ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

A Review on Genre Classification and Emotion Estimation of Audio Signals using Gaussian Process

Mugdha Magare[1], Prof. Ranjana Dahake[2]

[1]ME Student, Dept. of Computer Engineering, MET BKC, University of Pune, Nashik, Maharashtra, India

[2]Professor, Dept. of Computer Engineering, MET BKC, University of Pune, Nashik, Maharashtra, India

ABSTRACT: This paper highlights methods for two tasks of Music Information Retrieval (MIR). Genre Classification and Mood Estimation are the tasks that are surveyed in this paper. Genres are the categories in which music is generally classified. It's important to understand genre categorization for efficient music retrieval system. Techniques for genre classification automate the music retrieval systems to its users. On the other hand, human generally categorize music in terms of its emotional associations. Therefore, this paper also provides review of the methods that have been proposed for music emotion recognition. This MIR tasks needs features extraction. Here, we have focused on the method Gaussian Process (GP) and its models, Gaussian Process Classification and Gaussian Process Regression for both the tasks respectively.

KEYWORDS: Music Information Retrieval, Genres, Emotions, Features.

I. Introduction

Lots of music data is available to the users through the internet or any other sources. For users benefit, it is necessary to have efficient music information retrieval technology. Music Information Retrieval consists of tasks such as genre classification, artist identification, mood estimation, cover song identification, music annotation, melody extraction, etc. It gives efficient music search and recommendation systems, intelligent playlist generation, and other attractive applications. Here, we have considered only audio based musical signals only. As detailed work for these two tasks is possible with extracting features from audio based music signal.

Features are the characteristics of piece of music that are recorded at a particular instance of time. There are various feature types exists and their extraction algorithms too. Widely used features are as follows [1]:

- MFCC (mel frequency cepstral coefficients) – It represents short term power spectrum of a sound. It also gives amplitudes of the resulting spectrum.
- LSP (Line spectral pairs) – Speech related feature used for speech coding and for transmission over channel.
- TMBR (timbre feature) – It consists of spectral centroid, spectral flux, spectral rolloff, and zero crossings as scalar features. It shows character or quality of a musical sound, independent of pitch and loudness.
- SCF and SFM (spectral crest factor and spectral flatness measure) – These features are indicating spectral shape and used to differentiate between tone-line and noise-like sounds.
- CHR (chromagram) – This feature represents the spectrum distribution of the distinct semitones and provides information about the key and mode.

For example, chroma vectors are mostly used for some specific tasks, such as music transcription or music scene analysis. On the other hand, spectrum and its derivatives are also widely adopted for music pattern classification.

Generally, there are tools, which allow all of the above features to be extracted as well as any combination of them. When multiple features are calculated, they are stored in the form of single vector per frame of musical piece. The next step after feature extraction is genre classification.

Genre classification is a classical supervised classification task where given labeled data, i.e. songs with their true genre type coming from a finite set of categories i.e. genres. The goal is to predict the genre of an unlabeled music

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

piece.

Human categorization of music appears natural, yet it can be inconsistent, changing, and, in some cases may even seem arbitrary. Though human reactions or judgments are influenced by the audio signal, but also they are dependent on other factors such as artist fashion, dance styles, lyrics, social and political attachments, religious believes, etc. By the time new genres constantly coming into the picture while others are becoming forgotten or irrelevant. Therefore, it is difficult to come up with a commonly agreed set of music genres. Normally in music information retrieval, most popular ten types of genres are taken into consideration and those which are easily distinguishable types. Each genre classification system consists of minimum two blocks: feature extractor and classifier.

Various methods for building music genre classifiers have been utilized, like, support vector machines (SVM), compressive sampling models. In most of the cases, parametric models have been utilized. Those approaches include instances of supervised, semi-supervised and unsupervised methods.

