

VideoMule: A Consensus Learning Approach to Multi-Label Classification from Noisy User-Generated Videos

Chandrasekar Ramachandran, Rahul Malik, Xin Jin, Jing Gao,
Klara Nahrstedt, Jiawei Han

Department of Computer Science
University of Illinois at Urbana-Champaign
{cramach2,rmalik4,xinjin3,jinggao3}@illinois.edu, {klara,hanj}@cs.uiuc.edu

ABSTRACT

With the growing proliferation of conversational media and devices for generating multimedia content, the Internet has seen an expansion in websites catering to user-generated media. Most of the user-generated content is multimodal in nature as it has videos, audio, text (in the form of tags), comments and so on. Content analysis is a challenging problem on this type of media since it is noisy, unstructured and unreliable. In this paper we propose VideoMule, a consensus learning approach for multi-label video classification from noisy user-generated videos. In our scheme, we train classification and clustering algorithms on individual modes of information such as user comments, tags, video features and so on. We then combine the results of trained classifiers and clustering algorithms using a novel heuristic consensus learning algorithm which as a whole performs better than each individual learning model.

Categories and Subject Descriptors

H.51 [Multimedia Information Systems]: Video

General Terms

Algorithm

Keywords

Video Classification, Multimodal Information Processing

1. INTRODUCTION

The growth and increased proliferation of content-sharing websites like YouTube, Blinkx, Truveo and so on has resulted in a diverse collection of user-generated videos in the Internet. As an indicator of the enormous popularity of these websites, a recent 3-month study [3] estimated that YouTube currently controls 20% of all HTTP traffic or 10%

of all traffic on the web. This statistic is expected to grow over the next couple of years [1]. There are several common characteristics of the data in these content-sharing websites. Most of the content-sharing websites allow for seamless uploading of videos in standardized formats, they allow for tagging and commenting of these videos, sharing of videos between users and also embedding in a HTML page. In addition to this, many websites allow for rating and commenting of videos by the users. In this way, a typical document in a content-sharing website contains not only videos, but also their associated meta-data like video description, keywords, category, number of ratings, video responses, comments and so on.



Figure 1: Highlighted content shows the various useful metadata associated with a YouTube video. This is in addition to the semantic video features and audio content of this video.

The task of classifying such videos using multiple modes of information is an interesting and challenging problem. Essentially the predictive information for classifying this user-generated content is not restricted to the videos alone. The additional user-generated metadata can be used along with the video content to deliver better classification results. This forms a kind of consensus learning [10] where each mode of information generates some predictive results in the form of classifiers and clusters, and a combining algorithm then interprets these results, negotiates and generates a consensus viewpoint. Consensus learning assigns a set of input data to a group of classes by consolidating a set of underlying learn-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'09, October 19–24, 2009, Beijing, China.

Copyright 2009 ACM 978-1-60558-608-3/09/10 ...\$10.00.

ing models such that the output of the consolidated solution agrees as much as possible with the individual models. Figure 2 shows an example of consensus learning. Here different supervised and unsupervised learning models are built from various types of features. The labels obtained from these models are used as an input to the consensus learning algorithm.

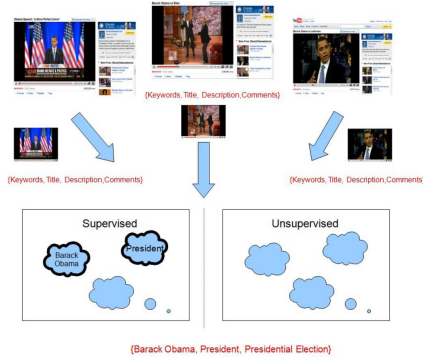


Figure 2: Consensus learning among different supervised and unsupervised models to assign multiple labels.

Currently this classification problem is challenging to be performed on user-generated videos because of the noisy and unstructured nature of the content. The semantic information submitted by users is also limited and sometimes absent. Additionally similar videos belonging to a particular category are annotated differently by different users. Some of the metadata such as YouTube comments are also irrelevant to a particular video as users may not even be discussing about the video.

In this paper, we propose *VideoMule*, a novel video classification algorithm which combines individual classification and clustering algorithms trained on textual metadata, audio and video through a heuristic consensus learning approach. Our video classification algorithm assigns multiple labels to the videos enabling a holistic semantic understanding as a particular video can belong to more than one category. We train individual learning algorithms (both supervised and unsupervised) on the videos and associated metadata, consolidate and negotiate between the individual models, and derive a strategy which assigns a particular video to multiple semantic labels.

