MULTIMODAL MOOD DETECTION BASED ON LYRICS AND AUDIO

Kenil Shah
V00903842
University of Victoria

Alina Bukatova
V00891033
University of Victoria

Radu Ionescu
V00891249
University of Victoria

ABSTRACT

The atmosphere of a song is largely conveyed through musical elements, but the lyrics which correlate to a song contribute a significant amount of information and have a crucial role in dissecting the meaning of a song as well, and understanding the mood of music based on these factors has a lot of potential for solving problems like genre identification and suggestion. In this project we will conduct a study into multimodal music emotion recognition, and strive towards designing a program which can accurately classify the emotion based on a predetermined emotion paradigm. The data which this program will consider as input for classification mainly includes a given song and its lyrics, as well as the corresponding MIDI file, whose possible inclusion will be determined as we further investigate its benefits and disadvantages. Moreover, this exploration of the music emotion recognition (MER) problem will utilize the multi-modal Music Information Retrieval Evaluation eXchange (MIREX) like emotion dataset - based on the emotion tags used in the MIREX Mood Classification Task.

1. INTRODUCTION

Since its conception, music has been a prime tool used for emotional expression, comprehension, and navigation. In recent years, the surge of music platforms and the increasingly extensive content that has accumulated across these platforms has gifted people the freedom to explore, experiment with, and be inspired by sounds that suit any desired mood and environment. This abundance, while beneficial to listeners and artists alike, has posed challenges for those conducting research in the fields of Music Information Retrieval (MIR) and MER due to how expansive the range has become for musical mood classification. The challenge of detecting the mood created by a musical piece is rather simple for human listeners, but algorithms designed to carry out the same action have so far only been able to achieve satisfying results to a certain extent; such investigations are both costly and time-consuming, and do not scale to the volume of music required to address real-world research questions. Many studies have solely considered the audio content of songs in order to classify music by mood (e.g. [3, 14, 16]), others have focussed on using song lyrics to achieve the same goal (e.g. [10, 12]), and some have done so with the use of MIDI files (e.g. [2]). Naturally, these studies have developed simultaneously emerging branches of research, where authors explore the intersection of two or three of these musical elements with multimodal approaches to music mood detection (e.g. [1, 2, 8, 9, 11]). In this project we will investigate this multimodal approach to design a program that considers audio, lyrics, and possibly MIDI file information in order to correctly classify the various moods of musical pieces.

2. PROCESS

Xiao, Kahyun & Downie in [1] presents a robust but flexible framework to undertake music mood classification using multiple information sources. We choose to apply this framework as it is flexible in that each component can be easily extended by changing the tools, algorithms and methods. The roadmap of this approach, which will be enumerated further in the next sections, is - Dataset construction → Feature Generation/Extraction → Feature Selection → Multimodal Combination → Classification → Evaluation and Analysis.

2.1 Dataset

For this project we are going to use the multi-modal MIREX like emotion dataset [2]. This dataset consists of 903 audio clips, 764 lyrics, 193 MIDIs. It uses MIREX mood tags. It contains 5 clusters with several emotional categories in each cluster.

2.2 Feature Generation

After choosing the dataset, the next step is to extract features from various modes of information (audio, lyrics, MIDI).

2.2.1 Audio

In the case of audio, we will generate standard audio features such as tempo, as well as spectral features such as MFCC's. To accomplish this task we use Marsyas [17], a software framework developed for audio processing in the field of MIR. It is also widely used in this field to study this task [1, 2, 10, 15]. As a backup we will also look into echonest API, as it can also help extract spectral features in addition to other percussive, harmonic, and structural features [13].

2.2.2 Lyrics

Music mood classification has hit a glass ceiling using single mode approaches [14]. Using lyrical features to compensate for these shortcomings is a common method, as lyrical features have shown to have strong correlation to audio features [14]. They also in some cases outperform audio features [9, 10]. We plan to use various feature types to form a comprehensive set of lyrical features.

