MULTIMODAL MUSIC EMOTION CLASSIFICATION USING ADABOOST WITH DECISION STUMPS

Dan Su, Pascale Fung, Nicolas Auguin

Human Language Technology Center
Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
dsu@ust.hk, pascale@ece.ust.hk, njlpauguin@ust.hk

ABSTRACT

We propose using AdaBoost with decision stumps to implement multimodal music emotion classification (MEC) as a more appropriate alternative to the conventional SVMs. By modeling the presence or absence of salient phrases in the lyric texts and seeking for proper thresholds for certain audio signal features, it exploits interdependencies between aspects from both modalities in the multimodal MEC system to make the final classification. It can especially prevent the "short text problem" in lyrics. Our accuracy reached an average of 78.19% for classifying 3766 unique songs into 14 emotion categories, with a statistically significant improvement over the audio-only and lyrics-only monomodal MEC systems. We also show that the proposed AdaBoost with decision stumps method performs statistically better on multimodal MEC than the well-known SVM classifier, which only has an average accuracy of 72.08%.

Index Terms— AdaBoost, Decision Stumps, Music, Emotion, Multimodal

1. INTRODUCTION

It is generally believed that music is composed, performed, or listened to with affect [1]. A computational music emotion classification (MEC) system has the potential to greatly enhance the user's experience with music as well as contribute to more effective music data storage and management for music service providers.

Traditional methods for building a computational MEC system are based on the audio content of music [2, 3, 4, 5]. Recently, some researchers have proposed using lyrics as complementary features to audio signals for MEC to address the so-called "glass ceiling" issue due to the "semantic gap" between the low-level music audio features and high-level human perception [6, 7, 8]. More recently, [9, 10, 11, 12] proposed multimodal systems by combining audio features and lyrics features together to improve overall MEC performance.

Yang et al. [9] applied statistical natural language processing methods to analyze lyrics and developed multi-modal fusion methods to combine audio and lyrics on a 4-class MEC task, with an average classification accuracy improvement from 46.6% to 57.1%. Laurier et. al. [12] combined the language model differences of the lyrics with the audio features into a multimodal system also on a 4-class MEC task, showing the effectiveness of their method. Hu et al. [10] examined a wide range of lyric text features, linguistic features and lyric text stylistic features. Classification accuracy from 63.8%, by using the best lyrics feature sets, to 67.5%, by using a multimodal system (by combining the best performing lyric features with audio features) was achieved.

While almost all the above mentioned works focused on investigating features, especially lyric features, and they all used SVMs as machine learning classifier, few of the work proposed investigating the effectiveness of other machine learning methods. Though the lyric features they investigated were some typical text features that are used in text categorization, unlike typical documents for text categorization, music lyrics are short [9]. Very often there are words in the test set that do not appear in the training set at all, causing the so-called "short text problem". The bag-of-words feature representation for the test lyrics may be sparse and not effectively represented, and SVMs classification performance then will be affected. In addition, they generally investigated 4-class MEC tasks (except [10]).

In this paper, we propose an AdaBoost algorithm with decision stumps as weak classifier to do multimodal MEC tasks. AdaBoost with decision stumps has been proved to work quite well in call routing and facial detection systems [13, 14, 15] as well as in anti-spam email filtering [16], where the calls and email texts are typically short. The AdaBoost algorithm has also been investigated for content-based audio classification tasks [17]. In a multimodal MEC system, the AdaBoost with decision stumps method can exploit interdependencies between aspects from both modalities by modeling the presence or absence of some salient phrases in lyrics and set proper thresholds for some continuous at-

tributes of the audio signal. We applied the proposed method on both monomodal and multimodal MEC systems and show that it outperforms traditional SVMs in multimodal MEC tasks. These experiments were conducted on a uniquely large dataset of 3766 Western songs of 14 music emotion categories.

This paper is organized as follows: Section 2 describes the methodology of applying AdaBoost with decision stumps to classify music emotion from lyrics and audio. In Section 3 we show the experimental setups as well as our results. Finally, we conclude in Section 4.

2. METHODOLOGY

AdaBoost is an aggregating machine learning method that combines many weak classifiers linearly, to form a single and accurate classifier. It has been found to work quite well empirically in call routing and facial detection systems [13, 14, 15].

