A Survey on the Approaches to Achieve Music Emotion Recognition (MER)

Varshanth R Rao
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, ON, N2L 3G1
varshanth.rao@uwaterloo.ca
CS 698: Intro to ML Fall 2017 Project Report

Abstract

The proliferate penetration & outreach of music in the lives of humans of all ages has driven the concept of Music Emotion Recognition (MER) to become vital for delivering critical insights to a variety of complex analytical applications. In this survey, we will transition the review from some of the early approaches of MER into the some of the state-of-the-art neural network based methods to achieve the same.

1 Introduction

Music and emotions have been proven to be tightly coupled and by tagging songs with the possible emotions it is associated to, it would become possible to apply this to various fields of study like psychological analysis, effect of music on social dynamics, situation based recommendations etc. MER is a field of affective computing which can be used for optimizing music retrieval (emotional state based storage & retrieval), delivering personalized context based music recommendations, development of efficient music therapies etc. MER has matured from a coarse classification problem into a granular regression problem. There are many proposed methods to solve MER through various machine learning & deep learning algorithms.

2 Related Works

MER has picked up momentum since its inception in 2007. [1] proposes an organized architecture for emotion & mood based tagging and clustering songs. They use a combination of audio analysis (based on timbre, intensity and rhythm) as well as mood detection from lyrics to categorize songs into the Valence-Arousal (VA) emotional quadrants. Their overall approach was intuitively strong but relied on hard coded weighting of features rather than utilizing powerful machine learning concepts. [2] approached MER as a multi-label classification problem. They used a modified version of the k Nearest Neighbors algorithm to cluster songs with similar multi-class labels (categorized into 6 super classes). The authors attribute the low accuracy level to the overlapping and subjective nature of the numerous labels. In [3], the authors feed the spectrogram of song segments into a Convolutional Neural Network (CNN) which then learns the feature maps which map segments to emotions. The paper is too mathematically condensed to understand the algorithm proposed. The approach of feature learning using spectrograms seems to be very novel and promising from their results.

Submitted to Introduction to Machine Learning (CS698, 2017 Fall). Do not distribute.

3 Survey

3.1 [4]: A Regression Approach to Music Emotion Recognition

3.1.1 Data Set

Curated data set of preprocessed & volume normalized 25s segments from 195 popular songs from Western, Chinese and Japanese albums distributed uniformly in each quadrant of VA planes, labeled in the evoking emotion's VA space.

Feature Space

1. ALL: 114 features consisting of 44 PsySound features, 30 Marsyas features, 12 Spectral Contrast features and 28 DWCH features
2. PsySound15: 15 recommended PsySound features
3. RReliefF: A feature selection algorithm applied to ALL and PsySound separately

3.1.2 Machine Learning Techniques

- Multiple Linear Regression (MLR): The baseline approach using a standard linear regression algorithm which is trained using the least squares estimation
- Support Vector Regression (SVR): SVR non-linearly maps the input feature vectors to a higher dimensional space using the kernelized approach
- BoostR: A non-linear regression technique in which a number of weak regression trees are trained to form range bounded predictions

The training and testing architecture is shown in Fig. 1a

3.1.3 Evaluation

Regression Metric Used: R^2 statistics which can be interpreted as the proportion of the underlying data variation that is explained by the fitted linear regression model

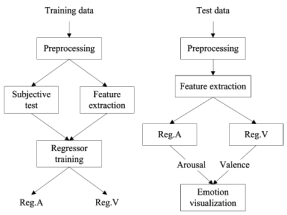
3.1.4 Significant Results & Interesting Conclusions

1. Transforming the data using PCA does not make significant difference to the prediction accuracy even though PCA reduces the Pearson correlation coefficient as shown in Fig. 1b
2. Feature selection greatly improves accuracy especially for valence
3. The upper bound for valence is low justifying that arousal is much easier to model than valence since valence is more subjective than arousal and may change over time
4. Group Wise MER which involves training VA regressors for each group of similar user personalities might yield a more consistent result localized to each group by reducing the inherent noise introduced by user differences

