Ranking Feature Sets for Emotion Models Used in Classroom Based Intelligent Tutoring Systems

David G. Cooper[1], Kasia Muldner[2], Ivon Arroyo[1],
Beverly Park Woolf[1], and Winslow Burleson[2]

[1] University of Massachusetts, Department of Computer Science,
140 Governors Drive, Amherst MA 01003, USA
dcooper@cs.umass.edu
[2] Arizona State University, School of Computing and Informatics,
Tempe AZ 85287, USA

Abstract. Recent progress has been made by using sensors with Intelligent Tutoring Systems in classrooms in order to predict the affective state of students users. If tutors are able to interpret sensor data with new students based on past experience, rather than having to be individually trained, then this will enable tutor developers to evaluate various methods of adapting to each student's affective state using consistent predictions. In the past, our classifiers have predicted student emotions with an accuracy between 78% and 87%. However, it is still unclear which sensors are best, and the educational technology community needs to know this to develop better than baseline classifiers, e.g. ones that use only frequency of emotional occurrence to predict affective state. This paper suggests a method to clarify classifier ranking for the purpose of affective models. The method begins with a careful collection of a training and testing set, each from a separate population, and concludes with a non-parametric ranking of the trained classifiers on the testing set. We illustrate this method with classifiers trained on data collected in the Fall of 2008 and tested on data collected in the Spring of 2009. Our results show that the classifiers for some affective states are significantly better than the baseline model; a validation analysis showed that some but not all classifier rankings generalize to new settings. Overall, our analysis suggests that though there is some benefit gained from simple linear classifiers, more advanced methods or better features may be needed for better classification performance.

1 Introduction

Student affect plays a key role in determining learning outcomes from instructional situations [1, 2]. For instance, learning is enhanced when empathy or support is present [3, 4]. While human tutors naturally recognize and respond to affect [5, 6], doing so is quite challenging for Intelligent Tutoring Systems (ITS), in part due to the lack of directly-observable information on a student's affect. A promising avenue for increasing model bandwidth, i.e., the quality and degree of information available to a student model, in terms of affect recognition

P. De Bra, A. Kobsa, and D. Chin (Eds.): UMAP 2010, LNCS 6075, pp. 135–146, 2010.

is sensing devices that capture information on students' physiological responses as they interact with adaptive systems. With the advent of inexpensive sensor technology, we have been able to deploy such sensing systems and use their output to infer information on student affect. Specifically, in the Fall of 2008 we performed a number of experiments in the classrooms of schools in both Western Massachusetts and Arizona, with a total of just under 100 students. In each experiment, students were queried about four emotional states (*confident*, *interested*, *frustrated*, and *excited*), providing the standard for validating our models. The study data was used to construct a number of linear classifiers for each emotional state, as we reported in [7]. The best classifiers for a given emotion obtained accuracies between 78% and 87% according to a leave-one-student-out cross-validation.

While these results are promising, it is important to validate the classifiers and verify that their performance generalizes to a new and/or larger population. This is particularly the case for our data, obtained from a classroom setting which involves a higher degree of noise and other distractions than standard controlled laboratory experiments. One aspect of validation involves verifying that our classifiers perform better than the baseline classifier (i.e., one that always outputs yes if the labels are yes most of the time, or no if the labels are no most of the time). In addition to validating our classifier performance, we also wanted to investigate if and how the sensors (or subsets of sensors) improved model performance over using only features from the tutor data (e.g. the number of hints requested). With an understanding of how each combination of sensor and tutor features predicts a given emotion, we can recommend which sensors to use for emotion recognition, and we can also rank the classifiers so that if some sensor data is unavailable, for instance due to an error, a comparable (or the next best) sensor set can be selected.

Thus, in this paper, we report on how we realized these objectives by utilizing a large data set for validation from experiments that we conducted in the Spring of 2009 with over 500 students. Our results show that our method is successful on three of our four target emotions: for each success, at least one linear classifier performs better than the baseline classifier and generalizes to a new and larger population.