Most of the users use genres or artist names when searching or categorizing music, but Music mostly listened because of its ability to communicate and trigger emotions in listeners. Thus, determining computationally the emotional content of music is attracting researchers towards it. Existing automatic systems for mood recognition are based on emotion representation which can be either categorical or dimensional. Categorical approaches involve finding emotional descriptors, usually adjectives, which can be arranged into groups. Given the perceptual nature of human emotion, it is difficult to come up with an intuitive and coherent set of adjectives and their specific grouping. To investigate consistent interpretation of mood categories, there are models designed to describe emotion using continuous multidimensional metrics defined on low- dimensional spaces. Most widely accepted is the Russell's two-dimensional Valence-Arousal (VA) space where emotions are represented by points in the VA plane [6].

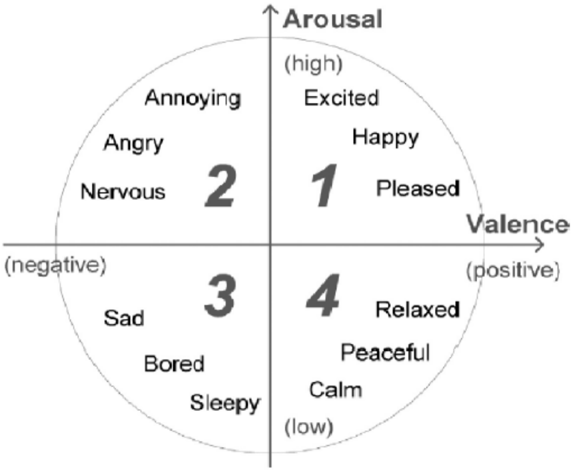

Figure 1: Two dimensional (Valence-Arousal) affective space of emotions.

Figure 1 show the space where some regions are associated with distinct mood categories. In music emotion recognition, the goal is to automatically find the point in the VA plane which is matching with the emotion induced by a given music piece. Valence and Arousal are continuous and independent parameters. They can be estimated separately using the same music feature sets and different regression models.

Regression models such as Multiple Linear Regression (MLR), Support Vector Regression (SVR), as well as Multi-Level Least Squares or regression trees have been successfully applied previously to music emotion estimation. Regression Model learning is supervised method and requires labeled training data. Finding consistent mood labels in terms of VA values is even more challenging than obtaining genre labels since emotion interpretation can be very subjective and varies among listeners. It requires music annotations by multiple experts, which is expensive, time consuming, and labor intensive.

Other than area of music information retrieval, there are many other unexplored research directions where Gaussian Process can be applied. Possibly it can be applied in speech processing and recognition where high performance is required.

II. LITERATURE SURVEY

Automatically extracting music information is gaining importance because of a need to organize the increasingly large numbers of music files available digitally on the Web. It is very likely that in the near future all recorded music in human history will be available on the Web.

Musical genres are labels created and used by humans for categorizing and differentiating music. Musical genres have no fix definitions. It also does not have boundaries as they arise through interaction between the public, marketing, historical, and cultural factors [2]. Genre hierarchies, typically created manually by human experts, are currently one of the ways used to structure music content on the Web. Automatic musical genre classification will automate this process and provide an important component for a complete music information retrieval system for audio signals. In addition it provides a framework for developing and evaluating features for describing musical content. Such features can be used for other music information retrieval tasks and form the foundation of most proposed audio analysis techniques for music.

Three feature sets for representing timbral texture, rhythmic content and pitch content of music signals were proposed and evaluated by G. Tzanetakis & P. Cook using statistical pattern recognition classifiers trained with large real-world audio collections [2]. The proposed features sets successfully testified thus can be used in other music information retrieval tasks. Another approach presented was Compressing Sampling based approach [3]. In that, they present CS-based classifier for music genre classification, with two sets of features, including short-term and long-term features of audio music. The proposed classifier generates a compact signature to achieve a significant reduction in the dimensionality of the audio music signals.

Another faster approach to extract features is investigated by M. Henaff et.al. [4] In this, they investigated a sparse coding method called Predictive Sparse Decomposition (PSD) that attempts to automatically learn useful features from audio data. Due to its faster nature, it is scalable to large scaled datasets.

