Modeling Students' Behaviors and Emotions while using an Intelligent Tutoring System for Algebra (Aplusix)

Background

An intelligent tutoring system (ITS) is a computer program that makes use of artificial intelligence to help students learn a target instructional domain (Conati, Gertner, & VanLehn, 2002). It maintains a model of student reasoning and learning and uses this model to adapt to a student's specific needs.

However, unlike human teachers, ITSs cannot sense students' affective state. They cannot tell if a student is frustrated, distressed, satisfied, or angry. Yet the recognition of affect is an important part of the teaching/learning process. Good teachers acknowledge students' emotional cues and temper responses accordingly, providing assurance, guidance, or praise as needed. Similarly, an intelligent system that aims to adapt to student needs must recognize student emotions and respond to them appropriately (Amershi, Conati, & Maclaren, 2006; Picard, 1997).

Objectives

The long-term goal of this research undertaking is to build an emotionally intelligent tutoring system. This is a multistage process that combines social science research with data mining, artificial intelligence, human-computer interaction, and educational technology. This first study will produce a classifier that correctly identifies a student's affective state, given a set of inputs. The researchers will gather data regarding student feelings and behaviors while interacting with an intelligent tutoring system for algebra, Aplusix. Using the observational data and the Aplusix log files, the researchers will establish correlations between behaviors and students' emotions. The data will then be used to train the classifier.

Scope and limitations

For this study, the researchers will record student behaviors and emotional responses when using an ITS. The scope of the behaviors observed were limited to on-task behavior, off-task behavior, inactivity, and gaming the system. "Gaming" refers to attempts to perform well in an educational task by "systematically taking advantage of properties and regularities of a system, rather than thinking about the material" and is negatively correlated with learning (see Baker, Corbett, Koedinger, & Wagner, 2004)

At the same time, the observers will note the whether the students express any one of six emotional states at the time of observation: boredom, confusion, delight, surprise, flow, frustration, and neutrality (D'Mello, Craig, Gholson, Franklin, Picard, & Graesser, 2005).

The data logger of Aplusix will record the students' interactions, e.g. past actions, time per action, outcomes, etc. which can be synchronized with the observational data. Researchers will mine the data to draw inferences about the students' cognitive or affective states.

Significance

Learning is an emotional experience. Learning episodes begin with curiosity and fascination. As the learning task increases in complexity, learners begin to experience confusion, frustration, and anxiety. These may lead to abandonment of the learning activity. Students who push beyond the negative emotions and achieve success are rewarded with satisfaction and pride. A good teacher helps students master the subject matter and marshal emotions towards a goal (Picard, 1997). A good ITS must do the same.

ITS research has been reasonably successful at guiding students through the cognitive aspects of a subject. However, research into creating emotionally aware ITSs has just begun. The assumption behind creating emotionally intelligent tutoring systems is that a system can interact more effectively with the student if it had information of the student's affective state (Conati, Chabbal, & Maclaren, 2003).

Research in this area is challenging because the business of emotions is highly ambiguous. Different people have different emotional reactions to the same situation (Conati, 2002). People tend to mask all but the strongest emotions (Amershi, Conati, & Maclaren, 2006). What emotions to detect, whether they are universal or culture-specific, and which of these are relevant to learning are still open to debate (Picard, 1997). Research in emotionally intelligent tutoring systems, though, does not attempt to answer all these questions. Rather, the goal of the field is to build models that enable ITSs to make principled decisions about how to react and possibly influence a student's emotional states (Conati & Zhou, 2004).

Review of Related Literature and Conceptual Framework

Since the 1960's, studies on emotion were left out of the scientific mainstream (Damasio, 2000). It was only in the 1990's that more and more researchers in various disciplines including computer science and engineering turned their attention to emotion, spawning, among other things, a field called affective computing. Affective computing is defined as computing that relates to, arises from, or deliberately influences emotions. It attempts to give computers the ability to recognize, express, and respond to emotions intelligently (Picard, 1997). By applying the principles of affective computing to ITSs, it is possible to design and develop emotionally intelligent tutoring systems.

An emotionally intelligent tutoring system first has to be able to recognize emotions. Based on facial expressions, body language, and content and tone of speech, a human teacher can reasonably judge a student's emotional state (Kort & Reilly, 2001). How can these same abilities be built into an ITS? Picard (1997) summarized the following design criteria of system that can recognize emotions:

1. Input: The system must accept a variety of input signals related to emotion.
2. Pattern recognition: The system must extract features from the data gathered and classify them into to significant categories, e.g. a smile versus a frown.
3. Reasoning: The system predicts the underlying emotions based on rules about how emotions are generated and expressed.
4. Learning: The system tunes the rules, based on individual nuances.
5. Bias: The system must have its own emotional state that influences recognition of ambiguous emotions.
6. Output: The system names or describes the emotion.

In terms of inputs, Bradley and Lang (2000) organized measurable data on emotions into three broad categories:

1. Overt acts or functional behavioral sequences
2. Emotional language including expressive communication
3. Physiological reactions in support of overt acts of emotion

ITS and affective computing researchers are challenged to capture these inputs and correlate them with the correct label.

A number of studies have attempted to arrive at models for judging student affective states. Because human beings are capable of hiding their emotions, some studies conducted at the University of British Columbia and the Massachusetts Institute of Technology use biometric sensors to monitor students' physiological responses. Data from skin conductance, heart rate, and electromyogram sensors synchronized with log files have been used to record student reactions to events in an educational game (Amershi, Conati, & Maclaren, 2006; Conati, Chabbal, and Maclaren, 2003). These studies established threshold values above which readings are considered indicative of an emotional response and correlated these with game events. Among the findings were that time between student actions is an important determinant of student affect. For example, short time intervals may imply impatience.