3.1.5 Analysis & Critiques

Strengths

1. The authors of the paper have put considerable effort into creating a minimally biased dataset with equal distribution by sampling from music of different genres & diverse AV values. Using the label consistency evaluation, they further measured the confidence range of the labeled AV values
2. Projection of the emotions to a continuous VA space, hence allowing the predicted emotional state to have more degrees of freedom



(a) System diagram of the proposed approach

R^2 STATISTICS FOR DIFFERENT COMBINATION OF DIFFERENT METHODS, DATA SPACES, AND FEATURE SPACES

Method	Data Space	Feature Space	R^2 statistics a	v
MLR	AV	Psy15	56.8%	10.9%
BoostR	AV	Psy15	55.3%	11.7%
SVR	AV	Psy15	57.0%	22.2%
SVR	PC	$RRF_{18,15}$	58.3%	28.1%
Test-retest [1]	N/A	N/A	80.5%	58.6%

(b) Important Results

Figure 1: Important Figures from [4]

Weaknesses

1. The song segment analysis was done only on the chorus segment of the song. The chorus might not represent the majority of the emotion expressed by the song. Quadrant specific regressors could have been trained to compute a set of probabilistic coefficients for the segments of a song which could be combined to compute a representative AV value
2. The authors do not consider the textual corpus while training the regressors
3. There are several drawbacks of using the R^2 statistic for evaluating regression models, as stated by [5] [6], the most impactful being that it can increase as the number of irrelevant predictors increases. Suggested metrics would be the MSE, Adjusted R^2 etc
4. PCA technique was not detailed sufficiently enough to make concrete conclusions of its accuracy with respect to using other data spaces

3.2 [7]: Multimodal Music Mood Classification using Audio & Lyrics

3.2.1 Data Set

- Audio & Text Corpus: 1000 songs selected from last.fm divided into 4 categories representing the VA space as happy, sad, angry and relaxed. Audio features include timbral, rhythmic and temporal descriptors taken from the evaluation task held by Music Information Retrieval Evaluation eXchange (MIREX). Lyrics corresponding to the songs were obtained from LyricWiki
- Dataspace
    1. TFIDF (Text Frequency, Inverse Document Frequency) weights matrix
    2. Latent Semantic Analysis (LSA) reduced TFIDF matrix which can be viewed as a latent representation of the words in the categorical space
    3. Language Model Differences (LMD) matrix in which the word space was organized per category in the decreasing order of the values of the compromising measure between absolute and relative difference. A set of n top discriminant words and their distances are then taken to represent each category
- Labels: 17 different annotators validated a synonym set from WordNet of the 4 broad emotion categories which were used as tags

3.2.2 Machine Learning Techniques

Sequential Minimal Optimization (SVM algorithm with polynomial kernel by WEKA for multi class classification), Random Forest, Multi-class logistic regression, and K Nearest Neighbors (for textual classification only). Elaborating on kNN, k most similar items from the annotated collection are retrieved and the category with most votes is used to label the test song segment. A small value of k provides less stability while a large value of k may result in weak powers of votes

3.2.3 Evaluation & Significant Results

Evaluation metric used is misclassification error

	Audio	Lyrics	Mixed
Angry	98.1%(3.8)	77.9%(10.3)	**98.3%(3.7)**
Happy	81.5%(11.5)	80.8%(11.2)	**86.8%(10.6)***
Sad	87.7%(11.0)	84.4%(11.2)	**92.8%(8.7)***
Relaxed	91.4%(7.3)	79.7%(9.5)	**91.7%(7.1)**

Figure 2: [7] Classification accuracies with audio features, lyrics with LMD and mixed feature space

- Audio Only: SVM achieved highest accuracy notably for the 'Angry' & 'Relaxed' category. This follows intuition since the arousal component in the music can be well classified by rhythmic descriptors. The standard deviations for the 'Happy' and 'Sad' categories were quite high indicating relatively lower confidence
- Lyrics Only
  1. TFIDF distance based document retrieval using Lucene fed to a kNN classifier. The best classification accuracy (60%) is moderate by itself
  2. LSA dimensionality reduction (to 30 dimensions) achieves nearly the same or worse performance as the kNN classifier
  3. LMD based approach with n=100 yielded the best results for text corpus based emotion classification with SVM yielded the best results as compared to other classifiers
- The multi-modal approach increased the classification accuracy for the 'Happy' and 'Sad' categories significantly while also reducing their standard deviation as shown in Fig. 2

