AN EMPIRICAL STUDY OF MULTI-LABEL CLASSIFIERS FOR MUSIC TAG ANNOTATION

Chris Sanden
Mathematics and Computer Science
University of Lethbridge
Lethbridge, AB Canada
sanden@cs.uleth.ca

John Z. Zhang
Mathematics and Computer Science
University of Lethbridge
Lethbridge, AB Canada
zhang@cs.uleth.ca

ABSTRACT

In this paper we study the problem of automatic music tag annotation. Treating tag annotation as a computational classification process, we attempt to explore the relationship between acoustic features and music tags. Toward this end, we conduct a series of empirical experiments to evaluate a set of multi-label classifiers and demonstrate which ones are more suitable for music tag annotation. Furthermore, we discuss various factors in the classification process, such as feature sets, frame sizes, etc. Experiments on two publicly available datasets show that the *Calibrated Label Ranking* (CLR) algorithm outperforms the other classifiers for a selection of evaluation measures.

1. INTRODUCTION

For the past decade, digital music collections have been growing enormously in volume, due to advances in technologies, such as storage capacity, network transmission, data compression, information retrieval, etc. The rapid rise in music downloading has created a major shift in the music industry away from physical media formats to electronic distributions. Large on-line music providers now offer millions of music catalogs. At present, these catalogs are commonly classified and accessed through *textual meta-data*, such as *genre*, *style*, *mood*, *artist*, etc. This classification scheme is referred to as *music tag annotation* and relies on human experts as well as amateurs to annotate the music [18].

While this meta-data is rich and descriptive, it is difficult to maintain and in many cases is not comprehensive, due to the ambiguity and subjectivity that is introduced in the annotation process [7]. Moreover, annotation by human experts is an involved process, in terms of financial and labor costs [4]. Therefore, manual annotation is insufficient and ineffective when facing large volumes of music. In *Music Information Retrieval* (*MIR*), automatic music tag annotation is an emerging area that aims to help automate the annotation process. The task of music tag annotation can be defined as follows [6]. Given a set of tags $T = \{t_1, t_2, ..., t_A\}$ and a set of music pieces $M = \{m_1, m_2, ..., m_R\}$, predict for each music piece $m_j \in M$ a tag annotation vector $A = (a_1, a_2, ..., a_A)$, where $a_i > 0$ if tag t_i has been associated with the piece, and $a_i = 0$, otherwise. These a_i describe the strength of the semantic associations between tags and the music piece and are typically referred to as *semantic weights*. Although these weights can be valuable in some applications, we focus on the binary association where a tag is either relevant to a music piece or not, i.e., its weight is mapped to {0,1} and can be interpreted as a class label. It is easy to see that a music piece can have multiple tags and therefore music tag annotation can be modeled as a multi-label classification process [6].

In our work, we study the problem of automatic music tag annotation by attempting to learn a relationship between acoustic features and music tags. We conduct a series of experiments on a set of multi-label classifiers which have shown promising results in other application domains including document classification, video annotation, functional genomics, etc. We demonstrate which classifiers are more suitable for music tag annotation using a set of evaluation measures. While some of these classifiers have been used for multi-label classification of music into emotions [13] and genres [9], we believe that it would be beneficial to explore their application in music tag annotation.

2. RELATED WORK

Automatic music tag annotation is an important problem in MIR with numerous applications, including music search, recommendation, organization, etc. It has received considerable attention as of recently and many related techniques have been proposed. One of the most important contributions to the problem is the work of Turnbull *et al.* [16], who propose, along with a dataset called *CAL500*, one of the first tag annotation systems based on a generative probabilistic model. This dataset has become a *de facto* benchmark for evaluating the performance of music tag annotation systems.

Hoffman *et al.* [5] present another probabilistic model, referred to as the *Codeword Bernoulli Average*, which attempts to predict the probability that a tag applies to a music piece. It is claimed that this model outperforms the one from Turnbull *et al.* [16] on the CAL500 dataset. In addition, Bertin-Mahieux *et al.* [2] propose *Autotagger*, a model that uses advanced ensemble learning schemes to combine the discriminative power of different classifiers. Ness *et al.* [6] describe how *stacked generalization* of the probabilistic outputs of a *Support Vector Machine* (*SVM*) can be used to improve the performance of automatic tag annotation.

