Genetic Search Feature Selection for Affective Modeling: a Case Study on Reported Preferences

Héctor P. Martínez
IT University of Copenhagen
Rued Langgaards vej 7, 2300
Copenhagen, Denmark
hpma@itu.dk

Georgios N. Yannakakis
IT University of Copenhagen
Rued Langgaards vej 7, 2300
Copenhagen, Denmark
yannakakis@itu.dk

ABSTRACT

Automatic feature selection is a critical step towards the generation of successful computational models of affect. This paper presents a genetic search-based feature selection method which is developed as a global-search algorithm for improving the accuracy of the affective models built. The method is tested and compared against sequential forward feature selection and random search in a dataset derived from a game survey experiment which contains bimodal input features (physiological and gameplay) and expressed pairwise preferences of affect. Results suggest that the proposed method is capable of picking subsets of features that generate more accurate affective models.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: User/Machine Systems—*Human factors*; I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems—*Games*

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Affective modeling, feature selection, genetic search, preference learning

1. INTRODUCTION

There is a growing interest for computer systems that can provide personalized affective experiences to the users. This is especially true for virtual worlds and video games given that the advances in game technology have made these environments capable of eliciting a plethora of emotional states [7] the inter-relationship of which is not trivial to be revealed. The first step towards developing affect-adaptive user experiences in game worlds (and beyond) is to create models that recognize the emotional state of the user based on the interaction with the system.

Human-machine interaction may provide access to large amounts of data from multiple modalities. The physiological state of the user, measured through different input biofeedback devices, contains useful information about the emotional state of the user [2, 10, 16] but the physical inputs introduced to the system, the actions performed on the virtual environment and their consequences (i.e. the interaction with the virtual environment) can also help in the prediction of ones emotional state [13, 11]. Given the large-scale streams of data that may be obtained via the interaction of the user with the system it is inappropiate to create computational models of emotion that exploit every single bit of such information. Hence, the information data sources are often reduced to a set of extracted *statistical features* (see [5] among others). Those indexes provide a representation of different characteristics of the user's input signals. Even thought the dimensionality is reduced via the extraction of statistical features not all the information embedded in them is necessarily relevant for building affective models of the user. Some studies have revealed that omitting unnecessary inputs improves the accuracy of the predictors (see [19, 17] among others). Additionally, it is desirable to keep the size of the model as small as possible both to make it real-time efficient and easier to analyze.

It is a rather complex task (if not impossible) to identify *a priori* which features are relevant for the prediction of an emotional state during a certain task. Exhaustive search is able find the best features but this is often infeasible due to the size of the solution space. Moreover, if the design of the predictor relies upon stochastic initial conditions (e.g. neuro-evolutionary learning), exhaustive search of the feature space does not guarantee that the best solution will be found. Therefore, automatic feature selection (FS) methods are necessary for dimensionality reduction of the search space if the designer of the model desires to maintain the physical meaning of all features that are extracted. The physical meaning of the parameters is particularly important for the analysis and the expressiveness of the affective model. Alternatively, one might use data pre-processing techniques such as principal component analysis [1] and fisher projection [9]; but those will transpose the parameter space in their effort to reduce it — thereby affecting any physical meaning of the features considered.

The problem of feature selection has been investigated on different research disciplines varying from statistics and data mining to user modeling. Jain et al. [8] perform a compar-

AFFINE'10, October 29, 2010, Firenze, Italy.
Copyright 2010 ACM 978-1-4503-0170-1/10/10 ...$10.00.

ative study of several automatic feature selection methods such as Sequential Backwards Selection (SBS) and Max-min. In the field of Affective Computing, Yannakakis et al. [19] apply two algorithms, n Best Individual feature Selection (nBest) and Sequential Forward Selection (SFS), to select the inputs of an entertainment predictor in a physical game from game and player features. To find the most relevant features from speech spectrograms to recognize emotions, SFS has been applied in [18, 6]. Picard et al. [15] utilize Sequential Forward Floating Selection (SFFS) to chose a set of physiological features to predict 8 different emotions. The nBest, SFS, SFFS and Perceptron Feature Selection (PFS) approaches are applied to select features for predictors of player experience in [14].

