Detecting Emotion in Music

Tao Li
Department of Computer Science
University of Rochester
Rochester, NY 14627
email: taoli@cs.rochester.edu

Mitsunori Ogihara
Department of Computer Science
University of Rochester
Rochester, NY 14627
email: ogihara@cs.rochester.edu

1 Introduction

Music is not only for entertainment and for pleasure, but has been used for a wide range of purposes due to its social and physiological effects. Traditionally musical information has been retrieved and/or classified based on standard reference information, such as the name of the composer and the title of the work etc. These basic pieces information will remain essential, but information retrieval based on these are far from satisfactory. Huron points out that since the preeminent functions of music are social and psychological, the most useful characterization would be based on four types of information: the style, emotion, genre, and similarity [Huron,2000].

The relation between musical sounds and their influence on the listener's emotion has been well studied. The celebrated paper of Hevner [Hevner,1936] studied this relation through experiments in which the listeners are asked to write adjectives that came to their minds as the most descriptive of the music played. The experiments substantiated a hypothesis that music inherently carries emotional meaning. Hevner discovered the existence of clusters of descriptive adjectives and laid them out (there were eight of them) in a circle. She also discovered that the labeling is consistent within a group having a similar cultural background. The Hevner adjectives were refined and regrouped into ten adjective groups by Farnsworth [Farnsworth,1958]. We hypothesize that emotion detection in music can be made by analyzing music signals. We approach to the problem using multi-label classification.

We cast the emotion detection problem as a *multi-label classification problem*, where the music sounds are classified into multiple classes simultaneously. That is a single music sound may be characterized by more than one label, e.g. both "dreamy" and "cheerful." We divide the process of emotion detection in music into two steps: *feature extraction* and *multi-label classification*. In the feature extraction step, we extract from the music signals information representing the music. The features extract should be *comprehensive* (representing the music very well), *compact* (requiring a small amount of storage), and *effective* (not requiring much computation for extraction). To meet the first requirement the design has to be made so that the both low-level and high-level information of the music is included. In the second step, we build a mechanism (an algorithm and/or a mathematical model) for identifying the labels from the representation of the music sounds with respect to their features.

A	cheerful,gay,happy	H	dramatic, emphatic
B	fanciful, light	I	agitated, exciting
C	delicate,graceful	J	frustrated
D	dreamy,leisurely	K	mysterious, spooky
E	longing, pathetic	L	passionate
F	dark, depressing	M	bluesy
G	sacred, spiritual		

Table 1: The adjective groups. The first ten of them are the Farnsworth groups and the last three are the additions.

2 The Music Data Used and Their Emotional Labels

A collection of 499 sound files was created from 128 music albums as follows: From each album the first four music tracks were chosen (three tracks from albums with only three music tracks). Then from each music track the sound signals over a period of 30 seconds after the initial 30 seconds were extracted in MP3. The collection covered four major music types, Ambient (120 files), Classical (164 files), Fusion (135 files), and Jazz (100 files).

The 499 files were labeled by a subject (a 39 year old, male). The ten adjective groups of Farnsworth [Farnsworth,1958] were used for the labeling. The subject was instructed to select for each track *all* adjective groups that match the sound with no limit to the number of groups chosen. Also, the subject was instructed to suggest a new adjective group if necessary. The subject added three new groups: mysterious, spooky; passionate; and bluesy, thereby increasing the total number of groups to thirteen. The subject was also asked to group the thirteen groups into "supergroups." He formed the six supergroups. Table 1 shows first few adjectives of each group. The six supergroups are $(A, B), (C, D), (E, L), (H, I, J), (G, K)$ and (F, M).

3 The Classification Method

In the traditional classification problem, classes are mutually exclusive by definition. In emotion detection in music, how-

ever, the disjointness of the labels is no longer valid, in the sense that a single music sound may be classified into multiple emotional categories. This stipulation seems to make the problem significantly more complicated. Unfortunately, the area is yet to be explored. The sparse literature on this subject is primarily geared toward text classification and, to our knowledge, no prior work exists in the music information retrieval domain.

We resort to the scarcity of literature in multi-label classification by decomposing the problem into a set of binary classification problems. In this approach, for each binary problem a classifier is developed using the projection of the training data to the binary problem. To determine labels of a test data, the binary classifiers thus develop are run individually on the data and every label for which the output of the classifier exceeds a predetermined threshold is selected as a label of the data. To build classifiers we used Support Vector Machines (SVM for short) and our implementation is based on the LIBSVM [1].