More recently, Shen *et al.* [11] propose a framework called *MMTagger* that combines advanced feature extraction techniques and high-level semantic concept modeling for music tag annotation. The proposed framework uses a multilayer architecture that gathers multiple *Gaussian mixture models* and SVMs. In addition, Zhao *et al.* [21] introduce a large-scale music tag recommender using *Explicit Multiple Attributes* based on tag semantic similarity and music content. Experiment results in the work show that the proposed recommender is more effective than existing ones and is at least as effective as other SVM-based approaches.

3. MULTI-LABEL CLASSIFICATION

Different from traditional single-label classification where each object belongs to only one class, multi-label classification deals with the problem where an object may belong to one or multiple classes simultaneously, i.e., objects are associated with a set of labels $Y \subseteq L$, where L ($|L| > 1$) is a set of disjoint class labels [14].

In our work, we evaluate the following multi-label classifiers for tag annotation. *Random k-Labelsets* (*RAkEL*), *Calibrated Label Ranking* (*CLR*), *Multi-label k-Nearest Neighbor* (*MLkNN*), *Backpropagation for Multi-Label Learning* (*BPMLL*), *Hierarchy of Multi-label Classifiers* (*HOMER*), *Instance Based Logistic Regression* (*IBLR*), and an adaptation of *kNN* using Binary Relevance (*BRkNN*). Moreover, we use a *Decision Tree* (*DT*) and *Support Vector Machine* (*SVM*) as base-level learning algorithms for CLR, RAkEL, and HOMER. A total of 10 multi-label classifiers are evaluated. For the sake of space and due to the nature of our work, we will not digress into the details of these classifiers. The interested reader is referred to [3, 12, 14, 15, 19].

In order to evaluate the performance of multi-label classifiers, a variety of evaluation measures are typically employed. However, as automatic music tag classification is relatively new in MIR, the evaluation measures used vary significantly. Furthermore, different classifiers may perform better under different evaluation measures. Therefore, it is desirable that multiple and contrasting evaluation measures are used in any multi-label classification experiment. We make use of the following measures which are commonly used in the multi-label classification literature: *Hamming Loss* (*HL*), *Average Precision* (*AP*), *Coverage* (*CO*), *Ranking Loss* (*RL*), *One-Error* (*OE*), *Macro F-Measure* (F_1), *Macro Precision* (*Precision*), and *Macro Recall* (*Recall*). The interested reader is referred to [14, 20] for details on them.

4. EXPERIMENT SETUP

In our experiments, the *Mulan* [1] open source library for multi-label learning is used to train and evaluate each of the 10 classifiers using default parameters, e.g., the number of neighbors is set to 10 for $MLkNN$ and $IBLR$, a linear kernel is used to train the SVM.

4.1 Dataset Selection

Our experiments are conducted on two publicly available datasets. The *Computer Audition Lab 500* dataset (*CAL500*) [16] is a collection of 500 Western songs recorded by 500 different artists. Each song is manually annotated with a subset of 174 tags, which are distributed across 6 attributes: Mood, Genre, Instrument, Song, Usage, and Vocal. All tags are manually generated under controlled experimental conditions and are therefore believed to be of high quality. For our experiments, we use the “hard” annotations provided with the CAL500 dataset which gives a binary value for all songs and tags indicating whether a tag applies to a song.

Magnatagatune is a collection of approximately 21,000 clips of music, each annotated with a combination of 188 different tags. The annotations are collected through an online game, referred to as “TagATune”, developed to collect tags for music and sound clips. Each clip, 29 seconds in length, is an excerpt of music provided by Magnatune.com and FreeSound.org. All of the tags in the collection have been verified, i.e. a tag is associated with a clip only if it is generated independently by more than two players. Moreover, only those tags that are associated with more than 50 clips are included in the collection. As discussed by Seyerlehner *et al.* [10], Magnatagatune is rather difficult to handle due to its size and skewed tag distribution and and has not been used as widely as CAL500.