All aforementioned studies utilize local search in the feature space. Our main hypothesis, instead, is that a global search feature selector can find feature sets that will yield more accurate computational models of user affect. For that purpose we present a genetic search [4] based feature selection method, namely *Genetic Feature Selection* (GFS), and compare its performance against a random search based method and a local search based method (SFS). The three feature selection methods are assessed on game survey datasets that contain self-reported pairwise preferences of affect and features from two modalities of input: physiological and gameplay. The features selected from the three methods are used as inputs of an artificial neural network (ANN) model that learns the mapping between the statistical features (physiological or gameplay or both) and the pairwise self-reported preferences of affect.

The paper is organized as follows. Section 2 presents the three feature selection methods compared in this study and the methodology applied to evaluate their results while Section 3 provides details on the dataset used in this paper. Experiments and conclusions derived are presented in Section 4 and Section 5, respectively.

2. FEATURE SELECTION

This section describes the method proposed, genetic feature selection, and reviews briefly the alternative feature selection algorithms used to compare against its performance: sequential forward feature selection and random feature selection (RFS). The three FS methods are tested for finding the subset of features that can predict emotional preferences more accurately (see Section 3); the method utilized to assess the quality of the feature sets found, neuro-evolutionary preference learning, is described in the first part of this section.

2.1 Neuro-evolution for Affective Preference Learning

We apply preference learning [3] to build affective models that predict users' self-reported emotional preferences based on the subsets of features selected by the FS algorithms.

In this study, the models are implemented as single layer perceptrons (SLPs) that are trained to map the selected features to an affective predictor of the reported pairwise emotional preferences, i.e. the pairwise preference relationship of the training data (e.g. A and B) is known (e.g. A is preferred to B or, otherwise, B is preferred to A) but the value of the target output is not (i.e. the magnitude of the preference). Thus, any gradient-based optimization algorithm is inapplicable to the training problem since the error function under optimization is not differentiable. In this paper we utilize the neuro-evolutionary preference learning apporach presented in [20] for training the SLP to approximate the function between the selected input features and the pairwise preferences.

2.2 Sequential Forward Feature Selection

Sequential forward feature selection is a bottom-up search procedure where one feature is added at a time to the current feature set. The feature to be added is selected from the subset of the remaining features such that the new feature set generates the maximum value of the performance function over all candidate features for addition. The search stops when adding a new feature does not yield an increase in performance.

This method has been successfully applied in dissimilar studies [19, 14] to select minimal subsets of features for affective preference prediction. In particular, SFS has been successful in selecting minimal and high-performing feature subsets on one of the two datasets used for the experiments presented here [11], which makes SFS an appropriate benchmark feature selection mechanism for this paper. This study extends experiments reported in [11] as it explores the use of genetic search for feature selection and investigates its impact on the learning process of reported preferences of affect which are linked to multimodal input data.

2.3 Random Feature Selection

Random FS selects features from the input set with a chance probability per feature. This method is used as a feature selection performance baseline for comparison purposes.

2.4 Genetic Feature Selection

The feature selection method proposed implements a generational genetic algorithm to search for the set of features that yields the most accurate preference predictor for the investigated affective state. According to the GFS mechanism, the whole set of input features are encoded as a bit string chromosome, c:

$$c = (g_1, g_2, ..., g_{N_F}) \tag{1}$$

where

$$g_i = \begin{cases} 1, & \text{if feature } i \text{ is included} \\ 0, & \text{if feature } i \text{ is not included} \end{cases} \tag{2}$$

and N_F is the total number of features existent in the input dataset.

A population of N_c chromosomes is initialized with all bits set to zero but one selected randomly; i.e. the first generation consists of sets of one randomly selected feature. The reason for initializing chromosomes with only one feature is because we desire to obtain minimal feature subsets which, nevertheless, yield high performing artificial neural network predictors of reported affect — serving as the input of the ANN model. Then, at each generation:

1. All chromosomes of the population are evaluated. For that purpose a preference model is trained via neuro-evolutionary preference learning for each chromosome and its performance is assessed via 3-fold cross validation. The fitness of each chromosome is the average 3-fold classification of the ANN trained on the feature set presented by the chromosome.