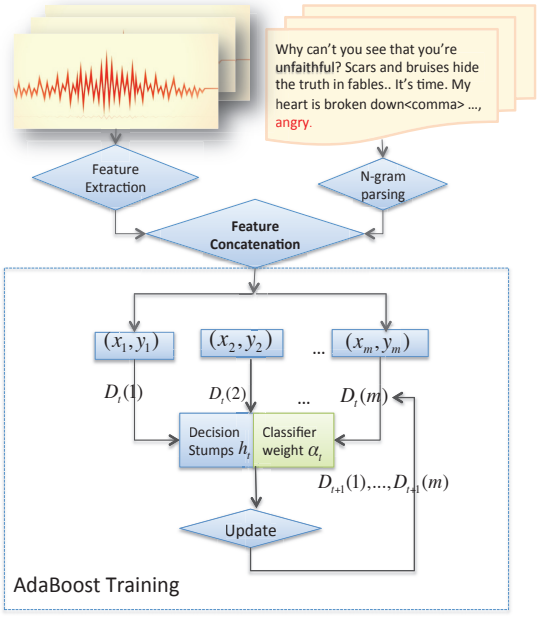

Fig. 1. Training Stage of AdaBoost for MEC from Lyrics

Fig 1 shows our proposed framework of training AdaBoost for multimodal MEC. Feature extraction from the audio files and corresponding lyric texts is performed, and the two feature sets are concatenated together into single multimodal feature vectors for each of the music pieces. For each emotion category, an AdaBoost classifier maintains a weight distribution over all input feature vectors in the training set. It is trained in a sequential way by repeatedly calling weak classifiers. At each iteration, a weak classifier is trained based on the training set and the weight distribution. The final classification is made by a linear combination of weak classifiers from each iteration.

For each music piece i, we use x_i to represent its multimodal feature vector, and $y_i \in \{+1, -1\}$ to represent the corresponding emotion label, where $+1$ indicates positive label of the music piece i and -1 negative, for a binary emotion classification task. Each input (x_i, y_i) will be assigned a weight $D_t(i)$ at AdaBoost learning iteration t. The weak classifier h_t and its corresponding weight α_t are trained based on the input feature sets (x_i, y_i) as well as the weight $D_t(i)$ of each music piece. The nonnegative weights α_t represent how important h_t is for an overall classification.

The weight distribution D_t is initially uniform. At the end of each iteration, it is updated by the following equation (1). The weights of the incorrectly classified music pieces are increased so that the weak classifier at next iteration $t+1$ will be forced to focus on classifying these particular music pieces.

$$D_{t+1}(i) = \frac{D_t(i)e^{(-\alpha_t y_i h_t(x_i))}}{Z_t} \quad (1)$$

with Z_t a normalization factor so that $\sum_{i=1}^{m} D_{t+1}(i) = 1$, as befits a distribution.

We choose decision stump as weak classifier h_t. It has a basic form of one-level decision tree (stump) using confidence-rated predictions. For our multimodal MEC task, it has two different stump forms. The stump is either a distinguished n-gram lyrics feature w, or a proper threshold value W for a certain continuous audio signal attribute. Two output values corresponding to the stump are also trained at each iteration. At the testing stage, a simple check for the absence or presence of the n-gram stump w (as we can see on Fig 2), or a check of the attribute value below or above the threshold is conducted on the test feature vector, and the corresponding output value will be assigned accordingly. If we note $w \in x$ when w occurs in the test feature vector x, and if we note $x(attr_t) > W_t$ when the value of $attr_t$ in the feature vector x is above the threshold, then we can formulate the decision stump classifier $h_t(x)$ at each iteration t in the following form:

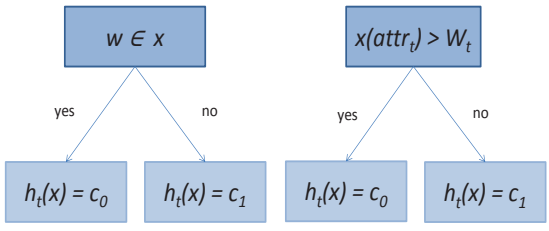

Fig. 2. Decision Stumps

$$h_t(x) = \begin{cases} c_0 & \text{if } w \in x \\ c_1 & \text{if } w \notin x \end{cases} \quad (2)$$

$$h_t(x) = \begin{cases} c_0 & \text{if } x(attr_t) > W_t \\ c_1 & \text{if } x(attr_t) <= W_t \end{cases} \quad (3)$$

where c_j is a real number, indicating the level of "confidence" in assigning the emotion label to x. c_j is chosen so that the normalization term Z_t is minimized for any particular term. The weight α_t corresponding to the weak classifier is computed at each iteration t, also intended to minimize the term Z_t, following the method described in [15]. h_t is then derived at each stage, according to the decision stumps.

Using T to denote the total iteration number, then the final overall classification can be represented as:

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \quad (4)$$

At the testing stage, for an input multimodal feature vector x, the sign of $f(x)$ will be the prediction of whether x belongs to the relative emotion category or not, and the magnitude of the prediction $|f(x)|$ is interpreted as a measure of "confidence" in the prediction.

We use icsiboost [1], an implementation of the AdaBoost.MH algorithm, a member of the boosting family of classifiers [18].

3. EXPERIMENTAL SETUPS AND RESULTS

3.1. Dataset & Evaluation Measure

Our music dataset consists of 3766 songs of Western music in 14 emotion categories (see Table 1). The emotion labels are created by experts from an online music guide.

Table 1. Emotion Categories and Song Distributions

Emotion	# of songs	Emotion	# of songs
sad	615	high	375
groovy	200	happy	401
lonely	332	sexy	315
energetic	339	romantic	187
angry	154	sleepy	156
nostalgic	131	funny	215
jazzy	54	calm	292

To train a binary classifier for each emotion category, we create each negative sample set by randomly selecting songs from other categories that do not appear in this set. We create positive and negative sample sets of the same size for each emotion category.

[1] http://code.google.com/p/icsiboost/

We extracted 30-second duration samples (30s-60s) from each music piece and converted them to 22,050 Hz and 16 bits format and mono channel PCM WAV files.

The lyrics of all songs were automatically collected from two websites: LyricsDB and LyricWiki[2].

We use the classification accuracy as the performance measure. For each emotion category, we show the average accuracy over a 10-fold cross validation.

3.2. Feature Sets

3.2.1. Audio Features

We used three toolkits Marsyas [19], PsySound [20] and openSMILE [21] to extract music audio-based, psychoacoustic-based, and speech emotion-based feature sets respectively.

We applied CfsSubsetEval [22], a correlation-based feature selection method that selects subsets of features that are highly correlated with the emotion labels while having low inter-correlation, paired with a Greedy Forward Search, to do a feature selection on each of the three feature sets. The selected features from each set were combined together as our final audio features.

3.2.2. Lyrics Features

N-grams The N-gram features are one of the best lyric text features. They were reported to perform better than lyric text stylistic features and linguistic features derived from sentiment lexicons [10]. We extracted unigrams, bigrams and trigrams from all lyrics texts of the training data, and constructed a bag-of-words (BOW) model, using Boolean representation.

We applied a Boolean representation for the n-grams because the key observation is the presence or the absence of certain salient words in lyric texts (which are critical for making the classification), and this is also the strategy that the AdaBoost with decision stumps method uses.

3.2.3. Multimodal Feature Concatenation

We concatenated the audio and lyrics features into a single feature vector before training the classifier. It yielded a truly multi-modal feature space.

3.3. Experimental Results

3.3.1. Multimodal MEC using AdaBoost with Decision Stumps

Table 2 shows the comparative results of using AdaBoost with decision stumps based on audio only, lyrics-only and multi-modal features combining audio and lyrics for MEC on the 14 emotion categories. It can be seen that the multimodal MEC

[2] http://lyrics.mirkforce.net/ http://www.lyricwiki.org/

performs the best for most emotion categories, and its average accuracy over the 14 emotion categories is the highest. Lyrics-only MEC shows its capacity on "nostalgic", which makes some sense since the music is "nostalgic" usually because of the lyrics. Audio-based MEC performs the best at "happy", "sexy", "groovy" and "funny". This can perhaps be explained by the fact that these emotions are expressed in music through audio signals more than through lyrics. The average improvement is statistically significant at a 99.9%-confidence level, according to a two-proportion z-test.