We begin by presenting the related work in Sect. 2, and then describing in Sect. 3 the setup and apparatus of the experiments used to collect the data. Section 4 outlines the method for constructing and validating the student emotion classifiers. Section 5 describes the comparison of classifiers. Section 6 summarizes the results, discusses the design of affective interventions based on the classifier output, and suggests future work on improving the classifiers.

2 Related Work

The results of a feature selection competition in 2004 suggest that feature selection can be very useful for improving classifiers [8]. In addition to using simple correlation coefficients as criteria for selection (as stepwise linear regression does), treed methods, wrapper and embedded methods have been used for feature selection. [9] compares features of a number of individual sensors used for detecting affective state with an ITS, but does not compare disparate sensors, nor are multiple sensors used in conjunction in a classifier. In this paper we use a method from [10] to compare and rank the different feature sets used in the linear classifiers as a way of ranking our features selected by stepwise linear regression.

There are a number of adaptive systems in existence that use real-time information about a student in order to address the student's affective state. Recent work includes [11], which discusses the use of electromyogram (EMG) data to improve an affective model in an educational game. This work does careful collection, cross-validation, and uses a pairwise t-test (a parametric test) for ranking the classifiers. [12] aimed to predict learners' affective states (boredom, flow/engagement, confusion, and frustration) by monitoring variations in the cohesiveness of tutorial dialogues during interactions with an ITS with conversational dialogs; here, both student self reports and independent judges were used to identify emotional states. The study compared the correlation between self-reports and independent judges, and used tutor and dialogue features automatically classify emotion with accuracies between 68% and 78%.

Other work, such as [13, 14], does not incorporate any sensor data to construct affective models. [13] uses Dynamic Bayesian Networks and Dynamic Decision Models specified by an expert to determine and respond to each student's affective state, while [14] uses self-reports to determine affective state and focuses on how affective feedback changes the student's experience. This work does use cross-validation and a parametric ranking for classifiers, but does not do a feature comparison or a validation with a separate population.

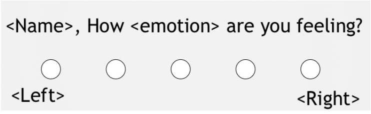
Much of this past research has focused on constructing models based on a fixed set of sensors or solely on expert knowledge. In contrast, our research compares the utility of different sensors as well as sensor and tutor interaction features in a variety of empirically-based models. Another difference relates to the source of the data: Since our data is obtained from actual schools rather than the laboratory, the ecological validity of our results is strengthened. Our features are ranked using non-parametric procedures and take an extra step of validating on a separate population in order to address the additional artifacts created by a classroom setting.

3 Data Collection: Sensors with Wayang Outpost in the Classroom

3.1 Setup

In the Fall of 2008 and the Spring of 2009 the geometry tutor Wayang Outpost was deployed with a set of sensors into real classroom environments [7, 15, 16]. The set of sensors included: a mouse that captured degree of pressure placed on its various points, a bracelet that measured skin conductance of the wrist, a chair that sensed the level of pressure on the chair back and seat, and a camera supplemented with software for facial emotion recognition.

These four sensors collected data on students' physiological responses while students worked with Wayang Outpost. Each student's physiological data and interactions with the tutor were logged. Subsequently, the interaction and sensor data were time-aligned and converted into tutor and sensor features, as described in [7]. At intervals of five minutes in the Fall, and three minutes in the Spring, students were presented with an emotional query about one of four affective states (*confident*, *interested*, *frustrated*, or *excited*) selected from a uniform random distribution. The queries were presented as shown in Fig. 1; to respond, students selected from the options shown in Table 1. The sensor and tutor features were used as predictors for the levels of the self-reported affective states.



<Name>, How <emotion> are you feeling?

○ ○ ○ ○ ○

<Left> <Right>

Fig. 1. An example of the Emotion query. Table 1 below has the values for each <> enclosed word, except for (<*Name*>), which is the name of the student.

