Using Feature Selection and Unsupervised Clustering to Identify Affective Expressions in Educational Games

Saleema Amershi[1], Cristina Conati[1], and Heather Maclaren[1]

[1]Department of Computer Science, University of British Columbia,
2366 Main Mall, Vancouver, BC, V6T 1Z4, Canada
{samershi, conati, maclaren}@cs.ubc.ca

Abstract. Educational games can induce a wide range of emotions, and so recognizing specific emotions may be valuable for an intelligent system that aims to adapt to varying student needs so as to improve learning. The long-term goal of this work is to understand how user affect impacts overall learning in an educational game. The main contribution of this paper is an investigation into the use of an unsupervised machine learning technique to help recognize meaningful patterns in biometric affective data. Results show that this method can identify interesting and sensible student reactions to different game events.

Keywords: biometrics, affective modeling, unsupervised learning

1 Introduction

Educational games can induce a wide range of emotions [4], and so recognizing specific emotions may be valuable for an intelligent system that aims to adapt to varying student needs so as to improve learning. The work presented here is an investigation into biometric expressions of affective reactions exhibited by students interacting with an educational game. Measuring the students' biometric expressions allows us to record affective reactions without interrupting the student's engagement during game-playing. The long-term goal of this work is to understand how student affect impacts overall learning in an educational game.

Other work on identifying biometric expressions of affect has looked at using supervised machine learning techniques to produce classifiers that map recorded data to affective labels (e.g., [12]) and to perform feature selection (e.g., [16]). Affective labeling of interaction events has involved methods such as video-annotations of specific emotions (e.g., [13]). However, we found in previous work [3] that most students tended to internalize all but the strongest feelings, leading to a low agreement rate amongst annotators. Collecting self-reports of emotion during engaging interactions [2] is also problematic since this method can only ask a very small number of highly directed questions in order to reduce disruption. Therefore students are unable to accurately report all of their current feelings.

We present an alternative approach to analyzing students' biometric expressions of affect that occur within an educational game. Instead of attempting to identify and label affective reactions to game events as they occur and then look for common

G. Rebolledo-Mendez, E. Martinez-Miron (Eds): Workshop on Motivational and Affective Issues in ITS. 8th International Conference on ITS 2006, pp 21-28, 2006.

features of biometric expression, we took a data mining approach in which we employed an unsupervised machine learning technique to automatically discover patterns of biometric expressions of emotions in unlabelled data. To do this we first computed a set of physiological and behavioral features for each event that occurred in the game, and then performed unsupervised feature selection and clustering to form an unbiased picture of the most common biometric expressions of emotion towards events within the game. Our analysis showed that only a few of the features initially determined were relevant in defining the clusters produced. Furthermore, clustering was able to identify several meaningful patterns of reactions within the data.

In this paper we give an overview of the study performed to collect data on game events and students' biometric expressions. We then describe our approach to the data analysis, including which features we extracted from the collected data, and the clustering method used. Finally, we present and discuss the results of using unsupervised clustering on our biometric data.

2 Introduction to PrimeClimb

Figure 1 shows a screenshot of PrimeClimb, a game designed to teach number factorization to 6th and 7th grade students. In the game, two players must cooperate to climb a series of mountains that are divided into numbered sectors. Each player can only move to a numbered sector that does not share any factors with the sector occupied by his partner. When a player makes a wrong move, she falls and starts swinging from the climbing rope. The game includes a pedagogical agent that can advise a student playing with a partner and can engage a student in a "practice climb" during which it climbs with the student as a climbing instructor. The goal of the pedagogical agent is to provide tailored support to help the player to learn number factorization while maintaining a high level of engagement. We are currently developing an affective model of the student so that the agent can include affective information when tailoring its responses [5]. Information about the student's current affective state, as well as which emotions are beneficial to student learning, will help the agent decide on appropriate interventions.