Starting with bag-of-words features evaluated on content words. Content words are all words except stop words. In particular, we are going to evaluate a combination of these bag-of-words features and their bigrams and trigrams, using two representational models: Boolean and TF-IDF weighting. Our decision to use these models is based on the excellent results they yield [8, 10]. As for using bigrams & trigrams, it is shown that bag-of-words using uni+bi+trigrams are very suitable for music emotion classification [13].

Next, we extract text stylistic features, which are compiled in [9]. This will be done in part using the jLyrics suites in jMIR [2, 7]. We will also employ Synesketch, a Java API, which uses NLP for textual emotion recognition [2]. We also hoped to use features generated from using English lexicons such as ANEW, but it was deemed to be impractical as lexicons such as ANEW are not readily available to use.

2.2.3 MIDI

We consider the methods used in [2] as this seems to be the only study in multimodal MER that employs features extracted from MIDI files. We employ the tools used in this research. Particularly, jSymbolic software [4], MIR MATLAB MIDI toolbox [5] and jMusic [6].

2.3 Feature Selection

Feature selection is an important component of hybrid MER approaches, as aggressive reduction in lyrical features have proven to give improved results [11]. Tentatively, we are going to employ SVM feature weighting. Other methods in consideration are F-scores, Chi-Square [1] and the relief algorithm [2]. The relief algorithm outputs a weight for each feature which is then ranked and optimal number of features determined experimentally by adding one feature at a time. As one would assume, SVM feature weighting gives the best results when using SVM classifiers.

2.4 Multimodal Combination

Hybrid approaches like ours need a method to fuse the multiple sources of information. We use two methods, namely EFFC (early fusion with feature concatenation), and LFLC (late fusion by linear combination) [15]. We decided to compare these two methods as studies employing them have shown very mixed results [1, 15].

2.5 Classification

We intend to explore and extend this section as we continue our efforts into this task. As a starting point we will start with SVM with grid parameter search as this classifier is shown to have the best accuracy in multiple studies employing the same approach as ours [1, 2, 10, 18, 19]. We will use Python's scikit-learn library.

3. TIMELINE

Feb 28 - Mar 6: Familiarize ourselves with the tools to use, requirement gathering for assigned sections and clean up dataset.

Mar 7 - 13: Complete code for feature generation.

Mar 14 - 20: Complete code for feature selection and multimodal combination

Mar 21 - 27: Complete code for classification.

Mar 28 - end of class: Evaluation and Analysis.

4. INITIAL RESULTS

4.1 Midi Features

Using jSymbolic [4], we extracted 278 features from 196 Midi files in our dataset, with a feature vector dimension of 1495. Due to the high dimensionality of our feature vector we decided to employ the following feature selection methods :-

1) Tree based (Random Forests) feature selection (TRF)
2) Recursive Feature Elimination with SVM estimator (RFE)
3) Anova F-scores

We ran an SVM classifier on the above three feature sets alongside a baseline. The accuracies are presented in table 1, figure 1, figure 2. All experiments were conducted with stratified 10-fold cross validation with 20 repetitions, on a 70-30 train-test dataset split.

In the case of RFE, we first employed recursive feature elimination with cross validation to find the optimal number of features. Then we employed

feature elimination until we were left with only the optimal number of features. In our case, our feature vector was pruned 1495 to only 30 optimal features. Top features selected by RFE were : *pitch skewness, melodic thirds, mean tempo, voice separation, initial tempo, medium rhythmic value offset.*
The low number of features combined with high training accuracy, leads us to believe that the model might be overfitting in this case and a more in-depth look is necessary.

As a separate experiment, we calculated the ANOVA F-scores of all the features, ranked them, and recursively added the top features five at a time to check the optimal number of features which gives us the best test accuracy. In this case, we found the optimal number of features to be 365, down from 1495.