The SVMs were trained using features extracted from the sounds. To extract features we used MARSYAS [Tzanetakis and Cook,2000]. The extracted features are divided into three different categories: timbral texture features, rhythmic content features, and pitch content features. The dimension of the final feature vector is 30.

The accuracy of the classifiers is measured using *precision*, *recall*, *break-even point* and *F1-measure*. Since the precision and the recall can be averaged over the classifiers with or without weighting, we use both *micro-averaged precision* P_{micro}, *micro-averaged recall* R_{micro}, *macro-averaged precision* P_{macro} and the *macro-averaged recall* R_{macro}. In addition, we also compute the Hamming accuracy (denoted by HA), which is defined to be the simple unweighted accuracy, that is the unweighted ratio of the total correct to the total input size. These are the performance measures that are widely used in information retrieval literature [Yang and Liu,1999].

4 Experiments

Our SVM-based multi-label classification method was tested for two problems: classification into the thirteen adjective groups and classification into the six supergroups. There was significant difference in the distribution of the positive data for some of the adjective groups (e.g., "bluesy" not appearing in the classical category). We constructed the supergroup classifier for each of the four styles. Due to the space limitation, we only include the results of all the thirteen adjective groups on all four styles. We divided the 499 sounds into training data and testing data by a random $50\% - 50\%$ split.

The accuracy measures on each of the thirteen classes are shown in Table 2. The overall accuracy for the two experiments are shown in Table 3. The breakeven point, i.e. the half-way point between the precision and the recall, was 46% in micro-avering and 43% in macro-avering. In our six-supergroup experiment the breakeven point was 50% in micro-avering and 49% in macro-averaging, so the overall accuracy was improved when the number of categories is reduced.

The overall low performance can be attributed to the fact that there were numerous borderline cases for which the labeler found it difficult to make decision. Also, the frequency of the labels was not equal across music types. We actually carried out another set of experiments, emotion detection for supergroups within each music type. We observed improvements especially, Performance stood out on supergroup 2 for classical and supergroup 4 for fusion. This may suggest that the use of genre information might improve emotion detection.

Our experiments show that emotion detection is a rather difficult problem and improvement of performance is the immediate issue. This can be resolved by: expanding the sound data sets, collecting labeling in multiple rounds to ensure confidence in labeling, using different sets of adjectives, incorporating style and genre information, and using different types of features.

[1] Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Group	TP	TN	FP	FN	P	R	HA
A	12	132	81	22	0.1290	0.3529	0.5830
B	3	189	44	11	0.0638	0.2143	0.7773
C	96	70	45	36	0.6809	0.7273	0.6721
D	53	106	64	24	0.4530	0.6883	0.6437
E	46	81	61	59	0.4299	0.4381	0.5142
F	43	102	73	29	0.3707	0.5972	0.5870
G	26	127	78	16	0.2500	0.6190	0.6194
H	28	156	40	23	0.4118	0.5490	0.7449
I	56	135	41	15	0.5773	0.7887	0.7733
J	10	178	47	12	0.1754	0.4545	0.7611
K	15	161	51	20	0.2273	0.4286	0.7126
L	13	144	70	20	0.1566	0.3939	0.6356
M	18	181	43	5	0.2951	0.7826	0.8057

Table 2: Accuracy measures on adjective group classification.

Measure	P_{micro}	R_{micro}	B_{micro}	F_{micro}
Values	0.3621	0.5893	0.4757	0.4486
Measure	P_{macro}	R_{macro}	B_{macro}	F_{macro}
Values	0.3247	0.5411	0.4329	0.4058

Table 3: Overall accuracy measures.

Acknowledgments

The authors thank Diane Cass for helping us in finding references. This work is supported in part by NSF grants EIA-0080124, DUE-9980943, and EIA-0205061, and in part by NIH grants RO1-AG18231 (5-25589) and P30-AG18254.

References

[Farnsworth,1958] Paul R. Farnsworth. *The social psychology of music*. The Dryden Press, 1958.

[Hevner,1936] Kate Hevner. Experimental studies of the elements of expression in music. *American Journal of Psychology*, 48:246–268, 1936.

[Huron,2000] D. Huron. Perceptual and cognitive applications in music information retrieval. In *International Symposium on Music Information Retrieval*,2000.

[Tzanetakis and Cook,2000] George Tzanetakis and Perry Cook. Marsyas: A framework for audio analysis. *Organized Sound*, 4(3):169–175, 2000.

[Yang and Liu,1999] Y. Yang and X. Liu. A re-examination of text categorization methods. In *SIGIR*, 1999.