Automatic Detection of User's Uncertainty in Problem Solving Task: A Multimodal Approach

Imène Jraidi, Maher Chaouachi, Claude Frasson
Université de Montréal
Dept. of Computer Science and Operations Research, QC, Canada
{jraidiim, chaouacm, frasson}@iro.umontreal.ca

Abstract

This paper presents a novel multimodal approach to automatically detect learner's uncertainty through the integration of multiple sensors. An acquisition protocol was established to record participants' electrical brain activity and physiological signals while interacting with a problem solving system specifically designed for uncertainty elicitation. Data were collected from 38 subjects using 8 sensors and two video feeds. Results from machine learning classifiers support the feasibility of our approach. 81% of accuracy was reached using Support Vector Machine (SVM) algorithm.

Keywords: uncertainty modeling, EEG, mental engagement, physiological sensors, intelligent systems, machine learning

Introduction

Endowing intelligent systems with abilities to assess users' cognitive and affective state is one of the most promising challenges to improve interaction methods, provide accurate support and enhance user performance (DuRousseau, Mannucci, and Stanley 2005). In this context, computer-based educational systems are focusing on identifying learner uncertainty as it is one of the most recurrently observed states during computer tutoring (Forbes-Riley, Rotaru, and Litman 2008) and due to its theorized relationship to learning (Craig et al. 2004). VanLehn et al. (2003) have shown that student uncertainty might warn learning impasses and thus cognitive difficulties but also signal predispositions to get more involved and engaged in learning. How can then a system automatically monitor user uncertainty? In most work so far, uncertainty modeling relies on acoustic-prosodic, lexical or discourse features extracted from utterance/dialogue based system interactions (D'Mello et al. 2008; Liscombe, Hirschberg, and Venditti 2005; Pon-Barry et al. 2006). However, we believe that such features could be insufficient, as they cannot always reflect user uncertainty. Besides, we believe that uncertainty encompasses cognitive factors as well as mental and emotional manifestations and is specific to each individual and context.

Nowadays intelligent systems integrate various gauges of cerebral and affective state through the use of nonintrusive electrophysiological sensors to accurately monitor user mental engagement and emotions (Picard 1997). In this paper we propose a multimodal approach to automatically assess learner uncertainty. This involves training machine learning techniques to model student uncertainty from electrophysiological parameters as well as cognitive and personal criteria. The hypothesis we establish is that these features can effectively predict learner uncertainty. An experimental study was conducted to validate our hypothesis. Our research questions were the following: can we predict learner uncertainty state? If so, can we model granularity levels of uncertainty?

The organization of this paper is as follows: In the first section, we present previous work done in fields similar to our own. In the second section, we detail our experimental methodology. In the third section we describe the features extracted for this study. In the fourth section, we present the obtained results and discuss them, in the last section, as well as present future work.

Previous Work

Significant research has been engaged in automatically recognizing uncertainty (D'Mello et al. 2008; Liscombe et al. 2005; Pon-Barry et al. 2006) and showing that adapting and responding to user uncertainty can greatly improve learning (Forbes-Riley, and Litman 2010; Pon-Barry et al. 2006). Pon-Barry et al. (2006) for example use linguistic cues (such as hedges, response latencies or filled-pause signals) extracted from human tutoring corpus through a frequency analysis to detect user uncertainty in a computer-based tutoring system. Liscombe et al. (2005) used acoustic-prosodic features to classify student certainness in a corpus collected from a speech-enabled intelligent tutorial system. Carberry, and Schroeder (2002) proposed an algorithm to recognize doubt by examining linguistic and contextual features of dialogue in conjunction with world knowledge. However in most of these studies, uncertainty modeling has been addressed without considering the arising mental state or affective reactions and that could be relevant in its assessment.

On the other side, the integration of neuro-physiological data in intelligent systems proved their effectiveness in assessing user state. Indeed, research in artificial intelligence is now accurately identifying user affective state through the use of non-intrusive sensors, analyzing signals like heart rate, skin conductivity or speech (Picard 1997). Besides, with the advent of portable and consumer oriented electroencephalogram (EEG), it is now possible to measure user mental state with a high time resolution and precision and develop systems that directly modulate their tasks to neural indexes of cognition. EEG engagement index developed at NASA (Pope, Bogart, and Bartolome 1995), is one of the most effective brainpower based mental indicators. It was used in a closed-loop system to modulate task allocation. Results have demonstrated that performance was improved when this index was used as a criterion for switching between manual and automated piloting mode (Pope et al. 1995). This index was also related to learner emotions as well as their performance in an educational context (Chaouachi et al. 2010). In this research we propose to integrate this indicator as engagement state is theoretically related to uncertainty (VanLehn et al. 2003). Our approach will also use affective indicators from physiological sensors as well as cognitive and personal criteria.

