Music emotion recognition using Gaussian Processes

Konstantin Markov, Motofumi Iwata
Human Interface Laboratory
The University of Aizu
Fukushima, Japan
{markov,s1180127}@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 music emotion recognition system developed at the University of Aizu for the Emotion in Music task of the MediaEval'2013 benchmark evaluation campaign. A set of standard feature types provided by the Marsyas toolkit was used to parametrize each music clip. Arousal and valence are modeled separately using Gaussian Process regression (GPR). We compared performances of the GPR and Support Vector regression (SVR) and found out that GPR gives better results than SVR for the static per song emotion estimation task. For the dynamic emotion estimation task GPR had some scalability problems and fair comparison was not possible.

1. INTRODUCTION

Gaussian Processes (GPs) [2] 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, i.e. the feature space and the V/A space. Previously, we have successfully applied GPs for music genre classification task [1] and this encouraged us to use GPs for music emotion estimation. Many previous studies [4] have focused on Support Vector regression (SVR) since in most cases it gives superior performance. In this study we compare GP regression with SVR and show that in certain cases GPR can significantly outperform SVR. In addition, GPR produces probabilistic predictions, i.e. it outputs a Gaussian distribution with mean which corresponds to the most probable target value and variance which shows the certainty of the prediction. As in the case of SVR, GPR also uses kernels, but in contrast, it allows kernels parameters to be learned from the training data.

Database used in this evaluation is described in detail in the Emotion in Music task overview paper [3].

2. GAUSSIAN PROCESS REGRESSION

Given input training data vectors $\mathbf{X} = \{\mathbf{x}_i\}, i = i, \ldots, n$ and their corresponding target values $\mathbf{y} = \{y_i\}$, general regression model relates them as: $y_i = f(\mathbf{x}_i) + \epsilon_i$ where $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ and $f()$ is an unknown nonlinear function. In GP, it is assumed that this function is normally distributed, i.e. the vector $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]$ has Gaussian distribution $\mathbf{f} \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$, where $\mathbf{K}$ is a kernel covariance matrix and the mean $\mathbf{m}$ is often set to zero. This assumption allows to express in closed form the predictive distribution of a test target y_* only in terms of training data and the input vector $\mathbf{x}_*$: $y_*|\mathbf{x}_*, \mathbf{y}, \mathbf{X} \sim \mathcal{N}(m_*, \sigma_*^2)$ where $m_* = \mathbf{k}_*^T(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y}$ and $\sigma_*^2 = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{k}_*$.

Covariance kernel parameters are learned by maximizing the marginal likelihood $p(\mathbf{y}|\mathbf{X}, \theta) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X}, \theta)d\theta$ w.r.t. θ which is known as maximum likelihood type II approximation.

3. SYSTEM DESCRIPTION

Dimensional music emotion recognition can be easily decomposed into two independent classical regression problems: one for the valence, and another for the arousal. Thus, our system consists of two regression modules and a common feature extraction module.

3.1 Feature extraction

Features are extracted only from the audio signal which is first downsampled to 22050 kHz. We tried various standard features tailored for music processing such as MFCC, Statistical Spectrum Descriptors (SSD), Chroma, Spectral Crest Factor (SCF), and Spectral Flatness Measure (SFM) separately as well as combinations of several of them. All feature vectors were calculated using the Marsyas toolkit with 512 samples frames with no overlap. For the dynamic emotion estimation task, first order statistics (mean and std) of the feature vectors are calculated for a window of about 1 sec. giving 45 vectors per musical clip. For the static emotion estimation, same statistics for these 45 vectors are calculated resulting in a single high dimensional feature vector per song. After extensive preliminary experimentation we found that the best performing combination of features for the per song emotion estimation is MFCC, SCF, and SFM. Adding SSD features did not have any noticeable effect, and Chroma features actually hurt the performance. We refer to this combination of features as *UoA features*. We have also experimented with features released by the benchmark organisers which we call *MediaEval features*.

3.2 GPR implementation

Valence and arousal are modeled by separate GPR. We used standard Gaussian likelihood function which allows exact inference to be performed. The GP mean was set to zero and only the type of covariance kernel was varied. We experimented with the following kernels:

- Linear (LIN): $k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T\mathbf{x}' + 1)/l^2$

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

- Squared Exponential (SE): $k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp(-(\mathbf{x} - \mathbf{x}')^T(\mathbf{x} - \mathbf{x}')/2l^2)$
- Rational Quadratic (RQ): $k(\mathbf{x}, \mathbf{x}') = \sigma^2(1+(\mathbf{x}-\mathbf{x}')^T(\mathbf{x}-\mathbf{x}')/2\alpha l^2)^{-\alpha}$
- Matérn 3 (MAT3): $k(\mathbf{x}, \mathbf{x}') = \sigma^2(1 + r)\exp(-r), r = \sqrt{3(\mathbf{x} - \mathbf{x}')^T(\mathbf{x} - \mathbf{x}')/l^2)}$

where σ and l are parameters learned from training data. Sums or products of several kernels are also valid covariance functions and often give better performance than single kernels.

