Multimodal automatic tagging of music titles using aggregation of estimators

Kaoutar El Ghali
LUTIN UserLab
30, avenue Corentin Cariou
Paris, France
kaoutar.elghali@gmail.com

Adil El Ghali
LUTIN UserLab
30, avenue Corentin Cariou
Paris, France
elghali@lutin-userlab.fr

Charles Tijus
LUTIN UserLab
30, avenue Corentin Cariou
Paris, France
tijus@lutin-userlab.fr

ABSTRACT

This paper presents the participation to the MusiClef 2012 Multimodal Music Tagging task. It expounds the approach that consists of an aggregation of estimators as a procedure to combine different sources of information.

Keywords

Music Tagging, Multimodal tagging, Aggregation of Estimators, Random Indexing, Classification

1. INTRODUCTION

The MusiClef is a benchmarking activity that offers an evaluation of different multimodal approaches for MIR tasks. The 2012 edition task consists in a automatic tag assignment. Having a set of tags that applies to a music title provides a good description for it. It is an equivalent approach to web mining approaches that most search engines adopt. Automatically assigning tags to music can then be used in a wide number of music information retrieval tasks, for example, music library management, recommender systems, playlist generation, retrieving music from textual queries, etc. These tasks require computers to hear to be achieved, meaning that computers should be able to recognize, organize, categorize sounds, to learn semantics and be able to apply it to sounds. The test collection is composed of 1355 songs performed by the 218 artists in the "Rolling Stone 500 Greatest Songs of All Times". The organizers provide audio features for the whole collection as well as web-pages, related to the artists and albums, and social tags retrieved from Lastfm. The vocabulary of tags to be assigned is composed of 94 tags representing both music genre and mood.

2. SYSTEM OVERVIEW

In this work the auto-tagging task of music titles is addressed as a classification problem, more precisely, multiple classification problems. Indeed, due to the large number of features compared to the number of samples, we chose to reduce the dimensionality of the problem by performing an aggregation procedure [8]. Hence, each type of data is treated separately (ie, collaborative data, content-based features, web-based data) to compute similarities to the tag list, the results are then aggregated in an higher classifier. In a clearer manner, each prediction of our system is made through a convex combination of 3 predictions weighted by a credibility criterion computed on a validation set.

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

2.1 Aggregation

Let Θ be the set of available features, $(\Theta_{audio}, \Theta_{text}, \Theta_{lastfm})$ a partition of Θ and $\{\overline{f}_{audio}, \overline{f}_{text}, \overline{f}_{lastfm}\}$ the family of estimators computed for each features subset on a portion of the training set $D_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ taking values in $\chi \times [-1, 1]^N$ where $N = 94$ is the number of tags. We index the family of estimators on $\Lambda = \{audio, text, lastfm\}$ and we note it $F(\Lambda)$. An aggregated is the real-value statistic of the form

$$\hat{f} = \sum_{\lambda \in \Lambda} \omega(\overline{f}_\lambda)\overline{f}_\lambda$$

where $(\omega(\overline{f}_\lambda))_{\lambda\in\Lambda}$ are computed on a validation set $D_l = \{(X_{n+1}, Y_{n+1}), ..., (X_{n+l}, Y_{n+l})\}$ and satisfy $\forall \lambda \in \Lambda, \omega(\overline{f}_\lambda) \geq 0$ and $\sum_{\lambda\in\Lambda} \omega(\overline{f}_\lambda) = 1$

We use an Aggregation with Exponential Weights (AEW) [2], for which the weights are defined as:

$$\forall f \in F(\Lambda), \omega(f) = \frac{e^{-lTA_l^{(\Phi)}(f)}}{\sum_{\lambda\in\Lambda} e^{-lTA_l^{(\Phi)}(\overline{f}_\lambda)}}$$

where $A_m^{(\Phi)} : \frac{1}{m}\sum_{i=1}^{m} \Phi(Y_i f(X_i))$ is the empirical Φ-risk function, $\Phi : x \rightarrow \log_2(1 + e^{-x})$ is the "Logit-Boosting" risk, and T is a temperature parameter. The excess of Φ-risk of an estimator f is defined as : $A^\Phi(f) - A^*$ where $A^* \equiv \min_{f:\chi\rightarrow[-1,1]^N}\{A^\Phi(f) = E[\Phi(Yf(X))]\}$ is the optimal risk. For the AEW estimator to minimize the excess of risk with an optimal speed, the strongest assumption is the independence hypothesis upon the feature set partition. In our case of study, it is fair to assume that it is true for audio features and web semantic features, whereas for social tags, further preprocessing is needed.

