Dynamic Music Emotion Recognition Using Kernel Bayes' Filter

Konstantin Markov
Human Interface Laboratory
The University of Aizu
Fukushima, Japan
markov@u-aizu.ac.jp

Tomoko Matsui
Department of Statistical Modeling
Institute of Statistical Mathematics
Tokyo, Japan
tmatsui@ism.ac.jp

ABSTRACT

This paper describes the temporal music emotion recognition system developed at the University of Aizu for the Emotion in Music task of the MediaEval 2015 benchmark evaluation campaign. The arousal-valence trajectory prediction is cast as a time series filtering task and is performed using state-space models. A simple and widely used example is the Kalman Filter, however, it is a linear parametric model and has serious limitations. On the other hand, non-linear and non-parametric approaches don't have such drawbacks, but often scale poorly with the number of training data and their dimension. One such method proposed recently is the Kernel Bayes' Filter (KBF). It uses only data Gram matrices and thus works (almost) equally well with data of both low and high dimension. In our experiments, we used the feature set provided by the organizers without any change. All the development data were clustered in six clusters based on the genre information available from the meta-data. For performance comparison, we build three more emotion recognition systems based on the standard Multivariate Linear regression (MLR), Support Vector machine regression (SVR) and Kalman Filter (KF). The results obtained from a 4-fold cross-validation on the development set show that all types of models, except KF, achieved very similar performance, which suggests that they may have reached the upper bound of the feature set discrimination power.

1. INTRODUCTION

Dynamic or continuous emotion estimation is more difficult task and there are several approaches to solve it. The simplest one is to assume that for a relatively short period of time emotion is constant and apply static emotion recognition methods. These include conventional regression methods as well as a combination of classification and regression where data are clustered in advance and for each cluster a separate regression model is built. Testing involves initial classification step or model selection procedure. A better approach is to consider emotion trajectory as a time varying process and try to track it or use time series modelling techniques involving state-space models (SSM). A popular and simple SSM is the Kalman filter (KF). It is a linear system and is quite fast since it requires just matrix multiplications and its complexity is linear in the number of data. However, the linearity assumption is a big drawback and KF performs poorly when the data relationship is non-linear.

Non-parametric non-linear kernel models [4] are becoming more and more popular in the Machine Learning community for their ability to learn highly non-linear mappings between two continuous data spaces. They extend the conventional kernel data mapping into high dimensional spaces to embedding data distributions in such spaces. This allows for Bayesian reasoning and developing inference algorithms which, however, involve only Gram matrices manipulations. Although, the complexity is $O(n^3)$ because of matrix inversions, it does not depend on data dimensionality, which is a big advantage compared to other non-linear methods based on Monte Carlo sampling [3].

The task and the database used in this evaluation are described in detail in the task overview paper [1].

2. KERNEL BAYES' FILTER

Details about the Kernel Baeys Filter can be found in [2]. Here we provide just the basic notation and the final update rules. During the KBF training truth values of both observations $\boldsymbol{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_T\}$ and corresponding state values $\boldsymbol{Y} = \{\boldsymbol{y}_1, \dots, \boldsymbol{y}_T\}$ are required. The prediction and conditioning steps of the standard filtering algorithms can be reformulated with respect to the kernel embeddings. The embedding of the predictive distribution $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$ is denoted as $\mu_{x_t|y_{1:t}}$ and is estimated as $\sum_{i=1}^{T} \alpha_i \phi(\boldsymbol{x}_i)$, where $\phi()$ is the feature map and α_t is updated recursively using

$$\begin{aligned} D^{t+1} &= diag((G + \lambda I)^{-1} \tilde{G} \alpha^t), \\ \alpha^{t+1} &= D^{t+1} K((D^{t+1} K)^2 + \beta I)^{-1}) D^{t+1} K_{:x^{t+1}} \quad (1) \end{aligned}$$

Here, G and K are the training states and observations Gram matrices, $\tilde{G}$ is a Gram matrix with entries $G_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_{j+1})$, and $K_{:x^{t+1}} = (k(\boldsymbol{x}_1, x_{t+1}), \dots, k(\boldsymbol{x}_T, x_{t+1}))$. The regularization parameters λ and β are needed to avoid numerical problems during matrix inversion.

There are few kernel functions that can be used with the KBF such as linear, rbf, and polynomial. Their parameters as well as the regularization constants λ and β comprise the set of hyper-parameters of a KBF system. Unfortunately, there is no algorithm for learning those hyper-parameters from data. They have to be set manually and as our experiments showed are critical for obtaining a good performance.

