Music Emotion Recognition

Gao Feng
School of Computer Science and Engineering
University of Aizu
Aizu-Wakamatsu City
Fukushima-ken, Japan
m5191104@u-aizu.ac.jp

Konstantin Markov
School of Computer Science and Engineering
University of Aizu
Aizu-Wakamatsu City
Fukushima-ken, Japan
markov@u-aizu.ac.jp

Jianguo Yu
School of Computer Science and Engineering
University of Aizu
Aizu-Wakamatsu City
Fukushima-ken, Japan
m5182104@u-aizu.ac.jp

ABSTRACT

In this paper, we describe our approaches for the MediaEvals' 2015 "Emotion in Music" task. Emotion analysis and recognition have become an interesting issue of research in the middle of the computer vision research area. Our methods consist of Multivariate Linear Regression (MLR), Support Vector Regression (SVR) and FeedForward Neural Networks (FFNN) for dynamic Arousal and Valence regression. In this paper, we first present the results by using the MLR and SVR, then present the results of FFNN. The recognition of music emotions using Deep Learning is one of the latest challenges in the field of speech processing. The simulation results show that recognition with FFNN is better than the traditional methods (MLR and SVR).

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Pattern RecognitionDeep Neural Network; D.2.8 [**Software Engineering**]: Music—*emotion recognition, deep learning*

General Terms

Deep Neural Network

Keywords

Emotion recognition, arousal, valence, linear regression, SVR, FFNN

1. INTRODUCTION

Emotion is a term for a psychological and physiological state associated with a broad variety of thoughts, feelings, and behaviors a [1]. Emotions are subjective experiences, or experienced from an individual point of view. Emotion is often associated with mood, temperament and personality. But in general emotions are short-term whereas moods are long-term and temperaments or personalities are very long-term. Human emotion can be of different types such as happiness, angry, fear, sadness, surprise, disgust, bored, shy etc. Recently, music and emotion recognition, which tries to recognize emotion from music signals, has received increasing attention. Music emotion recognition is a very challenging task of which extracting effective emotional features is an open question [2][3].

Nowadays, there are huge amount of speech and music data on the internet. Thus, availability of automatic systems which can estimate human emotions from speech and music can play an important role in developing new sophisticated applications and services in entertainment as well as in health care industries. A very important problem of automatic speech and music emotion recognition is that the emotion is not only subjective, but also difficult to quantify and analyze. Therefore, the establishment of flexible emotional models is highly demanded. A deep neural network (DNN) is a feed-forward neural network that has more than one hidden layer between its inputs and outputs. With sufficient training data and appropriate training strategies, FFNNs perform very well in many machine learning tasks [4].

2. RELATED WORKS

In the past ten years, there has been a rapid expansion of music information retrieval research towards automated systems with the processing of vast and easily-accessible digital music libraries. Recognition of emotions in music is still in processing, though it has received increasing attention recently [5]. Determining the emotional content of music audio computationally is a cross disciplinary endeavor spanning signal processing, machine learning, music theory and auditory perception. Computational systems for music mood recognition may be on the basis of emotion model, which remain an active topic of psychology research. Categorical and parametric models are supported through substantial prior research with human subjects. Both models are used in Music-IR systems, but the collection of "ground truth" emotion labels remains a particularly challenging problem regardless of the representation being used. The annual Music Information Research Evaluation eXchange (MIREX) is a community-based framework for evaluating Music-IR systems [6]. It included audio music mood classification as a task for the first time in 2007 [7]. The highest performing systems in this category demonstrate the improvement using solely acoustic features every year. But the emotion is not completely encapsulated within the audio alone(social context, for example, plays a prominent role), so approaches incorporating music metadata, such as tags and lyrics.

ICAIT '16, Oct. 6 – 8, 2016, Aizu-Wakamatsu, Japan.

