International Journal of Trend in Research and Development, Volume 2(3), ISSN: 2394-9333
www.ijtrd.com

Music Genre and Emotion Recognition Using Gaussian Processes

K. Vijaya Kumar [1], P Kalyanchakravarthi[2] , D suresh[3]
[1,2,3] Dept. Of ECE, GMRIT ,Rajam, India.

Abstract: Gaussian Processes (GPs) are Bayesian nonparametric models that are becoming more and more popular for their superior capabilities to capture highly nonlinear data relationships in various tasks, such as dimensionality reduction, time series analysis, novelty detection, as well as classical regression and classification tasks. In this paper, we investigate the feasibility and applicability of GP models for music genre classification and music emotion estimation. These are two of the main tasks in the music information retrieval (MIR) field. So far, the support vector machine (SVM) has been the dominant model used in MIR systems. Like SVM, GP models are based on kernel functions and Gram matrices; but, in contrast, they produce truly probabilistic outputs with an explicit degree of prediction uncertainty. In addition, there exist algorithms for GP hyper parameter learning—something the SVM framework lacks. In this paper, we built two systems, one for music genre classification and another for music emotion estimation using both SVM and GP models, and compared their performances on two databases of similar size. In all cases, the music audio signal was processed in the same way, and the effects of different feature extraction methods and their various combinations were also investigated. The evaluation experiments clearly showed that in both music genre classification and music emotion estimation tasks the GP performed consistently better than the SVM. The GP achieved a 13.6% relative genre classification error reduction and up to an 11% absolute increase of the coefficient of determination in the emotion estimation task.

I. INTRODUCTION

A lot of music data have become available recently either locally or over the Internet but in order for users to benefit from them, an efficient music information retrieval technology is necessary. Research in this area has focused on tasks such as genre classification, artist identification, music mood estimation, cover song identification, music annotation, melody extraction, etc. which facilitate efficient music search and recommendation services, intelligent playlist generation and other attractive applications.

The next step of the self-taught learning algorithm involves transformation of the labeled data into new feature vectors using the dictionary learned at the previous step. This is done using the same matrix factorization procedure as before with the only difference that the basis vectors matrix is kept fixed and only the activation matrix is calculated.[1] This way, each of the labeled data vectors is approximated by a linear combination of bases learned from a large amount of data. It is expected that the activation vectors will capture more information than the original labeled data they correspond to, since additional knowledge encapsulated in the bases is being used. [2][3]Finally, using labeled activation vectors as regular features, classical supervised classifier is trained for the task at hand. In this work, we used the standard Support Vector Machine (SVM) classifier.
In our experiments, we utilized two music databases: one as unlabeled music data and the other for the actual supervised classification task. We have published some preliminary experimental results on these databases, but this study provides a thorough investigation and comparison of the three matrix decomposition methods mentioned above.

Each genre classification system consists of minimum two blocks: feature extractor and classifier. Studies in music processing have investigated various feature types and their extraction algorithms. Carefully crafted music features such as Chroma vectors are mostly used for some specific tasks, for example, music transcription or music scene analysis. On the other hand, spectrum and its derivatives are also widely adopted for music pattern classification. Various methods for building music genre classifiers have been studied,

ranging from Support Vector Machines (SVM) to compressive sampling models. However, in most of the studies, parametric models have been utilized. Learning approaches include instances of supervised, semi-supervised, and unsupervised methods.

We have to note that, since each music piece in our experiments was represented by a single feature vector, this may not be classified as a large scale evaluation whereby SVMs could have practical advantage because of their lower computational complexity. In this regard, further research involving sparse GP learning and inference methods is necessary.

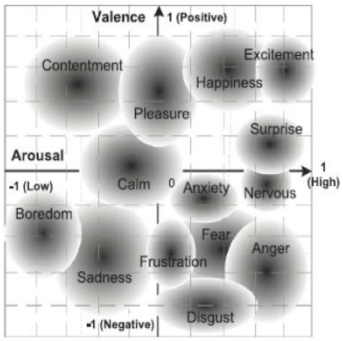

FIGURE 1: Two dimensional (Valence-Arousal) affective space of emotions. Different regions correspond to different categorical emotions

II. RELATED STUDIES

There are several studies where the semi-continuous learning framework has been used for music analysis and music information retrieval tasks. Based on a manifold regularization method, it has been shown that adding unlabeled data can improve the music genre classification accuracy rate. This approach is later extended to include fusion of several music similarity measures which achieved further gains in the performance.