3. EXPERIMENTS

Using the genre information available from the metadata, we divided all development clips into six clusters roughly corresponding to the following genres: *Classical, Electronic,*

MediaEval 2015 Workshop, September 14-15, 2015, Wurzen, Germany

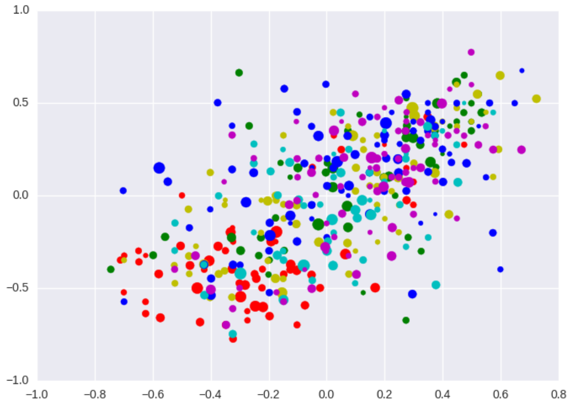

Figure 1: Mean A-V values distribution for the development data. Different colors represent different genres. Circle sizes are proportional to the A-V standard deviation.

Jazz-Blues, Rock-Pop, International-Folk, HipHop-SoulRB. The number of clusters was chosen such that the data distribution becomes as uniform as possible.

In order to visualize the relationship between clustered clips and their emotional content, we calculated arousal and valence statistics per clip and Figure 1 shows the distribution of mean AV vectors in the affect space. Different colors represent different genres/clusters and the circle size is proportional to the AV standard deviation. As can be seen, there are no clear grouping by genre, though some genres show more compact clouds than others. Both filtering systems, i.e. KF and KBF, were build using this clustering scheme where one model was trained for one genre and tested with the test data from the same genre only. Linear regression and SVR based systems were trained with no regard to genre clusters.

Since there is no validation data set available, we used 4-fold cross-validation approach to tune systems' parameters. The SVR and KBF models have hyper-parameters such as kernel function and regularization constants which cannot be learned from data. An unconstrained simplex search method was adopted to find optimum parameter setting, however, it does not guarantee global maximum and in the case of KBF, it turned out the initial point has a big impact on the final result.

4. RESULTS

Before the calculation of the correlation and RMSE performance measures, predicted arousal and valence values as well as the reference values were scaled to fit the range [-0.5,+0.5]. This similar to the way results were obtained during previous evaluations.

Table 1 shows the performance of the KBF for each genre as well as the total average. For some genres, the results are better, which may be due to differences in data distributions, but also because of a better hyper-parameter settings. Total averages from all the regression and state-space model based systems are summarised in Table 2.

The results using the official test data set are shown in Table 3. Due to time limitations, the KBF system uses reduced (to one forth) training set data which apparently has negative effect on the performance. Since the reference and predicted AV values are scaled to [-1.0, 1.0], direct comparison of the RMSE scores with those from previous tables is possible when they are divided by 2.

Table 1: Kernel Bayes' filter results on the development set.

Genre	Arousal		Valence	
	R	RMSE	R	RMSE
Classical	0.282	0.355	0.132	0.390
Electronic	0.306	0.347	0.265	0.355
Jazz-Blues	0.367	0.357	0.192	0.372
Rock-Pop	0.350	0.342	0.167	0.382
International	0.219	0.365	0.207	0.371
Hip-Hop, SoulRB	0.307	0.342	0.204	0.348
Average	0.305	0.351	0.194	0.369

Table 2: Averaged results of all systems on the development set.

Genre	Arousal		Valence	
	R	RMSE	R	RMSE
Regression				
Linear	0.269	0.341	0.184	0.357
SVM	0.283	0.340	0.214	0.351
Filters				
Kalman	0.113	0.390	0.068	0.393
Kernel Bayes'	0.305	0.351	0.194	0.369

Table 3: Averaged results on the test set.

Genre	Arousal		Valence	
	R	RMSE	R	RMSE
SVR	0.490	0.446	-0.019	0.542
KBR	0.419	0.498	-0.035	0.620

5. CONCLUSIONS

We described several systems developed at the University of Aizu for the MediaEval'2015 Emotion in Music evaluation campaign. Our focus is on the machine learning part of this very challenging task and, thus, we built and evaluated few systems based on conventional regression techniques as well as on a new non-parametric non-linear approach using Kernel Bayes' Filter state-space system. All used the feature set provided by the challenge organizers. Although the modelling techniques we utilized range from simple linear regression to sophisticated state-space Bayesian filter, there was a negligible difference in the performance. This suggests that the feature set may not have enough discriminating power to enable non-parametric non-linear models to show their advantages.

6. ACKNOWLEDGEMENT

Authors would like to thank Dr. Y.Nishiyama from the University of Electro-Communications, Tokyo, for sharing his Matlab Kernel mean toolbox (kmtb).

7. REFERENCES

[1] A. Aljanaki, Y.-H. Yang, and M. Soleymani. Emotion in music task at mediaeval 2015. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, September 2015.

[2] K. Fukumizu, L. Song, and A. Gretton. Kernel bayes' rule: Bayesian inference with positive definite kernels. *J. Mach. Learn. Res.*, 14(1):3753–3783, Dec. 2013.

[3] K. Markov, T. Matrui, F. Septier, and G. Peters. Dynamic speech emotion recognition with state-space models. In *Proc. EUSIPCO'2015*, pages 2122–2126, 2015.

[4] L. Song, K. Fukumizu, and A. Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *Signal Processing Magazine, IEEE*, 30(4):98–111, July 2013.