3.2.4 Analysis & Critiques

Strengths

1. The use of multi-modal approach for classification (both audio features and lyrics) maximizes the accuracy and reduces the standard deviation
2. Adaptation of the LMD based approach over the (inefficient) TFIDF document retrieval approach to cluster lyrics into emotionally similar categories

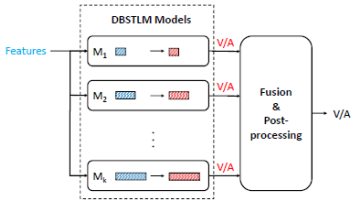
Weaknesses

1. Projection of labels to a discrete domain constricts the granularity of label space and may (undesirably) lead to accurate results
2. Text corpus is not filtered for stop-words which may form a chunk of categorically insignificant determinant words (e.g. were, today, then, need, but etc.)
3. The authors do not ensure the even distribution of the dataset across the VA quadrants, hence possibly localizing songs into clearly separable regions
4. The authors do not use any feature selection algorithm (proven to decrease noise) being applied to the audio features

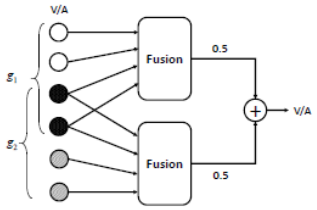
3.3 [8] DBLSTM Multi-scale Fusion for Dynamic Emotion Prediction in Music

3.3.1 Data Set

- MediaEval's Emotion in Music (EiM) dataset: Training set comprised of 431 audio excerpts (of effectively 30 seconds duration) and test set comprised of 58 full songs. The label space are AV values bounded in [-1,1]. 5 cross validation combinations were constructed each with 411 clips of train data and 20 clips of cross validation data randomly selected according to the genre distribution of the test data
- Feature Space (Extracted every 500ms): 260 features consisting of low level acoustic descriptors, their first order derivatives, Mel Frequency Cepstral Coefficients (MFCCs), vocal features and spectral features extracted using openSMILE toolbox



(a) Multi-scale DBLSTM based system framework



(b) Fusion Strategy: g1: models selected by RMSE first
g2: models selected by RMSE and data partition

Figure 3: Important Figures from [8]

3.3.2 Machine Learning Techniques

- Deep Bidirectional LSTM models to extract the temporal relation between the VA values of the fixed length song sequences
- Post Processing Units: Accepts a time continuous sequence VA values and derives its center point using either Triangle Smoothing Filter, SVR or MLR
- Fusion Units: Accepts a vector of VA values pertaining to a moment and outputs a single VA value for that moment using Average/MLR/Extreme Learning Machine (ELM) or ANN

The general architecture is shown in Fig. 3a

Salient Features of the Final Architecture (after Cross Validation)

1. 4 DBLSTM models were trained using sequences of 10,20,30 and 60 segments respectively
2. Cross validation was done with the 3 trials using each of the 5 data sets, hence yielding 15 different trained models
3. 6 different models were selected on the basis of lowest RMSE (RMSE first) and the combination of lowest RMSE and data set (Group first). 2 models were shared by a fusion architecture which considered an average of RMSE first and Group first selected models, as in Fig. 3b

3.3.3 Evaluation

Evaluation was done with RMSE, on the 4 sequence-length specific DBLSTM models so as to obtain the optimal test sequence length (lowest RMSE). Performance for fusion and post processing architectures were tested using this optimal test length sequence as baseline

3.3.4 Significant Results & Interesting Conclusions

1. Shorter sequences perform significantly better than longer sequences when trained on any sequence length model. Test sequence length of 10 yielded minimal RMSE on all the models
2. Models which were tested with a sequence length more than that which they are trained on performed worse as the sequence length increases
3. Triangle smoothing filter achieves the best result for arousal while SVR post processing is best for valence
4. All of the trained fusion techniques yielded lower RMSE than individual DBLSTM models which shows that multi-scale fusion captures structural information to improve on the temporal information gathered by the RNN
5. The sequential combination of fusion with ANN followed by post processing with SVR for valence and Triangle filter smoothing for arousal yielded the best performance for the experiment

