MULTI-LABEL CLASSIFICATION OF MUSIC INTO EMOTIONS

Konstantinos Trohidis
Dept. of Journalism & Mass Communication
Aristotle University of Thessaloniki
trohidis2000@yahoo.com

Grigorios Tsoumakas
Dept. of Informatics
Aristotle University of Thessaloniki
greg@csd.auth.gr

George Kalliris
Dept. of Journalism & Mass Communication
Aristotle University of Thessaloniki
gkal@auth.gr

Ioannis Vlahavas
Dept. of Informatics
Aristotle University of Thessaloniki
vlahavas@csd.auth.gr

ABSTRACT

In this paper, the automated detection of emotion in music is modeled as a multilabel classification task, where a piece of music may belong to more than one class. Four algorithms are evaluated and compared in this task. Furthermore, the predictive power of several audio features is evaluated using a new multilabel feature selection method. Experiments are conducted on a set of 593 songs with 6 clusters of music emotions based on the Tellegen-Watson-Clark model. Results provide interesting insights into the quality of the discussed algorithms and features.

1 INTRODUCTION

Humans, by nature, are emotionally affected by music. Who can argue against the famous quote of the German philosopher Friedrich Nietzsche, who said that "*without music, life would be a mistake*". As music databases grow in size and number, the retrieval of music by emotion is becoming an important task for various applications, such as song selection in mobile devices [13], music recommendation systems [1], TV and radio programs [1] and music therapy.

Past approaches towards automated detection of emotions in music modeled the learning problem as a single-label classification [9, 20], regression [19], or multilabel classification [6, 7, 17] task. Music may evoke more than one different emotion at the same time. We would like to be able to retrieve a piece of music based on any of the associated (classes of) emotions. Single-label classification and regression cannot model this multiplicity. Therefore, the focus of this paper is on multilabel classification methods.

A secondary contribution of this paper is a new multilabel dataset with 72 music features for 593 songs categorized into one or more out of 6 classes of emotions. The dataset is released to the public [2], in order to allow comparative experiments by other researchers. Publicly available multilabel datasets are rare, hindering the progress of research in this area.

[1] http://www.musicovery.com/

[2] http://mlkd.csd.auth.gr/multilabel.html

The primary contribution of this paper is twofold:

- A comparative experimental evaluation of four multilabel classification algorithms on the aforementioned dataset using a variety of evaluation measures. Previous work experimented with just a single algorithm. We attempt to raise the awareness of the MIR community on some of the recent developments in multilabel classification and show which of those algorithms perform better for musical data.
- A new multilabel feature selection method. The proposed method is experimentally compared against two other methods of the literature. The results show that it can improve the performance of a multilabel classification algorithm that doesn't take feature importance into account.

The remaining of this paper is structured as follows. Sections 2 and 3 provide background material on multilabel classification and emotion modeling respectively. Section 4 presents the details of the dataset used in this paper. Section 5 presents experimental results comparing the four multilabel classification algorithms and Section 6 discusses the new multilabel feature selection method. Section 7 presents related work and finally, conclusions and future work are drawn in Section 8.

2 MULTILABEL CLASSIFICATION

Traditional *single-label* classification is concerned with learning from a set of examples that are associated with a single label λ from a set of disjoint labels L, $|L| > 1$. In *multilabel* classification, the examples are associated with a set of labels $Y \subseteq L$.

2.1 Learning Algorithms

Multilabel classification methods can be categorized into two different groups [14]: i) *problem transformation* methods, and ii) *algorithm adaptation* methods. The first group

contains methods that are algorithm independent. They transform the multilabel classification task into one or more single-label classification, regression or ranking tasks. The second group contains methods that extend specific learning algorithms in order to handle multilabel data directly.

2.2 Evaluation Measures

Multilabel classification requires different evaluation measures than traditional single-label classification. A taxonomy of multilabel classification evaluation measures is given in [15], which considers two main categories: *example-based* and *label-based measures*. A third category of measures, which is not directly related to multilabel classification, but is often used in the literature, is ranking-based measures, which are nicely presented in [21] among other publications.

