How to Aggregate Multimodal Features for Perceived Task Difficulty Recognition in Intelligent Tutoring Systems

Ruth Janning
Information Systems and Machine Learning Lab
University of Hildesheim
janning@ismll.uni-hildesheim.de

Carlotta Schatten
Information Systems and Machine Learning Lab
University of Hildesheim
schatten@ismll.uni-hildesheim.de

Lars Schmidt-Thieme
Information Systems and Machine Learning Lab
University of Hildesheim
schmidt-thieme@ismll.uni-hildesheim.de

ABSTRACT

Currently, a lot of research in the field of intelligent tutoring systems is concerned with recognising student's emotions and affects. The recognition is done by extracting features from information sources like speech, typing and mouse clicking behaviour or physiological sensors. Multimodal affect recognition approaches use several information sources. Those approaches usually focus on the recognition of emotions or affects but not on how to aggregate the multimodal features in the best way to reach the best recognition performance. In this work we propose an approach which combines methods from feature selection and ensemble learning for improving the performance of perceived task difficulty recognition.

1. INTRODUCTION

Some research has been done in the area of intelligent tutoring systems to identify useful information sources and appropriate features able to describe student's emotions and affects. However, work on multimodal affect recognition in this area focuses more on engineering appropriate features for affect recognition than on the problem of aggregating the features from the different information sources in an good way. The usual approach is to use one classification model fed with one input vector containing the concatenated features (maybe reduced by feature selection) like in [3] or using standard ensemble methods on the features of the sources separately like in [4]. In this paper instead we propose to mixing up the different feature types and combining methods from feature selection and ensemble approaches to reach a classification performance improvement compared to using only either methods from feature selection or ensemble approaches. Feature selection methods can be used to reduce the number of features and find good combinations of features. They take advantage of statistical information like correlations. Ensemble methods like stacking use multiple learning models to obtain a better prediction performance.

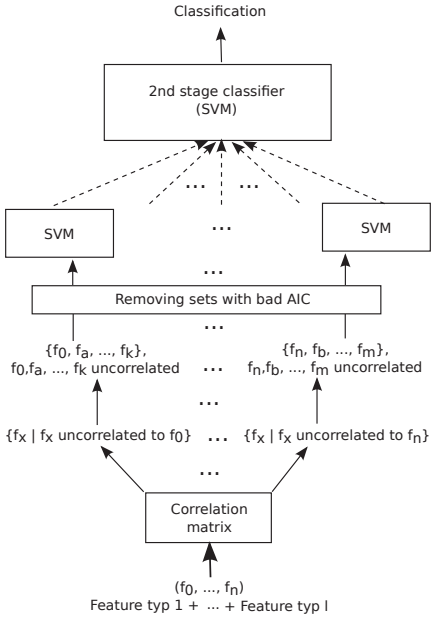

Figure 1: Multimodal feature aggregation approach.

Stacking learns to combine the classification decisions of several single classifiers by a further classifier which gets as input the outputs of the other classifiers.

2. MULTIMODAL FEATURE AGGREGATION

We propose to profit from the advantages of both feature selection methods and ensemble methods. Hence, we combine both (see fig. 1): In a first step the feature vectors of all l feature types are concatenated to reach one input feature vector $(f_0, \ldots, f_n)$. However, there could be dependencies between the different features. Hence, we create the correlation matrix reporting about the correlations between each pair of features. By means of this matrix we extract for each single feature f_y a set $uncorr_y$ containing all other features f_x not correlated to f_y. *Not correlated* means in this case that the correlation value $v_{x,y}$ of the pair (f_x, f_y) in the correlation matrix is near to 0.0, or more explicitly, $|v_{x,y}|$ is smaller than some positive thresh-

Table 1: Classification errors and F-measures.

(1)	SVM applied to amplitude features 31.25% (0.75, 0.59) SVM applied to articulation features 22.92% (0.81, 0.72)
(2)	SVM applied to all concatenated features 27.08% (0.77, 0.67)
(3)	SVM applied to most uncorrelated features 20.83% (0.81, 0.77)
(4)	Stacking applied to $uncorr_y$ sets 20.83% (0.83, 0,74)
(5)	Stacking applied to $uncorr2_y$ sets 16.67% (0.86, 0.80)
(6)	Stacking applied to $uncorr2_y$ sets with best AIC **8.33% (0.92, 0.91)**

