EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Multi-label classification of music by emotion

Konstantinos Trohidis[1*], Grigorios Tsoumakas[2], George Kalliris[1] and Ioannis Vlahavas[2]

Abstract

This work studies the task of automatic emotion detection in music. Music may evoke more than one different emotion at the same time. Single-label classification and regression cannot model this multiplicity. Therefore, this work focuses on multi-label classification approaches, where a piece of music may simultaneously belong to more than one class. Seven algorithms are experimentally compared for this task. Furthermore, the predictive power of several audio features is evaluated using a new multi-label feature selection method. Experiments are conducted on a set of 593 songs with six clusters of emotions based on the Tellegen-Watson-Clark model of affect. Results show that multi-label modeling is successful and provide interesting insights into the predictive quality of the algorithms and features.

Keywords: multi-label classification, feature selection, music information retrieval

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 [1], music recommendation systems [2], TV and radio programs[a], and music therapy.

Past approaches towards automated detection of emotions in music modeled the learning problem as a single-label classification [3,4] regression [5], or multi-label classification [6-9] task. Music may evoke more than one different emotion at the same time. Single-label classification and regression cannot model this multiplicity. Therefore, the focus of this article is on multi-label classification methods. The primary aim of this article is twofold:

- The experimental evaluation of seven multi-label classification algorithms using a variety of evaluation measures. Previous work experimented with just a single algorithm. We employ some recent developments in multi-label classification and show which algorithms perform better for musical data.
- The creation of a new multi-label 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[b], in order to allow comparative experiments by other researchers. Publicly available multi-label music datasets are rare, hindering the progress of research in this area.

The remaining of this article is structured as follows. Sections 2 reviews related work and Sections 3 and 4 provide background material on multi-label classification and emotion modeling, respectively. Section 5 presents the details of the dataset used in this work. Section 6 presents experimental results comparing the seven multi-label classification algorithms. Finally, conclusions are drawn and future work is proposed in Section 7.

2. Related work

This section discusses 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) multi-label classification.

2.1. Single-label classification

The four main emotion classes of Thayer's model were used as the emotion model in [3]. 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

* Correspondence: trohidis2000@yahoo.com
[1]Department of Journalism and Mass Communication, Aristotle University of Thessaloniki, Thessaloniki, 54124, Greece
Full list of author information is available at the end of the article

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 [4]. 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[c]. The top two submissions of the competition were based on support vector machines (SVM). The model of mood that was used in the competition included five clusters of moods proposed in [11], 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 multi-label classification approach could overcome this problem, by allowing the specification of multiple finer-grain emotion classes.

2.2. Regression

Emotion recognition is modeled as a regression task in [5]. 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 (with SVMs having the best performance) 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.

Furthermore, 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 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 [12].

2.3. Multi-label classification

Both regression and single-label classification methods suffer from the same problem: two different (clusters of) emotions cannot be simultaneously predicted. Multi-label classification allows for a natural modeling of this issue.

Li and Ogihara [6] used two emotion models: (a) the ten adjective clusters of Farnsworth (extended with three clusters of adjectives proposed by the labeler) and (b) a further clustering of those into six super-clusters. They only experimented with the BR multi-label classification method using SVMs as the underlying base single-label classifier. In terms of features, they used Marsyas [13] 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 three 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 [8], where the authors modified the k Nearest Neighbors algorithm in order to handle multi-label data directly. They found that the predictive performance was low, too. Recently, Pachet and Roy [14] used stacked binary relevance (2BR) for the multi-label classification of music samples into a large number of labels (632).

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

3. Multi-label 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 *multi-label* classification, the examples are associated with a set of labels $Y \subseteq L$.

3.1. Learning algorithms

Multi-label classification algorithms can be categorized into two different groups [15]: (i) *problem transformation* methods, and (ii) *algorithm adaptation* methods. The first group includes methods that are algorithm independent. They transform the multi-label classification task into one or more single-label classification, regression, or ranking tasks. The second group includes methods that extend specific learning algorithms in order to handle multi-label data directly.

We next present the methods that are used in the experimental part of this work. For the formal description of these methods, we will use $L = \{(\lambda j: j = 1...M\}$ to denote the finite set of labels in a multi-label learning

task and $D = \{(\boldsymbol{x}_i, Y_i), i = 1...N\}$ to denote a set of multi-label training examples, where $\boldsymbol{x}_i$is the feature vector and $Yi \subseteq L$ the set of labels of the i-th example.