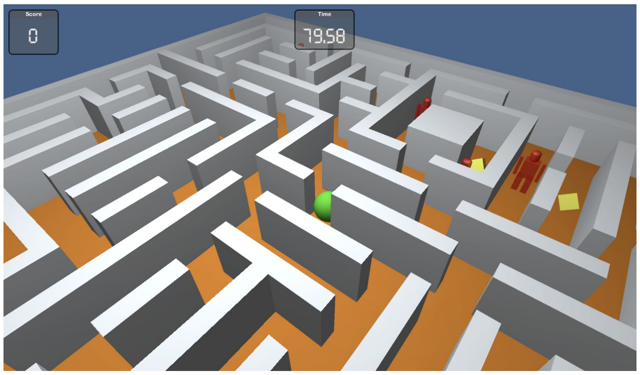


Figure 1: A camera configuration of the computer game MazeBall.

2. An *elitism* selection method chooses the best N_p individuals to be the parents of the next generation.

3. Pairs of parents are selected using a rank selection method that ranks the parents by their fitness and then selects two of them with a probability proportional to $\frac{1}{2+n}$ where n is the position in the ranking. A total of $N_c - N_p$ offspring are reproduced via uniform crossover with probability p_c. If crossover is not applied, the most-fit parent of the two is cloned to generate an offspring.

4. For each offspring, mutation occurs at each gene with probability p_m. The mutation scheme used flips the value of the selected gene which, in turn, suggests that the corresponding feature is either added (1) or removed (0) from the feature set. Finally, all offspring are inserted to the population.

The algorithm terminates after G_{max} generations are completed and the set of features corresponding to the highest performing preference predictor found across all generations is chosen. It is noteworthy that parent chromosomes are cloned to the new generation but their performance is reevaluated, i.e. a new ANN is trained on that feature set. Therefore, due to the non-deterministic nature of the neuroevolution, the fitness function of some individuals may fluctuate significantly from one generation to the next.

3. DATA COLLECTION

The data set used in this study was collected via a user survey experiment of 36 subjects playing the MazeBall game. MazeBall is a 3D prey/predator game where the player guides a ball through a maze. Golden tokens can be collected to increase the player's score, while red enemies that move around the maze decrease the player's score when they come in contact with the ball (see Figure 1).

The game is designed to study the impact of the virtual camera configuration on the players' affective state. Hence, each subject plays four pairs of 90 second-long games incorporating a different camera configuration/profile which is dependent on three parameters: *height* and *distance* of the camera from the player character and the speed between subsequent camera state transitions (*frame coherence*). During the game, blood volume pulse (BVP) and skin conductance (SC) signals are recorded from bio feedback sensors attached to the subject's left hand (see Figure 2). After each pair of games is completed, players are questioned to express their preferred game with regards to seven affective states, *anxiety, boredom, challenge, excitement, frustration, fun* and *relaxation*, via 4-alternative forced choice (4-AFC) questionnaire items. The 4-AFC scheme allows subjects to either express their *clear pairwise preference* — i.e. either the first or the second game is preferred (pairwise preference or 2-AFC); for instance, game A was more frustrating than game B — or to express their preference for both games equally (e.g. both games were equally frustrating) or neither game (e.g. neither game was frustrating). More details about the experimental protocol and the self-reported data can be found in [12].

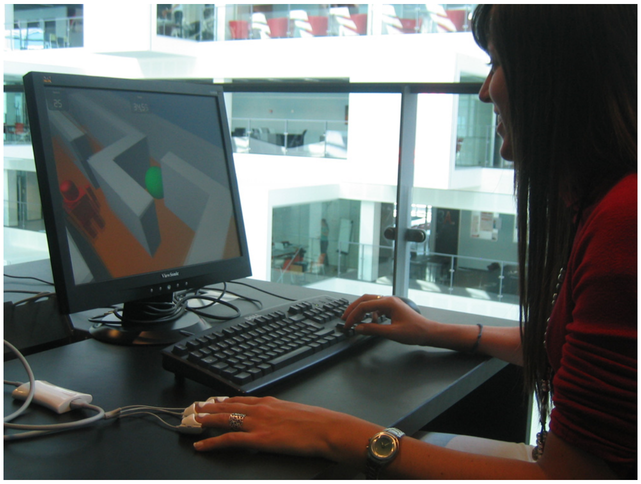


Figure 2: Subject playing Maze Ball. The game is controlled only with the arrow keys leaving one hand free to attach the biofeedback sensors.