Experimental Methodology

The experiment was thoroughly established to assess a multidimensional aspect of uncertainty by integrating multiple data sources. Experimental setup consisted of a computer-based problem solving system, a 6-channel EEG system, physiological sensors and two video feeds. All the data were synchronized using necessary time markers. This setup is important for our investigation to integrate the recorded signals with the rest of instrumental setup under specific (un)certainty states. Once learners were equipped with the equipment, a 5-minute baseline was recorded during which learners were instructed to relax, to establish a neutral state for the electrophysiological parameters.

The problem solving system developed for this study consists of a series of logical tasks that do not require particular perquisites but involve high level of mental engagement and concentration. These tasks imply inferential skill on information series and are typically found in brain training exercises or in tests of reasoning. The system is composed of 3 modules. Each module is concerned with specific forms of data: the first module deals with geometrical shapes, the second module with numbers and the third module with letters. Each module starts with a tutorial explaining the task and giving examples to get user accustomed with the types of problems. Then learners have to answer to 5 multiple-choice questions. Learners are asked to respond as quickly and efficiently as possible. They were informed that a correct answer is rewarded 4 points, -1 point is given for a bad answer, whereas 0 point is given for a no-answer. A fixed time limit of 80 seconds for each question was imposed. Failing to give an answer within the allowed time was considered as a no-answer.

One of the most important points in this research is to elicit uncertainty states to obtain an accurate mapping of the recorded parameters. Thus problem tasks were selected in a way that potentially can cause uncertainty. To choose the right answer, learners needed to perceive a logical rule. Without this rule, the learner is not able to be sure of his answer. Moreover, problems may have different difficulty levels and some of them can involve a second rule to decide between two answers that both match the first rule. For instance in the geometrical module, three shapes are successively presented in the interface. The first shape represents a black triangle, the second a white rectangle and the third a black pentagon. Learner is then asked to deduce the fourth element by choosing one answer among five possibilities. In this example the rule that one should deduce is to add a side in each shape and the correct answer would be a hexagon. Two hexagons (black and white) were included among the propositions and only one matches to the second rule that one should also deduce (i.e. alternating between the two colors) and the correct answer would be the white hexagon. Other questions were designed to systematically mislead learners. For instance in the number-based module, two perpendicular data series were presented. In the vertical series all the numbers are multiples of seven and in the horizontal series all the numbers are multiples of five. In this task one should deduce the missing intersection data element of both series and that should be multiple of both five and seven. But no such data is given among propositions. Hence while resolving a problem, learner can be either certain or uncertain about the accuracy of his reasoning and therefore his response can proceed from both states.

After each given answer, the system interacted with learners and prompted them to report how they answered to each question by choosing between the following: “I was certain about may response” or “I was uncertain about my response”. Furthermore to assess uncertainty granularity levels, learners are prompted to choose between the following: “I was certain at 50% or more” or “I was certain at less than 50%” whenever an uncertain response is reported. Hence three possibilities can be registered for each question: certain (Cert), uncertain (Uncert) and no-answer (No_Resp) with two possible granularity levels for Uncert namely certain at 50% or more (Low_Uncert) or at less than 50% (High_Uncert).

Further cognitive parameters were recorded during the task such as response time, and scores. Learners were also asked to fill in widely used information about their skill level on logical based problem solving. Non-cognitive personal variables were also measured. These included gender and scales on a personality test namely the Big Five Inventory (BFI) (John, Naumann, and Soto 2008). This test scales personality traits according to five dimensions namely openness, conscientiousness, extraversion, agreeableness, and neuroticism (OCEAN).

Electrophysiological Recordings

Signals were recorded from electroencephalogram (EEG), skin conductance (SC) and blood volume pulse (BVP) sensors. Data were digitized using the ProComp Infinity multi-channel data acquisition system.