One of the first applications of GPDM in audio signal processing was for speech phoneme classification. Although the absolute classification accuracy of the GPDM was not high, in certain conditions they outperformed the conventional hidden Markov model (HMM). In GPDM is used as a model for non-parametric speech representation and speech synthesis. Similar to GPDM is the GP based state- space model,it is essentially a non-linear Kalman filter and is very useful for time series processing[4]. Compared to some approximate Gaussian filters, such as the Extended Kalman filter (EKF) and the Unscented Kalman filter (UKL), it gives exact expected values in the prediction and filter steps.

III. GAUSSIANN PROCESSES

Gaussian processes are used to describe distributions over functions. Formally, the GP is defined as a collection of random variables any finite number of which has a joint Gaussian distribution. It is completely specified by its mean and covariance functions. For a real process f (x), the mean function m(x) and the covariance function k (x, x') are defined as

$$m(x) = E[f(x)]$$
$$k(x,x') = E[(f(x)-m(x))(f(x')-m(x'))] \quad (1)$$

Thus, the GP can be written as

$$f(x) \sim GP(m(x), k(x,x')) \quad (2)$$
$$X = \{x_i\} \in \square^d, i = 1,\ldots,n$$

A GP prior over function f (x) implies that for any finite number of inputs the vector of function has a multivariate Gaussian distribution is given by

$$f \sim N(m, K) \quad (3)$$

The covariance matrix K is given by

$$K = \begin{bmatrix} k(x_1,x_1)\ldots\ldots k(x_1,x_n) \\ k(x_2,x_1)\ldots\ldots k(x_2,x_n) \\ \\ k(x_n,x_1)\ldots\ldots k(x_n,x_n) \end{bmatrix} \quad (4)$$

and characterizes the correlation between different points in the process. For k (x, x'), any kernel function which produces symmetric and semi-definite covariance matrix can be used.

IV. GAUSSIAN PROCESS REGRESSION

Given input data vectors X = {xi}, i = 1,...,n and their corresponding target values y = {yi}, In the simplest regression task, y and x are related as

$$y = f(x) + \varepsilon \quad (5)$$

Given some new (test) input x∗, we can now estimate the unknown target y∗ and, more importantly, its distribution. Graphically, the relationship between all involved variables can be

represented as shown in Fig.(2). To find y∗, we first obtain the joint probability of training targets y and f∗ = f (x∗), which is Gaussian

$$p(y, f_* \mid x_*, X) = N\left(\mathbf{0}, \begin{pmatrix} K + \sigma_n^2 I & K_* \\ K_*^T & k(x_*, x_*) \end{pmatrix}\right)$$

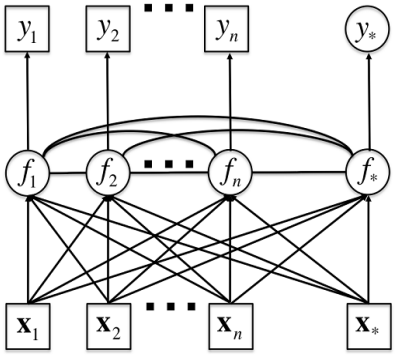

FIGURE 2: Graphical representation of observable x, y, (enclosed in squares), latent f, and unobservable y (enclosed in circles) variable relationships in Gaussian Process based regression task.

The conditional probability which is also Gaussian is obtained as

$$p(f_* \mid y, x_*, X) = N\left(f_* \mid \mu_{f_*}, \sigma_{f_*}^2\right)$$

Where mean and variance are

$$\mu_{f_*} = k_*^T (K + \sigma_n^2 I)^{-1} y$$

$$\sigma_{f_*}^2 = k(x_*, x_*) - k_*^T (K + \sigma_n^2 I)^{-1} k_*$$

It is worth noting that the mean μf∗ is a linear combination of the observed targets y. It can also be viewed as a linear combination of the kernel functions k (x∗, xi). On the other hand, the variance depends only on inputs X

$$p(y_* \mid y, x_*, X) = \int p(y_* \mid f_*) p(f_* \mid y, x_*, X) df_*$$

$$= N(y_* \mid \mu_{y_*}, \sigma_{y_*}^2)$$

Predictive mean and variance are

$$\mu_{y_*} = \mu_{f_*}$$

$$\sigma_{y_*}^2 = \sigma_{f_*}^2 + \sigma_n^2$$

Parameter learning:

Until now, we have considered fixed covariance function k (x, x'), but in general, it is parameterized by some parameter vector θ. This introduces hyper-parameters to GP, which are unknown and, in practice, very little information about them is available.[5] A Bayesian approach to their estimation would require a hyper-prior p (θ) and evaluation of the following

$$p(\theta \mid y, X) = \frac{p(y \mid X, \theta) p(\theta)}{p(y \mid X)} = \frac{p(y \mid X, \theta) p(\theta)}{\int p(y \mid X, \theta) p(\theta) d\theta} \quad \text{---------- (6)}$$