3. METHDOLOGY

3.1 Deep Neural Networks

Deep Neural Networks (DNNs) have recently achieved breakthrough results in almost every machine-learning task. They are a set of machine learning algorithms inspired by how the brain works. Unlike most traditional machine-learning algorithms, Deep Neural networks perform automatic feature extraction without human interference. [8] A simple DNN shows in Fig.1

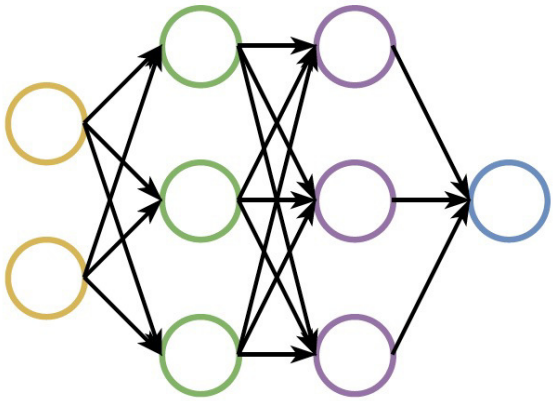


Figure 1: simple DNN

It is hard to understand the behavior of deep neural networks in general, but it is much easier to see what is happening when the data passed through a single layer: A mapping from the input space to the output space. It is a basally linear transformation followed by an activation function, whose mathematical description is Eq.(1), where $\vec{x}$ is the input, W is weights matrix, $\vec{b}$ is bias, a() is activation function and $\vec{y}$ is the output of this layer.

$$\vec{y} = a(W \times \vec{x} + \vec{b}) \tag{1}$$

The single layer actually transforms the data and create a new representation. The equation describe a process by the 5 space operations. First, change the dimensionality of the input space. Second, Rotate the input space. Third, Scale the input space. Fourth, Translate the input space. Fifth, "bend" the input space. The first 3 operations are done by W*$\vec{x}$, the 4th one is provided by $\vec{b}$, and the 5th operation is done by a() which gives nonlinearity to the layer.

3.2 Data Pre-Processing

Like other machine learning methods, pre-processing is needed before training.

Mean Subtraction

Mean subtraction is the most common way which can make it easy for the network to converge. It is just subtracting the mean across every individual feature in the data and it can center the cloud of data around the origin along every dimension.

Normalization

Normalization is the process of organizing the columns (attributes) and tables (relations) of a relational database to minimize data redundancy. There are two common ways of achieving this:

1. Divide each dimension by its standard deviation, once it has been a zero-centered.

2. Normalize each dimension so that the min and max along the dimension is -1 and 1.

Notes that the outputs should also correspond to the activation function of the output layer. For example, if we use the sigmoid activation function, then the range of outputs should be (0,1).

One-Hot Vector

If the task is classification, then instead outputting a most likely class, we also want to know the probabilities being other classes. Then, we need to convert our labels to one-hot vectors of size number of classes. For example, the class with index 12 would be the vector of all 0's and a 1 at position 12.

PCA and Whitening

PCA and Whitening are another form of preprocessing that also helps the convergence of FFNNs.

3.3 Activation functions

The mostly used activation functions are **ReLU, Sigmoid, Tanh**. ReLU has two benefits. First, it is fast than the other two. Second, it does not suffer from the vanishing gradient problem. ReLU's are faster to compute because 1) They supposedly do not require any normalization. 2) They do not require any exponential computation (such as those required in sigmoid or Tanh activations).

3.4 Loss Functions

We train a network by minimizing the error the current network makes. Therefore, first, we need to define the error. The function that measures the error is called the loss function. In our experiment, we just use the regression.

Regression

For the regression tasks, we hope the predictions and targets are as close as possible. The correlation coefficient, sometimes also called the cross-correlation coefficient, is a quantity that gives the quality of a least squares fitting to the original data. It is expressed as R^2. X means the targets and the Y means the predictions. Therefore, the loss function can be any type of distance between them, like root mean squared error(RMSE), y_i means the targets and $\bar{y}_i$ means the predictions. The correlation coefficient and RMSE are given by the following equation.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{2}$$

$$R^2 = \left(\frac{\mathbf{Cov}(X,Y)}{\sqrt{\mathbf{Var}(X)\mathbf{Var}(Y)}}\right)^2 \tag{3}$$

3.5 Updates

Once we have the loss function, we can update the parameters of FFNN by minimizing the error got from loss function.