4.2 Feature Sets and Extraction

Prior to classification, the music pieces must be parameterized based on a set of features and their changes over time. However, it is widely known that there is no accepted criteria as which features are best for music classification [1].

[1] http://mulan.sourceforge.net.

Therefore, we experiment with three different feature sets, to be described below, which are commonly used for music classification. The *Marsyas*[2] audio processing framework is used for the computation of the features.

The *Spectral* feature set, denoted FS_s, consists of spectral features, including Spectral Flatness Measure, Spectral Centroid, Spectral Crest Factor, Spectral Rolloff and Spectral Flux.

The *Timbral* feature set, denoted FS_t, consists of a combination of spectral, temporal and cepstral features. The following features are included: Zero Crossing Rate, Spectral Centroid, Spectral Rolloff, Spectral Flux, MFCC, Chroma.

The *Beat* feature set, denoted FS_b, extends FS_t by including rhythmic features that are derived by extracting periodic changes from a beat histogram.

Following a general practice in MIR [8], we model the audio signal as the statistical distribution of audio features computed on individual, short segments. This process yields a large number of feature vectors. Therefore, the feature vectors are then aggregated together using statistical methods. Although more elaborate representations have been proposed in the literature, the simplicity of using a single vector for classification is appealing [6]. Frame-level features in our experiment are compressed into a single set of song-level features by computing the mean and standard deviation across the feature vectors [6]. Furthermore, we investigate the effects of frame size on multi-label classification. For each <*feature set*, *classifier*> pair, we examine the classification performance as we adjust the frame size, f_r, represented as the number of samples collected during a certain time period, where $f_r \in \{256, 512, 1024, 2048, 4096\}$ with a 50% frame overlap [8].

5. RESULTS AND DISCUSSIONS

In this section, we present the results from our experiments. Following the practices used in [2, 5, 16], 10-fold cross validation is employed during the evaluation process.

5.1 CAL500

In the first set of experiments, we evaluate the multi-label classifiers using the CAL500 dataset. We find that for all feature sets, the Calibrated Label Ranking classifier using a Support Vector Machine, CLR_{SVM}, outperforms the other classifiers when $f_r \in \{1024, 2048, 4096\}$. Furthermore, we observe that CLR_{DT}, $BPMLL$, $MLkNN$ and $BRkNN$ also perform well over all of the frame sizes and feature sets.

When we analyze the performance of each classifier over the individual frame sizes, we find it difficult to select one that performs well for all of the classifiers. More specifically, we observe that the performance of each classifiers is not significantly affected by the variation in frame size. Despite this, we find that CLR_{SVM} performs the best when $f_r = 4096$. Table 1 shows a comparison of 5 classifiers, evaluated by HL, for the three feature sets when $f_r = 4096$; the value following ± gives the standard deviation.

Note that in the following tables, (↓) indicates better performance when the number is smaller while (↑) indicates better performance when the number is bigger.

	FS_s	FS_t	FS_b
CLR_{SVM}	0.125±0.004	0.127±0.004	0.128±0.004
$BPMLL$	0.211±0.009	0.218±0.008	0.217±0.010
$BRkNN$	0.130±0.004	0.134±0.003	0.136±0.004
$RAkEL_{\text{DT}}$	0.152±0.003	0.153±0.004	0.156±0.004
$MLkNN$	0.129±0.004	0.133±0.003	0.135±0.003

Table 1. Hamming Loss (↓) of the classifiers for the three feature sets, FS_s, FS_t, and FS_b, when $f_r = 4096$.

From the table we can see that HL of each classifier is better when FS_s is used. This is also observed for the other evaluation measures. Table 2 presents the performance of CLR_{SVM} for each of the feature sets as evaluated by HL, OE, CO, and AP; the best result for each measure is shown in bold face. We find that CLR_{SVM} performs the best, for a majority of the evaluation measures, when FS_s is used. While spectral features have shown promising results in various MIR classification tasks, the inclusion of rhythmic features has been shown to increase classification performance [17]. Further investigation into this result would be desirable.

