 and task 2 with task 2 ratings being significantly higher than ratings for task 1 ($Z = -2.536, p = 0.011$), suggesting that non-ML researchers also find HBP ML method with graphical visualisation significantly easier to understand than any other method. There were no other significant differences between tasks found for non-ML researchers group.
- *ML researchers*: For ML researchers as well, we found a statistically significant difference between the six tasks for the easiness of understanding, $\chi^2(5) = 26.976, p = 0.000; < 0.05$. Post-hoc tests showed that at Bonferroni adjusted alpha level of 0.008, task 2 was significantly easier to understand than task 1 ($Z = -2.699, p = 0.007$) as well as task 4 ($Z = -2.699, p = 0.007$). With the readjusted alpha level of 0.017 as mentioned before, the results showed that task 2 was also significantly easier to understand than task 3 ($Z = -2.508, p = 0.012$), task 5 ($Z = -2.410, p = 0.016$), and task 6 ($Z = -2.539, p = 0.011$). Task 5 was also found to be easier to understand than task 4 ($Z = -2.401, p = 0.016$). There were no other significant differences found between other tasks for this group. Like previous results, these results also suggest that even for the most technical ML researchers group, the task 2, which uses HBP ML method with graphical visualisations to reveal the internal states, is the easiest method to understand and that the methods that use graphical visualisations are overall rated easier than textual results.

5.2.2 Meaningfulness

- *Administrative staff*: As usual the Friedman test for the meaningfulness for the participants of administrative staff group showed a significant difference between the six tasks, $\chi^2(5) = 34.474, p = 0.000; < 0.05$. Post-hoc Wilcoxon tests at Bonferroni adjusted alpha level of 0.008 showed that task 2 was significantly more meaningful than task 1 ($Z = -2.680, p = 0.007$). With the readjusted alpha level of 0.017, the tests showed that task 2 was also significantly more meaningful than all other tasks. Task 5 was also found to be more meaningful than task 1 ($Z = -2.388, p = 0.017$). There were no other significant differences between tasks.
- *Non-ML researchers*: For non-ML-researchers, the ratings for the meaningfulness using Friedman test also showed a significant difference between the six tasks, $\chi^2(5) = 27.230, p = 0.000; < 0.05$. Post-hoc Wilcoxon pair-wise tests with Bonferroni corrected alpha level at 0.008 showed that task 2 was significantly more meaningful than task 1 ($Z = -2.677, p = 0.007$) as well as task 4 ($Z = -2.823, p = 0.005$). With the readjusted alpha level of 0.017 as mentioned before, it showed that task 2 was also significantly more meaningful than task 3 ($Z = -2.501, p = 0.012$), task 5 ($Z = -2.599, p = 0.009$), and task 6 ($Z = -2.448, p = 0.014$). Task 5 was also found to be significantly more meaningful than task 1 ($Z = -2.565, p = 0.010$) and task 4 ($Z = -2.410, p = 0.016$). There were no other significant differences found between tasks.
- *ML researchers*: For ML researchers, the results showed that there was a statistically significant difference between the six tasks for the meaningfulness, $\chi^2(5) = 33.238, p = 0.000; < 0.05$. Post-hoc analysis at Bonferroni adjusted alpha level of 0.008 showed that task 2 was significantly more meaningful than task 1 ($Z = -2.831, p = 0.005$), task 3 ($Z = -2.848, p = 0.004$), and task 4 ($Z = -2.829, p = 0.005$). With the readjusted alpha level of 0.017 task 2 was also found to be significantly more meaningful than task 5 ($Z = -2.388, p = 0.017$) and task 6 ($Z = -2.536, p = 0.011$). Task 5 was also rated significantly more meaningful than task 4 ($Z = -2.565, p = 0.010$). There were no other significant differences found between tasks.

These group-wise results for the meaningfulness also confirm our hypothesis and suggest that revealing the internal states update using graphical visualisations increases the meaningfulness of the real-time status update of ML process for all groups, and the best under the HBP ML method.