3 MUSIC AND EMOTION

Hevner [4] was the first to study the relation between music and emotion. She discovered 8 clusters of adjective sets describing music emotion and created an emotion cycle of these categories. Hevner's adjectives were refined and regrouped into ten groups by Farnsworth [2].

Figure 1 shows another emotion model, called Thayer's model of mood [12], which consists of 2 axes. The horizontal axis described the amount of stress and the vertical axis the amount of energy.

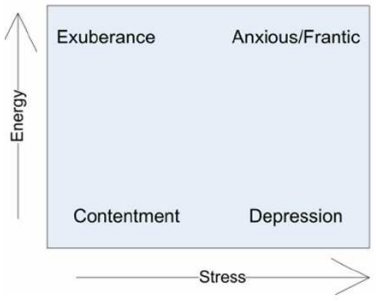

Figure 1. Thayer's model of mood

The model depicted in Figure 2 extends Thayer's model with a second system of axes, which is rotated by 45 degrees compared to the original axes [11]. The new axes describe (un)pleasantness versus (dis)engagement.

4 DATASET

The dataset used for this work consists of 100 songs from each of the following 7 different genres: Classical, Reggae, Rock, Pop, Hip-Hop, Techno and Jazz. The collection was created from 233 musical albums choosing three songs from each album. From each song a period of 30 seconds after the initial 30 seconds was extracted. The resulting sound clips were stored and converted into wave files of 22050 Hz sampling rate, 16-bit per sample and mono. The following subsections present the features that were extracted from each wave file and the emotion labeling process.

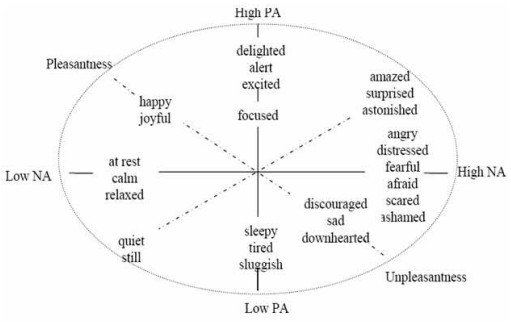

Figure 2. The Tellegen-Watson-Clark model of mood (figure reproduced from [18])

4.1 Feature Extraction

For the feature extraction process, the Marsyas tool [16] was used. The extracted features fall into two categories: rhythmic and timbre.

4.1.1 Rhythmic Features

The rhythmic features were derived by extracting periodic changes from a beat histogram. An algorithm that identifies peaks using autocorrelation was implemented. We selected the two highest peaks and computed their amplitudes, their BMPs (beats per minute) and the high-to-low ratio of their BPMs. In addition, 3 features were calculated by summing the histogram bins between 40-90, 90-140 and 140-250 BPMs respectively. The whole process led to a total of 8 rhythmic features.

4.1.2 Timbre Features

Mel Frequency Cepstral Coefficients (MFCCs) are used for speech recognition and music modeling [8]. To derive MFCCs features, the signal was divided into frames and the amplitude spectrum was calculated for each frame. Next, its logarithm was taken and converted to Mel scale. Finally, the discrete cosine transform was implemented. We selected the first 13 MFCCs.

Another set of 3 features that relate to timbre textures were extracted from the Short-Term Fourier Transform (FFT): Spectral centroid, spectral rolloff and spectral flux.

For each of the 16 aforementioned features (13 MFCCs, 3 FFT) we calculated the mean, standard deviation (std),

mean standard deviation (mean std) and standard deviation of standard deviation (std std) over all frames. This led to a total of 64 timbre features.