Music is composed to be emotionally expressive, and emotional associations provide an especially natural domain for indexing and recommendation in today's vast digital music libraries [5]. But such libraries require powerful automated tools, and the development of systems for automatic prediction of musical emotion presents a myriad challenges. The perceptual nature of musical emotion necessitates the collection of data from human subjects. The interpretation of emotion varies between listeners thus each clip needs to be annotated by a distribution of subjects.

E. Kim et al., made comparison in between state-of-the -art techniques for emotion recognition [5]. They have compared human annotations, contextual text information and content-based audio analysis. Still problems remain due to inherent ambiguities of human emotions. The VA plane which is generally used to point emotions on, it is proposed by James Russell [6]. It contains emotions placed on VA plane by their emotional characteristics.

M. Casey et.al. designed a CS-based classifier for music genre classification, with two sets of features, including short-time and long-time features of audio music. The proposed classifier generates a compact signature to achieve a significant reduction in the dimensionality of the audio music signals. They also outlines the problems of content-based music information retrieval and explores the state-of-the-art methods using audio cues (e.g., query by humming, audio fingerprinting, content-based music retrieval) and other cues (e.g., music notation and symbolic representation), and identifies some of the major challenges for the coming years [7].

Markov and Matsui have proposed a novel approach for genre classification and emotion estimation [1]. They have investigated the feasibility and applicability of Gaussian Process models for genre classification and emotion estimation. Gaussian Processes (GPs) are Bayesian nonparametric models.

III. SYSTEM FLOW

For genre classification, GTZAN song collection data set can be used. This database consists 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. For the music emotion estimation purpose, MediaEval'2013 database can be used. It consists of 1000 clips each is of 45 second long taken from different locations from 1000 different songs. This database has songs distributed in 8 genres as Blues, Electronic, Classical, Country, Pop, Jazz, Folk, and Rock.

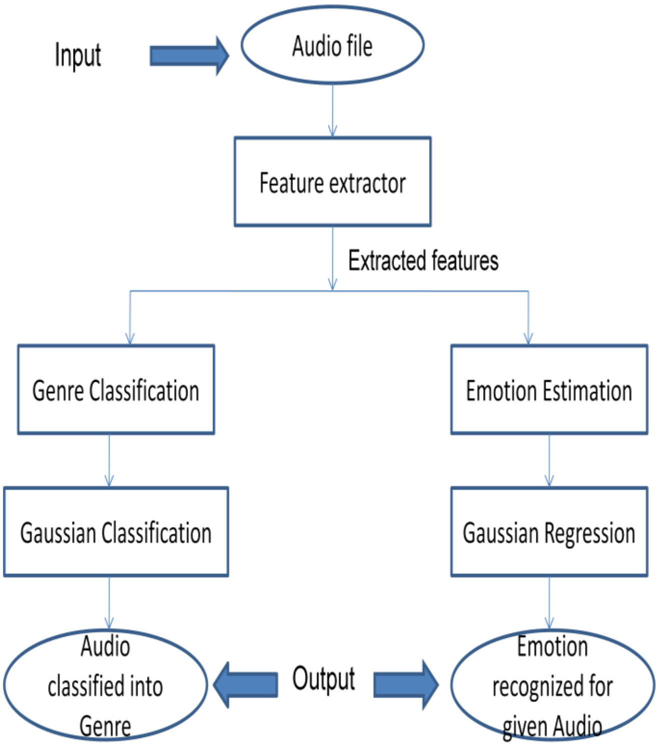

Figure 2: System Flow.

In this paper, used feature extraction methods in music signal processing studies are widely used and can be referred to as a standard set for such tasks. As shown in figure 2, both tasks can be performed after feature extraction. Thus, in this flow, first there is a feature extraction and then either of these tasks. Input is audio clip and output will be classified audio into genres and emotions recognized from input audio clips.

