A regressive boosting approach to automatic audio tagging based on soft annotator fusion

Rémi Foucard, Slim Essid, Mathieu Lagrange, Gael Richard

A REGRESSIVE BOOSTING APPROACH TO AUTOMATIC AUDIO TAGGING BASED ON SOFT ANNOTATOR FUSION

Rémi Foucard[(1)]*, Slim Essid*[(1)]*, Mathieu Lagrange*[(2)] *and Gaël Richard*[(1)]

[(1)] TELECOM ParisTech, CNRS-LTCI
37, rue Dareau
75014 Paris, France

[(2)] Ircam, STMS, CNRS, UPMC
1, place Igor Stravinsky
75004 Paris, France

remi.foucard@telecom-paristech.fr

ABSTRACT

Automatic tagging of music has mostly been treated as a classification problem. In this framework, the association of a tag to a song is characterized in a "hard" fashion: the tag is either relevant or not. Yet, the relevance of a tag to a song is not always evident. Indeed, during the ground-truth annotation process, several annotators may express doubts, or disagree with each other. In this paper, we propose to fuse annotators' decisions in a way to keep information about this uncertainty. This fusion provides us continuous scores, that are used for training a regressive boosting algorithm. Our experiments show that regression with this soft ground truth leads to a more accurate learning, and better predictions, compared to traditionally used binary classification.

***Index Terms*—** Machine learning, Music information retrieval, Regression analysis, Autotagging, Boosting

1. INTRODUCTION

Tags constitute a very useful tool for indexing multimedia documents. They consist in textual semantic labels describing any aspect of the document, helping database organization and structuring. Tags are widely used on social web services such as Flickr or Last.fm, where users are invited to associate such keywords to the documents they share, or appreciate.

Unfortunately, this user-based tagging method penalizes unpopular documents. Furthermore, the characteristics described by the tags are left as a free choice, and many tags potentially useful for data structuring may not be considered by the users. To avoid these problems, more precise and reliable tags can be obtained by consulting music experts, as done by the Music Genome Project, led by the online radio Pandora[1]. Every song of their catalog is annotated by professionals, who are asked to listen carefully to the songs, and consider many tags, as precise as *Jazz_Waltz_Feel*. The annotation yielded by this process is of high quality, but it appears also very long and costly to collect.

As an alternative, automatic tagging can be used for rapidly labeling a great number of documents. Many works on music "autotagging" have been produced in the last years [1, 2]. They make use of machine learning for building rules, deciding whether or not a tag is relevant for a new song.

Most of the time, this task is considered as a classification problem: a tag can only be present or absent. For a learning algorithm, positive and negative tag/song association can be respectively represented, by target scores 1 and 0. But as human beings, our response to a musical stimulus is more subtle than this binary description. Indeed, annotators may not be sure which class to choose, or may even give contradictory answers. In this paper, we resort to annotator fusion to express this uncertainty by considering continuous scores, translating a form of consensus among annotators, rather than binary scores. These scores are then mined by a regressive boosting algorithm.

The paper is organized as follows. Firstly, we describe in Section 2 automatic tagging algorithms and common techniques for building ground truth data. Then, in Section 3, we explain our annotator fusion method, and section 4 presents a regressive boosting method to properly exploit the obtained scores. Section 5 describes an experiment validating the approach. Finally, we conclude and suggest future works in the last section.

2. RELATED WORK

Autotagging is one of the main research areas in the field of music information retrieval. Consequently, many works have been published to address this problem [3].

Usually, the signal is cut into short overlapping frames, from which several descriptors are computed. For instance, common features are: Mel Frequency Cepstral Coefficients (MFCC) and their derivatives, chromagrams, spectral moments and zero-crossing rates. Thus, the signal is represented by a collection of feature vectors (one per frame).

This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation

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

Then, a "learning model" is used to infer a rule for deciding, from the features, whether a considered tag is present or not. Widely used learning models are: *Support vector machines* [4], *Gaussian mixture models* [2] and boosting [5]. All of these models are used for supervised learning, and need to be provided with ground-truth tag/song associations.

Several methods have been proposed for the labeling of training examples [6]. The first, and most accurate one is the survey: expert or non-expert annotators listen to the audio file and answer precise questions about the content. This procedure ensures that the annotators have considered all tags. Annotators can also contribute to the labeling through an annotation game. Less costly methods consist in automatically mining social tags or web documents.