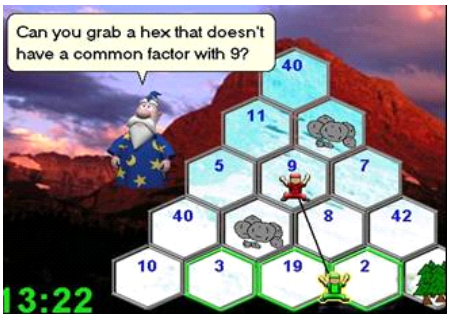

Fig. 1. The PrimeClimb educational game

Fig. 2. A student playing PrimeClimb wearing the sensors

3 Study Design

Thirty 6th and 7th grade students from two local schools interacted with PrimeClimb for approximately 10 minutes. The students played the game with an experimenter, who controlled the second climber and attempted to play as if she was the pedagogical agent accompanying the student in a "practice climb", i.e. making as few mistakes as possible. The pedagogical agent was autonomous and based its interventions on a Dynamic Bayesian network model of student learning [15].

Each biometric recording was synchronized with logs of the game events that could stimulate an emotional reaction (e.g., a climb, a fall, an agent intervention). The students' biometric expressions were recorded using 4 sets of sensors: Skin conductance (SC), heart rate (HR), and 2 electromyogram sensors on the forehead (EMG1 & EMG2). Our previous study [3] showed that using only one EMG sensor on the corrugator muscle enabled us to record muscle tension but was unable to distinguish between frowns and eyebrow raises. [7] showed comparing the signal to a second EMG sensor (EMG2), placed over the zygomatic muscle, should enable these two expressions to be distinguished. Figure 2 shows a student wearing the sensors.

4 Analysis of Study Data

Clustering is a class of unsupervised machine learning techniques used to automatically discover patterns in unlabeled data. Clustering operates on any measurable properties of the data (i.e., features). In this section we discuss the set of features we have chosen to perform clustering on, and then describe the feature selection and clustering method used. Finally, we present the results of our analysis.

4.1 Features

Initially, we identified 28 possibly influential features based on previous literature on user affect (e.g. [9,14]) and knowledge about the PrimeClimb application. This set included: *time to next event (seconds)*, *mean and standard deviation of the signal (EMG, SC and HR)*, *number of peak responses above the threshold*[1] *(EMG and SC)*, *mean and standard deviation of the peak responses (EMG and SC)*, *sum of peak response magnitudes and durations (EMG and SC)*, *sum of the areas under the peak responses (EMG and SC)*, and *average gradient of the peak responses (SC)*.

A new feature vector was computed for each event within the game (e.g. student falls, agent interventions). Physiological features were computed in the 4 second interval immediately after an event occurred in order to ascertain student reactions. [11] showed that this time is adequate for capturing responses in the biometric signals. The SC signals were normalized by subtracting the baseline SC value, and dividing by the SC range for the entire session. EMG and HR signals were normalized by subtracting the baselines for these signals. Because of time restrictions with the

[1] Threshold values were computed by 'mean + $k \cdot$ standard' (k=1 for SC and k=3 for EMG)

students who were taken out of class to participate in the study, we were unable to obtain baseline values during a rest period prior to playing the game. Instead, baseline values were obtained for each student at the time when SC was minimal over the entire session. Signal values were taken at this time because students tended to be excited at the start of the game, but would soon relax while playing (exhibited by a slight drop in the SC signal before increasing again as the interaction progressed).

The maximum level in the game reached by all students was level 3. Therefore, for our analysis we only used events that occurred between levels 1 and 3 to ensure that all of the students were exposed to the same game stimuli.

4.2 Method

Clustering was performed using an (average linked) hierarchical algorithm as defined in [8]. Initially, each feature vector begins in its own cluster. Then, between-cluster distances are computed in feature space, and nearby clusters are merged. This process is repeated until all of the clusters are eventually merged. The final result can be visualized using a *dendrogram* (e.g., Figure 3), where the height of each node represents the distance/dissimilarity between the two clusters connected at that node. Cutting the dendrogram at a given height results in a set of clusters separated by at least the corresponding distance. We used hierarchical clustering, as opposed to partition based clustering (e.g., *k*-means), because we cannot assume the clusters are isotropic, nor do we know the number of clusters in the data beforehand [10].

Feature selection is necessary in high dimensional data mining applications because natural groupings are often obscured by irrelevant features. In [8], the authors introduce an unsupervised feature selection procedure which can be used in conjunction with distance based, hierarchical clustering algorithms to detect subgroups of objects, each having (possibly different) relevant features. Feature relevance is determined by importance values. An importance value of *n* means that the range of values observed for that feature in the given cluster is approximately 1/*n*th of the overall range of values observed for that feature in the data. We considered a feature to be important to a cluster if its range of values was at most half the overall range of values (i.e., had an importance value of 2 or greater).