Backpropagation

Backpropagation is the key algorithm that makes training deep models computationally tractable. For modern neural networks, it can make training with gradient descent as much as ten million times faster, relative to a naive implementation.

Gradient Descent

Gradient Descent The method used in conjunction with Backpropagation for finding the minimum of loss function is Gradient descent. Generates update expressions of Eq.(4). It takes steps proportional to the negative of the gradient as , because we want to minimize the loss function.

$$param = param - learningrate \times gradient \qquad (4)$$

Depending on the Size of examples in each iteration, the name will also change:

1. Stochastic Gradient Descent (SGD): one example from training set in each iteration.

2. Mini-batch gradient descent: m examples from training set in each iteration and the gradient will be averaged over m examples.

Unlike vanilla Gradient descent that runs through all samples in the training set to do a single update for a parameter in a particular iteration, SGD often converges much faster. Note that sometimes people use the term SGD even when referring to mini-batch gradient descent. The size of the minibatch is a hyperparameter but it is not very common to cross-validate it. It is usually based on memory constraints. We use powers of 2 in practice because many vectorized operation implementations work faster when their inputs are sized in powers of 2. The smaller size tends to give more generalization because it will not fit the training set too well in each iteration. Gradient Descent also has a problem. Because it is a first-order optimization algorithm that finds a local minimum of a function, it can get stuck in local minima and fail to reach the global minima as shown in Fig.2.

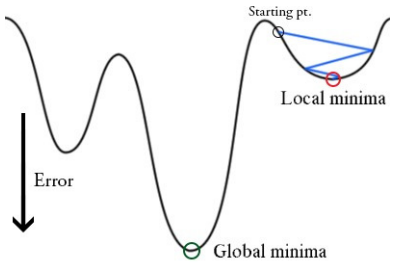
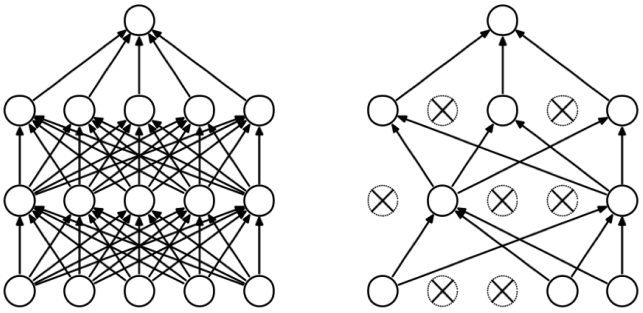


Figure 2: get stuck in local minima

The noise in the stochastic error surface is likely to bounce the network out of local minima, which is one of the reasons why SGD often converges much faster and better.

Dropout

Dropout is an extremely effective, simple regularization technique and recently introduced by Srivastava et al. In [9]. The key idea is to randomly drop units (along with their connections) from the neural network during training as indicated in Fig 3. This prevents units from coadapting too much. While training, dropout is implemented by only keeping a neuron active with some probability p (a hyperparameter), or setting it to zero otherwise.



(a) Standard Neural Net (b) After applying dropout.

