Music Emotion Tracking with Continuous Conditional Neural Fields and Relative Representation

Vaiva Imbrasaitė
Computer Laboratory
University of Cambridge
United Kingdom
Vaiva.Imbrasaite@cl.cam.ac.uk

Peter Robinson
Computer Laboratory
University of Cambridge
United Kingdom
Peter.Robinson@cl.cam.ac.uk

ABSTRACT

This working notes paper introduces the system proposed by the Rainbow group for the MediaEval Emotion in Music 2014 task. The task is concerned with predicting dynamic emotion labels for an excerpt of a song and for our approach we use Continuous Conditional Neural Fields and relative feature representation both of which have been developed or adapted by our group.

1. INTRODUCTION

The Emotion in Music task is concerned with providing dynamic arousal and valence labels and is described in the paper by Aljanaki *et al.*[1].

The use of relative feature representation has already been introduced to the field of dynamic music annotation and tested on MoodSwings dataset [4] by Imbrasaitė *et al.*[2]. They have shown substantial improvement over using standard feature representation with the standard Support Vector Regression (SVR) approach as well as comparable performance to more complicated machine learning techniques such as Continuous Conditional Random Fields.

Continuous Conditional Neural Fields (CCNF) have also been used for dynamic music annotation by Imbrasaitė *et al.*[3]. In our experiments we have achieved results that clearly outperformed SVR when using standard feature representation and produced similar results to using relative feature representation. It was suspected that the short extracts (only 15s) and little variation in emotion were the main reasons why the model was not able to achieve better results. In this paper we are applying the same techniques to a dataset that improves on both accounts with a hope of clearer results.

2. METHOD

2.1 Feature extraction and representation

In our system we used two feature sets. Both feature sets were extracted by OpenSMILE using a standard set of features. As CCNF can suffer when dealing with a large feature vector and fail to converge, we used a limited set of statistical descriptors extracted from the features limiting the total number of features to 150.

The first feature set was used as is, in the standard feature representation. For the second feature set we used a post-processing step to transform it into the relative feature representation—we calculated the average for each feature over each song and for each feature in each feature vector we used the average and the difference between the average and the actual feature to represent it. We thus doubled the size of the feature vector to 300. Relative feature representation is based on the idea of expectation in music. In the past we have shown [2] that using this feature representation can lead to substantially better results, improving the correlation coefficient by over 10% for both axes.

2.2 CCNF

Our CCNF model consists of a undirected graphical model that can model the conditional probability of a continuous valued vector $\mathbf{y}$ (for example emotion in valence space) depending on continuous $\mathbf{x}$ (in this case, audio features).

In our discussion we will use the following notation: $\boldsymbol{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is a set of observed input variables, $\mathbf{X}$ is a matrix where the i^{th} column represents $\boldsymbol{x}_i$,
$\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$ is a set of output variables that we wish to predict, $\mathbf{x}_i \in \mathcal{R}^m$ and $y_i \in \mathcal{R}$ (patch expert response), n is the length of the sequence of interest.

Our model for a particular set of observations is a conditional probability distribution with the probability density function:

$$P(\mathbf{y}|\boldsymbol{x}) = \frac{\exp(\Psi)}{\int_{-\infty}^{\infty} \exp(\Psi) d\mathbf{y}} \tag{1}$$

We define two types of features in our model: vertex features f_k and edge features g_k. Our potential function is defined as:

$$\Psi = \sum_i \sum_{k=1}^{K1} \alpha_k f_k(y_i, \mathbf{x}_i, \boldsymbol{\theta}_k) + \sum_{i,j} \sum_{k=1}^{K2} \beta_k g_k(y_i, y_j) \tag{2}$$

We constrain $\alpha_k > 0$ and $\beta_k > 0$, while $\boldsymbol{\Theta}$ is unconstrained. The model parameters $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \ldots \alpha_{K1}\}$,
$\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \boldsymbol{\theta}_{K1}\}$, and $\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots \beta_{K2}\}$ are learned and used for inference during testing

The vertex features f_k represent the mapping from the $\mathbf{x}_i$ to y_i through a one layer neural network, where $\boldsymbol{\theta}_k$ is the weight vector for a particular neuron k.

$$f_k(y_i, \mathbf{x}_i, \boldsymbol{\theta}_k) = -(y_i - h(\boldsymbol{\theta}_k, \mathbf{x}_i))^2 \tag{3}$$

$$h(\boldsymbol{\theta}, \mathbf{x}_i) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}_i}} \tag{4}$$

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

Table 1: Results for both the SVR and the CCNF models, using both the standard and the relative feature representation techniques