4.2 Emotion Labeling

The Tellegen-Watson-Clark model was employed for labeling the data with emotions. We decided to use this particular model because the emotional space of music is abstract with many emotions and a music application based on mood should combine a series of moods and emotions. To achieve this goal without using an excessive number of labels, we reached a compromise retaining only 6 main emotional clusters from this model. The corresponding labels are presented in Table 1.

Label	Description	# Examples
L1	amazed-surprised	173
L2	happy-pleased	166
L3	relaxing-calm	264
L4	quiet-still	148
L5	sad-lonely	168
L6	angry-fearful	189

Table 1. Description of emotion clusters

The sound clips were annotated by three male experts of age 20, 25 and 30 from the School of Music Studies in our University. Only the songs with completely identical labeling from all experts were kept for subsequent experimentation. This process led to a final annotated dataset of 593 songs. Potential reasons for this unexpectedly high agreement of the experts are the short track length and their common background. The last column of Table 1 shows the number of examples annotated with each label.

5 EMPIRICAL COMPARISON OF ALGORITHMS

5.1 Multilabel Classification Algorithms

We compared the following multilabel classification algorithms: binary relevance (BR), label powerset (LP), random k-labelsets (RAKEL) [15] and multilabel k-nearest neighbor (MLkNN) [21]. The first three are problem transformation methods, while the last one is an algorithm adaptation method. The first two approaches were selected as they are the most basic approaches for multilabel classification tasks. BR considers the prediction of each label as an independent binary classification task, while LP considers the multi-class problem of predicting each member of the powerset of L that exists in the training set (see [15] for a more extensive presentation of BR and LP). RAKEL was selected, as a recent method that has been shown to be more effective than the first two [15]. Finally, MLkNN was selected, as a recent high-performance representative of problem adaptation methods [21]. Apart from BR, none of the other algorithms have been evaluated on music data in the past, to the best of our knowledge.

5.2 Experimental Setup

LP, BR and RAKEL were run using a support vector machine (SVM) as the base classifier. The SVM was trained with a linear kernel and the complexity constant C equal to 1. The one-against-one strategy is used for dealing with multi-class tasks in the case of LP and RAKEL. The number of neighbors in MLkNN was set to 10.

RAKEL has three parameters that need to be selected prior to training the algorithm: a) the subset size, b) the number of models and c) the threshold for the final output. We used an internal 5-fold cross-validation on the training set, in order to automatically select these parameters. The subset size was varied from 2 to 5, the number of models from 1 to 100 and the threshold from 0.1 to 0.9 with a 0.1 step.

10 different 10-fold cross-validation experiments were run for evaluation. The results that follow are averages over these 100 runs of the different algorithms.

5.3 Results

Table 2 shows the predictive performance of the 4 competing multilabel classification algorithms using a variety of measures. We notice that RAKEL dominates the other algorithms in almost all measures.

	BR	LP	RAKEL	MLkNN
Hamming Loss	0.1943	0.1964	**0.1845**	0.2616
Micro F1	0.6526	0.6921	**0.7002**	0.4741
Micro AUC	0.7465	0.7781	**0.8237**	0.7540
Macro F1	0.6002	**0.6782**	0.6766	0.3716
Macro AUC	0.7344	0.7717	**0.8115**	0.7185
One-error	0.3038	0.2957	**0.2669**	0.3894
Coverage	2.4378	2.226	**1.9974**	2.2715
Ranking Loss	0.4517	0.3638	0.2635	**0.2603**
Avg. Precision	0.7378	0.7669	**0.7954**	0.7104

Table 2. Performance results

Table 3 shows the cpu time in seconds that was consumed during the training, parameter selection and testing phases of the algorithms. We notice that BR and MLkNN require very little training time, as their complexity is linear with respect to the number of labels. The complexity of LP depends on the number of distinct label subsets that exist in training set, which is typically larger than the number of labels. While the training complexity of RAKEL is bound by the subset size parameter, its increased time comes from

the multiple models that it builds, since it is an ensemble method. RAKEL further requires a comparatively significant amount of time for parameter selection. However, this time is still affordable (2.5 minutes), as it is only run offline.