Rest of the paper is organized as follows. In Section 2 we give an overview of the background research in this area, Section 3 gives a high-level system overview of *VideoMule*, Section 4 describes our algorithm while Section 5 provides experimental results and Section 6 concludes our work.

2. BACKGROUND AND RELATED WORK

Related work in the problem of multimodal video classification has been in: (a) approaches that train individual learning models on text, audio and video, and then combine those models using a single classifier, and (b) approaches which consider the entire video document (comprising audio, video and text) as a single feature vector.

Recently, people have begun working on combining different learning models. Research [8] has focused on video classi-

fication in web-shared videos. Here the authors propose an automatic mining framework which filters noisy tags, and uses a visual-based clustering approach to group together similar videos. In [9], a bipartite graph model was proposed for automatic topic discovery and tracking. A two step process with topic filtering and re-ranking is used in their bipartite graph model. However, their approach was focussed on single category and they focussed on mining structured data.

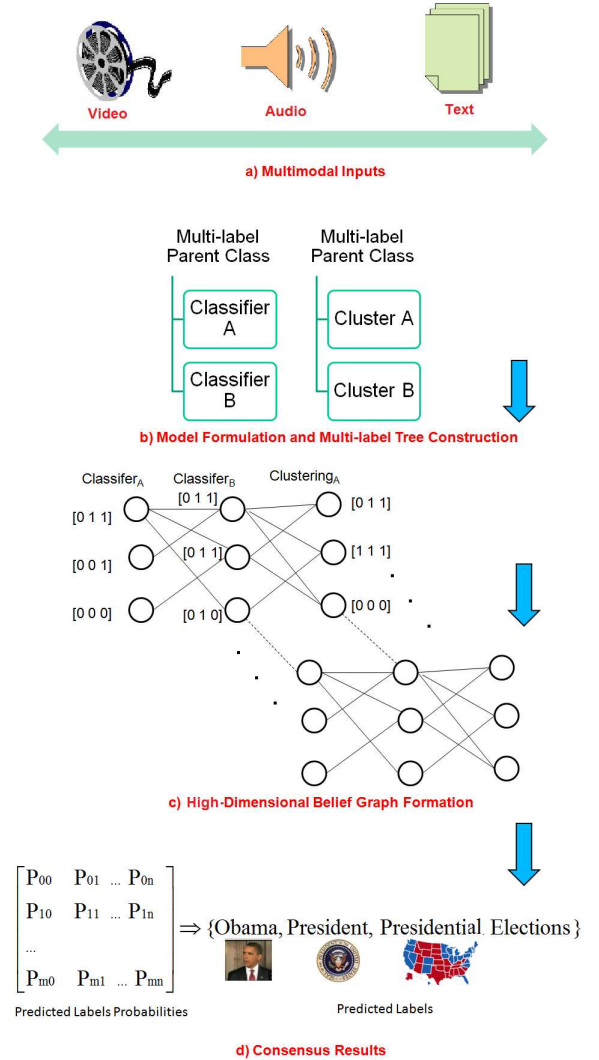


Figure 3: High-level system overview of *VideoMule*.

Algorithms that consider an entire document as a single feature vector, such as Multi-view learning [4] approach this problem from the aspect of co-training multiple complementary views. Other related work in this domain includes transfer learning ensemble, classifying webpages based on context and link information and so on. In [5], a more generic framework was proposed which considered heterogeneous sources for learning. This approach showed that a more consolidated framework would perform better than a mere combination of individual learning models. However, their methodology is quite different from ours.

3. SYSTEM OVERVIEW

Figure 3 gives a high-level system overview of VideoMule. This consists of four different steps: (a) Model formulation, (b) Multi-label tree construction, (c) High-dimensional belief-graph formation and (d) Consensus results.

Our objective in this work is to find the optimal multi-label class for a given test video through consensus learning. For training, the preliminary task is to extract features from text, audio and video content of a particular video. The feature vectors are then used as an input to several learning models, consisting of classifiers and clustering algorithms. Each learning model maps a particular video to a class label, or in the case of clustering assigns a cluster ID.

The learning models and the instances which are used for training are used for building a multi-label tree (in Figure 3 (b)). Each internal node of this tree represents a multi-label parent class containing a set of labels $L_n \subset L$, where L is the corpus of labels. Each leaf node represents the labels or cluster IDs. The task of the multi-label classifier is the prediction of one or more labels of its children in the tree.