5.2.3 Convincingness

- *Administrative staff*: For convincingness, the subjective ratings by the participants of administrative staff group showed a statistically significant difference between the six tasks, $\chi^2(5) = 33.852, p = 0.000; < 0.05$. Post-hoc analysis at Bonferroni adjusted alpha level of 0.008 showed that task 2 was significantly more convincing than task 1 ($Z = -2.818, p = 0.005$) as well as task 4 ($Z = -2.814, p = 0.005$). Task 5 was also found to be significantly more convincing than both task 1 ($Z = -2.701, p = 0.007$) and task 4 ($Z = -2.724$, $p = 0.006$) at that level. With the readjusted alpha level of 0.017, we also found both task 2 and task 5 to be significantly more convincing than task 1, task 3, and task 4, respectively. There were no other significant differences found between tasks.
- *Non-ML researchers*: The results for non-ML researchers' ratings also showed a statistically significant difference between the six tasks in the convincingness, $\chi^2(5) = 21.845, p = 0.001; < 0.05$. Post-hoc analysis at Bonferroni corrected alpha level of 0.008 showed task 2 to be significantly more convincing than task 4 only ($Z = -2.677, p = 0.007$). Using readjusted alpha level of 0.017, we found that task 2 was also significantly more convincing than task 1 ($Z = -2.536, p = 0.011$). Task 5 was also found to be significantly more convincing than task 4 ($Z = -2.565, p = 0.010$). There were no other significant differences found between tasks.
- *ML researchers*: The results for the ML researchers group showed a statistically significantly difference among the six tasks in the convincingness, $\chi^2(5) = 28.075, p = 0.000; < 0.05$. Post-hoc analysis with Wilcoxon signed-rank tests at Bonferroni adjusted alpha level of 0.008 showed that task 2 was significantly more convincing than task 1 ($Z = -2.699$, $p = 0.007$) and task 4 ($Z = -2.714, p = 0.007$). With readjusted alpha level of 0.017, task 2 was also found to be significantly more convincing than task 5 ($Z = -2.555, p = 0.011$) as well as task 6 ($Z = -2.456$, $p = 0.014$). There were no other significant differences found between tasks.

  Based on these group-wise results for the subjective ratings of convincingness, we can suggest that revealing the internal states update using graphical visualisations, especially with HBP ML method, increases the convincingness of ML prediction for all groups.

6 Discussion

In our previous study (Li et al., 2014), HBP showed better performance than Weibull in water pipe failure rate prediction. HBP was therefore considered as an advanced ML approach and Weibull was considered as a classical ML approach. Furthermore, from the study in this paper, we found that HBP method with the graph-based visualisations got higher ratings than any other task configurations in the three aspects of easiness, meaningfulness, and convincingness. Weibull with graph-based real-time status update got higher rates than HBP without graph-based real-time status update. This result showed that advanced ML approaches along with graphical presentations can help improve the easiness of understanding the ML data analysis process, meaningfulness of real-time status update, and convincingness of ML results. Even if classical ML approaches were used, users still gave higher ratings when graphical presentations were used than advanced ML approaches without graphical real-time status update. From this result, we can conclude that:

1. it is obvious that the ML approach used affects users' ratings on easiness, meaningfulness, and convincingness
2. the visualisation method used also greatly affects users' overall ratings on ML's easiness, meaningfulness, and convincingness of an ML approach, regard-less of what ML method is being used.

Differences were also shown in responses from different participant groups. For example, regarding the convincingness, administrative staff found that task 5 was significantly more convincing than task 1 and task 4. Task 5 was slightly more convincing than task 3 ($Z = -2.388$, $p = 0.017$). Non-ML researchers found that task 5 was significantly more convincing than task 4 only. However, ML researchers did not find any significant differences between task 5 and other tasks. These differences of responses maybe due to the ML knowledge background of participant groups. The ML knowledge may have contributed to the understanding of ML based data analysis, meaningfulness of real-time status update, and convincingness of ML results.

