MediaEval 2015: Music Emotion Recognition based on Feed-Forward Neural Network

Braja Gopal Patra, Promita Maitra, Dipankar Das and Sivaji Bandyopadhyay
Department of Computer Science & Engineering, Jadavpur University
Kolkata, India
{brajagopal.cse, promita.maitra, dipankar.dipnil2005}@gmail.com, sivaji_cse_ju@yahoo.com

ABSTRACT

In this paper, we describe the music emotion recognition system named as JU_NLP to find the dynamic valence and arousal values of a song continuously considered from 15 second to its end in an interval of 0.5 seconds. We adopted the feed-forward networks with 10 hidden layers to build the regression model. We used the correlation-based method to find out suitable features among all the features provided by the organizer. Then we applied the feed-forward neural networks on the above features to find out the dynamic arousal and valence values.

1. INTRODUCTION

Rapid growth of internet has been experienced all over the world since past ten years. It also expedites the process of purchasing and sharing digital music in the Web. Thus, such a large collection of digital music needs an automated process for their organization, management, search, playlists generation etc. People are more interested in creating music library that allows them to access songs in accordance with their moods compared to the title, artists and/or genre [1, 4]. People are also interested in creating music libraries based on several other psychological factors, for example, what songs they like or dislike (and in what circumstances); time of the day and their state of mind etc. [4]. Thus, the classification of music based on emotions is considered as one of the most important aspects in music industry.

Emotion in Music Task at MediaEval addresses the problem of automatic emotion prediction of music in a time frame of 0.5 second as we can observe the significant emotional changes during the discourse of a full length song. The organizers provided the annotated music clips for the Music Emotion Retrieval (MER) task. The music clips were annotated via crowdsourcing using Amazon's Mechanical Turk[1] (MTurk) [6]. They followed the dimensional representations of emotion because it is easier to describe emotions by positioning the content in comparison to a reference point [3]. The *Valence-Arousal* (V-A) representation has been selected for the annotation scheme.

2. FEED-FORWARD NEURAL NETWORK AND CORRELATION

Feed-Forward neural networks (also called the back-propagation networks and multilayer perceptron) are the most widely used models in several major application areas. Figure 1 illustrates a one-hidden-layer feed-forward neural network with inputs x_1, x_2,...,x_n and output ỹ. Each arrow in the Figure symbolizes a parameter in the network. The network is divided into multiple *layers* namely *input layer*, *hidden layer* and *output layer*. The *input layer* consists of just the inputs to the network. Then, the network follows a *hidden layer* which consists of any number of *neurons*, or *hidden units* placed in parallel. Each *neuron* performs a weighted summation of the inputs, which then passes a nonlinear *activation function* σ, also called the *neuron* function.

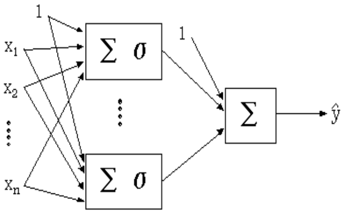

Figure1: A Feed-Forward neural network with one hidden layer and one output.

Mathematically, the functionality of a *hidden neuron* is described by: σ $(\sum_{j=1}^{n} w_j x_j + b_j)$, where the weights $\{w_j, b_j\}$ are symbolized with the arrows feeding into the neuron. The network output is formed by another weighted summation of the outputs of the neurons in the hidden layer [7]. This summation on the output is called the *output layer*. The gradient descent learning principle is used to update the weights as the errors are back propagated through each layer by the well-known back-propagation algorithm [5].

On the other hand, correlation is used to reduce the feature dimension. If we treat all the features and the class in a uniform manner, the feature-class correlation and feature-feature inter-correlations may be calculated as follows.

$$merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad (1)$$

where $merit_s$ is heuristic metric of a feature subset s containing k features, k is the number of components, $\overline{r_{cf}}$ is the average feature–class correlations, $\overline{r_{ff}}$ is the average feature-feature inter-correlation [8].

From the above equation, we can calculate how predictive one attribute is with respect to another. A collection of instances is considered *pure* if each instance is the same in contrast to the value of a second attribute; the collection of instances is *impure* (to some degree) if instances differ with respect to the value of the second attribute. To calculate the merit of a feature subset using above equation, feature-feature inter-correlations (the ability of one feature to predict another and vice versa) must be measured as well [8].

3. APPROACH

Subtask 1: In this subtask, a fixed feature set has been provided by the organizers and we have to implement the models of our choice to identify the valence and arousal for the clips captured in 0.5 second time interval. In this work, we employed two feed-forward neural networks based regression model, in order to map feature values to the arousal and valance scores.