Binary relevance (BR) is a popular problem transformation method that learns M binary classifiers, one for each different label in L. It transforms the original dataset into M datasets $D_{\lambda_j} : j = 1 \dots M$ that contain all examples of the original dataset, labeled positively if the label set of the original example contained $\lambda_j$and negatively otherwise. For the classification of a new instance, BR outputs the union of the labels $\lambda_j$that are positively predicted by the M classifiers.

BR is criticized, because it does not take into account label correlations and may fail to accurately predict label combinations or rank labels according to relevance with a new instance. One approach that has been proposed in the past in order to deal with the aforementioned problem of BR, works generally as follows: it learns a second (or Meta) level of binary models (one for each label) that consider as input the output of all first (or base) level binary models. It will be called 2BR, as it uses the BR method twice, in two consecutive levels. 2BR follows the paradigm of stacked generalization [16], a method for the fusion of heterogeneous classifiers, widely known as stacking. One of the earliest account of 2BR is [17], where 2BR was part of the SVM-HF method, a SVM based algorithm for training the binary models of both levels. The abstraction of SVM-HF irrespectively of SVMs and its relation to stacking was pointed out in [18]. 2BR was very recently applied to the analysis of musical titles [14].

Label powerset (LP) is a simple but effective problem transformation method that works as follows: it considers each unique set of labels that exists in a multi-label training set as one of the classes of a new single-label classification task. Given a new instance, the single-label classifier of LP outputs the most probable class, which is actually a set of labels.

The computational complexity of LP with respect to M depends on the complexity of the base classifier with respect to the number of classes, which is equal to the number of distinct label sets in the training set. This number is upper bounded by min (N, 2^M) and despite that it typically is much smaller, it still poses an important complexity problem, especially for large values of N and M. Furthermore, the large number of classes, many of which are associated with very few examples, makes the learning process difficult as well.

The random k-labelsets (RAkEL) method [19] constructs an ensemble of LP classifiers. Each LP classifier is trained using a different small random subset of the set of labels. This way RAkEL manages to take label correlations into account, while avoiding LP's problems. A ranking of the labels is produced by averaging the zero-one predictions of each model per considered label. Thresholding is then used to produce a classification as well.

Ranking by pairwise comparison (RPC) [20] transforms the multi-label dataset into $\frac{M(M-1)}{2}$ binary label datasets, one for each pair of labels (λ_i, λ_j), $1 \le i \le j \le M$. Each dataset contains those examples of D that are annotated by at least one of the two corresponding labels, but not both. A binary classifier that learns to discriminate between the two labels is trained from each of these datasets. Given a new instance, all binary classifiers are invoked, and a ranking is obtained by counting the votes received by each label.

Calibrated label ranking (CLR) [21] extends RPC by introducing an additional virtual label, which acts as a natural breaking point of the ranking into relevant and irrelevant sets of labels. This way, CLR manages to perform multi-label classification.

Multi-label back-propagation (BP-MLL) [22] is an adaptation of the popular back-propagation algorithm for multi-label learning. The main modification to the algorithm is the introduction of a new error function that takes multiple labels into account.

Multi-label k-nearest neighbor (ML-kNN) [23] extends the popular k nearest neighbors (kNN) lazy learning algorithm using a Bayesian approach. It uses the maximum *a posteriori* principle in order to determine the label set of the test instance, based on prior and posterior probabilities for the frequency of each label within the k nearest neighbors.

3.2. Evaluation measures

Multi-label classification requires different evaluation measures than traditional single-label classification. A taxonomy of multi-label classification evaluation measures is given in [19], which considers two main categories: *example-based* and *label-based measures*. A third category of measures, which is not directly related to multi-label classification, but is often used in the literature, is ranking-based measures, which are nicely presented in [23].

4. Emotions and music

4.1. Emotional models

Emotions that are experienced and perceived while listening to music are somehow different than those induced in everyday life. Many studies indicate the important distinction between one's perception of the emotion(s) expressed by music and the emotion(s) induced by music. Studies of the distinctions between perception and induction of emotion have demonstrated

that both can be subjected to not only the social context of the listening experience, but also to personal motivation [24]. There are different approaches as to how emotion can be conceptualized and described. The main approaches that exist in the literature are the categorical, the dimensional, and the prototype approach [25,26].