Algorithm *VideoMule*

inputs: m -dimensional feature vectors from K videos

output: A set S of multi-label classes, with a class S_i for video i

begin

1. For each item i belonging to feature vector of video K_i
 2. Set $T = MultiLabelTree()$
 3. Endfor
 4. Set node pointer to root of Tree T
 5. For each node n from T
 6. Do breadth-first search on current level
 7. Add n to high-dimensional belief graph G
 8. Endfor
 9. For each pair of nodes belonging to G
 10. Compute their similarity based on Jaccard coefficient
 11. Add to Group Similarity Matrix M
 12. Compute Diagonal Matrix D as sum of each row of M
 13. Use M and D to form the equation for matrix Q [5]
 14. Endfor
 15. Solve for Q as an optimization problem
 16. For each row Q_i of Q
 17. Set element S_i of output set S to Q_i
 18. Endfor
- end**
-

Figure 4: *VideoMule* algorithm for multi-label classification.

Each learning model is then used for partitioning the training videos into a set of multi-label parent classes with each parent class containing a set of videos. These multi-label parent classes are used for forming a High-dimensional belief graph G [2] as shown in Figure 3 (c). The belief graph provides an efficient way of passing messages among nodes and hence is used in many learning algorithms. Graph G generated in VideoMule has a set of labeled and unlabeled parent classes, with the unlabeled ones obtained from clustering. Each labeled node in G is represented as a binary distribution vector with 1 denoting that an instance belongs to this parent, and 0 otherwise. Each unlabeled node is assigned a conditional probability vector. The task of the consensus algorithm [5] is to average each unlabeled node's probability

Algorithm *MultiLabelTree*

input: Input set of instances X and λ learning models

output: Tree T

begin

1. Set tree pointer to root node
 2. Check if Tree is NULL. If Yes Then
 3. Create Root Node
 4. Assign x to Root
 5. Endif
 6. For each instance x of X do
 7. Forward x to the child nodes
 8. For each learning model λ
 9. Assign x to a label l based on λ
 10. Endfor
 11. Assign $\bigcup_{i=1}^L l$ to multi-label parent class S_x
 12. Recursively call *MultiLabelTree()*
 13. Endfor
- end**
-

Figure 5: *MultiLabelTree* algorithm for constructing a multi-label tree.

distribution vector based on its neighboring nodes' probabilities. The final probability distributions represent the multi-label classes for the training data.

4. ALGORITHM

The overall algorithm for multi-label classification of videos is shown in Figure 4. The inputs to this algorithm are m -dimensional feature vectors from K videos. The output is a set S of multi-label classes. Each learning model λ generates a set of classes (or clusters) which are used for building a multi-label tree T with n nodes and a set of labels $L_n \subset L$. The algorithm for creating T is shown in Figure 5. Each multi-label parent class consists of children which contain a subset of the labels of their parent. Each instance from the set of videos is mapped to T , with instances belonging to a common multi-label parent grouped together at a particular level of the tree.

Each node from T is mapped to an high-dimensional belief graph with probability distributions. Each node of the belief graph is assigned with a probability vector V as: $V = [v_0 v_1 v_2 \dots v_m]$. For labeled nodes their corresponding vector elements are assigned with 0 or 1 values depending on whether a particular instance belongs to that multi-label parent class. For unlabeled instances, the elements of the vector are set to 0 initially. The last step is to propagate the probability distributions from labeled to unlabeled instances until the belief graph becomes stable (or achieves consensus). The probability vectors obtained when the graph is stable denote the set S of multi-label classes.

Each element x of the probability distribution vector is of the form $P(y = x|x, S_x)$ where y is the target label for x given a multi-label parent class S_x . This propagation process is modeled as an optimization problem where the objective function is to converge the set of all V and is explained in Figure 4 and Figure 5. This objective function uses a Group Similarity Matrix M with each element of this matrix denoting the distance of a node from its neighboring nodes as measured by a Jaccard Coefficient [6]. In [5], the

objective function is used to generate matrix Q of probabilities. The elements of Q denote the results of negotiation among all models with knowledge from the supervised learning models. In this way, a consensus decision is made for the multi-label classification of videos.