	HL ↓	OE ↓	CO ↓	AP ↑
FS_s	**0.125±0.004**	0.102±0.037	**116.7±2.814**	**0.586±0.016**
FS_t	0.127±0.004	0.094±0.047	119.7±3.659	0.576±0.014
FS_b	0.128±0.004	**0.088±0.035**	121.6±4.018	0.567±0.013

Table 2. Classification performance (mean±std) of CLR_{SVM} on CAL500 for each feature set where $f_r = 4096$.

When we analyze HL of CLR_{SVM} for each of the feature sets over all of the frame sizes, we find it interesting that both FS_s and FS_t demonstrate good performance when $f_r \in \{1024, 2048, 4096\}$ while FS_b tends to perform better when $f_r \in \{256, 512, 1024\}$. This result is discussed further in Section 5.3.

Table 3 reports the experiment results of the top 5 multi-label classifier using FS_s and $f_r = 4096$ on CAL500. To make a clearer view of the relative performance between each classifier, a partial order "$\succ$" can be defined on the set of all classifiers for each evaluation measure, where A1

[2] http://marsyas.sness.net.

	HL ↓	OE ↓	CO ↓	RL ↓	AP ↑	F_1 ↑	Precision ↑	Recall ↑
CLR_{SVM}	**0.125±0.004**	**0.102±0.037**	**116.736±2.814**	**0.140±0.006**	**0.586±0.016**	**0.497±0.027**	**0.642±0.059**	0.124±0.009
CLR_{DT}	0.126±0.003	0.106±0.024	117.417±2.970	0.143±0.005	0.578±0.014	0.445±0.027	0.611±0.039	0.124±0.013
$BPMLL$	0.211±0.009	0.130±0.051	119.878±3.880	0.144±0.007	0.570±0.016	0.479±0.016	0.294±0.039	**0.469±0.026**
$BRkNN$	0.130±0.004	0.184±0.054	143.503±2.834	0.189±0.008	0.534±0.016	0.429±0.017	0.543±0.043	0.131±0.011
MLkNN	0.129±0.004	0.132±0.054	126.082±2.994	0.159±0.005	0.550±0.011	0.476±0.014	0.587±0.047	0.118±0.010

Table 3. Classification performance (mean±std) on CAL500 for FS_s where $f_r = 4096$.

≻ A2 means that the performance of classifier A1 is statistically better than that of classifier A2 on the specified measure. Following the practice used by Zhang and Zhou [20], a two-tailed paired t-test at 5% significance level is used to perform the comparison.

Note that the partial order "≻" only measures the relative performance between two classifiers A1 and A2 for a single evaluation measure. It is possible that A1 performs better than A2 in terms of some measure but worse than A2 in terms of other ones. In this case, it is hard to judge which classifier is superior. Therefore, in order to give an overall performance assessment of a classifier, a score is assigned to it which takes into account its relative performance with other classifiers on all evaluation measures. For each measure, for each possible pair of classifiers A1 and A2, if A1 ≻ A2 holds, then A1 is rewarded by a positive score +1 and A2 is penalized by a negative score -1. Based on the accumulated score of each classifier on all evaluation measures, a total order ">" is defined on the set of all classifiers [20]. Table 4 presents an example of this process; the accumulated score for each classifier is shown in parentheses.

Multi-label Classifier	
A1-$BPMLL$; A2-CLR_{DT}; A3-CLR_{SVM}; A4-$MLkNN$	
Hamming Loss	A2 ≻ A1, A3 ≻ A1, A3 ≻ A4, A4 ≻ A1
Coverage	A1 ≻ A4, A2 ≻ A4, A3 ≻ A4
Ranking Loss	A1 ≻ A4, A2 ≻ A4, A3 ≻ A4
Average Precision	A1 ≻ A4, A2 ≻ A4, A3 ≻ A1, A3 ≻ A4
Total Order	A3(6) >A2(4) >A1(-1) >A4(-9)

Table 4. Relative performance between four multi-label classification algorithms on the CAL500 dataset.