Most of the time, the annotation data is processed to obtain binary target scores. Thus, a label may only be present or absent. However, most people agree that music is a complex source of data, and thus the concepts behind each tag are seldom accurate and clear enough to be represented by a binary category.

There can be two sources of uncertainty in the ground-truth: the annotator's individual uncertainty, or inter-annotator disagreement. For instance, concepts such as emotion or mood are difficult to categorize, furthermore, the words describing emotions do not mean exactly the same for different people. This problem can be tackled by placing emotional states on a two-dimensional valence-arousal continuous space. Emotion recognition can then be formulated as a regression [7] or ranking [8] problem. The annotations consist in placing songs on the two-dimensional space, or ranking them. Target scores are obtained by averaging individual subject responses, thus yielding continuous scores. This formulation is very suitable for emotion recognition, but it does not draw categories, and would thus need further processing steps to be adapted to the autotagging framework.

In [9], the authors use the correlations between tags to draw ordered intermediate categories, which represent several confidence levels. However, these categories are built from the binary scores. Thus, the annotator uncertainty is only inferred, instead of directly used.

3. SOFT ANNOTATOR FUSION

The present study uses the CAL500 database, which has been annotated using the survey method. Every song has been annotated by at least three people. For many tags in the survey, the annotators can choose between several confidence levels. For instance, when annotating a particular song, each emotion concept can be rated from 1 to 5. As stated in the previous section, this way of producing the data is likely to generate annotator uncertainty, as well as inter-annotator disagreement.

We propose an annotator fusion outputting continuous scores, which should be more flexible and able to express doubts. Unfortunately, the overall soft scores provided with the database are not exhaustively documented and their building process from the individual annotations is, in some cases, difficult to understand. This led us to construct our own soft scores, based on the annotators' individual responses.

Firstly, every possible response is converted to a value $v \in [0, 1]$. Consecutive values are equally spaced, moreover $v = 0$ and $v = 1$ must always be possible answers. For instance, there are four possible responses for the tag *Instrument-Trumpet*: *None*, *Uncertain*, *Present* and *Prominent*. They are respectively mapped to: 0, 0.33, 0.67 and 1.

Then, for a given tag and a song, there are several ways of fusing the individual responses. For the CAL500 binary scores, a tag is considered as "positive" if 80% of test subjects agree that the tag is relevant [10]. Other fusion methods include: majority voting (*i.e.* the score corresponds to the most chosen category), or taking the (possibly thresholded) mean of the individual annotations. Because majority voting and thresholded mean calculation do not reflect uncertainty, we choose to average the individual scores (as done in [7, 8], but for a different kind of ground-truth data):

$$V_s = \frac{1}{K} \sum_{k=1}^{K} v_k, \tag{1}$$

where v_k is the value corresponding to the choice made by annotator k.

Alternatively, for the "negative" tags (*e.g. Emotion-NOT_happy*), the value is simply $V = 1 - P$, where P is the value associated with the corresponding "positive" tag.

To validate the soft scores obtained by this process, we measure their agreement with the binary ones provided by the Computer Audition Lab with CAL500. Cohen's kappa coefficient [11] gives such an evaluation, but needs two binary sets of annotations. So, in order to obtain comparable values we build new hard binary scores V_h, corresponding to the reconstructed soft scores V_s. The new binary scores are obtained by thresholding our soft values:

$$V_h = \begin{cases} 1 & \text{if} \quad V_s > t \\ 0 & \text{otherwise} \end{cases} . \tag{2}$$

The threshold giving the highest agreement ($t = 0.64$) leads to a mean Cohen's kappa of $\kappa = 0.80$ between the two sets of labels. According to [11], this value denotes a high agreement.

4. REGRESSIVE BOOSTING FOR SOFT MUSIC TAGGING

The soft scores obtained by annotator fusion will be used to train a regressive boosting system.