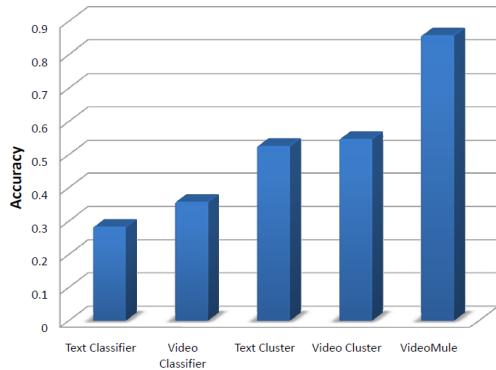


Figure 6: Accuracy of training on YouTube data.

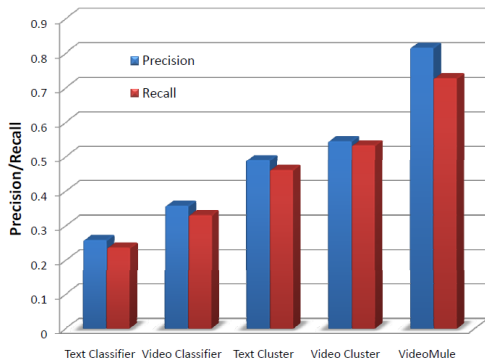


Figure 7: Precision and recall on YouTube data.

5. EXPERIMENTAL RESULTS

For our evaluation purposes we used videos obtained from YouTube by using the Google Data API ¹. We extracted an equal number of videos from each of the major categories in YouTube. We also extracted the title, description, comments, video responses, ratings, year, statistics and video category using this API. For extracting features from the video we used a spatio-temporal descriptor technique based on 3D gradients [7]. Each video has a 960-dimensional feature vector. For extracting the text features we used a ranking methodology based on the discriminative power of each keyword similar to TF-IDF [6]. We used the top-50 words by this ranking technique. Our evaluation methodology is based on comparing the precision, recall and accuracy of the VideoMule algorithm with individual classifiers and clustering algorithm on each mode of information.

Figure 6 shows the accuracy of training for VideoMule against individual text/video classifiers and clustering algorithms. For training using individual classifiers we used J48

¹<http://code.google.com/apis/youtube/overview.html>

decision tree classifiers packaged with Weka [6] since it provides support for multi-label classification. For clustering we used the k -means clustering algorithm [6]. The categories or labels were obtained from the ‘category’ section of each YouTube video, and they also provided the ground truth to verify our algorithm. From Figure 6 we can see that the accuracy values for VideoMule are much higher than the individual classifiers and clustering algorithms. A similar trend is seen in Figure 7 which shows the precision and recall values for the same set of algorithms.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have proposed VideoMule, a novel consensus learning approach to solving the problem of multi-label classification in noisy user-generated videos. We argue that the classification problem on such videos is a challenging task due to the noisy, unstructured and unreliable nature of such data. To aid in our classification process, we leverage the useful metadata present in such videos along with the video features to generate a consensus learning model. From our experimental results we show that VideoMule has a better accuracy, precision and recall than individual classifiers and clustering algorithms. In future work we plan to expand the evaluation process with data from other video-sharing websites. We also plan to compare the performance of VideoMule with other known multi-label classification algorithms.

7. REFERENCES

- [1] S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, and M. Aly. Video suggestion and discovery for youtube: taking random walks through the view graph. In *WWW*, pages 895–904, 2008.
- [2] C. M. Bishop. Chapter 8: Graphical models. In *Pattern Recognition and Machine Learning*. Springer., pages 359–418, 2006.
- [3] X. Cheng, C. Dale, and J. Liu. Understanding the characteristics of internet short video sharing: Youtube as a case study. In *CoRR abs*, 2007.
- [4] K. Ganchev, J. Graca, J. Blitzer, and B. Taskar. Multi-view learning over structured and non-identical outputs. In *UAI*, pages 204–211, 2008.
- [5] J. Gao, W. Fan, Y. Sun, and J. Han. Heterogeneous source consensus learning via decision propagation and negotiation. In *KDD*, pages 339–348, 2009.
- [6] J. Han and M. Kamber. Data mining: Concepts and techniques. In *Morgan Kauffman*, 2000.
- [7] A. Kläser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference 2008*.
- [8] L. Liu, Y. Rui, L.-F. Sun, B. Yang, J. Zhang, and S.-Q. Yang. Topic mining on web-shared videos. In *ICASSP*, pages 2145–2148, 2008.
- [9] L. Liu, L. Sun, Y. Rui, Y. Shi, and S. Yang. Web video topic discovery and tracking via bipartite graph reinforcement model. In *WWW*, pages 1009–1018, 2008.
- [10] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, pages 321–328, 2004.