Besides ratings of various aspects of easiness, meaningfulness, and convincingness, participants also gave interesting open comments on making ML useable further. It was interesting to find that different groups had different opinions. For example, in order to make ML-based data analysis more convincing, administrative staff suggested to use more graphical visualisations for extra information; non-ML researchers suggested to use static plotting but not dynamic plotting (that changed in response to user's selection) for the easy comparison, while ML researchers suggested to use higher refreshing rate to present real-time changes in graphical visualisatoins and show cost function changes with time. Some comments from different groups may seem conflicting with one another because of their knowledge background, for example, ML researchers focused more on cost functions, while administrative staff were more interested in number changes. However, they also had same views in some aspects, for example, all of groups suggested that more interactions with parameters and visualisations would make ML results more convincing.

Based on the results of this study, several guidelines are suggested for the design of user interfaces and interactions in ML-based applications:

- present real-time internal status update of ML processes to users when possible
- use graphical approaches to present ML results and internal status update of ML processes
- allow users to interact with parameters and ML results based on real-time status update.

7 Conclusions and future work

This paper focused on making a 'black-box' ML process transparent by presenting real-time internal status update of the ML process to users explicitly. A case study of water pipe failure history data analysis for future pipe failure prediction was presented to show the effectiveness of the proposed approach. A user study was performed to investigate the impact of revealing internal status update of ML to users on the easiness of understanding the data analysis process, meaningfulness of real-time status update, and convincingness of the ML results. The study showed that revealing the internal states of ML process can help to improve easiness of understanding the data analysis process, make real-time status update more meaningful, and make ML results more convincing. Finally, we also attempted to formulate guidelines/standards for the user interface designers of ML-based applications.

The future work of this research will focus on conducting more experiments using different datasets and ML methods in order to learn differences in datasets and ML methods for making ML useable. Interactions will also be analysed in making ML useable, such as interaction methods, parameters to be interactively changed in ML process. Various visualisation methods in real-time status update will be compared further in making ML useable. Objective evaluations based on user's responses such as Galvanic Skin Response (GSR) and eye-tracking will be investigated to find factors both from ML methods and user interfaces for making ML more useable. Emotional information (Hung et al., 2011) from video, text or other ways may also be investigated to find solutions for making ML more useable.

Acknowledgements

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence programme. The authors would like to thank Ronnie Taib for valuable comments on setting up the experiment. We would also like to thank numerous volunteer participants for the experiment. This work is partly supported by the Asian Office of Aerospace Research and Development (AOARD) under grant nos. FA2386-14-1-0007 AOARD 134144, and FA2386-14-1-0022 AOARD 134131.

References

Ankerst, M., Elsen, C., Ester, M. and Kriegel, H-P. (1999) 'Visual classification: an interactive approach to decision tree construction', in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '99*, ACM, New York, NY, USA, pp.392–396.

Becker, B., Kohavi, R. and Sommerfield, D. (2002) 'Visualizing the simple Bayesian classifier', in Fayyad, U., Grinstein, G.G. and Wierse, A. (Eds.): *Information Visualization in Data Mining and Knowledge Discovery*, pp.237–249.

Caragea, D., Cook, D. and Honavar, V.G. (2001) 'Gaining insights into support vector machine pattern classifiers using projection-based tour methods', in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '01*, pp.251–256.

Dai, J. and Cheng, J. (2008) 'HMMEditor: a visual editing tool for profile hidden Markov model', *BMC Genomics*, Vol. 9, Suppl. 1, p.S8.

Erra, U., Frola, B. and Scarano, V. (2011) 'An interactive bio-inspired approach to clustering and visualizing datasets', in *Proceedings of the 15th International Conference on Information Visualisation 2011, IV '11*, pp.440–447.

Fails, J.A. and Olsen Jr., D.R. (2003) 'Interactive machine learning', in *Proceedings of IUI2003, IUI '03*, pp.39–45.

Fiebrink, R. (2011) *Real-time Human Interaction with Supervised Learning Algorithms for Music Composition and Performance*, PhD thesis, Princeton University, Princeton, NJ, USA.

Fiebrink, R. and Trueman, D. (2012) 'End-user machine learning in music composition and performance', in *CHI 2012 Workshop on End-User Interactions with Intelligent and Autonomous Systems*, Austin, Texas.

Field, A. (2009) *Discovering Statistics Using SPSS (Introducing Statistical Method)*, 3rd ed., SAGE Publications Ltd., London, UK.