2.2 Semantic similarities

Models of vector representation for word semantics are a family of models representing the semantic similarity between words contingent on the text environment in which they appear. These models rely on Harris distributional hypothesis. The distribution of word co-occurrence in the corpus is collected, analyzed and transformed into a semantic space in which words are represented as vectors of a high dimension. There are different mathematical methods to achieve such the construction of a semantic space based on corpus. The used method Random Indexing [3]. The construction of a semantic space is achieved using the SemanticVectors library [10]. In this space, each document (*ie* web-page) is represented as a d-dimensional vector ($d = 512$). We then model each music title as the sum of all vectors representing documents related to it (*ie* web-pages regarding the music title performer and the album it is taken from) and each tag as the sum of the vectors representing the music titles it is assigned to in training set. Similarities are computed as the cosine between those vectors, those similarities are normalized so they lie in $[-1, 1]$.

2.3 Audio content-based similarities

The available audio features are the Mel-Frequency Cepstral Coefficients (MFCC). These features are based on the discrete cosine transform of the log amplitude Mel-frequency spectrum. Since MFCC are computed on successive frames of a fixed length, there number differs depending on the music title duration. The fundamental calculation here is to characterize each music title in a vectorial space so that all the corpus has comparable representations. To achieve this goal, we cluster, for each music title, its frames into groups through a K-means clustering [5] with a fixed number of clusters for all corpus $k = 16$. Each music title is then represented by a $\frac{k(k+1)}{2}$ dimensional vector composed of the norms of the cluster centroids and their distances to each other. The main idea here is that the occurrence of an event is of more interest than its timestamp in the context of automatic tagging. A neural network is then trained to compute the similarities to each tag in the list. The probabilities that a tag from the list is assigned to a certain music title are taken as similarity measures; as for the text similarities, those values are normalized to lie in $[-1, 1]$.

2.4 Collaborative filtering

Collaborative filtering relies on the assumption that data that occurs commonly in the same context is likely to have the same characteristics [1]. In the present work, we assume that artists and albums that share the same Lastfm tags are likely to be assigned to the same tag. This being said, the very large number of Lastfm tags compared to the number of available samples makes it difficult to recognize patterns in the tag assignment. We thus chose to cluster Lastfm tags into a smaller number ($m = 100$) of categories ($C_1, ...C_m$). Each music title X_i is then represented as a m-dimensional vector $\{x_{i1}, ..., x_{im}\}$ with : $\forall j \in [1, m], x_{ij} = \sum_{l \in L_i \cap C_j} \frac{n_{il}}{n_l}$, and where L_j is the set of Lastfm tags assigned to the music title X_i, n_l the total number of Lastfm users who assigned the tag l and n_{il} the number of Lastfm users who assigned the tag l to the music title X_i. As for the content-based data, we train a neural network to get the similarities of each music title to the provided list of tag.

2.5 Predicting tags

In order to predict the number of tags N_i assigned to a certain music title X_i, we chose to take as features meta data and the number of Lastfm tags. This choice is motived by the fact that the correlation between those features and the number of tags is high enough to assume that they are sufficient for predicting the number of tags. Based on a Classification And Regression Tree (CART) analysis, we build a decision tree for making those predictions. Finally, the predicted tags by our system for the music title X_i are the N_i highest values in $f(X_i)$.

3. CONCLUSION

In this paper we described a automatic tagging system of music title based on an aggregation procedure. Various works on automatic tagging of music titles have been conducted in the last years. They either rely on audio content data or contextual information. Mendel and Ellis [6] compare the multiple-instance learners to classifying 10-second song clips according to a tagged training set. Panagakis and Kotropoulos [7] propose a low-rank representation-based multi-label annotation based on weighted audio temporal modulation representations. Lamere [4] uses social tags as contextual knowledge to generate tags related to genre, mood, orchestration, etc. Some works try to combine both sources of information. Turnbull *et al.* [9] compare three algorithms for the combination of audio content and social source information. The present work expounds a novel approach for the combination of audio content and context information based on the aggregation of estimators, exploiting information of the audio, lastfm tags, and texts associated with the song to tag.

4. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE-TKDE*, 2005.

[2] A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Maching Learning*, 2008.

[3] P. Kanerva, J. Kristoferson, and A. Holst. Random indexing of text samples for latent semantic analysis. In *Proceedings of CogSci'00*. Erlbaum, 2000.

[4] P. Lamere. Social Tagging and Music Information Retrieval. *Journal of New Music Research*, 2008.

[5] B. Logan and A. Salomon. A music similarity function based on signal analysis. In *Proceedings of the IEEE Conference on Multimedia and Expo*, 2001.

[6] M. Mandel and D. Ellis. Multiple-instance learning for music information retrieval. In *ISMIR'08*, 2008.

[7] Y. Panagakis and C. Kotropoulos. Automatic music tagging by low-rank representation. In *Proceedings of ICASSP'12*, pages 497–550, Kyoto, Japan, 2012.

[8] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 2004.

[9] D. R. Turnbull, L. Barrington, G. Lanckriet, and M. Yazdani. Combining audio content and social context for semantic music discovery. In *Proceedings of SIGIR'09*, New York, NY, USA, 2009. ACM.

[10] D. Widdows and K. Ferraro. Semantic vectors a scalable opensource package and online technology management application. In *Proceedings of LREC'08*, 2008.