EEG recordings. EEG is a measurement of electrical brain activity produced by synaptic excitations of neurons. During the session, learners wore a stretch electro-cap and EEG signal was recorded from sites P3, C3, Pz and Fz as defined by the International 10-20 Electrode Placement System (Jasper 1958). Each site was referenced to Cz and grounded at Fpz. Two more active sites were used namely A1 and A2 typically known respectively as the left and right earlobe. This setup is known as “referential linked ear montage” and is depicted in figure 1. Roughly speaking, in this montage the EEG signal is equally amplified throughout both hemispheres. Moreover, the “linked-ear” setup yields a more precise and cleaner EEG signal by calibrating each scalp signal to the average of left and right earlobe sites (A1 and A2). For example, the calibrated C3 signal is given by (C3-(A1+A2)/2).

Each scalp site was filled with a non-sticky proprietary gel from Electro-Cap and impedance was maintained below 5 kilo Ohms. Any impedance problems were corrected by rotating a blunted needle gently inside the electrode until an adequate signal was obtained. The recorded sampling rate was at 256 Hz. Due to its weakness (in the order of micro volts), the EEG signal needs to be amplified and filtered. Besides, the electrical brain activity signal is usually contaminated by external noise such as environmental interference caused by surrounding devices. Such artifacts alter clearly the quality of the signal. Thus a 60-Hz notch filter was applied during data acquisition to remove these artifacts. In addition, the acquired EEG signal is easily suffering from noise caused by user body movements or frequent eye blinks. Thus a 48-Hz high pass and 1-Hz low pass de-noising filters were applied.

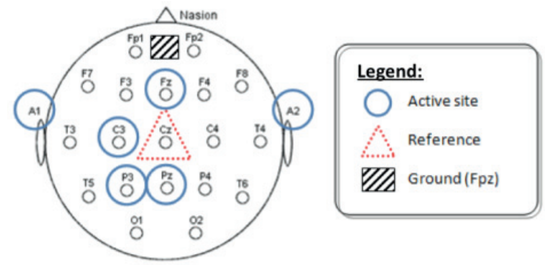

Figure 1: EEG Channel electrode placement

Physiological recordings. BVP and SC sensors were placed in the resting left hand fingers. Data were recorded at a sampling rate of 1024 Hz. SC measures changes in the resistance of the skin produced by sweat gland activity. A tiny voltage is applied through two electrodes strapped to the first and middle fingers on the palm side. This establishes an electric circuit and allows us to quantify the skin's ability to conduct electricity. BVP sensor was placed on the tip of the ring finger. It emits an infrared light and measures the amount of light reflected by the surface of the skin. This amount varies with the amount of blood present in the skin and thus with each heartbeat.

Participants

Thirty-eight learners (14 women) with a mean age of 27.31 ± 6.87 years, ranging from 19 to 47 years, were recruited for the experiment. Participation was compensated with 10 dollars. All participants were briefed about the experimental objectives and procedure and asked to sign a consent form.

Feature Extraction

A total of eleven features were extracted from the collected data to automatically learn uncertainty models: two EEG mental features, five physiological features and four additional cognitive and personal features.

EEG features. The engagement index was computed from EEG raw signal. As previously mentioned, this index reflects mental engagement level on a task (Pope et al. 1995). The engagement index is derived from the following ratio: (Beta / (Alpha + Theta)). An EEG power spectrum was calculated for each electrode site using a Fast Fourier Transformation (FFT) and the needed frequency bands were extracted, namely Theta (4-8 Hz), Alpha (8-13 Hz) and Beta (13-22 Hz). EEG band powers were then summed from the electrode sites P3, C3, Pz and Fz to compute the global ratio. The EEG engagement index at instant T was computed by averaging each engagement ratio within a 40s sliding window preceding instant T. This procedure was repeated every 2s and a new 40s sliding window was used to update the index. Two features were used from this index namely the *mean engagement* measured for each question and the *engagement variation* computed by subtracting the actual mean engagement values from the mean baseline.

Physiological features. SC signals were used to derive galvanic skin response (GSR) widely known to linearly vary with arousal ratings (Lang 1995). It increases as a person becomes more stressed. The *mean GSR* value in each question was then calculated. Also the *GSR variation* was recorded by subtracting the current GSR values from the baseline.