Accuracy	Baseline	TRF	RFE	F-scores
Training	30.42%	38.57%	68.53%	51.6%
Test	33.9%	35.59%	47.45%	64.4%

Table 1: Training and Test accuracies

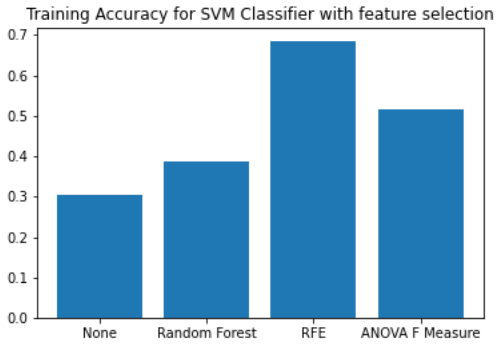

Figure 1: Training Accuracies

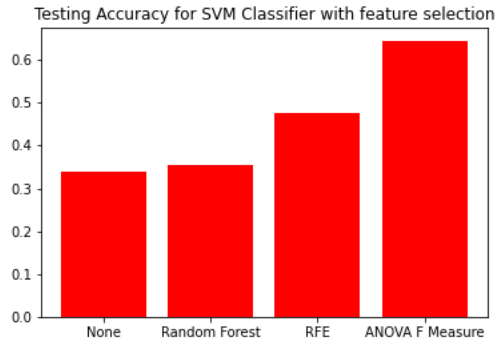

Figure 2: Test Accuracies

Feature selection is clearly shown to increase accuracy in our task. Next step in our experiments is going to be hyperparameter tuning for SVM and exploring other feature selection methods. We especially are keen to explore the chi-squared test for feature selection, as this test is shown to give good results on bag-of-words features which we aim to extract from our lyrics dataset.

4.2 Audio Features

Using liberosa we started by doing tempo analysis on all 903 audio files. After the tempo analysis was working we got the spectral features to work these features include :

1) MFCC
2) Time frequency representation
3) Frequency Domains
4) Spectrographs

We first ran these without doing any data cleansing meaning that all the music clips were mp3 not wav. After noticing some wierd discrepencies we formatted the mp3 files to wav this gave us a better set of features to work with.

There are a couple more features that we would like to see is Zero Crossing rate would be one of them as it can help with percussive vs pitched sound detection, monophonic pitch estimation, and voiced/unvoiced speech signal determination. Thus we are hoping by using it it would help with music genre classification and audio segmentation.

4.3 Lyric Features

For the textual processing portion of our program, we used a bag-of-words approach with Tf-Idf. This has yielded a feature matrix of 764 rows, corresponding to the number of lyric text files given in the MIREX dataset, and 9575 columns corresponding to the feature words that were extracted from these files.

Given these results we saw that this caused a significant dimensionality problem. One sub-solution that we implemented in our program, aimed to reduce the dimensionality problem, involved us preceding the model with an important preprocessing, which included word cleaning, stop words removal, stemming and lemmatization.

Since we used the term frequency-inverse inverse document frequency (Tf-Idf) approach, the values of words increased proportionally to count, but they were inversely proportional to the frequencies of the words in the corpus. We have initially focussed on capturing unigrams in our analysis to maintain a level of simplicity and understandability within the code, as well as proceeding with the results with the intention of comparing them to those given by capturing bigrams and trigrams. In order to drop some columns and reduce the matrix dimensionality, we are currently carrying out some feature selection, a process in which we treat each category as binary, and perform a Chi-Square test to determine if a feature and the binary target are independent, keeping only the features with a certain p-value from the test. Some of our results using this approach with a smaller subset of the data have yielded a feature reduction of about three times, but processing the entire dataset for training our program has created obstacles that have required revising our model and techniques and rewriting parts of our code.

In the next steps, we will further strengthen this portion of the program by following the results described above, and training a Naive Bayes algorithm classifier on the feature matrix, before testing it on the transformed test set.

5. Scenarios for completion

5.1 Basic/minimum scenario

For our basic scenario we believe to have the models trained for individual modes of information (Midi, lyrics, audio).

5.2 Expected Scenario

Our expected outcome for this project is to have tested a multimodal model trained from features extracted from all 3 modes of information (Midi, lyrics, audio) alongside tests conducted not only on the dataset but random songs introduced into the dataset.