The total order of all 10 multi-label classifiers on CAL500 is as follows: CLR_{SVM} (42) > CLR_{DT} (31) > $BPMLL$ (25) > $MLkNN$ (20) > $BRkNN$ (-1) > $RAkEL_{\text{SVM}}$ (-7) > $HOMER_{\text{SVM}}$ (-9) > $RAkEL_{\text{DT}}$ (-13) > $HOMER_{\text{DT}}$ (-35) > $IBLR$ (-53). It can be seen that CLR_{SVM} outperforms all the other classifiers on the CAL500 dataset. Furthermore, CLR_{DT}, $BPMLL$, $MLkNN$, and $BRkNN$ demonstrate good performance and outperform the remaining classifiers.

5.2 Magnatagatune

For the second set of experiments we evaluate the classifiers using the Magnatagatune dataset. We find that for all feature sets, CLR_{SVM} outperforms all the other classifiers when $f_r \in \{1024, 2048, 4096\}$. Furthermore, we observe that CLR_{DT}, $BPMLL$, $MLkNN$, and $BRkNN$, offer comparable performance over all of the frame sizes and feature sets.

	FS_s	FS_t	FS_b
CLR_{SVM}	0.022±0.002	0.021±0.002	0.021±0.001
$BPMLL$	0.073±0.003	0.074±0.002	0.022±0.002
$BRkNN$	0.021±0.002	0.021±0.002	0.022±0.002
$RAkEL_{\text{DT}}$	0.023±0.002	0.023±0.001	0.023±0.001
$MLkNN$	0.021±0.002	0.021±0.002	0.022±0.002

Table 5. Hamming Loss (↓) of the classifiers for the three feature sets, FS_s, FS_t, and FS_b, when $f_r = 2048$.

Once again, it is difficult to select a frame size that works well for all of the classifiers. We observe that each classifier performs differently, for each feature set, over the different frame sizes. In spite of this, CLR_{SVM} performs the best when $f_r = 2048$. Table 5 shows a comparison of 5 multi-label classifiers, as evaluated by HL, for the three feature sets when $f_r = 2048$. From the table, we find that HL is better for a majority of the classifiers when FS_t is used. It can be seen that HL of CLR_{SVM} and $BPMLL$ is better when FS_b is used. If we extend our analysis to include additional evaluation measures, we find that, on average, performance improves with the use of FS_t for a majority of classifiers. Table 6 presents the performance of CLR_{SVM} for each feature set.

	HL ↓	OE ↓	CO ↓	AP ↑
FS_s	0.022±0.002	0.423±0.028	40.6±4.709	0.479±0.017
FS_t	**0.021±0.002**	**0.403±0.050**	**38.9±4.687**	**0.505±0.027**
FS_b	**0.021±0.002**	0.413±0.037	40.7±5.031	0.495±0.024

Table 6. Classification performance (mean±std) of CLR_{SVM} on Magnatagatune for each feature set where $f_r = 2048$.

Table 7 presents the experiment results of the top 5 multi-label classifiers using FS_t and $f_r = 2048$ on Magnatagatune. We note that, for a majority of the evaluation measures, the performance of each classifier is better on Magnatagatune than on CAL500. We will discuss more on this

	HL ↓	OE ↓	CO ↓	RL ↓	AP ↑	F_1 ↑	Precision ↑	Recall ↑
CLR_{SVM}	**0.021±0.002**	**0.403±0.050**	**38.915±4.687**	**0.076±0.009**	**0.505±0.027**	0.350±0.029	**0.738±0.088**	0.018±0.002
CLR_{DT}	**0.021±0.002**	0.471±0.041	42.321±5.461	0.085±0.009	0.459±0.019	0.330±0.023	0.573±0.073	0.029±0.003
$BPMLL$	0.074±0.004	0.690±0.037	42.081±4.725	0.088±0.012	0.360±0.015	0.282±0.015	0.118±0.041	**0.268±0.030**
$BRkNN$	**0.021±0.001**	0.451±0.043	76.343±6.978	0.166±0.018	0.448±0.022	0.376±0.026	0.591±0.063	0.045±0.005
$MLkNN$	**0.021±0.002**	0.443±0.040	51.168±5.082	0.102±0.010	0.468±0.024	**0.390±0.041**	0.612±0.068	0.045±0.008

Table 7. Classification performance (mean±std) on Magnatagatune for FS_t where $f_r = 2048$.

in the following section.