old t, i.e. $uncorr_y := \{f_y\} \cup \{f_x \mid t > |v_{x,y}|\}$. The set $uncorr_y$ contains all features uncorrelated to f_y but between the features within this set there could still be correlations. Consequently, we compute for each feature f_y a set $uncorr2_y := \{f_y, f_a, \ldots, f_k\}$ where $f_y, f_a, \ldots, f_k$ all are uncorrelated. These sets $uncorr2_y$ are gained for each feature f_y by sequentially intersecting $uncorr_y$ with the sets belonging to the features within $uncorr_y$, or the intersection respectively. Different to feature selection, our goal is not to create one feature vector with reduced dimensionality but we aim at creating one feature vector per feature which will be fed into an own classifier, to consider each feature and to deliver as many input as needed for the ensemble method. Nevertheless, we remove some of the $uncorr2_y$ sets. The reason is that there is still some statistical information which we did not yet use: the quality of the models using these sets as input. Hence, for each set $uncorr2_y$ we compute the Akaike information criterion (AIC) – indicating the quality of a model. Subsequently, we remove the worse quarter of the sets. The remaining sets are fed into an support vector machine (SVM) each. In the next step we apply a stacking ensemble approach by feeding the outputs, i.e. the classification decisions, of the SVMs into a further SVM, which learns how to generate one common classification decision.

3. EXPERIMENTS

We prove our proposed multimodal feature aggregation approach by experiments with a real data set and multimodal low-level speech features. The data were gained by conducting a study in which the speech of ten 10 to 12 years old German students was recorded and their perceived task-difficulties were labelled by experts. During the study a paper sheet with fraction tasks was shown to the students and they were asked to explain their observations and answers. The acoustic speech recordings were used to gain two kinds of low-level speech features: *amplitude* and *articulation* features. The *amplitude features* ([1]) are taken from the raw speech data, or information about speech pauses respectively: ratio between (a) speech and pauses, (b) number of pause/speech segments and number of all segments, (c) avg. length of pause/speech segments and max. length of pause/speech segments, (d) number of all segments and number of seconds, and percentage of pauses of input speech data. The idea behind this kind of features is that depending on how challenged the student feels, the student makes more or less and shorter or longer speech pauses. The *articulation features* ([2]) are gained from an intermediate step of speech recognition which delivers information about vowels and consonants: ratio between (a) number of silence tags and number of all tags, (b) avg./min. length of vowels/obstruents/fricatives/silence tags and max./avg. length of vowels/obstruents/fricatives/silence tags. The idea behind this kind of features is that depending on how challenged the student is, the student shortens or lengthens vowels and consonants. The data collection resulted in 36 examples labelled with *over-challenged* or *appropriately challenged*, respectively 48 examples after applying oversampling to the smaller set of examples of class *over-challenged* to eliminate unbalance within the data. We conducted a 3-fold cross validation and we applied SVMs with an RBF-kernel and for each SVM used we conducted a grid search on each fold to estimate the optimal values for the hyper parameters. As baseline experiments we applied an SVM separately to both feature types. The classification test errors and F-measures (harmonic mean of *recall* and *precision*) for both classes (*over-challenged*, *appropriately challenged*) are reported in tab. 1, (1). An aggregation of both feature types only makes sense, if we can improve this results. A straight forward way to combine different feature types is to concatenate the features of all types and putting them into one feature vector which serves as input for one classification model. However, this approach does not deliver good results (see tab. 1, (2)) in cases where some features may be correlated and may disturb each other. Hence, one should restrict the input vector by considering the correlations. The results of using only features uncorrelated with most of the other features are shown in tab. 1, (3). As one can see considering correlations helps to improve the classification performance. But still there is space for improvement. Hence, in the following we combine ensemble methods with feature selection which takes into account correlations. In a first step we applied stacking ensemble to the outputs of SVMs applied to the $uncorr_y$ sets (see tab. 1, (4)). However, there could still be correlations within the $uncorr_y$ sets. Hence, as next step we computed for each feature the $uncorr2_y$ set and applied again stacking ensemble, resulting in a classification test error of 16.67 % (tab. 1, (5)). This result is already very good but there is one more statistical information to use: the AIC. We computed for each $uncorr2_y$ set the AIC, threw out the worst quarter of these sets and applied stacking to the remaining sets resulting in a very good classification test error of 8.33 % and F-measures $0.92, 0.91$ (tab. 1, (6)). In summary, the experiments have shown that our multimodal feature aggregation approach is able to improve the classification performance significantly.

4. REFERENCES

[1] R. Janning, C. Schatten, and L. Schmidt-Thieme. Feature analysis for affect recognition supporting task sequencing in adaptive intelligent tutoring systems. In *Proceedings of EC-TEL*, 2014.

[2] R. Janning, C. Schatten, L. Schmidt-Thieme, and G. Backfried. An svm plait for improving affect recognition in intelligent tutoring systems. In *Proceedings of ICTAI*, 2014.

[3] J. Moore, L. Tian, and C. Lai. Word-level emotion recognition using high-level features. In *CICLing*, 2014.

[4] S. Salmeron-Majadas, O. Santos, and J. Boticario. Exploring indicators from keyboard and mouse interactions to predict the user affective state. In *Proceedings of EDM*, 2014.