Multi-label classification of emotions in music

Alicja Wieczorkowska[1], Piotr Synak[1], and Zbigniew W. Raś[2,1,3]

[1] Polish-Japanese Institute of Information Technology, Koszykowa 86, 02-008 Warsaw, Poland
[2] University of North Carolina, Charlotte, Computer Science Dept., 9201 University City Blvd., Charlotte, NC 28223, USA
[3] Polish Academy of Sciences, Institute of Computer Science, Ordona 21, 01-237 Warsaw, Poland

Abstract. This paper addresses the problem of multi-label classification of emotions in musical recordings. The testing data set contains 875 samples (30 seconds each). The samples were manually labelled into 13 classes, without limits regarding the number of labels for each sample. The experiments and test results are presented.

1 Introduction

Music is always present in our lives and it is a tool by which composers can express their feelings [1]. Even, a study of the Psalms alone yielded an impressive role for music in the life of Biblical people. Often, music is associated with important moments of our life, brings to us memories and evokes emotions. It can keep soldiers brave, athletes motivated, can be even used in medical therapy, can bring us joy, sadness, excitement or calm. The popularity of the Internet and the use of compact audio formats with near CD quality, such as MP3, have given a great contribution to a tremendous growth of digital music libraries. This poses new and exciting challenges [18]. Any music database is really useful if users can find what they are looking for. Presently, most query answering systems associated with music databases can only handle queries based on simple categories such as author, title or genre. Some efforts have been made to search music databases by content similarity [8], [21]. In such systems, user can create a musical query through examples, e.g., by humming the melody he is searching for, or by specifying a song which is similar to what he is looking for, in terms of certain criteria such as rhythm, genre, theme, and instrumentation. To overcome search limitations resulting from a manual labelling, a clustering algorithm can be run on a musical database, yielding similar songs placed in the same cluster. Similarity relation is strictly dependent on the definition of a distance measure and what type of features is used to represent songs as vectors. Such systems are called automatic classification systems [15] and they may have hierarchical structures [17]. A challenging goal is to build a music storage and retrieval system with tools for automatic indexing of musical files taking into consideration the most dominating instruments played in a musical piece and

certain group of emotions they should invoke in listeners. This paper presents a preliminary results of building classifiers for automatic indexing of music by emotions. The labelling of musical pieces by sets of emotions was done by one of the authors who is a professional musician. Clearly, the type of emotions which music can invoke in each of us is rather a subjective measure.

So, with each emotions-based labelling done by a person, his/her ontology can be built. It can be achieved automatically by asking him/her a number of fixed questions and then designing an ontology graph [6] on the basis of received answers. Now, a clustering algorithm can be run on a collection of ontology graphs and then a representative ontology graph for each cluster can be chosen. Next, a separate classifier for automatic emotions-based indexing for each such a cluster can be built. Before any emotions-related query is answered, a user ontology graph first should be built and a nearest ontology cluster identified. For a simplicity reason, this paper omits that step and assumes that user has Western cultural and musical background, which in most cases should make his/her ontology similar to the corresponding ontology of the person who made the emotions-based labelling of our testing database.

The authors have already performed experiments with recognition of emotions in music data, using singular labelling [19], [20], [12], where single class labels have been assigned to each sample. In order to observe how multi-labelling influences correctness of automatic classification, the same parameterization and classification schemes have been followed, but multiple class labels allowed during data labelling and recognition.

2 Audio Parameters

The parameterization of music data can be based on various sound features, like loudness, duration, pitch, and more advanced properties, frequency contents and their changes over time. Descriptors originating from speech processing can be also applied, although some of them are not suitable to non-speech signal. Speech feature include prosodic and quality features, such as phonation type, articulation manned etc. [13]. General audio (music) features include low-level descriptors, such as structure of the spectrum time domain and time-frequency domain features, and also higher-level features, such as rhythmic content features [7], [10], [14].