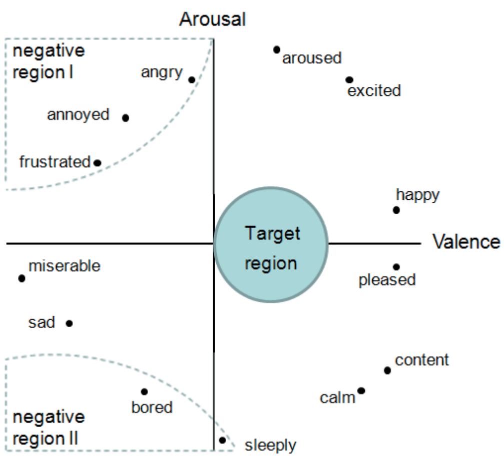

Figure 2: Russell's circumplex model of emotions with regions

From the BVP signal, heart rate (HR) was calculated by measuring the inverse of the inter-beat intervals (distance between successive pulse peaks). HR is extensively applied to understand the autonomic nervous system function and has shown a close correlation to valence (Lang 1995). Both *Mean HR* and *HR variation* were recorded for each entry. HR and GSR are jointly used to measure specific emotional activations and are widely used for emotional detection as emotions are characterized in terms of judged valence (negative to positive) and arousal (low to high). Physiological signals were thus analyzed according to the arousal/valence emotional space within the widely used Russell's circumplex model of emotions (Russell 1980). Two strategic emotional regions are defined during learning as depicted in figure 2 (Kaiser 2006). The first region involves negative emotions like frustration, boredom or angriness (negative region I and II) and should be avoided. The second region is the target emotional region specified by a slight positive valence and neutral arousal. This region is known to provide a maximum of efficiency and productivity (Kaiser 2006). We then extracted the *proportion of positive emotions* in the target region for each entry.

Additional features. Further parameters were considered namely question *response time*, as well as three personal criteria collected from learner self report: *skill level* on logical based problem solving (low, medium or high), *gender,* and scales on the BFI test. From the five personality traits (OCEAN), we have only considered the *conscientiousness scale* as a significant correlation was found between this trait and reported uncertainty levels. More precisely, we found that conscientiousness was positively correlated with the number of Cert responses ($r = 0.364$, $p < 0.05$), negatively correlated with the number of Uncert responses ($r = -0.399$, $p < 0.05$) and negatively correlated with the number of High_Uncert responses ($r = -0.501$, $p < 0.01$). Table 1 summarizes correlational results.

Table 1: Bi-variate correlation results

Correlation between conscientiousness personality trait and uncertainty levels (N = 38)		
	R	*p*
Cert	0.364*	0.0250
Uncert	-0.399*	0.013
High_Uncert	0.076	0.652
Low_Uncert	-0.501**	0.001
No_Resp	-0.019	0.908

* Correlation is significant at the 0.05 level (2 tailed).
** Correlation is significant at the 0.01 level (2 tailed).

Uncertainty Detection

Three approaches were considered to detect uncertainty namely: Naïve Bayes classifier, Decision Trees, and Support Vector Machines. The extracted features were fed as an input into these algorithms to automatically learn student uncertainty levels.

Naïve Bayes classifier

The Naïve Bayes classifier is a probabilistic learning algorithm that applies the Bayes' rule to compute the posterior probabilities of the different classes given the input attribute values with a 'naïve' class independence assumption. The algorithm assigns then a given sample to the class having the highest posterior probability according to the maximum a posteriori (MAP) decision rule (John, and Langley 1995). Despite its inaccurate assumption, it has been found that the Naïve Bayes classifier performs well and with comparable performance compared to other classification approaches (Han, and Kamber 2005).

Decision Tree

The decision tree classifier is a top-down divide-and-conquer approach that uses a tree like structure of decision to induce interpretable classification rules. Each node tests a particular input feature, the branches emerging from that node are possible test outcomes and terminal or leaf nodes represent the class value that will be returned. In this work, we use the well known C4.5 software, an extension of the ID3 decision tree induction algorithm (Quinlan 1986) which has been implemented as the J48 algorithm (Witten, and Frank 2005). At each node, the algorithm selects the attribute that most effectively splits the set of samples by maximizing the information gain. The algorithm then recurs to a smaller subset of samples to make classifications in the node's sub-tree (Safavian, and Landgrebe 1991). Given an unknown sample, the classifier routes it down the tree according to its attribute values tested in successive nodes tracing a path from the root to the leaf which holds the class prediction for that sample.