4.1. Features

In our system, audio data is represented in a bag-of-frames fashion, using the following set of features: the 15

Feature	Dim.	Description
Spectral Centroid	1	The centroid of the spectrum
Spectral Spread	1	Spread of the spectral energy
Spectral Skewness	1	Asymmetry of the spectrum
Spectral Kurtosis	1	"Peakedness" of the spectrum
Zero-crossing rate	1	Frequency of the signal sign change
Loudness	1	Perceived sound intensity
Sharpness	2	High frequency content
Timbral Width	1	Flatness of a loudness function
Volume	1	Perceived size of the sound
Spectral Dissonance	2	Roughness of spectrum components
Tonal Dissonance	2	Roughness of just tonal components
Pure Tonalness	1	Audibility of spectral pitches
Complex Tonalness	1	Audibility of virtual pitch
Multiplicity	1	Number of tones noticed
Tonality	1	Tonality of the song
Chord	1	Instantaneous chord
MFCC	13	Cepstral description
Chroma	12	Energy content for each note class

Table 1. Features used by the training systems.

psychoacoustic-related features recommended in [7] (loudness, tonal dissonance, ...)[2], completed by the common first 13 MFCC (dropping the energy), chroma, zero-crossing rate, and spectral spread, skewness and kurtosis. These features are presented in Table 1. These features are computed from half-overlapping 23 ms frames, and then temporally averaged over 2 s.

4.2. Regressive boosting

On these features, we apply two boosting algorithms. Boosting is a learning technique, training iteratively several complementary versions of a "weak" (performing badly) classifier. The best-known version of boosting is probably Adaboost, which is described in Algorithm 1. This version uses weights for putting emphasis on particular training examples. At each iteration r, the weights of the examples correctly classified by the weak classifier T_r, are decreased, thus putting the focus on the other examples in the following iterations.

Boosting is originally a classification algorithm, but has been generalized to handle regression with several differentiable loss functions [12]. In the case of squared error, there is no weighting system. Instead, at each iteration, the target values for regressor T_r are the prediction residuals:

$$res_{i,r} = y_i - \sum_{k=1}^{r-1} T_k(x_i) \tag{3}$$

where y_i is the target score for training example i, and x_i is the corresponding feature vector. The regressive boosting algorithm for squared error is presented in Algorithm 2.

During the test phase, a single score S_n is obtained for each song n, by averaging the algorithm predictions $H(x)$ corresponding to every frame of the song.

[2]These features have been extracted using Psysound (`http://www.psysound.org/`)

Algorithm 1 Adaboost algorithm.

initialize the example weights $w_i \longleftarrow \frac{1}{2m}, \frac{1}{2l}$, resp. for $y_i = 0, 1$, where m and l are the number of negative and positive examples, respectively

for $r = 1, \ldots, R$ **do**
  Fit a classifier $T_r(x)$ to the training data, using weights w_i
  // Compute weighted error rate
  $\epsilon_r \longleftarrow \frac{1}{\sum_i w_i} \sum_i w_i I(y_i \neq T_r(x_i))$
  // Coefficient associated with T_r
  $\alpha_r \longleftarrow \log \frac{1}{\beta_r}$, where $\beta_r = \frac{\epsilon_r}{1-\epsilon_r}$
  // Update the example weights
  for all examples x_i correctly classified by T_r **do**
    $w_i \longleftarrow w_i \beta_r$
  end for
end for

Output: $H(x) = I(\sum_r \alpha_r T_r(x) \geq \frac{1}{2} \sum_r \alpha_r)$

Algorithm 2 Regressive boosting algorithm for squared error.

initialize the example target values $m_i \longleftarrow y_i$

for $r = 1, \ldots, R$ **do**
  Fit a regressor $T_r(x)$ to the training data, with targets m_i
  // Update the example target values
  for all examples x_i **do**
    $m_i \longleftarrow m_i - T_r(x_i)$
  end for
end for

Output: $H(x) = \sum_r T_r(x)$

5. VALIDATION OF THE APPROACH

We conduct an experiment to demonstrate the usefulness of our soft-fused scores, compared to the binary ones, and the efficiency of our regression scheme. To this end, we run two tag prediction systems on the same audio data: one is trained on the binary labels V_h, and another one on the soft scores V_s.

5.1. Experimental framework

The experiment is done on the CAL500 database. This base contains 500 pop songs, with tags describing mood, instrumentation, genre, *etc*. We use the same 61 tags as in [10]. The tests are conducted with 10-fold cross-validation, keeping 450 songs for training, and 50 for testing. For complexity reduction, we only use 30 s of each song: between instants 30 s and 60 s.