According to the categorical approach, emotions are conceptualized as discrete unique entities. According to several discrete emotion theories, there is a certain basic number of emotion categories from which all the emotion states are derived such as happiness, sadness, anger, fear, and disgust [27-32]. Basic emotions are characterized by features having distinct functions, are found in all cultures, associated with distinct physiological patterns, experienced as unique feeling states and appear early in the development of humans [27,29-32]. In studies investigating music and emotion, the categorical model of emotions has been modified to better represent the emotions induced by music. Emotions such as disgust are often replaced with the emotion of tenderness, which is more suitable in the context of music.

While the categorical approach focuses on the distinct characteristics that distinguish the emotions from each other, in the dimensional approach, emotions are expressed on a two-dimensional system according to two axes such as valence and arousal. This type of model was first proposed by Izard [29,30] and later modified by Wundt [33].

The dimensional approach includes Russell's [34] circumplex model of affect, where all affective states arise from two independent systems. One is related to arousal (activation-deactivation) and the other is related to valence (pleasure-displeasure) and emotions can be perceived as varying degrees of arousal and valence. Thayer [35] suggested that the two dimensions of affect are presented by two-arousal dimensions, tension, and energetic arousal. The dimensional models have been criticized in the past by the lack of differentiation of neighborhood emotions in the valence and arousal dimensions such as anger and fear [36].

In our study, the Tellegen-Watson-Clark model was employed. This model (depicted in Figure 1) extends previous dimensional models emphasizing the value of a hierarchical perspective by integrating existing models of emotional expressivity.

It analyses a three-level hierarchy incorporating at the highest level a general bipolar happiness vs. unhappiness dimension, an independent positive affect versus negative affect dimension at the second order level below it, and discrete expressivity factors of joy, sadness, hostility, guilt/shame, fear emotions at the base. Similarly, a three-level hierarchical model of affect links the basic factors of affect at different levels of abstraction and integrates previous models into a single scheme. The key to this hierarchical structure is the recognition that the general bipolar factor of happiness and independent dimensions of PA and NA are better viewed as different levels of abstraction within a hierarchical model, rather than as competing models at the same level of abstraction. At the highest level of this model, the general bipolar factor of happiness accounts for the tendency for PA and NA to be moderately negatively correlated. Therefore, the hierarchical model of affect accounted for both the bipolarity of pleasantness-unpleasantness and the independence of PA and NA, effectively resolving a debate that occupied the literature for decades.

Over the years, a number of different dimensions have been proposed. Wundt [33] proposed a three-dimensional scheme with the three dimensions of pleasure-displeasure, arousal-calmness, and tension-relaxation. Schlosberg [37] proposed a three-dimensional model with three main dimensions expressing arousal, valence, and control. A similar model was proposed by

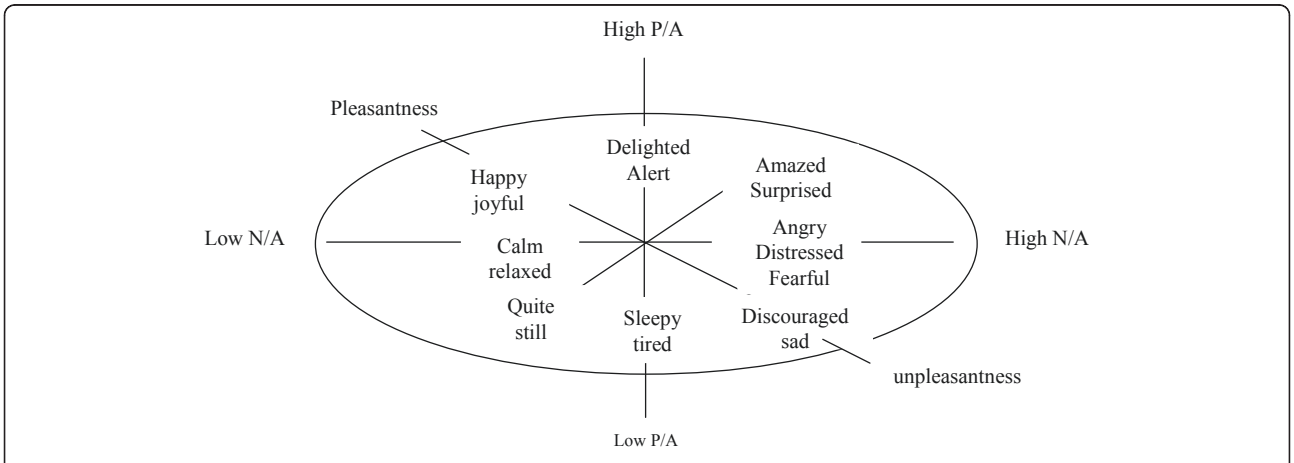

Figure 1 The Tellegen-Watson-Clark model of mood (figure reproduced from[51]).