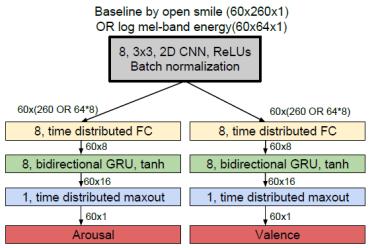


Figure 4: Stacked CNN & RNN architecture in [9]

3.3.5 Analysis & Critiques

Strengths

1. Projection of the emotions to a continuous space (VA values)
2. The use of bidirectional LSTM provides an advantage when there is a segment wise transition over time i.e. when the evoking emotion exhibits a shift during the sequence
3. The authors use the novel 'post processing' which can be considered as the statistical summary from a sequence and the 'fusion' step which extracts the hierarchy of VA values over the sequence length domain

Weaknesses

1. The authors do not consider textual corpus into making the predictions. With RNN, it becomes feasible to associate a sequence of lyrics with an emotional state, hence transforming it into a multi-modal approach
2. The post processing step could have been combined with the fusion step using a single non-expanding bottleneck ANN hence allowing the model to learn non linearities
3. The maximum sequence length was 60 segments which restricts the models which train on larger sequence lengths (e.g. 30 or 60) since there would be limited samples from the same song for the model to train on
4. The authors do use any feature selection algorithm (which may lower the number of parameters the network has to learn)

3.4 [9]: Stacked Convolutional and Recurrent Neural Networks for MER

3.4.1 Data Set

MediaEval's EiM dataset (whose feature set we address as the baseline feature set) and a competing feature space of the Raw Audio Feature which consists of 64 log mel-band energies

3.4.2 Machine Learning Techniques

Stacked CNN with RNN: A CNN fed to 2 parallel RNNs (for Arousal & Valence) as in Fig. 4.

Salient Features of the Final Architecture (after Cross Validation)

1. The CNN filters were used to extract the local shift invariant features from the audio
2. Time distributed layer construction i.e. inputs from CNN to the FC were time distributed i.e. Sequence Length * Flatten(Number of features * Number of filters). Hence both FC and maxout layer had their weights shared across time steps (the sequence is treated as a batch)
3. The bidirectional GRU was utilized to learn the temporal information from the time distributed output of the FC layer
4. Dropout was performed to avoid overfitting with a rate of 0.25 for baseline feature set and 0.75 for raw audio feature set

3.4.3 Evaluation

The baseline was set as the results of the DBLSTM based multi-scale fusion with post processing architecture in [8]. Performance was evaluated for sequence lengths of 10,20,30 and 60 segments for both the baseline feature set and the raw audio feature set in terms of the lowest RMSE and standard deviation achieved

3.4.4 Significant Results & Interesting Conclusions

1. Shorter sequences perform significantly better than longer sequences using the log mel-band energy features. This is concurrent with intuition as shorter sequences of sound tend to be associated with more granular emotions and the longer the sequence gets, the more probable an emotional state transition can occur

2. The proposed architecture with the baseline feature set contains $\sim$30k parameters (400 times fewer than [8]) and with just the log mel-band features the architecture contains $\sim$10k parameters (1200 times fewer than [8])

3. The proposed architecture with the log mel-band features performs very similar to [8]. This proves the authors hypothesis that the network can learn the information from the first and second order derivatives and first order statistics from the raw features

3.4.5 Analysis & Critiques

Strengths

1. Projection of the emotions to a continuous space (AV Values)

2. The use of stacked convolutional and recurrent layers provides a more efficient mechanism (both in terms of performance and number of parameters) than the previous complex DBLSTM based multi-scale fusion architecture

3. The use of bidirectional GRUs provides an advantage when predicting the VA value of the audio sequence to identify segment wise transitions over time i.e. when the evoking emotion (hence its VA values) exhibits a shift during the sequence

4. The authors record a highly significant result for the field of MER with their discovery that raw audio features can be used by the network to learn complex statistical features which otherwise would have to be extracted and incorporated into the feature set explicitly. This drastically reduces the number of parameters the neural network has to learn