	Arousal				Valence			
	rho	range	RMSE	range	rho	range	RMSE	range
Baseline	0.18	+/-0.36	0.27	+/-0.12	0.11	+/-0.34	0.19	+/-0.11
Basic SVR	0.129	+/-0.325	0.146	+/-0.062	0.073	+/-0.267	0.100	+/-0.055
Basic CCNF	0.116	+/-0.632	0.139	+/-0.068	0.063	+/-0.593	0.102	+/-0.064
Relative SVR	0.148	+/-0.326	0.147	+/-0.064	0.074	+/-0.290	0.099	+/-0.062
Relative CCNF	0.181	+/-0.604	0.118	+/-0.069	0.066	+/-0.530	0.098	+/-0.062

The number of vertex features $K1$ is determined experimentally during cross-validation, and in our experiments we tried $K1 = \{5, 10, 20, 30\}$.

The edge features g_k represent the similarities between observations y_i and y_j. This is also affected by the neighborhood measure $S^{(k)}$, which allows us to control the existence of such connections.

$$g_k(y_i, y_j) = -\frac{1}{2} S^{(k)}_{i,j} (y_i - y_j)^2. \quad (5)$$

In our linear chain CCNF model, g_k enforces smoothness between neighboring nodes. We define a single edge feature, i.e. $K2 = 1$. We define $S^{(1)}$ to be 1 only when the two nodes i and j are neighbors in a chain, otherwise it is 0.

2.2.1 Learning and Inference

We are given training data $\{\boldsymbol{x}^{(q)}, \mathbf{y}^{(q)}\}_{q=1}^{M}$ of M song samples, together with their corresponding dimensional continuous emotion labels. The dimensions are trained separately, but all the parameters ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$) for each dimension are optimised jointly.

We convert the Eq.(1) into multivariate Gaussian form. It helps with the derivation of the partial derivatives of log-likelihood, and with the inference.

For learning we can use the constrained limited memory Broyden-Fletcher-Goldfarb-Shanno algorithm for finding locally optimal model parameters. We use the standard Matlab implementation of the algorithm. In order to make the optimisation both more accurate and faster we used the partial derivatives of the $\log P(\mathbf{y}|\mathbf{x})$, which are straightforward to derive and are similar to those of CCRF [2].

A more thorough description of the model as well as the code to reproduce the results can be found at http://www.cl.cam.ac.uk/research/rainbow/projects/ccnf/

3. RESULTS

In order to get a better understanding of where CCNF stands in terms of performance, we decided to compare it to another standard approach used in the field. We used Support Vector Regression (SVR) model with the Radial Basis Function kernel in the same way we used CCNF—we trained a model for each axis, using 2-fold cross-validation to pick the best parameters for training. The experimental design was identical to the one used in our previous paper [3], which makes the results comparable not only to the baseline method in this challenge, but also between several datasets.

There are several interesting trends visible from the results (see Table 1). First of all, CCNF combined with the relative feature representation clearly outperforms all the other methods for the arousal axis, as well as the baseline method. Secondly, the spread for correlation for CCNF model is twice as big as the one for SVR, while there is little difference between the spread for RMSE for the different methods. In fact, there is little difference in performance between the different methods and the different representations used for the valence axis.

4. FURTHER INSIGHTS

We found it interesting to compare the results achieved with this dataset to those achieved with the MoodSwings dataset. This shows how much of an impact the dataset has on the performance and even the ranking of different methods. In our previous work CCNF clearly outperformed SVR with the standard feature representation, while the results with the relative feature representation were comparable between the two models. With this dataset, we would have to draw very different conclusions—with the standard representation the results were comparable, if not better for SVR, while there was a clear difference between the two when using the relative feature representation for the arousal axis, with CCNF clearly outperforming SVR. This maybe due to the fact that there are more training (and testing) samples in this dataset, the extracts are longer and, possibly, better suited to the task.

The valence axis is still proving problematic. The fact that quite heavyweight techniques are not able to outperform simple models with small feature vectors seems to be indicating that we are approaching the problem from a wrong angle. Improving results for the valence axis should be the top priority for our future work.

5. REFERENCES

[1] A. Aljanaki, Y.-H. Yang, and M. Soleymani. Emotion in music task at mediaeval 2014. In *MediaEval Workshop*, 2014.

[2] V. Imbrasaite, T. Baltrusaitis, and P. Robinson. Emotion tracking in music using continuous conditional random fields and relative feature representation. In *Proc. of ICME*. IEEE, 2013.

[3] V. Imbrasaitė, T. Baltrušaitis, and P. Robinson. Ccnf for continuous emotion tracking in music: Comparison with ccrf and relative feature representation. In *Proc. of ICME*. IEEE, 2014.

[4] J. A. Speck, E. M. Schmidt, B. G. Morton, and Y. E. Kim. A comparative study of collaborative vs. traditional music mood annotation. *Proc. of ISMIR*, 2011.