Table 2. Multimodal and Monomodal MEC using AdaBoost with Decision Stumps (Acc %)

Emotion (# of songs)	Audio	Lyrics	Multimodal
sad(615)	75.94	70.57	**77.40**
high(375)	74.22	74.74	**77.85**
groovy(200)	**84.00**	75.50	82.50
happy(401)	**72.82**	68.75	71.57
lonely(332)	75.44	70.35	**76.36**
sexy(315)	**75.57**	69.00	75.42
energetic(339)	78.76	72.44	**80.68**
romantic(187)	74.31	73.20	**75.12**
angry(154)	84.58	81.66	**86.38**
sleepy(156)	84.01	81.58	**84.07**
nostalgic(131)	74.17	**81.02**	75.69
funny(215)	**76.74**	70.00	75.32
jazzy(54)	75.33	73.94	**78.50**
calm(292)	77.00	70.35	**77.88**
average	77.35	73.94	**78.19**

3.3.2. AdaBoost vs SVMs

We also compared the performance of the proposed AdaBoost with decision stumps method with SVMs on multimodal MEC; the results are shown in Table 3.

The audio feature sets are the same for the two sets of systems, and we varied the lyrics feature representation using Boolean, frequency and tf*idf for the SVM-based multimodal MEC system since they are the three most often used weighting methods in SVM-based MEC systems [10]. We used linear kernels and default settings for SVMs because other non-linear kernels may cause over-fitting on the multimodal feature vector with high dimensionality, and it is mentioned that other non-linear kernels can not improve the performance further, while they require a longer training time [10].

From the comparative results, we can see that the proposed method performs better at all of the 14 emotion categories. The improvement is statistically significant at a 99.9%-confidence level, according to a two-proportion z-test.

For a multimodal MEC system, the performance is largely affected by the lyrics feature representations. And traditional bag-of-words lyrics feature representations may degrade the performance of some classifiers like SVMs because of the "short text problem". The lyrics feature vectors are often very sparse and are not effectively represented for there are not enough overlapped terms between the training and the testing lyrics. But the proposed AdaBoost with decision stumps method does not care whether there are enough overlapped terms in the test lyrics, which are in the training set vocabulary; it only cares about whether the salient phrases selected from the training lyrics set appeared in the test lyrics or not. So it can prevent the "short text problem" to some extent.

Table 3. Multimodal MEC using AdaBoost outperforms SVM in all emotion categories (Acc %)

Emotion	Multimodal_SVM			Multi_AdaBoost
(# of songs)	tf*idf	Freq	Bool	Bool
sad(615)	68.3	68.2	68.3	**77.40**
high(375)	70.0	72.20	73.2	**77.85**
groovy(200)	73.0	73.50	72.8	**82.50**
happy(401)	69.7	68.6	70.0	**71.57**
lonely(332)	69.1	70.2	68.5	**76.36**
sexy(315)	71.3	68.6	71.4	**75.42**
energetic(339)	68.4	68.4	67.1	**80.68**
romantic(187)	70.9	71.9	69.5	**75.12**
angry(154)	78.9	74.4	78.2	**86.38**
sleepy(156)	75.3	70.2	76.6	**84.07**
nostalgic(131)	76.7	76.0	74.8	**75.69**
funny(215)	70.5	71.6	69.8	**75.32**
jazzy(54)	75.0	70.4	76.9	**78.50**
calm(292)	71.7	69.9	72.1	**77.88**
average	72.06	70.98	72.08	**78.19**

4. CONCLUSION

In this paper, we proposed to apply the AdaBoost with decision stumps method to classify music pieces into emotion categories, by exploiting information contained in both lyric texts and audio features. We showed that we gain a statistically significant improvement over the audio-only and lyrics-only monomodal MEC systems. This method also performs statistically better on multimodal MEC systems (with an average accuracy of 78.19%) than the widely used SVM classifier (which only reaches an average accuracy of 72.08%) for a 14-class MEC task on a 3766-song data set. In addition, our proposed method is not constrained by the number of emotion categories.