Weaknesses

1. The authors do not consider textual corpus into making the predictions (multi-modality absent)

2. The fact the authors have used a dropout rate of 0.75 (to avoid rampant overfitting) when considering the raw audio feature set makes it evident that they have not tuned/constructed the optimal architecture separately for the raw audio feature set. It cannot be directly implied that the model with the raw audio feature set will perform better until empirically proven so

3. The CNN used has a single layer with 8, 3x3 filters, having a local receptive field of size 3x3. This effectively implies that each filter would only be able to characterize 3 timesteps, which is insufficient considering that the average sequence length (out of 10,20,30 and 60) is 30. By training 'n' $(1 > n \geq 4)$ different depth CNN models for the 4 different sequence lengths, with the number of CNN layers directly proportional to the input sequence length, the receptive field feeding into the RNN would have been effectively comparable to the sequence length

4. With 3x3 filter, horizontally, they consider only 3 features. With different sized filters, more features will be considered for convolution possibly resulting in performance change

4 Survey Analysis

[9] produces the current state of the art MER approach by simplifying the work done by [8]. As stated in the related works section, the approach taken by [3] seems very promising by working on the raw spectrogram of audio for feature selection. From the results of [7], it is very evident that a multi-modal approach is more superior than using a text-only or audio only approach. Since emotions are subjective and very volatile, an important open problem is to address a mechanism which captures and localizes this volatility. The authors of [4] suggest group-wise MER which can be used for localizing the accuracy of a label. It is also evident that emotions are very complex and to effectively predict on this complexity, we must explore combinations of various machine learning and deep learning architectures as the authors have done in [8] and [9].

5 Conclusion

Since music is one of the most influential aspects of all humans, MER is an extremely potent avenue to exploit. The struggles in this field stresses the importance of a standardized dataset and a unified well defined label space. Effectively capturing an emotion requires a descriptive data representation, use of feature selection, acceptance of multi-modal approaches and adoption of innovative and justifiable architectures. Reliable datasets must be composed to evaluate the effectiveness of competing approaches. It is also very evident that a song may not represent a rigidly unitary emotion but rather a sequence of emotions. Further research can be done in developing a new label space which represents emotion transitions and composite emotion values. In this survey we presented a brief summary of 4 papers which approached MER differently, discussed their significant contributions and posed a critical analysis for each of them in terms of their strengths and weaknesses.

References

[1] P. Singh, A. Kapoor, V. Kaushik, and H. B. Maringanti, "Architecture for automated tagging and clustering of song files according to mood," *CoRR*, vol. abs/1206.2484, 2012.

[2] A. Wieczorkowska, P. Synak, and Z. W. Raś, *Multi-Label Classification of Emotions in Music*, pp. 307–315. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

[3] X. Liu, Q. Chen, X. Wu, Y. Liu, and Y. Liu, "CNN based music emotion classification," *CoRR*, vol. abs/1704.05665, 2017.

[4] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 448–457, Feb 2008.

[5] J. Frost, "blog.minitab.com: Five reasons why your r-squared can be too high," Feb. 2016.

[6] raegtin, "stats.stackexchange.com: Is r^2 useful or dangerous," July 2011.

[7] C. Laurier, J. Grivolla, and P. Herrera, "Multimodal music mood classification using audio and lyrics," in *2008 Seventh International Conference on Machine Learning and Applications*, pp. 688–693, Dec 2008.

[8] X. Li, J. Tian, M. Xu, Y. Ning, and L. Cai, "Dblstm-based multi-scale fusion for dynamic emotion prediction in music," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, July 2016.

[9] M. et. al, "Stacked convolutional and recurrent neural networks for music emotion recognition," *Accepted for Sound and Music Computing*, 2017.

[10] Y. H. Yang and J. Y. Liu, "Quantitative study of music listening behavior in a social and affective context," *IEEE Transactions on Multimedia*, vol. 15, pp. 1304–1315, Oct 2013.

[11] Y. E Kim, E. M Schmidt, R. Migneco, B. G Morton, P. Richardson, J. Scott, J. A Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," 01 2010.