Where the likelihood p (y|X, θ) is actually the GP marginal likelihood over function values f

$$p(y \mid X, \theta) = \int p(y \mid f) p(f \mid X, \theta) df \quad \text{---------- (7)}$$

However, the evaluation of the integral in Eq. (6) can be difficult and as an approximation we may directly maximize Eq. (7) w.r.t. the hyper-parameters θ. This is known as maxi- mum likelihood II (ML-II) type hyper-parameter estimation. Since both the GP prior f |X ∼N (0, K) and the likelihood y|f∼N (f, σ2 n I) are Gaussians, the logarithm of Eq. (7) can be obtained analytically

$$\log p(y \mid X, \theta) = -\frac{1}{2} y^T K_y^{-1} y - \frac{1}{2} \log \mid K_y \mid - \frac{n}{2} \log 2\pi \quad \text{-- (8)}$$

Where $K_y = K + \sigma_n^2 I$ is the covariance matrix of noisy targets y.

V. GAUSSIAN PROCESS CLASSIFICATION

For binary classification, given training data vectors xi ∈ Rd with corresponding labels yi∈ {−1, +1}, we would like to predict the class membership probability of a test point x∗. This is done using an unconstrained latent function f (x) with GP prior and mapping its value into the unit interval [0, 1] by means of a sigmoid shaped function. Common choice for such function is the logistic function or the cumulative density function 8 of the standard Gaussian distribution[6][7]. When the sigmoid is point symmetric, the likelihood p (y|x) can be written as sig(y · f (x)).

Sigmund function is nothing but the cumulative distribution function.

A) Parameter learning:

As in the case of Gaussian Process regression, kernel function parameters can be learned by marginal likelihood p (y|X,θ) maximization. However, in this case, the likelihood p (y|f) is no longer Gaussian and analytic solution does not exist. Again, Laplace or EP approximation can be used. For the maximization, good candidates are gradient

based methods, such as the conjugate gradient optimization or the BFGS algorithm.

B) Related to SVM:

For the soft margin support vector machine, the optimization problem is defined as

$$\min_{w,w_0} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} (1 - y_i f_i) \text{ ---------- (9)}$$

Although it is possible to give probabilistic interpretation to the SVM outputs by wrapping them with sigmoid function, this is a rather ad hoc procedure which also requires tuning of the sigmoid parameters.

VI. EXPERIMENTS WITH MUSIC EMOTION RECOGNITION

In this study, we assume that music emotion recognition is to estimate the Valence-Arousal (VA) values for a song, or a clip as in our case, given its feature representation. Separate Gaussian Process regression (GPR) and Support Vector regression (SVR) models are independently trained using the same training data and corresponding reference VA values. [8][1]The models' performance is measured in terms of R2 measure. It is widely used to describe the goodness of fit of a statistical model and is defined as

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

.

VIII. EXPERIMENTS WITH MUSIC GENRE CLASSIFICATION

In these experiments, we again compared the Gaussian Processes and Support Vector Machines, but in the classification task. We kept the same amount of data, feature extraction methods and cross-validation type of evaluation as in the previous regression task.

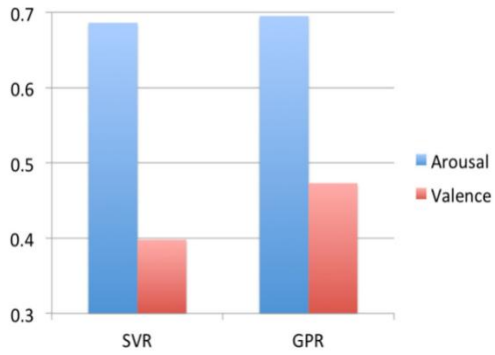

FIGURE 4. Gaussian Process (GPR) and Support Vector machine regression (SVR) best performance comparison in terms of R2 for both the Arousal and Valence prediction tasks.

A) Database and Feature Extraction:

We used the popular GTZAN song collection which consisted of 30 second long music clips belonging to one of the following 10 genres: Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae, and Rock. There were 100 clips per genre and 1000 clips in total.

All 1000 clips were processed in the same way as the MediaEval'2013 data for music emotion estimation and exactly the same features were extracted as well. Again, each music clip was represented by a single feature vector consisting of two level statistics of the frame level features, as depicted in Fig.3.

B) SVM And GP Classification Evaluation:

Since the SVM and GP are binary classifiers, in both cases, multiclass classification is simulated by one-versus-others setting. As in the music emotion experiments, SVM cost parameter C was manually optimized and the RBF scale was set to its default value.

