UNICAMP-UFMG at MediaEval 2012: Genre Tagging Task*

Jurandy Almeida[1], Thiago Salles[2], Éder F. Martins[2], Otávio A. B. Penatti[1],
Ricardo da S. Torres[1], Marcos A. Gonçalves[2], and Jussara M. Almeida[2]

[1]RECOD Lab, Institute of Computing, University of Campinas (UNICAMP), Campinas, SP – Brazil, 13083-852
{jurandy.almeida, penatti, rtorres}@ic.unicamp.br

[2]Dept. of Computer Science, Federal University of Minas Gerais (UFMG), Belo Horizonte, MG – Brazil, 31270-010
{tsalles, ederfm, mgoncalv, jussara}@dcc.ufmg.br

ABSTRACT

Developed in the context of Genre Tagging Task at MediaEval 2012, this work consists in automatically assigning genre tags to a set of Internet videos. We approach this task from the classification point of view and focus on different learning strategies: video similarity for processing visual content and an ensemble of classifiers for text-processing. In this paper, we describe the proposed framework and report the results obtained on the official submission runs.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]

General Terms

Algorithms, Experimentation, Performance

Keywords

multimedia information retrieval, video genre classification, meta-learning approach, stacked generalization

1. INTRODUCTION

Current solutions to predict category labels for video data are usually based on learning strategies. Most of those research works have focused on using a single classifier. Recently, combining individual predictions from a set of classifiers has been established as an effective way to improve classification performance.

In this paper, we present an approach for genre tagging which relies on different learning strategies for processing textual and visual contents. For text-processing, we use a meta-learning scheme for combining classifiers, known as *stacked generalization* [4]. In contrast, a simple, yet effective, *histogram of motion patterns* (HMP) [1] is used for processing visual information.

This work is developed in Genre Tagging Task at MediaEval 2012 and its goal is to automatically assign tags to Internet videos using features derived from speech, audio, visual content or associated textual or social information. Details about data, task, and evaluation are described in [3].

2. THE PROPOSED FRAMEWORK

The proposed framework exploits both textual and visual information of shared media. Due to varying character of the videos, the metadata and transcripts are available in different languages. In our framework, we consider only metadata in English and use only title, tags, and description as textual features. To avoid trivial recommendations (e.g., plurals and other simple variations of the same word), we apply the Porter Stemming algorithm[1] to remove the affixes of each word in each collected feature. Next step is the removal of stop words. Finally, we represent each video using the Bag-of-Words (BoW) model [2].

To encode visual features, we adopt the *histogram of motion patterns* (HMP) [1]. For each frame of an input sequence, motion features are extracted from the video stream. After that, each feature is encoded as a unique pattern, representing its spatio-temporal configuration. Finally, those patterns are accumulated to form a normalized histogram. The extracted features (textual and visual) are then classified with the following methods:

K-Nearest Neighbors (KNN): a nonlinear classifier that assigns to a test example the majority class among those of its k nearest neighbor training examples in the vector space. We here use cosine similarity to determine the nearest neighbors of a test example, defined as $sim(\vec{x}_i, \vec{x}_j) = \frac{\langle \vec{x}_i, \vec{x}_j \rangle}{\|\vec{x}_i\| \cdot \|\vec{x}_j\|}$, where $\langle \cdot, \cdot \rangle$ is the inner product between vectors.

Naïve Bayes (NB): a probabilistic learning method that aims at inferring a model for each class by assigning to a test example the class associated with the most probable model that would have generated it. The predicted class y for the test example x is thus the one which maximizes the posterior class probabilities $P(y|x)$, defined as $P(y|x) \propto P(y) \cdot P(x|y)$, where$P(y)$ is the class prior probability and $P(x|y)$ denotes the conditional probability of observing x given y.

Random Forests (RF): a variation of decision tree's bagging, in which an ensemble of de-correlated decision trees is learned. A series of random procedures, such as bootstrapping of the input data and random selection of features to compose decision nodes, is applied in order to reduce the correlation between trees.

Support Vector Machine (SVM): a classifier that aims at finding an optimal separating hyperplane between the positive and negative training documents, maximizing the distance (margin) to the closest points from either class. Since the SVM classifier is naturally a binary classifier, we follow the common one-against-one approach to adapt binary SVM for a multi-class classification problem, as the explored datasets are composed of more than two classes.

Stacked generalization (stacking): an ensemble that consists of a two-stage process [4]. In the first stage (the so-called level-0), a series of classifiers are learned using

*We thank FAPEMIG, FAPESP, CAPES, and CNPq for funding.

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

[1] http://tartarus.org/~martin/PorterStemmer/

Experiment	Input	Classifier	Explored Features	MAP
run 1	audio/visual only, no ASR	KNN	HMP	0.1238
run 4	everything allowed, no uploader ID	Stacking	BoW	0.2112

Table 1: Obtained results on official submitted runs.