	BR	LP	RAKEL	MLkNN
Training	0.77	3.07	6.66	0.51
Parameter selection	0	0	151.59	0
Testing	0.00	0.02	0.03	0.06

Table 3. Average training, parameter selection and testing cpu time in seconds

Concerning the test time, we notice that BR is the fastest algorithm, followed by LP and RAKEL. MLkNN is the most time-consuming algorithm during testing, as it must calculate the k nearest neighbors online after the query.

Table 4 shows the classification accuracy of the algorithms for each label (as if they were independently predicted), along with the average accuracy in the last column. We notice that based on the ease of predictions we can rank the labels in the following descending order L4, L6, L5, L1, L3, L2. L4 is the easiest with a mean accuracy of approximately 87%, followed by L6, L5 and L1 with mean accuracies of approximately 80%, 79% and 78% respectively. The hardest labels are L2 and L3 with a mean accuracy of approximately 73% and 76% respectively.

	BR	LP	RAKEL	MLkNN	Avg
L1	0.7900	0.7906	0.7982	0.7446	0.7809
L2	0.7115	0.7380	0.7587	0.7195	0.7319
L3	0.7720	0.7705	0.7854	0.7221	0.7625
L4	0.8997	0.8992	0.9031	0.7969	0.8747
L5	0.8287	0.8093	0.8236	0.7051	0.7917
L6	0.8322	0.8142	0.8238	0.7422	0.8031

Table 4. Accuracy per label

Based on the results, one can see that the classification model performs better for emotional labels such as label 4 (quiet) rather than label 2 (happy). This can be interpreted due to the fact that emotions such as quietness can be easily perceived and classified by humans in a musical context, as it is more objective than more difficult and abstract emotional labels such as happiness, which is more subjective.

6 EMPIRICAL FEATURE EVALUATION

Multilabel feature selection has been mainly applied to the domain of text categorization, due to the typically high dimensionality of textual data. The standard approach is to consider each label separately, use an attribute evaluation statistic (such as χ^2, gain ratio, etc) for each label, and then combine the results using an averaging approach. Two averaging approaches that appear in the literature are *max*, which considers the maximum score for each feature across all labels and *avg*, which considers the average of the score of each feature across all labels, weighted by the prior probability of each label. The disadvantage of these approaches, similarly to BR, is that they do not consider label correlations.

We propose a different approach in this paper, which has not been discussed in the literature before, to the best of our knowledge. At a first step, we apply the transformation of the LP method in order to produce a single-label classification dataset. Then, we apply a common attribute evaluation statistic. We argue that this approach could be more beneficial than the others, as it considers label correlations.

In order to evaluate our hypothesis we compare the Hamming loss of the MLkNN algorithm (known to suffer from the curse of dimensionality) using the best 1 to 71 features according to the three feature selection approaches using the χ^2 statistic. Figure 3 shows the results.

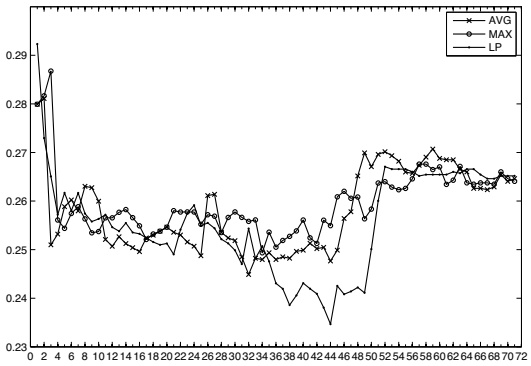

Figure 3. Hamming loss of the MLkNN classifier using the best 1 to 71 features as ordered by the χ^2 feature selection method, using the *max* and *avg* averaging approaches and the proposed method.