Mehrabian [38]. He tried to define a three-dimensional model with three basic principles related to pleasure, arousal, and dominance.

Finally, the prototype approach is based on the fact that language and knowledge structures associate with how people conceptualize information [39]. The prototype approach combines effectively the categorical and dimensional approaches providing the individual contents of emotions and the hierarchical relationship among them.

5. 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 s after the initial 30 s was extracted.

The resulting sound clips were stored and converted into wave files of 22,050 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.

5.1. Feature extraction

For the feature extraction process, the Marsyas tool [13] was used, which is a free software framework. It is modular, fast, and flexible for rapid development and evaluation of computer audition applications and has been commonly used for music emotion classification and MIR tasks.

The extracted features fall into two categories: rhythmic and timbre. We select the categories of temporal and spectral features due to the highly correlation with valence and arousal dimensions of emotion. For example, songs with fast tempo are often perceived as having high arousal. Often fluent and flowing rhythm is usually associated with positive valence whereas firm rhythm is associated with negative valence. On the other hand, high arousal often correlates with bright timbre and vice versa low arousal with soft timbre.

(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 BPMs (beats per minute) and the high-to-low ratio of their BPMs. In addition, three features were calculated by summing the histogram bins between 40 and 90, 90 and 140, and 140 and 250 BPMs, respectively. The whole process led to a total of eight rhythmic features.

(2) Timbre features: Mel frequency cepstral coefficients (MFCCs) are used for speech recognition and music modeling [40]. 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 three features related to timbre textures were extracted from the short-term Fourier transform (FFT): spectral centroid, spectral rolloff, and spectral flux. This kind of features model the spectral properties of the signal such as the amplitude spectrum distribution, brightness, and the spectral change.

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 features and 8 rhythmic features.

5.2. Emotion labeling

The Tellegen-Watson-Clark model was employed for labeling the data with emotions. We decided to use this particular model because it presents a powerful way of organizing emotions in terms of their affect appraisals such as pleasant and unpleasant and psychological reactions such as arousal. It is also especially useful for capturing the continuous changes in emotional expression occurring during a piece of music.

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 six main emotional clusters from this model. The corresponding labels are presented in Table 1.

The sound clips were annotated by a panel of experts of age 20, 25, and 30 from the School of Music Studies in our University. All experts had a high musical background. During the annotation process, all experts were encouraged to mark as many emotion labels as possible induced by music. According to studies of Kivy [41], listeners make a fundamental attribution error in that they habitually take the expressive properties of music for what they feel. This argument is strongly supported by other studies [42] in which listeners are instructed to

Table 1 Description of emotion clusters

Label	Description	# of 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

describe both what they perceived and felt in response to different music genres. Meyer [43] argues that when a listener reports that he felt an emotion, he describes the emotion that a passage of music is supposed to indicate rather than what he experienced. Taking this notion into account, we instructed the subjects to label the sound clips according to what they felt rather than what the music produced.

Only the songs with completely identical labeling from at least two 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 indicates the number of examples annotated with each label. Out of the 593 songs, 178 were annotated with a single label, 315 with two labels, and 100 with three labels.

6. Empirical comparison of algorithms

6.1. Multi-label classification algorithms

We compare the following multi-label classification algorithms that were introduced in Section 2: BR, 2BR, LP, RA*k*EL, CLR, ML-*k*NN, and BP-MLL.

The first two approaches were selected as they are the most basic approaches for multi-label classification tasks. RA*k*EL and CLR were selected, as two recent methods that have been shown to be more effective than the first two. Finally, ML-*k*NN and BP-MLL were selected, as two recent high- performance representatives of problem adaptation methods. Apart from BR, none of the other algorithms have been evaluated on music data in the past, to the best of our knowledge.

6.2. Experimental setup

LP, BR, RA*k*EL, and CLR were run using a SVM as the base classifier. The SVM was trained with a linear kernel and the complexity constant C equal to 1. An SVM with the same setup was used for training both the first level and the meta-level models of 2BR. The one-against-one strategy was used for dealing with multi-class tasks in the case of LP and RA*k*EL. RA*k*EL was run with subset size equal to 3, number of models equal to twice the number of labels (12), and a threshold of 0.5 which corresponds to a default parameter setting. 10-fold cross-validation was used for creating the necessary meta-data for 2BR. The number of neighbors in ML-*k*NN was set to 10 and the smoothing factor to 1 as recommended in [23]. As recommended in [22], BP-MLL was run with 0.05 learning rate, 100 epochs and the number of hidden units equal to 20% of the input units.