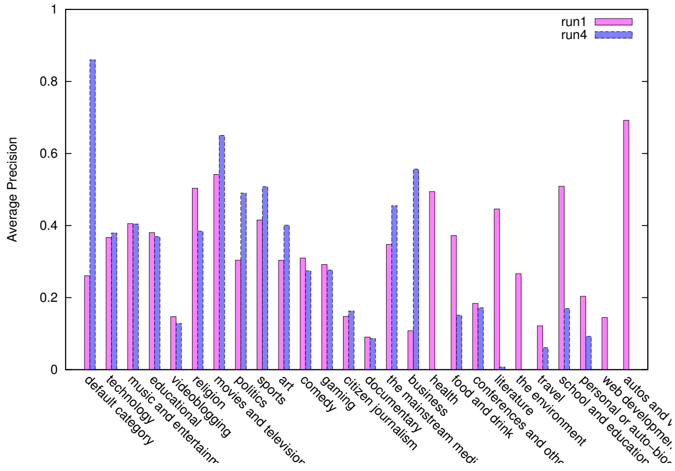

Figure 1: AP per class achieved in each run.

the original training set. Specifically, by means of a cross-validation procedure, each classifier is learned and then applied to classify the training examples, generating a set of outputs (the scores associated with each possible class). These outputs are then aggregated, in order to produce a new (transformed) training set. That is, the original feature space is mapped to a new space composed by the generated class scores, and the new training set consists of each training example represented in such a transformed space. This new training set will be subsequently used to learn the final stacked classifier. In order to classify a set of unseen examples, the level-0 classifiers are trained using the entire training set. The original test set is then classified, and a set of scores is generated. In the second stage, called level-1, such class scores output is then used to feed the previously learned stacked classifier, generating the final predictions.

3. EXPERIMENTS & RESULTS

For this task, we performed one submission for two different runs. Our learning strategies were based on the kind of features allowed in each run. For runs where textual features were allowed, we applied the previously described stacked generalization approach, taking advantage of learners highly effective for text classification (high dimensional sparse data). Specifically, we considered as level-0 learners the NB, KNN and RF classifiers. For the level-1 stage, we employed a non-linear SVM classifier, with a RBF kernel. All the learned classifiers were calibrated through cross validation on the training data.

Considering runs in which only visual features were permitted, we preferred to place our bets on more complex non-linear classification models, since the problem is characterized by a dense data matrix with the number of visual features smaller than the number of examples. Specifically, we applied a single KNN classifier, with parameter k optimized on the training set. The optimal value found was $k = 5$, defining complex non-linear class boundaries. The idea here was to test how well video similarity in terms of its motion sequence would fit our purposes of predicting their genre.

We evaluated our strategies considering the available resources and the aforementioned classifiers. The development set was used for training and is composed by $5,288$ videos distributed among 26 genres. For the runs considering the visual features, the average dimension was 3703.59 (from a total of $4,888$ features), while for the runs considering textual metadata the average dimension was 31.03 (from a vocabulary of $54,796$ terms). We ran a 5-fold cross validation procedure to evaluate the accuracy of each strategy (measured by Mean Average Precision—MAP).

Table 1 presents the results obtained on the official submission runs, for 9550 videos of the test set. As we can observe, when considering the visual features we achieved a MAP of 0.1238, by applying a single KNN classifier with complex class boundaries, as discussed previously. The runs considering textual metadata achieved a MAP of 0.2112, here considering the more elaborated stacking approach, which makes use of all discussed classifiers.

Figure 1 presents the Average Precision (AP) per class achieved in each of the submitted runs. Notice that our learning strategy for visual features provides a good discriminative power on video genres like (the number in the brackets is AP): “health” (0.494), “food and drink” (0.372), “literature” (0.446), “school and education” (0.509), “autos and vehicles” (0.692), while at the left end are video genres whose contents are more reflected with textual information, such as “default category” (0.860) and “business” (0.557).

Those results indicate that the combination of both textual and visual features may be promising, for it may reduce misclassifications of single modalities and, therefore, improve the classification performance.

4. CONCLUSIONS

In our approach, we explored textual information found in video metadata (e.g., title, tags, and description) and visual properties. We used different learning strategies for each data modality: video similarity for processing visual content and an ensemble of classifiers for text-processing. Obtained results demonstrate that the proposed framework is promising. By combining textual and visual features, we believe that we can make a contribution to better results. Future works include the investigation of learning strategies for combining features from different modalities and considering other information sources, such as ASR transcripts, to include more features semantically related to each category.

5. REFERENCES

[1] J. Almeida, N. J. Leite, and R. da S. Torres. Comparison of video sequences with histograms of motion patterns. In *ICIP*, pages 3673–3676, 2011.

[2] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, 2008.

[3] S. Schmiedeke, C. Kofler, and I. Ferrané. Overview of mediaeval 2012 genre tagging task. In *MediaEval*, 2012.

[4] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.