The features applied in this research allowed to characterize musical data start with the description of sound timbre. Further, the extension of the parameterization to the observation of time series of these features is planned. In this paper, we followed the parameterization scheme used in [20], [12] to check how multi-labelling may influence classification results. The audio data represent Western music, recorded stereo with 44100 Hz sampling frequency and 16-bit resolution. The analyzing frame of 32768 samples (with Hanning windowing) was applied, taken from the left channel, in order obtain precise spectral bins and describe longer time fragment. The spectral components

have been calculated up to 12 kHz and no more than 100 partials (spectral components), since it is assumed that this range covers sufficient (from the point of view of the perception of emotions) amount of the spectrum elements.

The set of the following 29 audio descriptors was calculated for the analysis window [20], [12]:

- *Frequency*: dominating fundamental frequency of the sound
- *Level*: maximal level of sound in the analyzed frame
- $Tristimulus1, 2, 3$: Tristimulus parameters calculated for *Frequency*, given by [11]:

$$Tristimulus1 = \frac{A_1^2}{\sum_{n=1}^{N} A_n^2} \tag{1}$$

$$Tristimulus2 = \frac{\sum_{n=2,3,4} A_n^2}{\sum_{n=1}^{N} A_n^2} \tag{2}$$

$$Tristimulus3 = \frac{\sum_{n=5}^{N} A_n^2}{\sum_{n=1}^{N} A_n^2} \tag{3}$$

  where A_n denotes the amplitude of the n^{th} harmonic, N is the number of harmonics available in spectrum, $M = \lfloor N/2 \rfloor$ and $L = \lfloor N/2 + 1 \rfloor$
- *EvenHarm* and *OddHarm*: Contents of even and odd harmonics in the spectrum, defined as

$$EvenHarm = \frac{\sqrt{\sum_{k=1}^{M} A_{2k}^2}}{\sqrt{\sum_{n=1}^{N} A_n^2}} \tag{4}$$

$$OddHarm = \frac{\sqrt{\sum_{k=2}^{L} A_{2k-1}^2}}{\sqrt{\sum_{n=1}^{N} A_n^2}} \tag{5}$$

- *Brightness*: brightness of sound - gravity center of the spectrum, defined as

$$Brightness = \frac{\sum_{n=1}^{N} n\, A_n}{\sum_{n=1}^{N} A_n} \tag{6}$$

- *Irregularity*: irregularity of spectrum, defined as [5]

$$Irregularity = \log\left(20 \sum_{k=2}^{N-1} \left| \log \frac{A_k}{\sqrt[3]{A_{k-1} A_k A_{k+1}}} \right| \right) \tag{7}$$

- $Frequency1, Ratio1, ..., 9$: for these parameters, 10 most prominent peaks in the spectrum are found. The lowest frequency within this set is chosen as $Frequency1$, and proportions of other frequencies to the lowest one are denoted as $Ratio1, ..., 9$

- $Amplitude1, Ratio1, ..., 9$: the amplitude of $Frequency1$ in decibel scale, and differences in decibels between peaks corresponding to $Ratio1, ..., 9$ and $Amplitude1$. These parameters describe relative strength of the notes in the music chord.

3 Multi-Label Classification

The classes representing emotions can be labelled in various ways [4], [7], [13]. For instance, Dellaert et al. [4] classified emotions (in speech) into 4 classes: happy, sad, anger, and fear. Tato et al. [13] used 2-dimensional space of emotions, i.e. quality vs. activation, and 3 classes regarding levels of activation: high (angry, happy), medium (neutral), and low (sad, bored). Li and Ogihara [7] used 13 classes, later grouped into 6 super-classes. Authors followed Li and Ogihara's classification scheme in our previous research [19], [20], with only one class assigned to each sample. In this paper, we wanted to check how multi-labelling influences the correctness of recognition, so again the same 13 classes for labelling the data with emotions [7] have been used. They are given below:

1. frustrated,
2. bluesy, melancholy,
3. longing, pathetic,
4. cheerful, gay, happy,
5. dark, depressing,
6. delicate, graceful,
7. dramatic, emphatic,
8. dreamy, leisurely,
9. agitated, exciting, enthusiastic,
10. fanciful, light,
11. mysterious, spooky,
12. passionate,
13. sacred, spiritual.

The data have been grouped into the following 6 super-classes [7]:

1. frustrated, agitated, exciting, enthusiastic, dramatic, emphatic,
2. bluesy, melancholy, dark, depressing,
3. longing, pathetic, passionate,
4. cheerful, gay, happy, fanciful, light,
5. delicate, graceful, dreamy, leisurely,
6. mysterious, spooky, sacred, spiritual.