4.3 Results

Feature selection using the 28 dimensional feature vectors showed that only the following features were found to be relevant for clustering: *time to next event* and the *mean and standard deviation of the signal (EMG1, EMG2, SC and HR)*. The features found to be unimportant all correspond to peaks in the signals. Few peaks were found for any signal during the 4 seconds subsequent to an event, which could indicate that threshold values used were too high. In [9] the authors used these features to detect a single strong emotion, driver stress. In an educational game scenario, students may experience a range of (possibly conflicting) emotions [4] that are unlikely to be as pronounced as driver stress. The features found to be relevant may be less sensitive to this limited arousal. Further investigation is necessary to evaluate this hypothesis. For

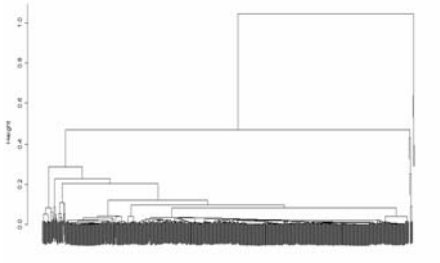

Fig. 3. Dendrogram produced for student climbs using the 28D feature vectors

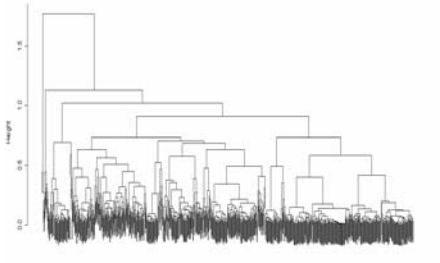

Fig. 4. Dendrogram produced for student climbs using the reduced, 9D feature vectors

the rest of our analysis, we used only the 9 features found to be relevant in this initial analysis because these lead to clearer cluster separation (see Figures 3 and 4).

We performed a separate feature selection and cluster analysis for each type of game event: Student Climbs, Student Falls, Experimenter Climbs, and Agent Interventions. We cut the dendrograms at a dissimilarity level of 0.5 as this seemed to result in compact and well separated clusters (see Figure 4 for example). We analyzed clusters containing 10 or more feature vectors assuming that meaningful clusters would contain at least this many similar student reactions to any specific game event. Here we discuss the most interesting findings for each analysis.

Student Climbs. 772 feature vectors were computed for this analysis, producing 18 clusters. For each cluster, a different (and sometimes overlapping) set of features were found to be important.

Cluster 2 (14 feature vectors) and Cluster 12 (61 feature vectors) were both highly clustered on the mean of the SC signal. The SC values for these clusters were lower than the other clusters (see Figure 5). The mean and standard deviations of the EMG1 and EMG2 signals were also shown to be important for Cluster 12. The values for these signals were centered about the baseline, with low standard deviation. These reactions could be interpreted as the students having low emotional arousal [11] following a climb, possibly indicating disengagement from the game. Alternatively, the student is simply feeling calm and in control due to making a successful move.

Clusters 13 and 14 (23 and 16 feature vectors respectively) exhibited high arousal, as indicated by high SC values compared to the other groups (see Figure 5). These clusters were also highly clustered on the mean of the HR signal, with values slightly higher than the baseline. Both clusters showed a higher EMG1 signal mean than EMG2, signifying more tension along the corrugator muscle, i.e. a possible frowning expression. These results suggest that the students may have been frustrated.

In contrast, Cluster 15 (113 feature vectors) showed a higher EMG2 signal mean than EMG1. In combination, this suggests that the students may have been raising their eyebrows, possibly indicating surprise that a move had been successful. The SC signal mean was not found to be highly relevant for this cluster because of the evident SC signal variance (see Figure 5). This finding is unlike Clusters 13 and 14, where the mean SC values were shown to be important and high. This appears to support results from previous studies (e.g., [6]) that suggested that the SC signal should be higher for emotions with negative valence than for emotions with positive valence.