Table 4 compares SVM and GP based classification systems' performance for various feature sets. The GP model was trained using SE covariance, zero mean and ERF likelihood functions. These results clearly show that GP consistently outperforms the SVM classifier in all cases. Again the best performance is achieved with the full feature set: MFCC+TMBR+SCF+SFM+CHR+LSP.

CASE	FEATURES	DIMS	SVM	GP
1	MFCC	52	65.8 6.3	68.7 6.3
2	1+TMBR	68	70.2 5.5	72.2 6.0
3	2+SCF+SFM	260	74.9 3.1	76.4 3.1
4	3+CHR+LSP	388	76.5 3.5	78.3 3.7

TABLE 4. Music genre classification accuracy (%). The SVM kernel function is RBF. The GP classifier uses SE covariance, ERF likelihood and zero mean. Results are given as mean and STD values of 10-fold cross-validation.

COVARIENCE	LIKELIHOOD	
	ERF	LOGISTICS
LIN	75.8 3.0	75.9 3.1
SE	78.3 3.7	78.3 3.1
RQ	78.7 3.4	78.7 3.3
MAT3	78.6 3.2	78.6 3.1
LIN+RQ	78.9 3.2	79.0 3.4
SE+RQ	78.9 3.2	79.3 3.0
LIN+RQ	78.5 3.4	79.3 3.3

TABLE 5. GP music genre classification accuracy (%). results are given as mean and STD values of 10-fold cross-validation.

A comparison between the two GP likelihood functions with respect to various covariance kernels and their combinations is given in Table 5. It seems that the Logistic function is slightly better, especially in the composite kernels case. The absolute difference between GP and SVM best results of 79.3% and 76.5% is 2.8%, which corresponds to 13.6% relative error reduction.

CONCLUSIONS

In this paper, we described and evaluated two systems based on Gaussian Process models for music genre and emotion recognition, respectively. In each of these tasks, Support Vector Machine is currently considered as the state-of-the-art model and therefore we used it for comparison. The GP and SVM have many common characteristics. They are both non-parametric, kernel based models, and their implementation and usage as regressions or binary classifiers are the same. However, GP are probabilistic Bayesian predictors which in contrast to SVM produce Gaussian distributions as their output.

Another advantage is the possibility of parameter learning from the training data. On the other hand, SVM provide sparse solution, i.e. only ‘‘support’’ vectors are usedfortheinference,which can be apluswhenworkingwith large amount of data. The evaluation experiments carried out using the MediaEval’2013 music database for emotion estimation and GTZAN corpus for genre classification have shown that GP models consistently outperform the SVM, especially in the classification task.

We have extended the GP application field into the area of music information retrieval, but there are many other unexplored research directions where GP can become viable alternative to the current state-of-the-art methods. One such direction is speech processing and recognition where high performance temporal sequences discrimination and non- linear dynamical system modeling are demanded.

REFERENCES

[1] Konstantin Markov(Member IEEE) AND Tomoko Matsui(Member IEEE)"Music Genre and Emotion Recognition Using Gaussian Processes", IEEE ACCESS, practical innovations:open solutions *june 25, 2014*.

[2] K. Markov and T. Matsui, ‘‘High level feature extraction for the selftaught learning algorithm,’’ EURASIP J. Audio, Speech, Music Process., vol. 2013, no. 6, Apr. 2013.

[3] G. Tzanetakis and P. Cook, ‘‘Musical genre classification of audio signals,’’ IEEE Trans. Speech Audio Process., vol. 10, no. 5, pp. 293–302, Jul. 2002.

[4] J. C. Lena and R. A. Peterson, ‘‘Classification as culture: Types and trajectories of music genres,’’ Amer. Soc. Rev., vol. 73, no. 5, pp. 697–718, Oct. 2008.

[5] T. Li and M. Ogihara, ‘‘Detecting emotion in music,’’ in Proc. Int. Soc. Music Inform. Retr. Conf. (ISMIR), vol. 3. Oct. 2003, pp. 239–240.

[6] G. Tzanetakis, ‘‘Marsyas submissions to MIREX 2007,’’ in Proc. Int. Soc. Music Inform. Retr. Conf. (ISMIR), 2007.

[7] Y.-H. Yang and H. H. Chen, ‘‘Machine recognition of music emotion: A review,’’ ACM Trans. Intell. Syst. Technol., vol. 3, no. 3, pp. 40:1–40:30, May 2012.

[8] Y.-H. Yang and H. H. Chen, ‘‘Prediction of the distribution of perceived music emotion s using discrete samples,’’IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 7, pp. 2184–2196, Sep. 2011.