In previous experiments, each sample was labelled by a single class label (i.e. class number). In the presented research, the number of allowed labels per sample is not limited to one (as in Li and Ogihara's experiments). Number

Table 1. Number of objects representing emotion classes in 875-element database of 30-second audio samples

Class	No. of objects	Class	No. of objects
Agitated, exciting, enthusiastic	304	Fanciful, light	317
Bluesy, melancholy	214	Frustrated	62
Cheerful, gay, happy	62	Longing, pathetic	147
Dark, depressing	41	Mysterious, spooky	100
Delicate, graceful	226	Passionate	106
Dramatic, emphatic	128	Sacred, spiritual	23
Dreamy, leisurely	151		

of objects representing each class is given in Table 1. As one can see, some classes are represented by more objects than the others.

Multi-label classification is often performed in text mining and scene classification where documents or images can be labelled with several labels describing their contents [2,9,3]. Such a classification poses additional problems, including the selection of the training model, and set-up of testing and evaluation of results.

The use of training examples with multiple labels can follow a few scenarios:

- *MODEL-s* - the simplest model, assuming labelling of data by using **s**ingle, the most likely label,
- *MODEL-i* - **i**gnoring all the cases with more than one label, but in such a model there are no data for training,
- *MODEL-n* - **n**ew classes are created for each combination of labels occurring in the training sample, but in this model the number of classes easily becomes very large, especially, if the number of labels is only limited by the number of available labels, and for such a huge number of classes the data may easily become very sparse, so some classes may have very few training samples,
- *MODEL-x* - the most efficient model, in our opinion; in this case **cross**-training is performed, where samples with many labels are used as positive examples (and not as negative examples) for each class corresponding to the labels.

In our experiments, we decided to follow the *MODEL-x*. The testing and evaluation set-up is described in the next section.

4 Experiments and Results

The experiments on automatic recognition of emotions in music data have been performed on a database of 875 audio samples 30 seconds each. The

database, created by Dr. Rory A. Lewis from the University of North Carolina at Charlotte, contains excerpts from songs and classic music pieces. The data were originally recorded in .mp3 format, and later converted to .au format for parametrization purposes. Each audio sample in the audio database is represented by a single vector of parameters, according to formulas described in Section 2.

For the classification purpose, a modified version of k-NN (k nearest neighbors) algorithm was applied. The modification aimed at taking multiple labels into account. Therefore, the modified k nearest neighbors algorithm returns corresponding sets of labels for each neighbor of a tested sample. Then, the histogram presenting the number of appearances of each class label in the neighborhood is calculated and next normalized by the number of all labels (including repetitions) returned by the algorithm for a given testing sample. Therefore, a number $p \in [0, 1]$ is assigned to each label from the k-neighborhood (see Figure 1). This p can be considered as a probability measure. The output of the algorithm is a set of labels with assigned probability p, and only the labels exceeding some threshold level (chosen experimentally) are taken into account.

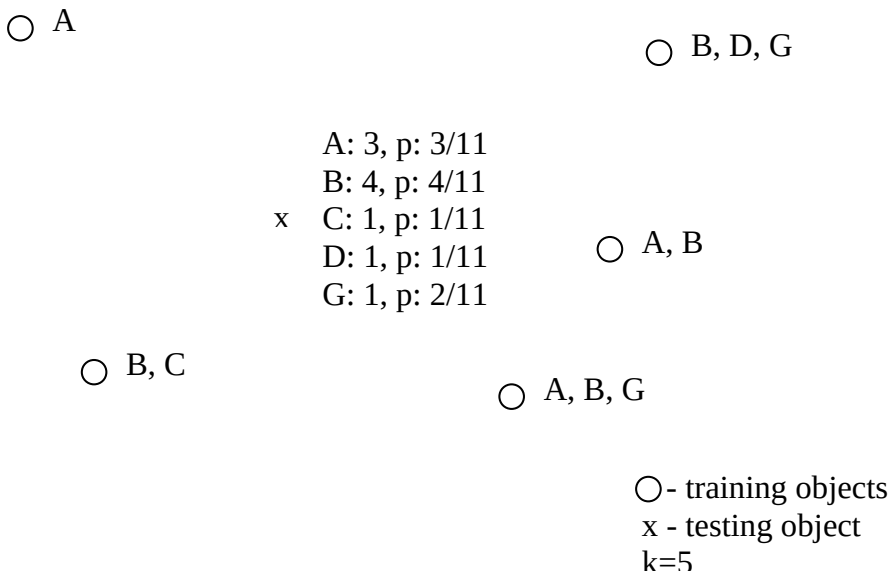

Fig. 1. Assignment of the measure p for the implemented version of the multi-class k-NN