Support Vector Machines

Support vector machine is a linear machine learning system working in a high k-dimensional feature space formed by an implicit processing of an n-dimensional input data X into a k-dimensional space (k>n) through the use of a nonlinear mapping $\phi(X)$. This allows constructing hyperplanes that linearly separate data normally only separable with non-linear rules in the input space into classes. The algorithm searches for maximal margin hyperplanes creating decision boundaries with the highest possible margin or separation between classes. In this work we use the Sequential Minimal Optimization (SMO) algorithm (Platt 1999) for training support vector machines.

Performance Evaluation

To evaluate the classifier performance, we use a K-fold cross validation technique (Efron, and Tibshirani 1993) where the input data set is divided into K subsets. The classifier is trained on K-1 subsets and evaluated on the remaining subset. This process is repeated K times, the accuracy estimates are averaged to yield the overall classifier accuracy. This study employed the Weka software (Witten, and Frank 2005), a collection of machine learning algorithms intended for data mining tasks, for the three algorithms.

Experimental Results and Discussion

Our first objective was to create and train a model to predict the state of uncertainty by taking as an input electrophysiological data as well as personal and cognitive parameters. Thus we first trained a binary classifier to predict responses formulated from an uncertain reasoning from those resulting from certainness (Uncert, Cert). Then, we extended our analysis to predict uncertainty in a more detailed granularity level (High_Uncert, Low_Uncert, Cert).

A total of 570 entries (15 questions x 38 participants) were collected through this experiment and were classified as follows: 323 for Cert responses, 189 for Uncert (103 High_Uncert and 86 Low_Uncert) and 58 No_Resp. Although the small proportion of the no-answers (10.17%), one question can be raised: should these samples be included with uncertain responses indicating that the learner was so uncertain about his answer that he did not take the risk to respond? Or these merely indicate that the learner did not have the time to respond even if he knew the correct answer?

Table 2: Classifier accuracy results

1st dataset (No_Resp included)			
Classes	*DT*	*NB*	*SMO*
Cert, Uncert	77.72%	76.15%	79.13%
Cert, Uncert, No_Resp	69.65%	69.23%	72.99%
Cert, Low_Uncert, High_Uncert, No_Resp	63.16%	62.63%	63.34%
2nd dataset (No_Resp excluded)			
Classes	*DT*	*NB*	*SMO*
Cert, Uncert	78.71%	78.32%	81.64%
Cert, Low_Uncert, High_Uncert	73.24%	70.50%	73.44%

We hence considered two separate datasets. In the first dataset, No_Resp samples were either included with the Uncert samples or gathered in a separate class. In the second dataset, No_Resp samples were discarded. Results of classification accuracies from Decision Tree, SVM and Naïve Bayes classifier are listed in table 2. Prediction performance was evaluated using a 20-fold cross validation.

As presented in table 2, the SVM classifier has shown the highest prediction rates in all cases with accuracies ranging from 63.34% for the 4-class model (Cert, Low_Uncert, High_Uncert, No_Resp) to 81.64 % for the binary model (Cert, Uncert) excluding the no-answers from the training set. Indeed, we noticed that merging the

No_Resp examples in the Uncert category slightly decreases the quality of the model to 79.13%, which suggests that trained models are clearly sensitive to the introduced parameters and since the no-answers can involve both uncertainty and certainty state, a bias is introduced in the model.

Table 3 shows the details of classification accuracy for the best classifier namely the SVM algorithm (Kappa statistic = 0.67) among the 2 classes Cert and Uncert for the second dataset (No_Resp exluded).

Table 3: SVM detailed accuracy by class

Classes	*Precision*	*Recall*	*F-Measure*
Uncert	0.757	0.741	0.749
Cert	0.85	0.861	0.855

Conclusion

We have presented in this paper a multimodal approach to automatically assess students' uncertainty from their electrophysiological activity as well as cognitive and personal criteria. Our research questions were the following: can these features efficiently predict the state of uncertainty and if so, can we model uncertainty granularity levels?