Table 1. The mapping of tags to text in Fig. 1 above

<emotion>	<Left>	<Right>
confident	I feel anxious	I feel very confident
interested	I am bored	I am very interested
frustrated	Not frustrated at all	Very frustrated
excited	I'm enjoying this a lot	This is not fun

The Fall 2008 data collection involved 93 students using the Wayang Tutor. Of the 93 students 85 of them had at least one working sensor connected to them while using the tutor. Students used the tutor as part of a class, and class sizes ranged from three to twenty-five students with one teacher in the classroom and between one and three experimenters. The students had between two and five sessions with Wayang Outpost, based on teacher preference and availability of the student. The student ages were 15-16, 18-22, and 22-24. These data were used as our training set.

The Spring 2009 data collection involved over 500 students using the Wayang Tutor. 304 of the students were connected to at least one working sensor. The Spring collection differed from the Fall collection as follows: (1) The students in the Spring were from different schools; (2) The ages of Spring students were 13-14, and 15-16; (3) The camera sensor in the Spring had upgraded software. The Spring data was used purely for validation of the Fall Data.

3.2 Tutor and Sensor Features

We considered nine tutor features and forty sensor features as potential predictors for the emotion classifiers (see Table 2). The forty sensor features are based on four ways of summarizing ten specific features: the mean, the standard deviation, the min value, and the max value over the course of a problem. Since the sensor and tutor logging happens asynchronously, their data are interpolated in a piecewise constant fashion with the constraint that only data from the past is used to predict missing sensor or tutor values. The tutor logs when a problem is opened and closed, creating boundaries for summarizing the interpolated sensor data (i.e. to compute each feature, we use data over the span of a single problem). When there is an emotional query after a problem, the result becomes the affective state label for that problem. For each student and for each emotion there are between two and five affective-state labels. For more detail on the full specification of these features see [7].

Table 2. Features used for each problem that includes an affective state label in order to train the emotion classifiers (features are shown in abbreviated form). The nine tutor features are shown on the left and the ten sensor features are on the right. Features used in a classifier that is significantly better than the baseline ($p < 0.05$) are in **bold**.

Tutor feature	Definition
Solv. on 1st	1st attempt correct
Sec. to 1st	time to 1st attempt
Sec. to solv.	time to a correct
# incorrect	responses
# hints	requested
LC	learning companion
Group	which LC (Jake, Jane, or none)
Time in session	same day
Time in tutor	all days

Sensor feature	Definition
Agreeing	camera mental states
Concentrating	
Thinking	
Interested	
Unsure	
Mouse	sum of pressure
Sit Forward	movement in chair
Seat change	
Back change	
Skin conductance	value from wrist

4 Method

The current standards for evaluating affective classifiers do not address our need to rank classifiers for the purpose of actionable affect detection. Though each individual step in our method has been established and tested, the combination of these steps yields a more robust test for the classifiers constructed. The use of our classifiers in a classroom environment necessitates our method described in the rest of this section and summarized in Table 3.

4.1 Collection

The data collection described in Sect. 3 is the first step in our methodology for building affect classifiers. The key parts of the data collection are that the emotion labels are made at the time of the experience, and the training and validations sets are taken from distinct populations using the same basic setup, allowing the validation results to be more likely to generalize. Here, the Fall collection is our training data set and the Spring collection is our validation data set.

Table 3. Our affect detection method summarized

1. Data Collection
   - in situ self-reports of emotion
   - training and validation sets from different population
2. Feature Selection
   - remove central self-report values
   - use step-wise linear regression to select features and train classifiers
3. Cross-validation (leave-one-student-out)
   - compute the mean accuracy, sensitivity, and specificity per student
4. Classifier ranking
   - parametric and nonparametric ranking using $p < 0.05$
5. Validation
   - run steps 3 and 4 on validation set using classifiers from step 2

4.2 Predictor Selection

Once the data were collected and summarized as described in Sect. 3.2, we used the entire set of labeled training data to create a subset of predictors using a combination of tutor and sensor features. For each combination of features, a subset of the data set that was not missing data for the features was selected. Then stepwise linear regression was performed in R to select the 'best' subset of features from those available. The subset of features was stored as a formula for use in training the classifiers and performing cross validation.