3.1 Extracted Features

Blood Volume Pulse and Skin Conductance were collected in real-time at 32 Hz. Heart rate (HR) is computed every 5 seconds by extrapolating the inter-beat time intervals detected in the BVP signal. A total of 42 statistical features, namely *physiological features*, are extracted from the signals including, for example, average and standard deviation of the three signals and heart rate variability measures such as standard deviation of the inter-beat time intervals.

In addition, several game metrics are logged during play including the elements of the game state and the player's inputs (keystrokes) and a total of 41 statistical features are extracted such as distance measures to the closest pellet and enemy, and reaction time measures. These features are referred as *gameplay features* in the remainder of this paper. For a full description of the features extracted from both modalities of input, the reader is referred to [12, 11].

4. EXPERIMENTS

Three datasets are generated to test our hypotheses, the first includes only physiological features, the second only gameplay, and the third is the complete dataset of both modalities. The reported affective states contain a different number of samples: 97, 83, 86, 92, 90, 90 and 54, respectively, for challenge, excitement, anxiety, fun, relaxation, frustration and boredom. Different sample sizes indicate different numbers of self-reported *clear preferences* for the various affective states.

The three feature selection methods presented in Section 2, random, sequential forward and genetic, are applied to the three datasets. Their performance in selecting appropriate feature sets for each affective state and the size of the selected subsets are compared.

This section presents comparative studies first with the two single-modality datasets independently and then with the bimodal inputs. The first comparison investigates the impact of feature selection on dissimilar datasets whereas the second comparative analysis explores the benefits of genetic search on the training of preference models built on bimodal inputs.

To minimize the effect of the random initialization of genetic search, every algorithm is run three times and the highest performing set of features is chosen. The input data samples are distributed randomly into three folds (for the purpose of performance cross-validation) before each run; note that the same distribution is used for all three algorithms.

GFS uses a population of 50 individuals and 20% of them are used as parents of the next generation ($N_p = 10$); the algorithm stops after 10 generations are completed. These parameters have been assigned low values to reduce the time required to perform the test. The probabilities that crossover and mutation occur are 0.4 and 0.01 respectively. These parameters have been selected after preliminary sensitivity analysis experiments for maximizing GA performance.

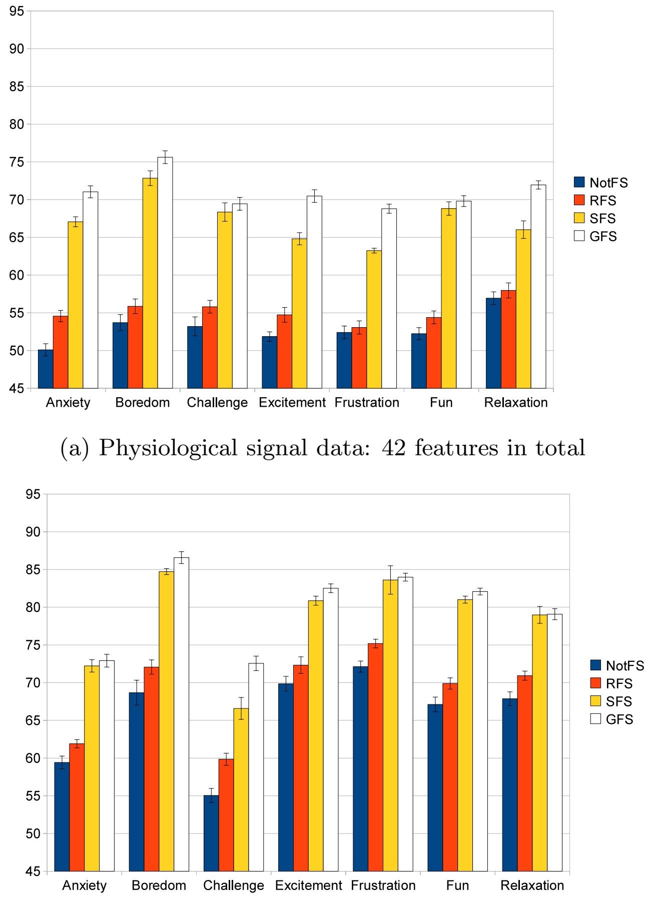
4.1 Impact on preference learning

Figure 3illustrates the average performance of twelve runs of RFS, SFS and GFS on physiological and gameplay feature sets, respectively, and the average performance of 12 models trained on the two complete datasets.