An experiment was conducted during which learners interacted with a problem solving system and were asked to respond to a series of logical tasks. EEG, BVP and SC sensors were used to record participants' electrophysiological signals. Machine learning techniques were used to classify uncertainty: up to 81% of accuracy was reached. Results suggest that this approach can be further extended to handle uncertainty levels (73%) reinforcing our belief that electrophysiological sensors could be a reliable alternative for intelligent systems to assess user state. In our future work we are planning to develop an intelligent agent to implement appropriate pedagogical strategies according to model predictions. This agent will use associations between user's actions and detected uncertainty level to thoroughly adapt problem level and support to the learner.

Acknowledgments

We would like to acknowledge the CRSNG (Conseil de Recherches en Sciences Naturelles et en Génie du Canada) for funding this work.

References

Carberry, S., and Schroeder, L. 2002. Toward recognizing and conveying an attitude of doubt via natural language. *AAI* 16: 495 - 517.

Chaouachi, M., Chalfoun, P., Jraidi, I., and Frasson, C. 2010. Affect and Mental Engagement: Towards Adaptability for Intelligent Systems. *23rd International FLAIRS Conference*. AAAI Press, Florida, USA.

Craig, S., Graesser, A., Sullins, J., and Gholson, B. 2004. Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media* 29: 241-250.

D'Mello, S., Craig, S., Witherspoon, A., McDaniel, B., and Graesser, A. 2008. Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction* 18: 45-80.

DuRousseau, D. R., Mannucci, M. A., and Stanley, J. P. 2005. Will augmented cognition improve training results? *1st International Conference on Augmented Cognition*, Las Vegas, NV.

Efron, B., and Tibshirani, R. 1993. *An Introduction to the Bootstrap*. Chapman and Hall.

Forbes-Riley, K., and Litman, D. 2010. Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. *Comput. Speech Lang.* 25: 105-126.

Forbes-Riley, K., Rotaru, M., and Litman, D. J. 2008. The relative impact of student affect on performance models in a spoken dialogue tutoring system. *User Modeling and User-Adapted Interaction* 18: 11-43.

Han, J., and Kamber, M. 2005. *Data Mining: Concepts and Techniques, Second Edition.* San Francisco, CA: Morgan Kaufmann Publishers.

Jasper, H. H. 1958. The ten-twenty electrode system of the International Federation. *Electroencephalography and Clinical Neurophysiology*: 371-375.

John, G. H., and Langley, P. 1995. Estimating continuous distributions in Bayesian classifiers. *11th Conference on Uncertainty in Artificial Intelligence*, pp. 338-345.

John, O. P., Naumann, L. P., and Soto, C. J. 2008. Paradigm Shift to the Integrative Big-Five Trait Taxonomy: History, Measurement, and Conceptual Issues. In: O. P. John, R. W. Robins, and Pervin, L. A., (Eds), *Handbook of personality: Theory and research*. NY: Guilford Press.

Kaiser, R. 2006. Prototypical development of an affective component for an e-learning system. Master Thesis, Department of Computer Science, University of Rostock, Germany.

Lang, P. J. 1995. The emotion probe: Studies of motivation and attention. *American Psychologist* 50: 372-385.

Liscombe, J., Hirschberg, J., and Venditti, J. J. 2005. Detecting Certainness in Spoken Tutorial Dialogues. *9th European Conference on Speech Communication and Technology*, pp. 1837-1840, Lisbon.

Picard, R. 1997. *Affective Computing*. MIT Press.

Platt, J. C. 1999. Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods*. MIT Press.

Pon-Barry, H., Schultz, K., Bratt, E. O., Clark, B., and Peters, S. 2006. Responding to Student Uncertainty in Spoken Tutorial Dialogue Systems. *Int. J. Artif. Intell. Ed.* 16: 171-194.

Pope, A. T., Bogart, E. H., and Bartolome, D. S. 1995. Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology* 40: 187-195.

Quinlan, J. R. 1986. Induction of Decision Trees. *Mach. Learn.* 1: 81-106.

Russell, J. 1980. A circumplex model of affect. *JPSP* 39: 1161-1178.

Safavian, S. R., and Landgrebe, D. 1991. A survey of decision tree classifier methodology. *IEEE Trans. Sys. Man Cyb.* 21: 660-674.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., and Baggett, W. B. 2003. Why Do Only Some Events Cause Learning During Human Tutoring? *Cognition and Instruction* 21: 209 - 249.

Witten, I. H., and Frank, E. D. 2005. *Data Mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.