Similarly as the CAL500 dataset, the partial order "$\succ$" and the total order "$>$" are also defined on the set of all classifiers. The total ordering for all 10 multi-label classifiers on Magnatagatune is as follows (the accumulated score for each classifier is shown in parentheses): CLR_{SVM} (37) $>$ $MLkNN$ (28) $>$ CLR_{DT} (24) $>$ $BRkNN$ (22) $>$ $BPMLL$ (-1) $>$ $RAkEL_{\text{DT}}$ (-7) $>$ $HOMER_{\text{SVM}}$ (-11) $>$ $RAkEL_{\text{SVM}}$ (-21) $>$ $IBLR$ (-31) $>$ $HOMER_{\text{DT}}$ (-40). It can be seen that CLR_{SVM} outperforms all of the multi-label classification algorithms on the Magnatagatune dataset. Furthermore, $MLkNN$, CLR_{DT}, $BRkNN$, and $BPMLL$ perform well for a selection of evaluation measures.

5.3 Discussions

Base Classifier: From our experiments presented above, we observe that using a SVM as the base-level learning algorithm for CLR, $RAkEL$, and $HOMER$ offers improvements over using a decision tree. This result is observed for both of the datasets. Table 8 reports the experimental results of CLR, $RAkEL$, and $HOMER$ on the CAL500 dataset using a SVM and DT as base classifiers. It would be interesting to explore alternative base-level learning algorithms for music tag annotation.

	HL ↓	OE ↓	AP ↑
CLR_{DT}	0.126±0.003	0.106±0.024	0.578±0.014
CLR_{SVM}	0.125±0.004	0.102±0.037	0.586±0.016
$HOMER_{\text{DT}}$	0.196±0.007	0.808±0.061	0.355±0.020
$HOMER_{\text{SVM}}$	0.159±0.004	0.581±0.051	0.427±0.015
$RAkEL_{\text{DT}}$	0.151±0.003	0.283±0.045	0.473±0.010
$RAkEL_{\text{SVM}}$	0.125±0.004	0.239±0.048	0.424±0.013

Table 8. Classification performance (mean±std) of CLR, $RAkEL$, and $HOMER$ on CAL500 using a SVM and DT as base classifiers.

Feature Set: We find it interesting that, on average, classification using FS_s and FS_t tends to demonstrate good performance when $f_r \in \{1024, 2048, 4096\}$ while using FS_b results in better performance when $f_r \in \{256, 512, 1024\}$. This might be explained by the notion that the smaller frame captures better rhythmic information over the entire music piece. Furthermore, a large frame may be more likely to capture the long-term nature of the music, including melodic, and harmonic composition, which could lead to improved classification accuracy. While we find small improvements in classification performance using different frame sizes, we observe large differences in performance between the best feature set and worst feature set for a selection of evaluation measures and classifiers. For example, the performance of $BPMLL$ on Magnatagatune, as evaluated by AP, varies from 0.04% using FS_b to 37% using FS_t. In addition, we find that the best classification performance is achieved on CAL500 and Magnatagatune using FS_s and FS_t, respectively. However, it is important to note that there is no accepted criteria as which features are best for music classification [1]. Therefore, our observation in the experiments reported in this work may not be conclusive.

Datasets: For a majority of the evaluation measures, it can be seen that the classifiers perform better on Magnatagatune, compared to CAL500. For example, CLR_{SVM} achieves a Hamming Loss of 0.0211 on the former and 0.1247 on the latter. One possible explanation for this observation is that the average number of tags for each instance in Magnatagatune is less than CAL500, i.e., each music piece in Magnatagatune is annotated with approximately 3 tags while each music piece in CAL500 is annotated with approximately 26 tags. We also observe that classification performance varies for each dataset depending on individual feature sets. For instance, classification using FS_s performs the best on CAL500 while using FS_t demonstrates the best performance on Magnatagatune; we note that using FS_s shows the worst classification performance on Magnatagatune. This leads us to believing that the spectral features used in our experiment tend to give rise to better performance over longer pieces of music while using timbral features performs better on shorter music. Whether this is true in general needs further investigation.