4.3 Cross Validation

For each set of features determined by the feature selection, we performed leave-one-student-out cross-validation on linear classifiers for each affective state. During the cross-validation, we calculated the mean accuracy, sensitivity, and specificity for each test student. We also performed the same cross-validation on a linear classifier with a constant model, which we used as our baseline. This step differs from [7] in two ways: 1) The mean was taken across each test student instead of across tests. 2) We calculated sensitivity and specificity in addition to accuracy.

Though the cross-validation described above provides a general indication of the performance of each classifier, the information is not sufficient to enable appropriate pedagogical action selection by an ITS for *new* populations of students. Thus, we validated that the classifiers are generalizable and so can be used with a new population without having to be retrained. We also ranked the classifiers according to how sensors and features impact accuracy, allowing us to make informed decisions about sensor selection (e.g. if some sensors become unavailable, to select the next best alternative).

4.4 Classifier Ranking

A number of alternative techniques exist for classifier comparison. One is to use classifier accuracy, which identifies the overall performance of a classifier, but does not express accuracy on positive vs. negative instances. To do so, the following two measures can be used: (1) sensitivity, also referred to as the true positive rate, which provides information about the accuracy of a positive response; (2) specificity, the true negative rate, which provides information about the accuracy of negative responses.

Since the purpose of our classifiers is to help an ITS make decisions of how to appropriately respond to student emotion, one approach would be to only make a decision when there is confidence in the prediction. So, if one classifier has very good sensitivity relative to the baseline, then the ITS would act when the classifier reports a positive result. Similarly, if a classifier has a very good specificity relative to the baseline, then the ITS would act when the classifier reports a negative result.

In order to compare our classifiers' accuracy, sensitivity, and specificity for each affective state, we first performed a one-way analysis of variance (ANOVA), with classifier as the independent variable and either accuracy, sensitivity, or specificity as the dependent variable. When there was a significant difference between classifiers, we performed Tukey's HSD test to rank the differences in the means.

There is some question about the soundness of the ANOVA and Tukey's HSD test for these comparisons because the design is not balanced (not every student had all sensors available), and the responses are not normally distributed. So, in addition to the ANOVA, a Kruskal-Wallace test was performed; when there was a significant difference between classifiers, a Nonparametric Multiple Comparison Procedure (NPMC) for an unbalanced one-way layout was performed, as described in [10].

We conducted both parametric and non-parametric tests because the parametric tests are known to be robust to violations of the assumptions, so performing both was a way to verify the findings. Here, for all tests, we only report results with significant differences.

4.5 Validation with Follow-on Data

As mentioned above, we used the Spring data set to validate the classifiers trained on the Fall Data set (the Spring data set was not used to inform any of the training). The validation consisted of the following three steps. First, for each feature set selected by the feature selection step, a linear classifier was created using the entire subset described in Sect. 4.2. Second, each classifier was tested on the relevant subset of data from the Spring data set. Third, the accuracy, sensitivity, and specificity values and rankings were compared to the cross-validated values and rankings to determine how the classifiers generalized to a new and larger population.

5 Results

The classifier sets were designed to compare the performance of (1) a classifier using just tutor features vs. (2) one using features from one sensor in addition to the tutor features vs. (3) a classifier using all of the available features. The collection, feature selection, and cross validation results from the training data (Fall 2008) are described in [7]; however, a couple of important details are needed here. First, although the feature selection has the option of using both tutor data and other sensor data, sometimes it only selected tutor data. Table 4 shows the results of the feature selection. Second, we extended the cross-validation results to include sensitivity and specificity. Third, we modified the grain size, in that the samples in this work are on a per student rather than per test basis. The ranking and validation results are discussed below.