Ten different 10-fold cross-validation experiments were run for evaluation. The results that follow are averages over these 100 runs of the different algorithms. Experiments were conducted with the aid of the Mulan software library for multi-label classification [44], which includes implementations of all algorithms and evaluation measures. Mulan runs on top of the Weka [45] machine learning library.

6.3. Results

Table 2 shows the predictive performance of the seven competing multi-label classification algorithms using a variety of measures. We evaluate the seven algorithms using three categories of multi-label evaluation measures, namely example-based, label-based, and ranking-based measures. Example-based measures include hamming loss, accuracy, precision, recall, F1-measure, and subset accuracy. These measures are calculated based on the average differences of the actual and the predicted sets of labels over all test examples.

Label-based measures include micro and macro precision, recall, F1-measure, and area under the ROC curve (AUC). Finally, ranking-based measures include one-error, coverage, ranking loss, and average precision. Table 2 shows the predictive performance of the seven competing multi-label classification algorithms.

6.3.1. Example-based

As far as the example-based measures are concerned, RA*k*EL has a quite competitive performance, being best in Hamming loss, second best in accuracy behind LP, best in the combination of precision and recall (F_1), and second best in subset accuracy behind LP again.

Table 2 Predictive performance of the seven different multi-label algorithms based on a variety of measures

	BR	LP	RA*k*EL	2BR	CLR	ML-*k*NN	BP-MLL
Hamming Loss	0.1943	0.1964	0.1849	0.1953	0.1930	0.2616	0.2064
Accuracy	0.5185	0.5887	0.5876	0.5293	0.5271	0.3427	0.5626
Precision	0.6677	0.6840	0.7071	0.6895	0.6649	0.5184	0.6457
Recall	0.5938	0.7065	0.6962	0.6004	0.6142	0.3802	0.7234
F_1	0.6278	0.6945	0.7009	0.6411	0.6378	0.4379	0.6814
Subset acc.	0.2759	0.3511	0.3395	0.2839	0.2830	0.1315	0.2869
Micro prec.	0.7351	0.6760	0.7081	0.7280	0.7270	0.6366	0.6541
Micro rec.	0.5890	0.7101	0.6925	0.5958	0.6103	0.3803	0.7189
Micro F_1	0.6526	0.6921	0.6993	0.6540	0.6622	0.4741	0.6840
Micro AUC	0.7465	0.8052	0.8241	0.7475	0.8529	0.7540	0.8474
Macro prec.	0.6877	0.6727	0.7059	0.6349	0.7036	0.4608	0.6535
Macro rec.	0.5707	0.7018	0.6765	0.5722	0.5933	0.3471	0.7060
Macro F_1	0.6001	0.6782	0.6768	0.5881	0.6212	0.3716	0.6681
Macro AUC	0.7343	0.8161	0.8115	0.7317	0.8374	0.7185	0.8344
One-error	0.3038	0.3389	0.2593	0.2964	0.2512	0.3894	0.2946
Coverage	2.4378	1.9300	1.9983	2.4482	1.6914	2.2715	1.7664
Ranking loss	0.2776	0.1867	0.1902	0.2770	0.1456	0.2603	0.1635
Avg. precis.	0.7378	0.7632	0.7983	0.7392	0.8167	0.7104	0.7961

Example-based measures evaluate how well an algorithm calculates a bipartition of the emotions into relevant and irrelevant, given a music title. LP models directly the combinations of labels and manages to perform well in predicting the actual set of relevant labels. RA*k*EL is based on an ensemble of LP classifiers, and as expected further improves the good performance of LP. One of the reasons for the good performance of LP is the relatively small number of labels (six emotional clusters). As mentioned in Section 2, LP has problems scaling to large numbers of labels, but RA*k*EL does not suffer from such scalability issues.

6.3.2. Label-based

As far as the micro and macro averaged measures are concerned, LP and RA*k*EL again excel in the combination of precision and recall (F_1) achieving the first two places among their competitors, while BP-MLL immediately follows as third best. The macro F_1 measure evaluates the ability of the algorithms to correctly identify the relevance of each label, by averaging the performance of individual labels, while the micro F_1 measure takes a more holistic approach by summing the distributions of all labels first and then computing a single measure. Both measures evaluate in this case the retrieval of relevant music titles by emotion.