We train one Adaboost classification system on the recreated binary labels V_h, and one regression system using the soft ground truth V_s. Each of them will be trained with 500 boosting iterations. We use decision stumps (decision trees with two leaves) as weak classifiers, as done in [1].

To compare the prediction accuracy of the two systems

Annotator fusion method	MAP	AUC
Binary	0.46	0.67
Soft	0.50	0.71

Table 2. Performance on CAL500 with binary and soft annotator fusions.

on the test set, we measure their ability to predict the binary ground-truth. We use two different ranking metrics to evaluate this output. Ranking metrics evaluate the list of examples ranked by predicted score S_n. This list is compared against the binary ground truth. A perfect ranking would put all positive songs at the top. Our first measure is the Mean Average Precision (MAP). It can be obtained by moving down the ranked list, and averaging the precision obtained at every truly positive example. We also use the Receiver Operating Characteristic (ROC) curve. This curve represents the correct detection rate with respect to the false alarm rate, computed at each element in the ranking. The Area Under the ROC Curve (AUC) will be our second measure.

5.2. Results

The performance of the two systems is presented in Table 2. We can clearly see that the regressive system delivers better predictions than the classification system. Cross-validated paired t-tests [13] have shown that the difference between the two systems is significant, with more than 99% confidence. This means that the information about annotation uncertainty, brought by the soft scores, is actually useful to learning systems.

It is important to notice that the regressive system does not require more annotation data than the other one: only the processing of the annotations differs between the two systems. And the results show that there is indeed a loss of useful information when annotations are processed in a binary way.

6. CONCLUSION

In this paper, we have described a way of fusing annotations that preserves information about the uncertainty of the tag/song association. We have also proposed to use regressive boosting for learning the scores obtained by this fusion. Our tests show that the soft scores, combined with regressive learning, lead to a better learning of the tags.

Future work may include exploitation of the tag correlations, which has been proved to bring useful information for audio tagging [9]. Indeed, in the present study, tags have been considered as independent concepts. However, tags such as *Song-Very_danceable* and *Usage-At_a_party* are expected to appear together many times. This correlation could be exploited by methods such as multivariate regression.

7. REFERENCES

[1] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl, "Aggregate features and ADABOOST for music classification," *Machine Learning*, vol. 65, no. 2-3, pp. 473–484, 2006.

[2] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic Annotation and Retrieval of Music and Sound Effects," *TASLP*, vol. 16, no. 2, pp. 467–476, Feb. 2008.

[3] T. Bertin-Mahieux, D. Eck, and M. Mandel, "Automatic Tagging of Audio: The State-of-the-Art," in *Machine Audition: Principles, Algorithms and Systems*, Wenwu Wang, Ed. IGI Publishing, 2010.

[4] C. Xu, N. Maddage, X. Shao, F. Cao, and Q. Tian, "Musical genre classification using support vector machines," in *ICASSP*, 2003, pp. 429–432.

[5] R. Foucard, S. Essid, M. Lagrange, and G. Richard, "Multi-scale temporal fusion by boosting for music classification," in *ISMIR*, 2011.

[6] D. Turnbull, L. Barrington, and G. Lanckriet, "Five Approaches to Collecting Tags for Music," in *ISMIR*, 2008, pp. 225–230.

[7] Y. Yang, Y. Lin, Y. Su, and H. Chen, "A Regression Approach to Music Emotion Recognition," *TASLP*, vol. 16, no. 2, pp. 448–457, 2008.

[8] Y. Yang and H. Chen, "Ranking-Based Emotion Recognition for Music Organization and Retrieval," *TASLP*, , no. 99, 2010.

[9] Y. Yang, Y. Lin, A. Lee, and H. Chen, "Improving Musical Concept Detection by Ordinal Regression and Context Fusion," in *ISMIR*, 2009, pp. 147–152.

[10] L. Barrington, M. Yazdani, D. Turnbull, and G. Lanckriet, "Combining Feature Kernels for Semantic Music Retrieval," in *ISMIR*, 2008, pp. 614–619.

[11] J. Landis and G. Koch, "The measurement of observer agreement for categorical data.," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 3 edition, 2009.

[13] T.G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Computation*, vol. 10, no. 7, 1998.