6. CONCLUSION

In this paper we present our initial attempts on automatic music tag annotation. In our work, we conduct a series of experiments, on a set of multi-label classifiers, exploring the effects of different feature sets and frame sizes on tag annotation. The results offer insight into which classifiers and features are more suitable for this task. We find that the Calibrated Label Ranking (CLR) classifier consistently performs well for a selection of evaluation measures when using spectral and timbral features.

Further investigation is needed into the selection of classifier parameters. Recall that each classifier is trained using default parameters. It would be interesting to explore the influence of these parameters on tag annotation performance. In addition, it would be interesting and beneficial to compare our results to existing results in the literature based on a set of common measures.

7. REFERENCES

[1] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl. Aggregate features and AdaBoost for music classification. *Machine Learning: Special Issue on Machine Learning in Music*, 65(2-3):473–484, 2006.

[2] T. Bertin-Mahieux, D. Eck, F. Maillet, and P. Lamere. Autotagger: A model for predicting social tags from acoustic features on large music databases. *J. New Music Research*, 37(2):115–135, 2008.

[3] W. Cheng and E. Hüllermeier. Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, 76(2-3):211–225, 2009.

[4] A. Eronen. *Signal Processing Methods for Audio Classification and Music Content Analysis*. PhD thesis, Tampere University of Technology, Tampere, Finland, 2009.

[5] M. Hoffman, D. Blei, and P. Cook. Easy as CBA: A simple probabilistic model for tagging music. In *Proc. Int'l Conf. Music Information Retrieval*, pages 369–374, 2009.

[6] S. R. Ness, A. Theocharis, G. Tzanetakis, and L. G. Martins. Improving automatic music tag annotation using stacked generalization of probabilistic svm outputs. In *Proc. ACM Int'l Conf. Multimedia*, pages 705–708, 2009.

[7] F. Pachet. Content management for electronic music distribution. *Communications of the ACM*, 46(4):71–75, 2003.

[8] F. Pachet and P. Roy. Improving multi-label analysis of music titles: A large-scale validation of the correction hypothesis. *IEEE Trans. Audio, Speech & Language Processing*, 17(2):335–343, 2009.

[9] C. Sanden and J. Z. Zhang. Enhancing multi-label music genre classification through ensemble techniques. In *Proc. ACM Int'l Conf. Research and development in Information*, pages 705–714, 2011.

[10] K. Seyerlehner, G. Widmer, M. Schedl, and P. Knees. Automatic music tag classification based on block-level features. In *Proc. Sound and Music Computing Conf.*, 2010.

[11] J. Shen, W. Meng, S. Yan, H. Pang, and X. Hua. Effective music tagging through advanced statistical modeling. In *Proc. ACM Int'l Conf. Information Retrieval*, pages 635–642, 2010.

[12] E. Spyromitros, G. Tsoumakas, and I. Vlahavas. An empirical study of lazy multilabel classification algorithms. In *Proc. Hellenic Conf. Artificial Intelligence*, pages 401–406, 2008.

[13] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas. Multilabel classification of music into emotions. In *Proc. Int'l Conf. Music Information Retrieval*, pages 325–330, 2008.

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

[15] G. Tsoumakas, I. Katakis, and I. Vlahavas. Effective and efficient multilabel classification in domains with large number of labels. In *Proc. ECML/PKDD Workshop on Mining Multidimensional Data*, 2008.

[16] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Trans. Audio, Speech, and Language Processing*, 16(2):467–476, 2008.

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

[18] K. West. *Novel Techniques for Audio Music Classification and Search*. PhD thesis, University of East Anglia, UK, 2008.

[19] M-L. Zhang and Z-H. Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Trans. Knowl. and Data Eng.*, 18:1338–1351, 2006.

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

[21] Z. Zhao, X. Wang, Q. Xiang, A. M. Sarroff, Z. Li, and Y. Wang. Large-scale music tag recommendation with explicit multiple attributes. In *Proceedings of the international conference on Multimedia*, pages 401–410, 2010.