We notice that all approaches have similar performance for most of the horizontal axis (number of features retained), apart from the section from 36 to 50 features, where the proposed method leads to better results. It is in this area that the best result is achieved for MLkNN, which is a Hamming loss of approximately 0.235. This is an indication that taking label correlations may be fruitful, especially for selecting important features beyond those that are correlated directly with the labels.

7 RELATED WORK

We discuss past efforts on emotion detection in music, mainly in terms of emotion model, extracted features and the kind

of modeling of the learning problem: a) single label classification, b) regression, and c) multilabel classification.

7.1 Single-label Classification

The four main emotion classes of Thayer's model were used as the emotion model in [9]. Three different feature sets were adopted for music representation, namely intensity, timbre and rhythm. Gaussian mixture models were used to model each of the four classes. An interesting contribution of this work, was a hierarchical classification process, which first classifies a song into high/low energy (vertical axis of Thayer's model), and then into one of the two high/low stress classes.

The same emotion classes were used in [20]. The authors experimented with two fuzzy classifiers, using the 15 features proposed in [10]. They also experimented with a feature selection method, which improved the overall accuracy (around 78%), but they do not mention which features were selected.

The classification of songs into a single cluster of emotions was a new category in the 2007 MIREX (Music Information Retrieval Evaluation eXchange) competition. The top two submissions of the competition [3] were based on support vector machines. The model of mood that was used in the competition, was 5 clusters of moods proposed in [5], which was compiled based on a statistical analysis of the relationship of mood with genre, artist and usage metadata. Among the many interesting conclusion of the competition, was the difficulty to discern between certain clusters of moods, due to their semantic overlap. A multilabel classification approach could overcome this problem, by allowing the specification of multiple finer-grain emotion classes.

7.2 Regression

Emotion recognition is modeled as a regression task in [19]. Volunteers rated a training collection of songs in terms of arousal and valence in an ordinal scale of 11 values from -1 to 1 with a 0.2 step. The authors then trained regression models using a variety of algorithms (again SVMs perform best) and a variety of extracted features. Finally, a user could retrieve a song by selecting a point in the two-dimensional arousal and valence mood plane of Thayer.

Futhermore, the authors used a feature selection algorithm, leading to an increase of the predictive performance. However, it is not clear if the authors run the feature selection process on all input data or on each fold of the 10-fold cross-validation used to evaluate the regressors. If the former is true, then their results may be optimistic, as the feature selection algorithm had access to the test data. A similar pitfall of feature selection in music classification is discussed in [3].

[3] http://www.music-ir.org/mirex/2007

7.3 Multilabel Classification

Both regression and single-label classification methods suffer from the same problem: No two different (clusters of) emotions can be simultaneously predicted. Multilabel classification allows for a natural modeling of this issue.

Li and Ogihara [6] used two emotion models: a) the 10 adjective clusters of Farnsworth (extended with 3 clusters of adjectives proposed by the labeler) and b) a further clustering of those into 6 super-clusters. They only experimented with the BR multilabel classification method using SVMs as the underlying base single-label classifier. In terms of features, they used Marsyas [16] to extract 30 features related to the timbral texture, rhythm and pitch. The predictive performance was low for the clusters and better for the super-clusters. In addition, they found evidence that genre is correlated with emotions.

In an extension of their work, Li and Ogihara [7] considered 3 bipolar adjective pairs (Cheerful vs Depressing), (Relaxing vs Exciting), and (Comforting vs Disturbing). Each track was initially labeled using a scale ranging from -4 to +4 by two subjects and then converted to a binary (positive/negative) label. The learning approach was the same with [6]. The feature set was expanded with a new extraction method, called Daubechies Wavelet Coefficient Histograms. The authors report an accuracy of around 60%.

The same 13 clusters as in [6] were used in [17], where the authors modified the k Nearest Neighbors algorithm in order to handle multilabel data directly. They found that the predictive performance was low, too.

Compared to our work, none of the three aforementioned approaches discusses feature selection from multilabel data, compares different multilabel classification algorithms or uses a variety of multilabel evaluation measures in its empirical study.