6.3.3. Ranking-based

A first clear pattern that can be noticed is the superiority of CLR, as far as the ranking measures are concerned. Based on the pairwise comparisons of labels, it ranks effectively relevant labels higher than irrelevant labels. Therefore, if the goal of a music application was to present an ordered set of emotions for a music title, then CLR should definitely be the algorithm to employ. Such an application for example, could be one that recommends emotions to human annotators, in order to assist them in their labor intensive task. The good probability estimates that CLR obtains for the relevance of each label through the voting of all pairwise models, is also indicated by the top performance of CLR in the micro and macro averaged AUC measures, which are probability based. BP-MLL is also quite good in the ranking measures (apart from one-error) and in the micro and macro averaged AUC measures, which indicates that it also computes good estimates of the probability of relevance for each label.

6.3.4. Label prediction accuracy

Table 3 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 (quiet-still), L6 (angry-fearful), L5 (sad-lonely), L1 (amazed-surprised), L3 (relaxing-calm), and L2 (happy-pleased). L4 is the easiest with a mean accuracy of approximately 88%, followed by L6, L5, L1, and L3 with mean accuracies of approximately 81, 80, 79, and 77% respectively. The hardest label is L2 with a mean accuracy of approximately 72%.

Table 3 Accuracy of the seven multi-label classification algorithms per each label

	BR	LP	RA*k*EL	2BR	CLR	ML-*k*NN	BP-MLL	Avg.
L1	0.7900	0.7907	0.7976	0.7900	0.7954	0.7446	0.7871	0.7851
L2	0.7115	0.7380	0.7584	0.7113	0.7137	0.7195	0.7161	0.7241
L3	0.7720	0.7705	0.7804	0.7661	0.7735	0.7221	0.7712	0.7651
L4	0.8997	0.8992	0.9019	0.9002	0.8970	0.7969	0.8923	0.8839
L5	0.8287	0.8093	0.8250	0.8283	0.8295	0.7051	0.7894	0.8022
L6	0.8322	0.8142	0.8275	0.8320	0.8325	0.7422	0.8054	0.8123

Based on the results, one can see that the classification model performs better for emotional labels such as L4 (quiet-still) rather than L2 (happy-pleased). This is not at all in agreement with past research [46,47] claiming that the happy emotional tone tend to be among the easiest one to communicate in music.

An explanation for this result is that happiness is a measure of positive valence and high activity. Expressive cues describing the happiness emotion are fast tempo, small tempo variability, staccato articulation, high sound level, bright timbre, fast tone attacks, which are more difficult to model using the musical features extracted. On the other hand, quiet emotion is just a measure of energy corresponding to the activity dimension only, thus it can be more successfully described and represented by the features employed.

7. Conclusions and future work

This article investigated the task of multi-label mapping of music into emotions. An evaluation of seven multi-label classification algorithms was performed on a collection of 593 songs. Among these algorithms, CLR was the most effective in ranking the emotions according to relevance to a given song, while RA*k*EL was very competitive in providing a bipartition of the labels into relevant and irrelevant for a given song, as well as retrieving relevant songs given an emotion. The overall predictive performance was high and encourages further investigation of multi-label methods. The performance per each different label varied. The subjectivity of the label may be influencing the performance of its prediction.

Multi-label classifiers such as CLR and RA*k*EL could be used for the automated annotation of large music 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.

Interesting future work directions are the incorporation of features based on song lyrics [48,49] as well as the experimentation with hierarchical multi-label classification approaches [50], based on a hierarchical organization of emotions.

Endnotes

[a]http://www.musicovery.com/
[b]http://mulan.sourceforge.net/datasets.html
[c]http://www.music-ir.org/mirex/2007

List of abbreviations

AUC: area under the ROC curve; BP-MLL: multi-label back-propagation; BR: binary relevance; CLR: calibrated label ranking; *k*NN: *k* nearest neighbors; LP: label powerset; ML-*k*NN: multi-label *k*-nearest neighbor; RA*k*EL: random *k*-labelsets; RPC: ranking by pairwise comparison; SVM: support vector machine.

Author details

[1]Department of Journalism and Mass Communication, Aristotle University of Thessaloniki, Thessaloniki, 54124, Greece [2]Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, 54124, Greece

Competing interests

The authors declare that they have no competing interests.