In order to describe the quality of such a classification, both omitting the correct labels and false identification of incorrect labels must be taken into account. Therefore, a measure $m \in [0, 1]$ was assigned to each tested

sample, instead of binary measure (correct-incorrect) used in regular, single-label classification. The following measure was used:

$$m = \frac{(\sum_{i=1}^{I} p(c_i))(1 - \sum_{j=1}^{J} p(f_j))}{n} \tag{8}$$

where c_i - correctly identified label, f_j - falsely identified label, and n - number of labels originally assigned to the sample.

The averaged measure m for all test objects represents the total accuracy obtained for the entire test set.

Standard CV-5 procedure was applied for testing, i.e. 80% of data was used as the training data, the remaining 20% was used as the test set, and then this procedure was repeated 5 times and the results were averaged. The results of classification for 13 classed was 27.1%, obtained for $k = 13$, i.e. it exceeded the results of previous experiments (the previous result was 20%). For 6 super-classes, the obtained accuracy was 38.62%, obtained for $k = 15$. These results seem to be low, but one must remember that these result are comparable to the results obtained in subjective tests for human listeners, and the emotions are always subjective and vague.

5 Conclusions

This research is a continuation of our previous work, where authors performed a trial of automatic identification of emotions in music audio data for the same set of samples, but with a single class label assigned to each sample. However, for many reasons, even for a human listener, it is sometimes quite difficult to classify a given music sample. The experts asked for cross-test recognition of emotions reported the need for the use of multiple labels, and the recognition accuracy was law (in a case of both computer and human recognition), so the authors decided to continue the experiments with multi-labelling. The available list of labels represents 13 emotional classes, which is quite a big number to choose from. First of all, a long list of emotions is inconvenient for a quick browsing and identifying the appropriate one(s). Additionally, although emotions associated with a given sample may remain stable, often they can also change even for the same listener. Also, some of the emotions are very similar, unnecessarily making their list longer. On the other hand, some emotions (not used in our classification) can also be perceived during listening. For all the above reasons, it is rather recommended to use a small number of basic emotion which can be graded on a scale either continuous (in a graphical form, for instance in 2D or 3D space) or discrete (with a few possible values). The authors are considering limiting the number of classes to a few and use 2 or 3-dimensional space in which the level of perceived emotion can be graded on a discrete scale (or for convenience in a continuous scale which later has to be discretized). The extension of the list of attributes is also planned. First of all, the changes in time should be observed, i.e. how

the audio data evolve, and also rhythm and tempo should be included in a feature vector, since they both influence perception of emotions in music.

6 Acknowledgements

This research was supported by the grant 3 T11C 002 26 from Ministry of Scientific Research and Information Technology of the Republic of Poland, by the National Science Foundation under grant IIS-0414815 and by the Research Center at the Polish-Japanese Institute of Information Technology, Warsaw, Poland.

The authors would like to express numerous thanks to Doctor Rory A. Lewis from the University of North Carolina at Charlotte for elaborating the initial audio database for research purposes, and to Dorota Weremko and Jarosław Kąkolewski for technical help.