Table 4. These are the results of the feature selection. The baseline classifier for each emotion is just a linear model trained on a constant. The classifier names are the concatenation of an abbreviated emotion and the contributing sensor features. If there are no sensor features, then Tutor comes after the emotion, and when there is more than one classifier with the same feature set a letter is added to disambiguate the names. Names in **bold** are for classifiers that performed significantly better than the baseline for that emotion in at least one way.

Classifier name	Features
confBaseline	constant
confTutorA	Solv. on 1st + Hints Seen
confTutorM	# Incorrect + Solv. on 1st + Session
confSeat	# Incorrect + Solv. on 1st + sitForward Std Dev.
intBaseline	constant
intMouse	Group + # Hints + mouse Std Dev + mouse Max
intCamera	Group + # Hints + interestedMin
excBaseline	constant
excTutor	Group + # Incorrect
excCamera	interested Mean + # Incorrect
excCameraSeat	netSeatChangeMean + interestedMin + sitForwardMean

5.1 Classifier Ranking

Accuracy had a significant main effect on both the *interested* and *excited* affective states, but not for the *confident* and *frustrated* states. For the *interested* state, the classifier using the mouse and tutor features is significantly better than the baseline with a mean of 83.56% vs. 42.42%, according to both Tukey's HSD and NPMC tests. For the *excited* state, the classifiers with the tutor features were significantly better than the baseline with a mean of 73.62% vs. 46.31%.

As far as sensitivity is concerned, there is a significant main effect for *confident*, *interested*, and *excited* affective states using both parametric and nonparametric tests. However for *confident*, no classifier performed better than the baseline. For *interested*, both the camera and tutor, and mouse and tutor features were better than the baseline. For *excited*, the camera with seat sensors, camera sensors, and tutor only performed better than the baseline.

For specificity, there is only a significant main effect for *confident*, with TutorA, TutorM, and Seat classifiers performing better than the baseline. The details of these results are shown in Table 5.

Table 5. Classifier ranking using cross-validation data ($p < 0.05$)

Confident	Tukey HSD	NPMC
Specificity	$(confTutorA \sim confTutorM \sim confSeat) > confBaseline$	$(confTutorA \sim confTutorM) > confBaseline$

Interested	Tukey HSD	NPMC
Accuracy	$intMouse > intBaseline$	$intMouse > intBaseline$
Sensitivity	$(intCamera \sim intMouse) > intBaseline$	$(intCamera, intMouse) > intBaseline$

Excited	Tukey HSD	NPMC
Accuracy	$excTutor > excBaseline$	$excTutor > excBaseline$
Sensitivity	$(excTutor \sim excCamera \sim excCameraSeat) > excBaseline$	$(excTutor \sim excCamera \sim excCameraSeat) > excBaseline$

Given these results, our findings suggest that the tutor could generate interventions more reliably when it detects interest and excitement. If the tutor wanted to intervene when the student is *interested*, then using the mouse and tutor features or the camera and tutor features would be most appropriate. If the tutor wanted to intervene when the student is *excited* then either the camera with seat features, camera features, or tutor features classifier would all be appropriate.

It may be more relevant to intervene when a student is not *interested* or not *excited*, or not *confident*. Our results do not provide information on which features to use to predict low interest or low excitement, but to detect lack of confidence, we could use either the TutorA, TutorM, or Seat features trained on *confident*. The corresponding features are shown in Table 4.

5.2 Validation with Follow-on Data

In order to verify that our classifier ranking generalizes to new data sets, we tested the classifiers by training them with all of the Fall data and testing them with the Spring data. Performance results of the significantly ranked classifiers from the cross-validation done above are compared to the validation set and shown in Table 6. Since the data are from an entirely separate population, it is likely that the overall performance will degrade somewhat; however, if each classifier's performance is similar, then that will provide evidence that the classifiers should be preferred as they were ranked during the cross-validation phase.

When comparing mean accuracy for the training vs. test sets, there is a general drop in accuracy of between 2% and 15%, though in some cases, there is a much larger difference of up to 37%. The larger differences suggest that some of the features do not generalize well to new populations.