5. ACKNOWLEDGMENTS

This research is partially supported by the grant RDC R5437 from Velda Limited.

6. REFERENCES

[1] J.A. Sloboda and P.N. Juslin, "Music and emotion: Theory and research," *Emotions in everyday listening to music. In. Juslin, P., Sloboda, JA, editors*, 2001.

[2] T. Li and M. Ogihara, "Detecting emotion in music," in *Proceedings of the International Symposium on Music Information Retrieval, Washington DC, USA*, 2003, pp. 239–240.

[3] Y.H. Yang, C.C. Liu, and H.H. Chen, "Music emotion classification: a fuzzy approach," in *Proceedings of the 14th annual ACM international conference on Multimedia*. ACM, 2006, pp. 81–84.

[4] L. Lu, D. Liu, and H.J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 5–18, 2006.

[5] H. Chen and Y. Yang, "Prediction of the distribution of perceived music emotions using discrete samples," *IEEE Transactions on Audio, Speech, and Language Processing*, , no. 99, pp. 1–1, 2011.

[6] H. He, J. Jin, Y. Xiong, B. Chen, W. Sun, and L. Zhao, "Language feature mining for music emotion classification via supervised learning from lyrics," *Advances in Computation and Intelligence*, pp. 426–435, 2008.

[7] M. Van Zaanen and P. Kanters, "Automatic mood classification using tf* idf based on lyrics," *ISMIR*, pp. 75–80, 2010.

[8] Rada Mihalcea and Carlo Strapparava, "Lyrics, music, and emotions," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea, July 2012, pp. 590–599, Association for Computational Linguistics.

[9] Y.H. Yang, Y.C. Lin, H.T. Cheng, I.B. Liao, Y.C. Ho, and H. Chen, "Toward multi-modal music emotion classification," *Advances in Multimedia Information Processing-PCM 2008*, pp. 70–79, 2008.

[10] X. Hu and J.S. Downie, "Improving mood classification in music digital libraries by combining lyrics and audio," in *Proceedings of the 10th annual joint conference on Digital libraries*. ACM, 2010, pp. 159–168.

[11] B. Schuller, F. Weninger, and J. Dorfner, "Multi-modal non-prototypical music mood analysis in continuous space: Reliability and performances," in *Proceedings 12th International Society for Music Information Retrieval Conference, ISMIR*, 2011, pp. 759–764.

[12] C. Laurier, J. Grivolla, and P. Herrera, "Multimodal music mood classification using audio and lyrics," in *Seventh International Conference on Machine Learning and Applications, 2008. ICMLA'08.* IEEE, 2008, pp. 688–693.

[13] G.D. Fabbrizio, D. Dutton, N.K. Gupta, B. Hollister, M. Rahim, G. Riccardi, R. Schapire, and J. Schroeter, "AT&T help desk," in *Seventh International Conference on Spoken Language Processing*, 2002.

[14] G. Tur, D. Hakkani-Tur, L. Heck, and S. Parthasarathy, "Sentence simplification for spoken language understanding," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011*. IEEE, 2011, pp. 5628–5631.

[15] R.E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine learning*, vol. 39, no. 2, pp. 135–168, 2000.

[16] X. Carreras and L. Marquez, "Boosting trees for anti-spam email filtering," *arXiv preprint cs/0109015*, 2001.

[17] G. Guo, H.J. Zhang, S.Z. Li, et al., "Boosting for content-based audio classification and retrieval: an evaluation," in *IEEE International Conference on Multimedia and Expo*, 2001.

[18] Benoit Favre, Dilek Hakkani-Tür, and Sebastien Cuendet, "Icsiboost," http://code.google.come/p/icsiboost, 2007.

[19] G. Tzanetakis and P. Cook, "Marsyas: A framework for audio analysis," *Organised sound*, vol. 4, no. 3, pp. 169–175, 1999.

[20] D. Cabrera et al., "Psysound: A computer program for psychoacoustical analysis," in *Proceedings of the Australian Acoustical Society Conference*, 1999, vol. 24, pp. 47–54.

[21] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[22] M.A. Hall, *CORRELATION-BASED FEATURE SELECTION FOR MACHINE LEARNING*, Ph.D. thesis, The University of Waikato, 1999.