Received: 17 January 2011 Accepted: 18 September 2011
Published: 18 September 2011

References

1. M Tolos, R Tato, T Kemp, Mood-based navigation through large collections of musical data, in *Proceedings of the 2nd IEEE Consumer Communications and Networking Conference (CCNC 2005)*, 71–75, (3-6 January 2005)
2. R Cai, C Zhang, C Wang, L Zhang, W-Y Ma, Musicsense: contextual music recommendation using emotional allocation modeling, in *Proceedings of the 15th International Conference on Multimedia*, 553–556 (2007)
3. L Lu, D Liu, H-J Zhang, Automatic mood detection and tracking of music audio signals. IEEE Trans Audio Speech Lang Process. **14**(1), 5–18 (2006)
4. Y-H Yang, C-C Liu, H-H Chen, Music emotion classification: a fuzzy approach, in *Proceedings of ACM Multimedia 2006 (MM'06)*, 81–84 (2006)
5. Y-H Yang, Y-C Lin, Y-F Su, H-H Chen, A regression approach to music emotion recognition. IEEE Trans Audio Speech Lang Process. **16**(2), 448–457 (2008)
6. T Li, M Ogihara, Detecting emotion in music, in *Proceedings of the International Symposium on Music Information Retrieval*, USA, 239–240
7. T Li, M Ogihara, Toward intelligent music information retrieval. IEEE Trans Multimedia **8**(3), 564–574 (2006)
8. A Wieczorkowska, P Synak, ZW Ras, Multi-label classification of emotions in music, in *Proceedings of the 2006 International Conference on Intelligent Information Processing and Web Mining (IIPWM'06)*, 307–315 (2006)
9. K Trohidis, G Tsoumakas, G Kalliris, I Vlahavas, Multilabel classification of music into emotions, in *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, USA (2008)
10. E Schubert, *Measurement and Time Series Analysis of Emotion in Music*. PhD thesis, University of New South Wales, (1999)
11. X Hu, JS Downie, Exploring mood metadata: relationships with genre, artist and usage metadata, in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, 67–72 (2007)
12. R Fiebrink, I Fujinaga, Feature selection pitfalls and music classification, in *Proceedings of the International Conference on Music Information Retrieval (ISMIR 2006)*, 340–341 (2006)
13. G Tzanetakis, P Cook, Musical genre classification of audio signals. IEEE Trans Speech Audio Process. **10**(5), 293–302 (2002). doi:10.1109/TSA.2002.800560
14. F Pachet, P Roy, Improving multilabel analysis of music titles: A large-scale validation of the correction approach. IEEE Trans Audio Speech Lang Process. **17**(2), 335–343 (2009)
15. G Tsoumakas, I Katakis, I Vlahavas, Mining Multi-Label Data, Data Mining and Knowledge Discovery Handbook, Part 6, O. Maimon L Rokach (Ed.), Springer, 2nd edition, pp. 667–685, (2010)
16. DH Wolpert, Stacked generalization. Neural Netw. **5**, 241–259 (1992). doi:10.1016/S0893-6080(05)80023-1
17. S Godbole, S Sarawagi, Discriminative methods for multi-labeled classification, in *Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2004)*, 22–30 (2004)
18. G Tsoumakas, I Katakis, Multi-label classification: an overview. Int J Data Warehousing Mining **3**(3), 1–13 (2007)
19. G Tsoumakas, I Vlahavas, Random k-labelsets: An ensemble method for multilabel classification, in *Proceedings of the 18th European Conference on Machine Learning (ECML 2007), Poland*, 406–417 (2007)
20. E Hüllermeier, J Fürnkranz, W Cheng, K Bringer, Label ranking by learning pairwise preferences. Artif Intell. **172**(16-17), 1897–1916 (2008). doi:10.1016/j.artint.2008.08.002
21. J Fürnkranz, E Hüllermeier, EL Mencia, K Brinker, Multilabel classification via calibrated label ranking. Mach Learn. **73**(2), 133–153 (2008). doi:10.1007/s10994-008-5064-8
22. M-L Zhang, Z-H Zhou, Multi-label neural networks with applications to functional genomics and text categorization. IEEE Trans Knowl Data Eng. **18**(10), 1338–1351 (2006)
23. M-L Zhang, Z-H Zhou, ML-KNN: a lazy learning approach to multi-label learning. Pattern Recog. **40**(7), 2038–2048 (2007). doi:10.1016/j.patcog.2006.12.019
24. PN Juslin, P Lukka, Expression, perception and induction of musical emotions: a review and questionnaire study of every day listening. J New Music Res **33**, 217–238 (2004). (2004). doi:10.1080/0929821042000317813
25. PN Juslin, JA Sloboda (Eds.), *Music and Emotion: Theory and Research*. (Oxford University Press, New York, 2001)
26. T Eerola, JK Vuoskoski, A comparison of the discrete and dimensional models of emotion in music. Psychol Music. **39**(1), 18–49 (2011). doi:10.1177/0305735610362821
27. P Ekman, An argument for basic emotions, Cognition Emotion. **6**, 169–200 (1992)
28. PR Farnsworth, A study of the Hevner adjective list. J Aesth Art Crit. **13**, 97–103 (1954). doi:10.2307/427021
29. CE Izard, *The Emotions* (Plenum Press, New York, 1977)
30. R Plutchik, *The Psychology and Biology of Emotion* (Harper Collins, New York, 1994)
31. J Panksepp, A critical role for affective neuroscience in resolving what is basic about basic emotions. Psychol Rev. **99**, 554–560 (1992)
32. K Oatley, Best laid schemes. *The Psychology of Emotions* (Harvard University Press, MA, 1992)
33. WM Wundt, *Outlines of Psychology* (Wilhelm Engelmann, Leipzig, 1897) (Translated by CH Judd)
34. JA Russell, A circumplex model of affect. J Soc Psychol. **39**, 1161–1178 (1980)
35. RE Thayer, *The Biopsychology of Mood and Arousal* (Oxford University Press, 1989)
36. D Tellegen, D Watson, LA Clark, On the dimensional and hierarchical structure of affect. Psychol Sci. **10**(4), 297–303 (1999). doi:10.1111/1467-9280.00157
37. H Schlosberg, Three dimensions of emotion. Psychol Rev. **61**(2), 81–88 (1954)
38. A Mehrabian, Pleasure-arousal-dominance: a general framework for describing and measuring individual. Curr Psychol. **14**(4), 261–292 (1996). doi:10.1007/BF02686918
39. R Rosch, Cognition and categorization, in *Principles of categorization*, ed. by Rosch E, Loyd BB (Hillsdale, NJ, 1978), pp. 27–48
40. 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)
41. P Kivy, *Sound Sentiment. An Essay on the Musical Emotions* (Temple University Press, Philadelphia, PA, 1989)
42. MR Zentner, S Meylan, KR Scherer, Exploring 'musical emotions' across five genres of music, in *Proceedings of 6th International Conference of society for Music Perception and Cognition (ICMPC)* (2000)
43. LB Meyer, *Emotion and Meaning in Music* (University of Chicago Press, Chicago, 1956)