Table 6. This shows validation results of all classifiers that performed better than the baseline classifier during training. All values are the mean value per student. Fall specifies the training set based on the leave-one-student-out cross-validation, and Spring specifies the results of the classifiers trained on the training set (Fall Data), and tested on the validation set (Spring Data). Values in **bold** are significantly better ($p < 0.05$) than the baseline.

model	Accuracy		Sensitivity		Specificity	
	Fall	Spring	Fall	Spring	Fall	Spring
confBaseline	65.06%	62.58%	72.22%	76.13%	55.56%	44.14%
confTutorA	70.49%	65.49%	47.07%	46.04%	**90.43%**	**84.88%**
confTutorM	68.64%	67.53%	52.31%	52.26%	**82.41%**	**80.68%**
confSeat	65.70%	67.13%	54.63%	60.17%	**79.26%**	**70.32%**
intBaseline	42.42%	78.30%	0.00%	0.00%	81.82%	100.00%
intMouse	**83.56%**	63.34%	**29.73%**	5.09%	90.54%	81.60%
intCamera	69.44%	57.65%	**52.08%**	**12.11%**	64.58%	68.53%
excBaseline	46.31%	74.31%	0.00%	0.00%	96.15%	100.00%
excTutor	**73.62%**	62.99%	**36.54%**	**12.45%**	87.88%	77.28%
excCamera	66.33%	51.53%	**38.67%**	**28.39%**	72.00%	52.24%
excCameraSeat	70.67%	43.34%	**32.00%**	**15.97%**	83.00%	54.07%

Results of ranking the classifiers on the validation data are shown in Table 7. Note that the accuracy rankings no longer hold, and the mouse classifier for the *interested* affective state is no longer significantly better than the baseline.

Table 7. Classifier ranking using validation data from the Spring of 2009. All differences indicated by '>' are significant with $p < 0.01$.

Confident	Tukey HSD	NPMC
Specificity	$(confCameraA \sim confTutorA \sim confTutorM) > (confSeat \sim confTutorW) > confBasline$ $confCameraB > confTutorW > confBaseline$	$(confCameraA \sim confTutorA \sim confTutorM) > (confSeat \sim confTutorW) > confBasline$ $confCameraB > confTutorW > confBaseline$

Interested	Tukey HSD	NPMC
Sensitivity	$intCamera > intBaseline$	$intCamera > intBaseline$

Excited	Tukey HSD	NPMC
Sensitivity	$((excCamera > excTutor) \sim excCameraSeat) > excBaseline$	$excCamera > excCameraSeat > excTutor > excBaseline$

6 Discussion

In this paper we describe a method for discovering actionable affective classifiers for Intelligent Tutoring Systems (ITS). Though the method was used with specific sensors, features, ITS and classifiers based on linear models, each of these could conceivably be swapped out for another system.

Our results identify a clear ranking for three classifiers designed to detect low student confidence, one classifier to detect interest, and three classifiers for detecting excitement. For not *confident*, two different sets of tutor only features performed better than the tutor and seat features, so it is unlikely that there would be a time that we would use the classifier with the seat sensor.

Now that we have actionable classifiers for three affective states, our ITS will be able to leverage the results to make a decision. For instance, the ITS could intervene whenever the classifier detects low student confidence, in order to help the student gain self efficacy. This intervention will have to also take into account other emotions detected, e.g., the detection of high excitement and/or high interest may change the type of intervention that is most appropriate.

Future work will involve implementing these various affect-based interventions, and evaluating their impact on student learning, affect and motivation. We also plan to explore how we can design classifiers for affect recognition that perform better than the baseline for the subset of affective states that our classifiers performed poorly on. One approach for doing so that we plan to implement is to identify more complex features based on the sensor data than those currently used. A more complete set of affective classifiers will likely improve the ITS interventions. For example, if we had a classifier that had good sensitivity for confidence, then that classifier could be used to stop interventions relating to low confidence.