It is apparent that random search finds sets that, on average, perform significantly worse than the other two methods in both input sets and across all seven affective states investigated. Moreover, the feature sets selected by genetic search appear to outperform sequential feature selection in all the case studies depicted in Figure 3(a) and Figure 3(b). A two tailed paired t-test reveals that the difference in performance between GFS and SFS is statistically significant (significance is 5% in this paper) in 4 of the states in the physiological input set (anxiety, p-value= 0.0005; excitement, p-value= 0.0003; frustration, p-value= 0.00001; relaxation, p-value= 0.00004) and in challenge (p-value = 0.01) and fun (p-value = 0.03) in the gameplay dataset. For all affective states in both datasets, the models built on features selected automatically outperform the models trained on all the features (see NotFS in Figure 3(a) and Figure 3(b)). The models generated via feature selection are more accurate than NotFS even when features are selected randomly (RFS).

Figure 4 shows the average number of features selected by the three algorithms. RFS selects approximately half of the available features which is expected given that all features can be selected with chance probability. GFS selects small subsets of features but slightly larger than SFS which selects minimal sets by following bottom-up search.

By looking at the difference in the performance between the models built on physiological and gameplay features, it is clear that the gameplay features are better sources of information for the prediction of the reported affective states on the MazeBall datasets. Gameplay features appear to be better predictors of affective states reported in a 3D game environment like MazeBall in which distances to enemies and pellets, time spent on various locations in the maze and visible areas of the maze seem to yield better estimators than heart rate or skin conductance signal features for all emotional states investigated



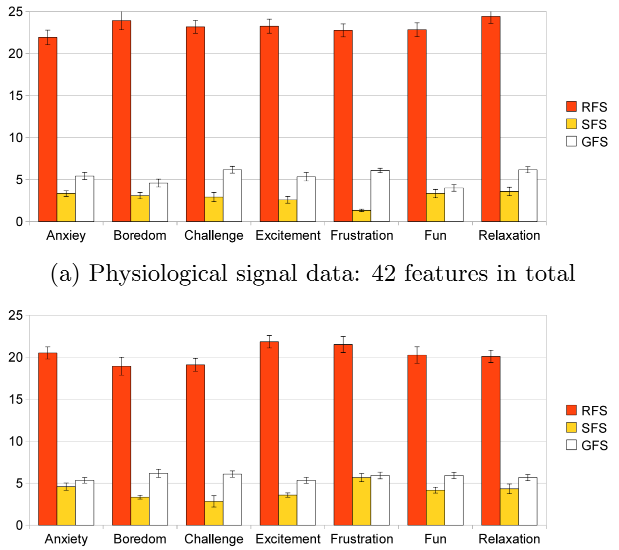
(a) Physiological signal data: 42 features in total

(b) Gameplay data: 41 features in total

Figure 3: Average performance and standard error of 12 runs of random feature selection (RFS), sequential forward feature selection (SFS) and genetic feature selection (GFS) and average performance and standard error of 12 models trained on all features (NotFS).

4.2 Impact on bimodal input data

This section presents experiments for investigating the impact of genetic search on bimodal feature sets and the construction of computational models of reported affect on both modalities of input. For that purpose the three FS algorithms run 12 times on the complete input dataset, i.e. both the physiological and the gameplay features are now considered. Figure 5 shows the average performances obtained across the affective states investigated and the average performance of 12 models trained on the complete set of features.



(a) Physiological signal data: 42 features in total

(b) Gameplay data: 41 features in total

Figure 4: Average number of selected features and standard error of 12 runs of random feature selection (RFS), sequential forward feature selection (SFS) and genetic feature selection (GFS).