4. RESULTS

First, we present our results on the development data obtained after 7-fold cross validation. In addition to GPR, results from SVR using the same conditions are given in the following tables. In the SVR case, the parameter C was manually optimized using a grid search in the range [0.01,100], and kernel parameters are set to their default values (using LIBSVM package) since they cannot be learned.

Table 1 shows the result for the static emotion estimation in terms of R^2 metrics for both *MediaEval* and *UoA* feature sets. The last row of each feature set type shows the best performing combination of GPR covariance kernels.

Table 1: R^2 results of SVR and GPR on development data.

Algorithm	Kernel	Valence	Arousal
MediaEval features			
SVR	Linear	0.112	0.300
	RBF	0.017	0.028
GPR	LIN	0.132	0.565
	SE	0.142	0.590
	RQ	0.150	0.562
	MAT3	0.143	0.590
	LIN+RQ	0.170	0.581
UoA features			
SVR	Linear	0.314	0.604
	RBF	0.367	0.653
GPR	LIN	0.322	0.603
	SE	0.375	0.656
	RQ	0.430	0.662
	MAT3	0.395	0.668
	SExRQ	0.437	0.671

In Table 2, we summarize results of the dynamic emotion estimation task where Kendal τ measure is calculated after pooling all arousal or valence estimates from all songs together. We have to mention that since in this task the amount of data was 40 times bigger, we ran into some scalability problems with the GPR implementation and had to resort to approximations of the kernel matrix using much less data which, of course, decreased the performance noticeably.

Table 3 presents the results of the UoA submission runs: two for the static and one for the dynamic emotion estimation tasks. They are obtained using GPR with corresponding best performing kernels. Direct comparison with Tables 1 and 2 is possible only for RSQ lines and it can be seen that in contrast to *MediaEval*, *UoA* features give similar results.

Table 2: Kendal τ results of SVR and GPR on development data using UoA features.

Algorithm	Kernel	Valence	Arousal
SVR	Linear	0.288	0.512
	RBF	0.346	0.530
GPR	LIN	0.289	0.508
	SE	0.327	0.515
	RQ	0.339	0.521
	MAT3	0.333	0.519
	SE+MAT3	0.340	0.523

Table 3: Official results on the test data obtained using GPR.

Features (kernel)	Measure	Valence	Arousal
Per song estimation			
MediaEval (LIN+RQ)	RSQ	-0.128	-0.408
	MSE	0.026	0.043
	MAE	0.134	0.172
	SE-std	0.031	0.054
	AE-std	0.094	0.116
UoA (SExRQ)	RSQ	0.404	0.695
	MSE	0.014	0.009
	MAE	0.095	0.079
	SE-std	0.020	0.013
	AE-std	0.070	0.055
Dynamic estimation			
UoA (SE+MAT3)	rho-avg	0.025	0.101
	rho-std	0.020	0.216
	MSE-avg	0.009	0.037
	MSE-std	0.010	0.036
	MAE-avg	0.076	0.152
	MAE-std	0.044	0.078

5. CONCLUSIONS

We described the UoA emotion recognition system for the "Emotion in Music" task of the MediaEval'2013 benchmark evaluation which is based on the Gaussian Process regression algorithm. Compared to the Support Vector regression, GPR has several advantages, such as truly probabilistic prediction, and ability to learn hyperparameters from data. Performance wise, the GPR achieved better results for the static per song emotion estimation, but failed for the dynamic emotion estimation due to some scalability problems.

6. REFERENCES

[1] K. Markov and T. Matsui. Music genre classification using gaussian process models. In *Proc. IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, 2013.

[2] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. The MIT Press, 2006.

[3] M. Soleymani, M. Caro, E. M. Schmidt, C. Sha, and Y. Yang. 1000 songs for emotional analysis of music. In *Proceedings of the ACM multimedia 2013 workshop on Crowdsourcing for Multimedia*, CrowdMM. ACM, ACM, 2013.

[4] Y.-H. Yang and H. Chen. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology*, 3(3):40:1–40:30, May 2012.