Acknowledgments. We thank Sharon Edwards and Sarah English for their coordination of the school studies. This research was funded by awards from the National Science Foundation, 0705554, IIS/HCC *Affective Learning Companions: Modeling and Supporting Emotion During Teaching*, Woolf and Burleson (PIs) with Arroyo, Barto, and Fisher and the U.S. Department of Education to Woolf, B. P. (PI) with Arroyo, Maloy and the Center for Applied Special Technology (CAST), *Teaching Every Student: Using Intelligent Tutoring and Universal Design To Customize The Mathematics Curriculum.* Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

1. Beebe, S.A., Ivy, D.K.: Explaining student learning: An emotion model (1994)
2. Kort, B., Reilly, R., Picard, R.: An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion. In: IEEE International Conference on Advanced Learning Technologies, Proceedings, pp. 43–46 (2001)
3. Graham, S., Weiner, B.: Theories and principles of motivation. In: Berliner, D., Calfee, R. (eds.) Handbook of Educational Psychology, vol. 4, pp. 63–84. Macmillan, New York (1996)
4. Zimmerman, B.J.: Self-efficacy: An essential motive to learn. Contemporary Educational Psychology 25, 82–91 (2000)

5. Lepper, M.R., Woolverton, M., Mumme, D.L., Gurtner, J.L.: Technology in education. In: Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors, pp. 75–105. Lawrence Erlbaum Associates, Inc., Mahwah (1993)
6. Derry, S.J., Potts, M.K.: How tutors characterize students: a study of personal constructs in tutoring. In: ICLS '96: Proceedings of the 1996 international conference on Learning sciences, International Society of the Learning Sciences, pp. 368–373 (1996)
7. Cooper, D.G., Arroyo, I., Woolf, B.P., Muldner, K., Burleson, W., Christopherson, R.: Sensors model student self concept in the classroom. In: Houben, G.-J., McCalla, G., Pianesi, F., Zancanaro, M. (eds.) UMAP 2009. LNCS, vol. 5535, pp. 30–41. Springer, Heidelberg (2009)
8. Guyon, I., Gunn, S., Ben-Hur, A., Dror, G.: Result analysis of the nips 2003 feature selection challenge. In: Advances in Neural Information Processing Systems (2004)
9. D'Mello, S., Picard, R.W., Graesser, A.: Toward an affect-sensitive autotutor. IEEE Intelligent Systems 22(4), 53–61 (2007)
10. Munzel, U., Hothorn, L.A.: A unified approach to simultaneous rank test procedures in the unbalanced one-way layout. Biometrical Journal 43(5), 553–569 (2001)
11. Conati, C., Maclaren, H.: Modeling user affect from causes and effects. In: Houben, G.-J., McCalla, G., Pianesi, F., Zancanaro, M. (eds.) UMAP 2009. LNCS, vol. 5535, pp. 4–15. Springer, Heidelberg (2009)
12. D'Mello, S.K., Craig, S.D., Graesser, A.C.: Multimethod assessment of affective experience and expression during deep learning. Int. J. Learn. Technol. 4(3/4), 165–187 (2009)
13. Hernandez, Y., Arroyo-Figueroa, G., Sucar, L.: Evaluating a probabilistic model for affective behavior in an intelligent tutoring system, pp. 408–412 (July 2008)
14. Robison, J., McQuiggan, S., Lester, J.: Evaluating the consequences of affective feedback in intelligent tutoring systems, pp. 1–6 (2009)
15. Arroyo, I., Cooper, D.G., Burleson, W., Woolf, B.P., Muldner, K., Christopherson, R.: Emotion sensors go to school. In: Dimitrova, V., Mizoguchi, R., du Boulay, B., Graesser, A.C. (eds.) AIED, vol. 200, pp. 17–24. IOS Press, Amsterdam (2009)
16. Arroyo, I., Woolf, B.P., Royer, J.M., Tai, M.: Affective gendered learning companions. In: Dimitrova, V., Mizoguchi, R., du Boulay, B., Graesser, A.C. (eds.) AIED, vol. 200, pp. 41–48. IOS Press, Amsterdam (2009)