Mota and Picard's (2003) study used a chair outfitted with sensors to record subjects' posture while using an educational game. An adult observer labeled behaviors and affective states (high, medium, and low interest; taking a break; bored). The data from the sensors synchronized with the observations were then used to train a neural network. The trained neural network was subsequently able to recognize student emotions with an accuracy of 82.3%. If these assessments were then fed back into the educational game, the game could intervene when student interest begins to wane.

Other studies detected student emotions without using biometrics. De Vicente and Pain (2002) used expert opinion to arrive at rules to diagnose student motivation. They showed a series of experts videos of individual children using a tutoring system. Using a specially-designed software package, the experts could play back recordings of student interactions and then rate the student's emotional state. From the exercise, the researchers arrived at a set of 85 motivational diagnosis rules, e.g. mouse movement was not random and the student performed the task quickly therefore the student was confident and highly satisfied. Based on these rules, the researchers concluded that it is possible to make diagnose a student's motivational state based solely on the information provided by the student's interaction with the tutoring system.

Information on student achievement and log files of student behavior with an exploratory learning environment enabled researchers to associate behaviors with low learning outcomes and high learning outcomes (Amershi & Conati, 2006). After training, the resulting detector was able to identify behaviors that were detrimental to learning, even after seeing only 10% of a student's actions. If incorporated in an ITS, the model could enable the ITS to choose the most appropriate interface or intervention based on a learner's classification.

Baker, Corbett, and Koedinger (2004) used human observation to detect student misuse of an ITS. Observers noted when students "gamed" the system, e.g. when students re-read stories they already knew or asked for help even before attempting to address the problem.

From these data sets, researchers built models that could reliably detect differences in how students choose to use an intelligent tutoring system. They incorporated these models into intelligent agents, giving the agents the capability to exhibit positive emotions when students used an ITS properly and negative emotions when students harmfully gamed the system (Baker, Corbett, Koedinger, Evenson, Roll, Wagner, Naim, Raspat, Baker, & Beck, 2006). They also tried to develop a generalized detector that could be used in another ITS without retraining (Baker, Corbett, Koedinger, & Roll, 2006). They found that it is possible to build a generalized detector by collecting data from a broad set of curricular lessons and training the detector with these sets. A detector trained with data from a single curricular lesson will over-fit the detector to that lesson. The resulting detector will not be as accurate with other curricular lessons.

This study will be limited to Picard's (1997) recognition design criteria. It will not attempt to build an ITS as part of this project's scope.

Methodology

The data collection methodology is adapted from the methodology in Baker, Corbett, Koedinger, and Wagner (2004). Behavior and emotion data will be collected from high school students from the schools within the Metro Manila/Rizal area. The list of schools has yet to be finalized but may include the following:

- Assumption Antipolo
- Ateneo de Manila
- College of the Holy Spirit
- Kostka School
- St. Paul's Pasig
- Xavier School

Prior to the laboratory exercise, the target students will be asked to answer a 10-item pre-test. They will have 20 minutes to complete it.

During the laboratory exercise itself, ten students will be observed using the software in a computer laboratory at any one time. Upon entering the computer laboratory, each student was directed to a pre-assigned computer. One researcher will brief the students on the purpose of the activity and on the use of the software. After the orientation, they will be asked to complete a questionnaire indicating their ages; grade average from last year; computer access, skills, and aptitude; their interest in problem solving; and prior experience with Aplusix. The students will then be asked to use the software for 30 minutes.

Two to three observers will observe each student sequentially. (The observers have already been trained. They were graduate students of Dr. Rodrigo during the first semester of 2006-2007. During this time, they conducted a similar experiment and committed to continuing the research.) Each member of the team will observe the same student at the same time. Each observation will last 20 seconds, after which the observers will move on to the next student. In a 30-minute time span, the observers will be able to collect 80 observations in total or 8 observations per student.

To synchronize the movements of the teams, a timed PowerPoint presentation with slides numbered 1 to 80 will be flashed on the wall of the computer laboratory. The numbers 1 to 80 represent the observation period. Each team member will glance at the slide on the wall to verify whether it is time to move to the next student.

Observers will record their observations regarding behavior and emotion on pre-printed checklists. With regards to behavior, the observers will indicate whether the student observed was

1. On-task and working on the tutor
2. On-task and engaged in conversation with the teacher or with another student
3. Off-task and engaged in conversation with another student
4. Off-task and engaged in a solitary activity
5. Inactive
6. Gaming the system

With regards to emotion, the observers will indicate whether the student expresses any of the following categories of emotion (D'Mello, Craig, et al. 2005):

1. Boredom
2. Confusion
3. Delight
4. Surprise
5. Flow
6. Frustration
7. Neutrality

The observers will put a check mark beside the first behavior and emotion they detected. If, during the 20-second window, the student exhibits more than one behavior or emotion, the observers only record the first one observed.

At the end of the observation period, the students were given a ten-problem, written post-test. The students will be given 20 minutes to answer the post-test.

At the end of data collection, the observation data will then be synchronized with the Aplusix log files. Analyses of the data will include findings on:

- Relationships between emotions and behaviors (both observed and logged)
- Relationships between student achievement (based on self-reported grade point averages and post-test results) and emotions or behaviors