8 CONCLUSIONS AND FUTURE WORK

The task of multi-label mapping of music into emotions was investigated. An evaluation of four multi-label classification algorithms was performed on a collection of 593 songs. Among these algorithms, RAKEL was the most effective and is proposed for emotion categorization. The overall predictive performance was high and encourages further investigation of multilabel methods. The performance per each different label varied. The subjectivity of the label may be influencing the performance of its prediction.

In addition, a new multilabel feature ranking method was proposed, which seems to perform better than existing methods in this domain. Feature ranking may assist researchers working on feature extraction by providing feedback on the predictive performance of current and newly designed individual features. It also improves the performance of multilabel classification algorithms, such as MLkNN, that don't

take feature importance into account.

Multilabel classifiers such as RAKEL could be used for the automated annotation of large musical collections with multiple emotions. This in turn would support the implementation of music information retrieval systems that query music collections by emotion. Such a querying capability would be useful for song selection in various applications.

Future work will explore the effectiveness of new features based on time frequency representation of music and lyrics, as well as the hierarchical multilabel classification approach, which we believe has great potential in this domain.

9 REFERENCES

[1] Rui Cai, Chao Zhang, Chong Wang, Lei Zhang, and Wei-Ying Ma. Musicsense: contextual music recommendation using emotional allocation modeling. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 553–556, 2007.

[2] P Farnsworth. *The social psychology of music*. The Dryden Press, 1958.

[3] R. Fiebrink and I. Fujinaga. Feature selection pitfalls and music classification. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR 2006)*, pages 340–341, 2006.

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

[5] X Hu and J.S. Downie. Exploring mood metadata: relationships with genre, artist and usage metadata. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, pages 67–72, 2007.

[6] T. Li and M. Ogihara. Detecting emotion in music. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 239–240, Washington D.C., USA, 2003.

[7] T. Li and M. Ogihara. Toward intelligent music information retrieval. *IEEE Transactions on Multimedia*, 8(3):564–574, 2006.

[8] B. Logan. Mel frequency cepstral coefficients for music modeling. In *Proceedings of the 1st International Symposium on Music Information Retrieval (ISMIR 2000)*, Plymouth, Massachusetts, 2000.

[9] 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*, 14(1):5–18, January 2006.

[10] E. Schubert. *Measurement and Time Series Analysis of Emotion in Music*. PhD thesis, University of New South Wales, 1999.

[11] A. Tellegen, D. Watson, and L.A. Clark. On the dimensional and hierarchical structure of affect. *Psychological Science*, 10(4):297–303, July 1999.

[12] R.E. Thayer. *The Biopsychology of Mood and Arousal*. Oxford University Press, 1989.

[13] M. Tolos, R. Tato, and T. Kemp. Mood-based navigation through large collections of musical data. In *2nd IEEE Consumer Communications and Networking Conference (CCNC 2005)*, pages 71–75, 3-6 Jan. 2005.

[14] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.

[15] G. Tsoumakas and I. Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. In *Proceedings of the 18th European Conference on Machine Learning (ECML 2007)*, pages 406–417, Warsaw, Poland, September 17-21 2007.

[16] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, July 2002.

[17] A. Wieczorkowska, P. Synak, and Z.W. Ras. Multi-label classification of emotions in music. In *Proceedings of the 2006 International Conference on Intelligent Information Processing and Web Mining (IIPWM'06)*, pages 307–315, 2006.

[18] D. Yang and W. Lee. Disambiguating music emotion using software agents. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR'04)*, Barcelona, Spain, 2004.

[19] 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 (TASLP)*, 16(2):448–457, February 2008.

[20] Y.-H. Yang, C.-C. Liu, and H.-H. Chen. Music emotion classification: A fuzzy approach. In *Proceedings of ACM Multimedia 2006 (MM'06)*, pages 81–84, Santa Barbara, CA, USA, 2006.

[21] M-L Zhang and Z-H Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.