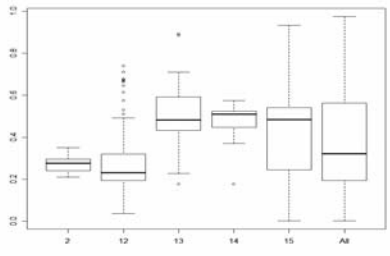

Fig. 5. Feature values along *mean SC signal* dimension

Student Falls. Only 141 student falls were found within the data set producing 7 clusters. Only two of the clusters showed meaningful results: Cluster 3 (22 feature vectors) and Cluster 5 (13 feature vectors). Cluster 3 exhibited similar characteristics as the 'calm' groups (Clusters 2 and 12) described in the previous section. Cluster 5 was analogous to the 'frustrated' groups (Clusters 13 and 14) from the previous section. Interestingly, no cluster exhibited expressions similar to those found for the cluster that indicated a 'surprised' reaction (Cluster 15 in the previous section).

Experimenter Climbs. 602 feature vectors were computed for the 602 successful experimenter moves. 18 clusters were found. Three different clusters showed characteristics of frustration, Clusters 7, 9 and 11 (10, 44 and 21 feature vectors respectively). All of these were highly clustered on the mean and standard deviation of the SC signal, all being high compared to the rest of the data. These clusters also showed a higher EMG1 signal mean then EMG2, suggesting tension along the corrugator muscle.

Cluster 12 (48 feature vectors) exhibited characteristics of low arousal defined by high importance along the SC and EMG signal dimensions, with low values in all cases. Here, it is interesting that the mean time to the next game event was higher for this cluster than the mean for the entire data set (7.46 seconds and 4.62 seconds respectively). Further examination of what the students were doing during this period is needed to determine whether this may be an indication of student disengagement.

Analysis for experimenter climbs also reveals several clusters in which only the time until the next game event was determined to be important. These were Clusters 2, 3, 4 and 5 (62, 47, 114 and 143 members respectively). The mean times were 4, 5, 2, and 3 seconds respectively. For each cluster, the signal values along each of the other dimensions were highly variable. Inspection of the dendrogram shows that these clusters would have been grouped together at a cut height of 0.6 (see Figure 6).

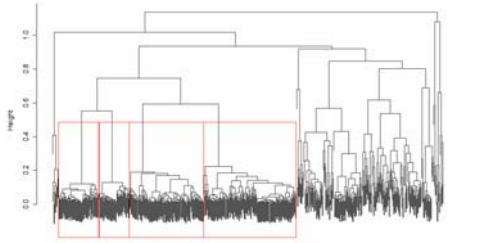

Fig.6. Dendrogram produced for Experimenter Climbs (Clusters 2-5 are outlined)

Agent Interventions. 114 agent interventions occurred corresponding to this many feature vectors. 12 clusters were produced. Clusters 2 (20 members) and 12 (10 members) were characterized by low arousal and EMG means centered at the baseline with low variance. Cluster 2 showed low mean and standard deviation for the HR signal, while Cluster 12 showed low mean and standard deviation for the SC signal.

Again, a large (50 member) cluster was found that was only highly clustered on the time to the next event, and variable along the other dimensions. The mean time was low compared to the rest of the data (5.42 seconds compared to 9.91 seconds).

5 Discussion

Significantly more successful moves were made by both the student and the experimenter, in comparison to the other game event types. It is expected that more interesting results would be obtained for these data sets as unsupervised clustering tends to work better with more available data. It is likely that very few failed moves were found because we were only looking at events in levels 1 to 3. More failed student moves may be obtained using all the levels.

Overall, feature selection determined that the only relevant biometric features were the mean and standard deviation of EMG, SC and HR. Analysis of the agent interventions and experimenter climbs also showed time to be an important factor. However, in evaluating these results, it is important to consider what the next event in the game was. For example, very short time intervals after experimenter climbs followed by a student move could suggest a student was impatient. Thus it may be beneficial to experiment with additional features, such as the type of the next event.

Hierarchical clustering could have been affected by the feature normalization method chosen. In this work, features were normalized by subtracting baseline values with baselines computed at the time when the SC signal was lowest. And so feature values at any given time are relative to this baseline. Alternatively, we could have computed feature values at a given time relative to values at the previous time step.