References

1. Bernstein., L. (1959) The Joy of Music, New York, Simon and Schuster.
2. Boutell, M., Shen, X., Luo, J., Brown, C. (2003) Multi-label Semantic Scene Classification. Technical Report, Dept. of Computer Science, U. Rochester
3. Clare, A., King, R.D. (2001) Knowledge Discovery in Multi-label Phenotype Data. Lecture Notes in Computer Science **2168** 42–53.
4. Dellaert, F., Polzin, T., Waibel, A. (1996) Recognizing Emotion in Speech. Proc. ICSLP 96 **3** 1970–1973.
5. Fujinaga, I., McMillan, K. (2000) Realtime recognition of orchestral instruments. Proceedings of the International Computer Music Conference, 141–143.
6. Guarino, N. (Ed.) (1998) Formal Ontology in Information Systems, IOS Press, Amsterdam.
7. Li, T., Ogihara, M. (2003) Detecting emotion in music. 4th International Conference on Music Information Retrieval ISMIR, Washington, D.C., and Baltimore, MD. Available at http://ismir2003.ismir.net/papers/Li.PDF
8. Logan, B. and Salomon, A. (2001) A Music Similarity Function Based on Signal Analysis, IEEE International Conference on Multimedia and EXPO (ICME 2001).
9. McCallum, A. (1999) Multi-label Text Classification with a Mixture Model Trained by EM. AAAI'99 Workshop on Text Learning.
10. Peeters, G. Rodet, X. (2002) Automatically selecting signal descriptors for Sound Classification. ICMC 2002 Goteborg, Sweden
11. Pollard, H. F., Jansson, E. V. (1982) A Tristimulus Method for the Specification of Musical Timbre. Acustica **51** 162–171
12. Synak, P. and Wieczorkowska, A. (2005). Some Issues on Detecting Emotions in Music, in: D. Slezak, J. Yao, J. F. Peters, W. Ziarko, X. Hu (Eds.), Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. 10th International Conference, RSFDGrC 2005, Regina, Canada, August/September 2005, Proceedings, Part II. LNAI 3642, Springer, 314-322

13. Tato, R., Santos, R., Kompe, R., Pardo, J. M. (2002) Emotional Space Improves Emotion Recognition. 7th International Conference on Spoken Language Processing ICSLP 2002, Denver, Colorado
14. Tzanetakis, G., Cook, P. (2000) Marsyas: A framework for audio analysis. Organized Sound **4(3)** 169-175. Available at http://www-2.cs.cmu.edu/ gtzan/work/pubs/organised00gtzan.pdf
15. Tzanetakis, G., Essl, G. and Cook, P. (2001) Automatic Musical Genre Classification of Audio Signals, 2nd International Conference on Music Information Retrieval (ISMIR 2001)
16. Wieczorkowska, A. A. (2005) Towards Extracting Emotions from Music, in: L. Bolc, Z. Michalewicz, T. Nishida (Eds), Intelligent Media Technology for Communicative Intelligence, Second International Workshop, IMTCI 2004, Warsaw, Poland, September 2004, Revised Selected Papers. LNAI 3490, Springer, 228-238.
17. Wieczorkowska, A. A., Ras, Z.W., Tsay, L.-S. (2003) Representing audio data by FS-trees and Adaptable TV-trees, in **Foundations of Intelligent Systems**, Proceedings of ISMIS Symposium, Maebashi City, Japan, LNAI, Springer-Verlag, No. 2871, 2003, 135-142
18. Wieczorkowska, A. A., Ras, Z.W. (Eds.) (2003) Music Information Retrieval, Special Issue, Journal of Intelligent Information Systems, Kluwer, Vol. 21, No. 1, 2003
19. Wieczorkowska, A., Synak, P., Lewis, R., Ras, Z. W. (2005) Extracting Emotions from Music Data, in: M.-S. Hacid, N. V. Murray, Z. W. Ras, S. Tsumoto (Eds.), Foundations of Intelligent Systems. 15th International Symposium, ISMIS 2005, Saratoga Springs, NY, USA, May 25-28, 2005, Proceedings. LNAI 3488, Springer, 456-465.
20. Wieczorkowska, A., Synak, P., Lewis, R., Ras, Z. W. (2005) Creating Reliable Database for Experiments on Extracting Emotions from Music, in: M. A. Klopotek, S. Wierzchon, K. Trojanowski (Eds.), Intelligent Information Processing and Web Mining. Proceedings of the International IIS: IIPWM'05 Conference held in Gdansk, Poland, June 13-16, 2005. Advances in Soft Computing, Springer, 395-402.
21. Yang, C. (2001) Music Database Retrieval Based on Spectral Similarity, 2nd International Conference on Music Information Retrieval (ISMIR 2001), Poster.