44. G Tsoumakas, E Spyromitros-Xioufis, J Vilcek, I Vlahavas, Mulan: A Java Library for Multi-Label Learning. J Mach Learn Res. **12**, 2411–2414 (2001)
45. IH Witten, E Frank, Data Mining: Practical Machine Learning Tools and Techniques. (Morgan Kaufmann, 2011)
46. A Gabrielsson, PN Juslin Emotional expression in music performance: between the performers intention and the listeners experience. Psychol Music. **24**(1), 68–91 (1996). doi:10.1177/0305735696241007
47. L Krumhansl, An exploratory study of musical emotions and psychophysiology. Can J Exp Psychol. **51**, 336–353 (1997)
48. C Laurier, J Grivolla, P Herrera, Multimodal music mood classification using audio and lyrics, in *Proceedings of the International Conference on Machine Learning and Applications*, USA (2008)
49. Y-H Yang, Y-C Lin, H-T Cheng, I-B Liao, Y-C Ho, H-H Chen, Toward multi-modal music emotion classification, in *Proceedings of the 9th Pacific Rim Conference on Multimedia (PCM 2008)*, pp. 70–79 (2008)
50. C Vens, J Struyf, L Schietgat, S Džeroski, H Blockeel, Decision trees for hierarchical multi-label classification. Mach Learn. **73**(2), 185–214 (2008). doi:10.1007/s10994-008-5077-3
51. D Yang, W Lee, Disambiguating music emotion using software agents. in *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR'04)*, Barcelona, Spain, (2004)

doi:10.1186/1687-4722-2011-426793
Cite this article as: Trohidis *et al.*: **Multi-label classification of music by emotion.** *EURASIP Journal on Audio, Speech, and Music Processing* 2011 **2011**:4.