IV. Details of Gaussian Process

Gaussian Processes (GP) are used to describe distributions over functions. 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 function. For real process *f* (*x*), the mean function *m*(*x*) and the covariance function *k*(*x*, *x*') are defined as,

$$m(x) = \mathrm{E}[f(x)] \qquad \text{eq. (1)}$$

$$k(x, x') = \mathrm{E}[(f(x) - m(x))(f(x') - m(x'))] \qquad \text{eq. (2)}$$

Thus, the GP can be written as,

$$f(x) \sim GP(m(x), k(x, x')) \qquad \text{eq. (3)}$$

A GP prior over function *f* (*x*) implies that for any finite number of inputs $X = \{x_i\} \epsilon R^d$, $i = 1,...,n$, the vector of function values $f = [f(x_1), ..., f(x_n)]^T = [f_1, ..., f_n]^T$ has a multivariate Gaussian distribution using eq. (1) and (2).

$$f \sim N(m, K) \qquad \text{eq. (4)}$$

where,

The mean **m** is often assumed to be zero. *N* is the multivariate Gaussian distribution. The covariance matrix **K** has the following form,

DOI: 10.15680/IJIRCCE.2015. 0312015

$$K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & \cdots & k(x_2, x_n) \\ \vdots & & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix}$$

For emotion estimation purpose, Gaussian process regression is used. And for genre classification purpose, gaussian process classification is used.

A] Gaussian Process Classification

For binary classification, given training data vectors $x_i \epsilon R^d$ with corresponding labels $y_i \epsilon \{-1,+1\}$, here 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. It is carried out using logistic function.

Let $X = [x_1, \ldots, x_n]$ be the training data matrix, $y = [y_1, \ldots, y_n]^T$ be the vector of target values, and $f = [f_1, \ldots, f_n]^T$ with $f_i = f(x_i)$ be the vector of latent function values. Given the latent function, the class labels are assumed independent Bernoulli variables and therefore the likelihood can be factorized as shown in eq. (5),

$$p(y \mid f) = \prod_{i=1}^{n} p(y_i \mid f_i) = \prod_{i=1}^{n} sig(y_i f_i) \qquad \text{eq. (5)}$$

B] Gaussian Process Regression

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

$$y = f(x) + \varepsilon \qquad \text{eq. (6)}$$

Where, the latent function $f(x)$ is unknown and ε is often assumed to be a zero mean gaussian noise, i.e. $\varepsilon \sim N(0,\sigma^2_n)$ as shown in eq. (6). Putting GP prior over $f(x)$ allows marginalizing it out, which means that we do need to specify its form and parameters. It makes models very flexible, since $f(x)$ can be any non-linear function of unlimited complexity.

Thus, this work of classification and regression using Gaussian process is based on [1].

V. Conclusion

In this paper, we give the review on implementation of Gaussian Process for music genre classification and emotion estimation purpose. To carry out the model, for genre classification, here Gaussian Process classification is utilized. And for emotion estimation task, Gaussian Process Regression is used. Gaussian Process can be said as much more effective in the field of music information retrieval. Its predictions are truly probabilistic and also it is a nonparametric method. It is also helpful in parameter learning from training data. Gaussian gives better results in music information retrieval than state of the art techniques. Gaussian may also be applied in speech recognition area.

References

1. K. Markov and T. Matsui, “Music Genre and Emotion Recognition using Gaussian Process”, IEEE Access, pp. 688-697, June 2014.
2. 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.
3. K. Chang, J.-S. Jang and C. Iliopoulos, “music Genre Classification via Compressive Sampling”, Proc. Int. Soc. Music Information Retrieval (ISMIR), pp. 387-392, 2010.
4. M. Henaff, K. Jarrett, K. Kavukcuoglu, Y. LeCun, “Unsupervised Learning of Sparse Features for Scalable Audio Classification”, Proceedings of International Society for Music Information Retrieval (ISMIR), pp. 681-686, 2011.
5. E. Kim et al., “Music Emotion Recognition: A state of the art review”, Proceedings of 11th International Society for Music Information Retrieval Conference (ISMIR), pp. 255-266, 2010.
6. J. A. Russell, “A Circumplex Model of Effect”, Journal of Personality and Social Psychology, vol. 16, no. 6, pp. 1161-1178, Dec. 1980.
7. M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, “Content-based music information retrieval : Current directions and future challenges”, Proceeding of IEEE, Vol. 96, no. 4, pp. 668-696, Apr. 2008.