Figure 3: Only apply dropout during training

4. DATASET

430 songs have been selected from Free Music Archive(FMA). Last year they filtered out the songs with low agreement to provide a cleaner development set. The extracted 45 seconds excerpts are all re-encoded to have the same sampling frequency, i.e, 44100Hz. Since at the start of the dynamic annotations the annotations were not stable, we discarded the first 15 seconds and the dynamic annotations of the last 30 seconds are provided. The 45 seconds excerpts are extracted from random(uniformly distributed) starting point a song. The dynamic (continuous) annotations were collected at a sampling rate which varied by browsers and computer capabilities. Therefore, we resampled the annotations and generated the averaged annotations with 2Hz sampling rate. To combine the annotations collected for the whole song, on nine points scale, we took the average across all annotators and rounded. The songs were annotated by crowdworkers (annotators) on Amazon Mechanical Turk. Each song was annotated once for arousal and once for valence separately. The crowdworkers were asked were asked to annotate the emotion music intends to induce and not the crowdworkers' own emotion. They had more than 1700 songs from which they selected 430 songs which had the best agreement and changes in their emotional levels. This way they provide a better set for my systems.

The database includes 430 songs. 344 songs for training and 86 songs for testing. They were labeled with arousal and valence values.The features have 260 dimensions, were extracted with openSMILE toolkit. The features include Root Mean Square energy, Zero crossings, Mel Frequency Cepstrum Coefficient, Spectral Flux, etc.

5. APPROACH AND EXPERIMENTS

Music emotion recognition system is mainly based on the study of the psychology. It can be divided into two-representation method: Categorical approach and Dimensional approach [10]. For categorical approach, SVM can be used to train a model, and then classify the categories. But there are two problems with the approach: granularity and ambiguity. Too many categories will lead to many similar categories which are nearly same from each other. However, few categories cannot lead to an effective way to distinguish different emotions. Ambiguity refers to whether the adjective emo-

tional categories used are easy or difficult to distinguish from each other. For dimensional approach, the most widely used is 2-dimensional emotion plane. The problems of granularity and ambiguity can be solved by dimensional approach. But many experiments show that the dimensions of them are not independent from each other, so this method is not perfect. In this paper, we proposed a novel deep neural network to music emotion recognition.

In Table 1, we report the performances(R^2 and RMSE) of three approaches calculated individually for each music piece. All the results are obtained by the same dataset. The analysis of the results obtained this year indicate that all our runs performed better than the algorithms of SVR and MLR. In our experiments, the results showed significant improvement over MLR and SVR. Then, we present the results of the experiments for evaluating the proposed algorithms. We use root mean square error(RMSE) and correlation coefficient(R^2) to evaluate the performance of models. The input dimension is 260 and the output dimension is 2(arousal and valence). The results are as follows:

Table 1: The performance of MLR, SVR and FFNN

Approaches	RMSE-a	RMSE-v	R^2-a	R^2-v
MLR	0.16	0.17	0.35	0.26
SVR	0.17	0.19	0.36	0.25
FFNN(hl=3)	0.25	0.26	0.68	0.66

6. CONCLUSION

In this paper, we present a novel method for emotion recognitions. The method use deep neural network to categorical approach and dimensional approach. From the numerical experiments, the method was proven to be effective for music emotion recognitions. However, there is a still considerable room to improve and show the extant distance between the human brain and the computer in our future work.

7. REFERENCES

[1] Yi-Hsuan Yang and Homer H Chen, *Music emotion recognition*, CRC Press, 2011.

[2] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[3] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[4] Felix Weninger, Florian Eyben, and Björn Schuller, "On-line continuous-time music mood regression with deep recurrent neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5412–5416.

[5] Eduardo Coutinho, George Trigeorgis, Stefanos Zafeiriou, and Björn Schuller, "Automatically estimating emotion in music with deep long-short term memory recurrent neural networks," 2015.

[6] J Stephen Downie, "The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research," *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247–255, 2008.

[7] XHJS Downie, Cyril Laurier, and MBAF Ehmann, "The 2007 mirex audio mood classification task: Lessons learned," in *Proc. 9th Int. Conf. Music Inf. Retrieval*, 2008, pp. 462–467.

[8] Jürgen Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[9] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[10] Konstantin Markov, Tomoko Matsui, Francois Septier, and Gareth Peters, "Dynamic speech emotion recognition with state-space models," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 2077–2081.