Analysis was only done on the clusters obtained by dendrogram cuts of height 0.5. However, as indicated by the visible subclusters within the clusters found for Experimenter Climbs, variable dendrogram cuts may reveal more meaningful clusters.

6 Conclusions and Future Work

This paper describes the feature selection and cluster analysis on affective reactions to events in the PrimeClimb educational game. Unsupervised clustering was able to identify several interesting patterns of reactions to game events that may help in the understanding how user affect impacts learning. This work demonstrates the benefits of this machine learning technique compared to manual labeling of affective data.

Our first next step is to conduct a more detailed analysis of the affective data, addressing the issues that arose in this work. We also plan to annotate video footage (e.g., frowning and eyebrow raises) in order to compare the performance of the unsupervised method to manual labeling techniques.

In an effort to reach our long term goal of understanding how affect impacts learning in an educational game, we intend to examine how the biometric expressions that we found correlate with learning outcomes computed from pre-tests and post-tests that our student participants also completed during the study. If there are strong correlations between occurrences of biometric expressions and learning outcomes, then we may be able to use the clusters to construct a classifier as in [1] that could be used by the pedagogical agent to provide help. Alternatively, we could incorporate our findings within the Dynamic Bayesian Network (DBN) model of user affect that we are currently developing to inform the pedagogical agent [5]. This DBN is able to account for both *causes* and *effects* of a student's emotions. Causal factors, such as user goals, may better explain some of the effects detected by hierarchical clustering.

References

1. Amershi, S., Conati, C.: Automatic Recognition of Learner Groups in Exploratory Learning Environments. To appear in Intelligent Tutoring Systems (2006)
2. Conati, C.: How to Evaluate models of User Affect? Tutorial and Research Workshop on Affective Dialogue Systems. Kloster Irsee, Germany (2004)
3. Conati, C., Chabbal, R., Maclaren, H.: A Study on Using Biometric Sensors for Monitoring User Emotions in Educational Games. Workshop on Modeling User Affect and Actions: Why, When and How, User Modeling. Johnstown, PA, USA (2003)
4. Conati, C., Klawe, M.: Socially Intelligent Agents in Educational Games. In: Dautenhahn, K., Bond, A., Canamero, D., Edmonds, B. (eds.): Socially Intelligent Agents - Creating Relationships with Computers and Robots. Kluwer Academic Publishers (2002)
5. Conati, C., Maclaren, H.: Data-driven Refinement of a Probabilistic Model of User Affect. User Modeling (2005)
6. Ekman, P., Levenson, W., Friesen, W.V.: Autonomic nervous system activity distinguishes among emotions. Science, Vol. 221 (1983) 1208-1210
7. Fernandez, R., Scheirer, J., Picard, R.: Expression glasses a wearable device for facial expression recognition. Tech. Rep. 484, MIT Media Lab (1999)
8. Friedman, J.H., Meulman, J.J.: Clustering Objects on Subsets of Attributes. Tech. Rep. Stanford University (2002)
9. Healey, A.H., Picard, R.W.: Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. IEEE Trans. On Intelligent Transportation Systems 6 (2005) 156-166
10. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. ACM Computing Surveys 31 (1999) 264-323
11. Lang, P. J., Greenwald, M.K., Bradley, M.M., Hamm, A.O.: Looking at pictures: Affective, facial, visceral, and behavioral reactions. Psychophysiology 30 (1993) 261-273
12. Lisetti, C., Nasoz, F.: Using Noninvasive Wearable Computers to Recognize Human Emotions from Physiological Signals. J. Applied Signal Processing 11 (2004) 1672-1687
13. Kapoor, A., Picard R.: Multimodal Affect Recognition in Learning Environments. Multimedia. Singapore (2005)
14. Mandryk, R.L., Inkpen, K.M., Calvert, T.W.: Using Psychophysiological Techniques to Measure User Experience with Entertainment Technologies. J. Behavior and Info. Tech. (Special Issue on User Experience) 25 (2006) 141-158
15. Manske, M., Conati, C.: Modeling Learning in Educational Games. AI in Education. (2005)
16. Vyzas, E., Picard, R.: Affective Pattern Classification. AAAI Fall Symposium Series: Emotional and Intelligent: The Tangled Knot of Cognition (1998)