Frank, E. and Hall, M. (2003) 'Visualizing class probability estimators', in Lavrač, N., Gamber-ger, D., Todorovski, L. and Blockeel, H. (Eds.): *Knowledge Discovery in Databases: PKDD 2003, Lecture Notes in Computer Science*, Springer, pp.168–179.

Henelius, A., Puolamäki, K., Boström, H., Asker, L. and Papapetrou, P. (2014) 'A peek into the black box: exploring classifiers by randomization', *Data Min. Knowl. Discov.*, Vol. 28, Nos. 5–6, pp.1503–1529.

Hung, J.C., Lee, M.F. and Wang, Y.B. (2011) 'Using emotional classification model for travel information system', *Int. J. Comput. Sci. Eng.*, Vol. 6, No. 4, pp.283–293.

Ibrahim, J.G., Chen, M-H. and Sinha, D. (2005) 'Bayesian survival analysis', in *Encyclopedia of Biostatistics*, John Wiley & Sons, Ltd.

Jakulin, A., Možina, M., Demšar, J., Bratko, I. and Zupan, B. (2005) 'Nomograms for visualizing support vector machines', in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD '05*, ACM, New York, NY, USA, pp.108–117.

Kulesza, T., Stumpf, S., Wong, W-K., Burnett, M.M., Perona, S., Ko, A. and Oberst, I. (2011) 'Why-oriented end-user debugging of naive Bayes text classification', *ACM Trans. Interact. Intell. Syst*., Vol. 1, No. 1, pp.2:1–2:31.

Li, Z., Zhang, B., Wang, Y., Chen, F., Taib, R., Whiffin, V. and Wang, Y. (2014) 'Water pipe condition assessment: a hierarchical beta process approach for sparse incident data', *Mach. Learn*., Vol. 95, No. 1, pp.11–26.

Nakagawa, S. (2004) 'A farewell to Bonferroni: the problems of low statistical power and publication bias', *Behav. Ecol*., Vol. 15, No. 6, pp.1044–1045.

Paiva, J.G., Florian, L., Pedrini, H., Telles, G. and Minghim, R. (2011) 'Improved similarity trees and their application to visual data classification', *IEEE Trans. Vis. Comput. Graph*., Vol. 17, No. 12, pp.2459–2468.

Patel, K., Fogarty, J., Landay, J.A. and Harrison, B. (2008) 'Examining difficulties software developers encounter in the adoption of statistical machine learning', in *Proceedings of the 23rd National Conference on Artificial Intelligence*, Chicago, Vol. 3, pp.1563–1566.

Perneger, T.V. (1998) 'What's wrong with Bonferroni adjustments', *BMJ*, Vol. 316, No. 7139, pp.1236–1238.

Rothman, K.J. (2010) 'Curbing type I and type II errors', *Eur. J. Epidemiol*., Vol. 25, No. 4, pp.223–224.

Swamy, M.K., Reddy, P.K., Kiran, R.U. and Reddy, M.V. (2012) 'Temporality-based user interface design approaches for desktop and small screen environment', *Int. J. Comput. Sci. Eng*., Vol. 7, No. 1, pp.52–64.

Talbot, J., Lee, B., Kapoor, A. and Tan, D.S. (2009) 'EnsembleMatrix: interactive visualization to support machine learning with multiple classifiers', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09*, pp.1283–1292.

Tarng, W., Lin, C.M., Liu, Y.T., Tong, Y.N. and Pan, K.Y. (2011) 'A virtual reality design for learning the basic concepts of synchrotron light source', *Int. J. Comput. Sci. Eng*., Vol. 6, No. 3, p.175.

Wagstaff, K. (2012) 'Machine learning that matters', in *Proceedings of ICML2012*, pp.529–536.

Witten, I.H., Frank, E. and Hall, M.A. (2011) *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers, Burlington, MA, USA.

Zhou, J., Li, Z., Wang, Y. and Chen, F. (2013) 'Transparent machine learning – revealing internal states of machine learning', in *Proceedings of IUI2013 Workshop on Interactive Machine Learning*.