[1] www.mturk.com/

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

Both of the feed-forward neural networks use the same set of feature values but are respectively trained on the arousal or valence score. Each of the feed-forward neural networks is employed with 10 neurons in the hidden layers. We divided the whole training set into 5 parts to reduce the computation time, i.e. we trained our system using around 5000 instances at a time. Then, we tuned our system using a single portion as training and another portions for testing. We calculated the Mean Square Error (MSE) for each of the training sets. Finally, we tested the whole test dataset using five trained modules and got five sets of results. These five sets of results are combined using average and inverse weighted average technique. In the average technique, we simply took the average of all the five results. Whereas, the inverse weighted average is calculated as the equation below,

$$\text{Output}_{weighted} = \frac{\sum_i \frac{y_i}{\delta_i}}{\sum_i \frac{1}{\delta_i^2}} \qquad (2)$$

Where y_i is the i^{th} output and δ_i is the MSE of the i^{th} module. From the equation, we can see that we gave less priority to the result, which has derived from the module having the maximum MSE.

Finally, the Root-Mean-Square Error (RMSE) is used to evaluate the MER systems. We also reported the Pearson's correlation (r) of the prediction and the ground truth. The final RMSE and 'r' for the above two systems (named as baseline feature system with our model average and weighted average) were given in the Table 1.

Subtask 2: In this subtask, a fixed regression model has been provided by the organizers to develop the MER systems based on the different features of our choice. According to the literature, we found that most of the important features are provided by the organizer as the baseline features. So, we focused to find the important feature rather than finding any extra feature. Thus, we used the correlation based feature reduction technique to reduce the baseline feature set given by the organizers as per the formula of correlation in equation 1.

We found 70 and 114 numbers of important features using this correlation formula for the arousal and valence, respectively. Later on, we used the restricted correlation and found 24 and 70 numbers of important features for the arousal and valence, respectively. We also selected 28 important features for valence.

Subtask 3: In this subtask, we implemented the feed-forward neural network on the derived features using correlation in order to build the MER system. We built five systems for difference sets of arousal and valence features described in subtask 2. The RMSE and 'r' values for the above five models were shown in Table 1.

4. CONCLUSION

We used feed-forward neural network to develop a regression based system to find the dynamic arousal and valence for analyzing the emotion in music. The correlation method is used to reduce the feature dimension in order to find suitable features for both arousal and valence. The best model yields minimum RSME of 0.2622 and 0.2913 for arousal and valence respectively using 70 best features, but the 'r' value for the arousal was high as compared to other system for arousal. In future, we want to explore deep learning neural networks for music emotion recognition.

5. ACKNOWLEDGMENTS

The first author is supported by Visvesvaraya Ph.D. Fellowship funded by Department of Electronics and Information Technology (DeitY), Government of India. The authors are also thankful to the organizers A. Aljanaki, Y. Yang and M. Soleymani for their support and help.

6. REFERENCES

[1] B. G. Patra, D. Das, and S. Bandyopadhyay, Unsupervised Approach to Hindi Music Mood Classification, *Mining Intelligence and Knowledge Exploration*, *R. Prasath and T. Kathirvalavakumar (Eds.):* LNAI 8284, pp. 62–69, Springer International Publishing, 2013.

[2] B. G. Patra, D. Das, and S. Bandyopadhyay. 2013. Automatic Music Mood Classification of Hindi Songs. In *Proceedings of the 3rd Workshop on Sentiment Analysis where AI meets Psychology (SAAIP-2013)*, Nagoya, Japan, pp. 24-28.

[3] M. Soleymani, M. N. Caro, E. M. Schmidt, C. Sha, and Y. Yang. 1000 Songs for Emotional Analysis of Music. In *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, pp. 1-6, ACM, 2013.

[4] N. Duncan, and M. Fox. Computer–aided music distribution: The future of selection, retrieval and transmission, *First Monday*, 10(4), 2005.

[5] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Cognitive modeling*, *5*(3):1988.

[6] A. Aljanaki, Y. Yang, and M. Soleymani. Emotion in Music Task at MediaEval 2015. In *MediaEval 2015 Workshop*, September 14-15, 2015, Wurzen, Germany

[7] Mathematica Neural Networks- Train and Analyze Neural Networks to Fit Your Data, *Wolfram Research Inc.*, First Edition, September 2005, Champaign, Illinois, USA

[8] M. A. Hall. Correlation-based feature selection for machine learning. *PhD dissertation,* The University of Waikato, 1999

Table 1: Regression results with different models (BaF: Baseline Feature, OM: Our Model, OF: Our Feature, X: Not Available)

	Arousal				**Valence**			
	RMSE	**Range**	**r**	**Range**	**RMSE**	**Range**	**r**	**Range**
BaF+OM (Average)	0.2689	±0.1073	0.4678	±0.2307	0.3538	±0.1612	-0.0082	±0.3671
BaF+ OM (Weighted Average)	0.2702	±0.1062	0.4671	±0.2282	0.3646	±0.1627	-0.0074	±0.3543
OF (24) + OM	0.2829	±0.1011	0.2787	±0.2531	X	X	X	X
OF (70) + OM	**0.2622**	±0.0899	0.3929	±0.2489	**0.2913**	±0.1452	-0.0037	±0.0281
OF (114) + OM	X	X	X	X	0.3799	±0.1666	-0.0376	±0.3312
OF(28) + OM	X	X	X	X	0.4300	±0.1801	-0.0180	±0.3113