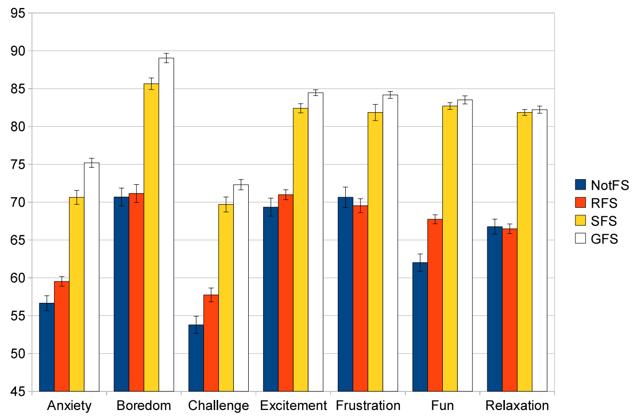


Figure 5: Bimodal input data: average performance and standard error of 12 runs of random feature selection (RFS), sequential forward feature selection (SFS) and genetic feature selection (GFS) on a set of 83 features (42 physiological and 41 game play features in total) and average performance and standard error of 12 models trained on all 83 features (NotFS).

Unsurprisingly, random search performs significantly worse than both SFS and GFS in the construction of affective models for all seven affective states. More importantly, in the comparison between GFS and SFS, the former finds, on average, feature sets that generate better-performing preference models for all 7 affective states investigated. This difference is statistically significant for anxiety (p-value = 0.002), boredom (p-value = 0.003), challenge (p-value = 0.05), excitement (p-value = 0.01) and frustration (p-value = 0.02) Models trained on features selected by SFS or GFS predict the affective preferences more accurately than models trained on all the features (NotFS); even RFS yields higher performance than NotFS in most cases.

The size of the subsets of features selected from the bimodal dataset follows the same trend observed earlier within the unimodal sets: GFS, on average, selects more features than SFS (see Figure 6).

Observing the average performances across the three input datasets it is apparent that GFS is able to find better subsets of features on the unified dataset of physiological and gameplay features than on the two unimodal input sets. In particular, the performance of GFS on the bimodal dataset is significantly higher for all affective states when compared to the corresponding GFS performance on the physiological input dataset; challenge is the only affective state for which such a performance improvement is not observed. Moreover, the predictions of reported relaxation and excitement, supported by GFS, are significantly better when built on bimodal compared to the predictors built on gameplay data.

SFS, just as GFS, performs significantly better in the unified dataset than in the physiological dataset. On the other hand, the difference of performance achieved with SFS between the bimodal and the gameplay set of features, although not statistically significant, is negative for both anxiety and frustration.

5. CONCLUSIONS AND DISCUSSION

In this paper we propose a feature selection method that performs global search on the attribute/feature space of user input data and selects feature subsets that guide preference learning algorithms for the construction of accurate models of affect. The impact of the genetic search feature selector on the accuracy of the affective models is investigated both in comparison to other feature selection mechanisms in different datasets but also with respect to the multimodality of the input data. Thus, the performance of the GA-based feature selection algorithm is assessed on three datasets containing self-reported pairwise preferences of affect collected via a game survey experiment: two unimodal data input sets (physiological and gameplay) and one unified dataset (bimodal). The genetic feature selection method is compared against the random search and the sequential forward feature selection methods.

Results obtained show that GFS is able to select feature subsets that yield more accurate preference models than RFS and SFS. Even though SFS has proved to be a rather effective hill-climbing method for constructing computational models of affect trained on pairwise preference data, results obtained in this paper are not entirely surprising given the superiority of global genetic search over hill-climbing and random search in rough search landscapes. In particular, the performance improvement in this initial study is statistically significant in 4 out of 7 affective states on models built on physiological data, 2 out of 7 on gameplay models and 5 out of 7 on models built on bimodal feature sets. We expect that a more careful adjustment of the parameters of the genetic search will provide higher improvements.

The experiments also reveal that the gameplay features incorporate more relevant information for an affective predictor than the physiological features in the dataset used. Moreover, on average, the subsets of features found by GFS

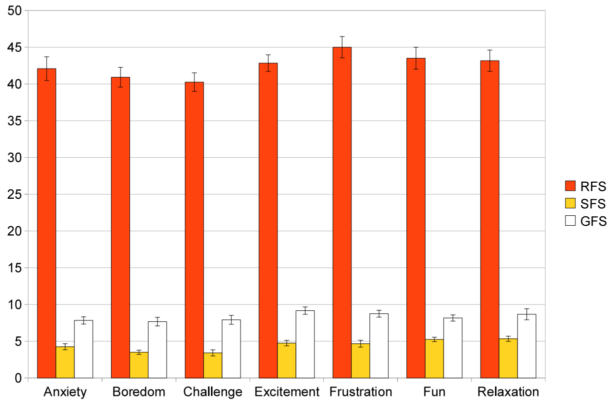


Figure 6: Average number of selected features and standard error of 12 runs of random feature selection (RFS), sequential forward feature selection (SFS) and genetic feature selection (GFS) on a set of 83 features (42 physiological and 41 game play features in total).

that combine the two modalities yield higher performances than the unimodal sets.