5.3 Stretch Goal

The stretch goal for us is to create a simple app, where any user can input a combination of midi, lyrics and audio files, and the app will automatically detect the mood based on the five clusters we are experimenting with.

6. REFERENCES

[1] Hu, Xiao & Choi, Kahyun & Downie, J.. (2016). A framework for evaluating multimodal music mood classification. Journal of the Association for Information Science and Technology. 68. n/a-n/a. 10.1002/asi.23649.

[2] Panda R., Malheiro R., Rocha B., Oliveira A. & Paiva R. P. (2013). “”. 10th International Symposium on Computer Music Multidisciplinary Research – CMMR'2013, Marseille, France.

[3] Salamon, Justin & Gómez, Emilia. (2011). Melody Extraction from Polyphonic Music: MIREX 2011.

[4] Mckay, Cory & Fujinaga, Ichiro. (2006). JSymbolic: A feature extractor for MIDI files. International Computer Music Conference, ICMC 2006.

[5] Eerola, Tuomas & Toiviainen, Petri. (2004). Mir in matlab: The MIDI toolbox.

[6] Brown, Andrew R. & Sorensen, Andrew C.(2000) Introducing jMusic. In Australasian Computer Music Conference, 2000-07-05 - 2000-07-08.

[7] Cory Mckay. 2010. Automatic music classification with jmir. Ph.D. Dissertation. McGill University, CAN. Order Number: AAINR68492.

[8] Xiao Hu and J. Stephen Downie. 2010. Improving mood classification in music digital libraries by combining lyrics and audio. In Proceedings of the 10th annual joint conference on Digital libraries (JCDL '10). Association for Computing Machinery, New York, NY, USA, 159–168. DOI:https://doi.org/10.1145/1816123.1816146

[9] Hu, X., & Downie, J.S. (2010). When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis. ISMIR.

[10] Hu, Xiao et al. “Lyric Text Mining in Music Mood Classification.” ISMIR (2009).

[11] C. Laurier, J. Grivolla and P. Herrera, "Multimodal Music Mood Classification Using Audio and Lyrics," *2008 Seventh International Conference on Machine Learning and Applications*, 2008, pp. 688-693, doi: 10.1109/ICMLA.2008.96.

[12] He, Hui & Jin, Jianming & Xiong, Yuhong & Chen, Bo & Sun, Wu & Zhao, Ling. (2008). Language Feature Mining for Music Emotion Classification via Supervised Learning from Lyrics. 426-435. 10.1007/978-3-540-92137-0_47.

[13] Mcvicar, Matt & Freeman, Tim & Bie, Tijl. (2011). Mining the Correlation between Lyrical and Audio Features and the Emergence of Mood.. Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011. 783-788.

[14] Aucouturier, Jean-julien & Pachet, Francois. (2004). Improving Timbre Similarity: How high's the sky?. J. Negat. Results Speech Audio Sci. 1.

[15] Yang, yi-hsuan & Lin, Yu-Ching & Cheng, Heng-Tze & Liao, I-Bin & Ho, Yeh-Chin & Chen, Homer. (2008). Toward Multi-modal Music Emotion Classification. 70-79. 10.1007/978-3-540-89796-5_8.

[16] Hu, Xiao & Downie, J. & Laurier, Cyril & Bay, Mert & Ehmann, Andreas. (2008). The 2007 MIREX audio mood classification task: Lessons learned. ISMIR 2008 - 9th International Conference on Music Information Retrieval. 462-467.

[17] Tzanetakis, G., & Lemstrom, K. (2007). Marsyas-0.2: a case study in implementing music information retrieval systems. Intelligent Music Information Systems.

[18] Yang, Y.-H. et al: A regression approach to music emotion recognition. IEEE Trans.
Audio, Speech and Language Processing, Vol. 16, No. 2 (2008) 448–4

[19] Cheng, H.-T. et al: Automatic chord recognition for music classification and retrieval. Proc. ICME (2008) 1505–150