This paper proposed a method for enhancing the performance of affective preference models without reducing their expressiveness. For that purpose, we focused on automatic feature selection that reduces the size of the input models while keeping the physical meaning of the input data. The results presented here, in agreement with other studies (see [11, 19, 17] among others), show that reducing the dimensionality of the input by omitting the non-relevant features improves substantially the performance of the affective preference models. A detailed analysis of the models found by GFS will be reported in a future study.

6. REFERENCES

[1] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211. ACM, 2004.

[2] S. Fairclough. Fundamentals of physiological computing. *Interacting with computers*, 21(1-2):133–145, 2009.

[3] J. Fürnkranz and E. Hüllermeier. Preference learning. *Künstliche Intelligenz*, 19(1):60–61, 2005.

[4] D. Goldberg. *Genetic Algorithms in Search and Optimization*. Addison-wesley, 1989.

[5] E. Haapalainen, P. Laurinen, H. Junno, L. Tuovinen, and J. Röning. Methods for classifying spot welding processes: A comparative study of performance. *Innovations in Applied Artificial Intelligence*, pages 412–421, 2005.

[6] L. He, M. Lech, N. Maddage, and N. Allen. Stress and emotion recognition using log-Gabor filter analysis of speech spectrograms. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–6, 2009.

[7] E. Hudlicka. Affective game engines: motivation and requirements. In *Proceedings of the 4th International Conference on Foundations of Digital Games*, pages 299–306. ACM, 2009.

[8] A. Jain and D. Zongker. Feature selection: evaluation, application, and small sampleperformance. *IEEE transactions on pattern analysis and machine intelligence*, 19(2):153–158, 1997.

[9] K. Kim, S. Bang, and S. Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and biological engineering and computing*, 42(3):419–427, 2004.

[10] R. Mandryk, K. Inkpen, and T. Calvert. Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & Information Technology*, 25(2):141–158, 2006.

[11] H. P. Martinez, K. Hullett, and G. N. Yannakakis. Extending Neuro-evolution Preference Learning through Player Modeling. In *Proceedings of 2010 IEEE Conference on Computational Intelligence and Games (CIG)*, Copenhagen, Denmark, August 2010.

[12] H. P. Martinez, A. Jhala, and G. N. Yannakakis. Analyzing the Impact of Camera Viewpoint on Player Psychophysiology. In *Proceedings of the Int. Conf. on Affective Computing and Intelligent Interaction*, pages 394–399, Amsterdam, The Netherlands, September 2009. IEEE.

[13] S. McQuiggan, S. Lee, and J. Lester. Predicting user physiological response for interactive environments: an inductive approach. In *Proceedings of the 2nd Artificial Intelligence for Interactive Digital Entertainment Conference*, pages 60–65, 2006.

[14] C. Pedersen, J. Togelius, and G. Yannakakis. Modeling Player Experience for Content Creation. *Narrative*, 12:13, 2010.

[15] R. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, pages 1175–1191, 2001.

[16] P. Rani, N. Sarkar, and C. Liu. Maintaining optimal challenge in computer games through real-time physiological feedback. In *Proceedings of the 11th International Conference on Human Computer Interaction*, pages 184–192, 2005.

[17] J. Togelius, T. Schaul, J. Schmidhuber, and F. Gomez. Countering poisonous inputs with memetic neuroevolution. In *Parallel Problem Solving From Nature 10*, 2008.

[18] D. Ververidis and C. Kotropoulos. Automatic speech classification to five emotional states based on gender information. *Proc. Eusipco, Vienna*, pages 341–344, 2004.

[19] G. Yannakakis and J. Hallam. Game and player feature selection for entertainment capture. In *Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Games (CIG 2007)*, 2007.

[20] G. N. Yannakakis. Preference Learning for Affective Modeling. In *Proceedings of the Int. Conf. on Affective Computing and Intelligent Interaction (ACII09)*, Amsterdam, The Netherlands, September 2009.