Predicting Media Interestingness via Biased Discriminant Embedding and Supervised Manifold Regression

Yang Liu[1,2], Zhonglei Gu[1], Tobey H. Ko[3]
[1]Department of Computer Science, Hong Kong Baptist University, HKSAR, China
[2]Institute of Research and Continuing Education, Hong Kong Baptist University, Shenzhen, China
[3]Department of Industrial and Manufacturing Systems Engineering, University of Hong Kong, HKSAR, China
csygliu@comp.hkbu.edu.hk,cszlgu@comp.hkbu.edu.hk,tobeyko@hku.hk

ABSTRACT

In this paper, we describe our model designed for automatic prediction of media interestingness. Specifically, a two-stage learning framework is proposed. In the first stage, supervised dimensionality reduction is employed to discover the key discriminant information embedded in the original feature space. We present a new algorithm dubbed biased discriminant embedding (BDE) to extract discriminant features with discrete labels and use supervised manifold regression (SMR) to extract discriminant features with continuous labels. In the second stage, SVM is utilized for prediction. Experimental results validate the effectiveness of our approaches.

1 INTRODUCTION

Predicting the interestingness of multimedia content has long been studied in the psychology community [1, 6, 7]. More recently, we witness an explosion of multimedia content due to the accessibility of low cost multimedia creation tools, the automatic prediction of media interestingness thus started to attract attention in the computer science community because of its many useful applications to content providers, marketing, and managerial decision-makers.

In this paper, we propose to use dimensionality reduction to extract low-dimensional features for MediaEval 2017 Predicting Media Interestingness Task. Specifically, we propose a new algorithm called biased discriminant embedding (BDE) for discrete labels and utilize supervised manifold regression (SMR) [4] for continuous labels.

2 DIMENSIONALITY REDUCTION

2.1 Biased Discriminant Embedding

Given the data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$, where $\mathbf{x}_i \in \mathbb{R}^D$ denotes the feature vector of the i-th image or video, and label vector $\mathbf{l} = [l_1, l_2, ..., l_n]$, where $l_i \in \{0, 1\}$ denotes the corresponding label of $\mathbf{x}_i$, with 1 for interesting and 0 for non-interesting, biased discriminant embedding (BDE) aims to learn a $D \times d$ transformation matrix $\mathbf{W}$, which maximizes the *biased* discriminant information in the reduced subspace. The motivation for proposing the *biased* discrimination is that in media interestingness prediction, we are probably more interested in the *interesting* class than the *non-interesting* one. The objective function of BDE is given as follows:

$$\mathbf{W} = \arg\max_{\mathbf{W}} tr\left(\frac{\mathbf{W}^T\mathbf{S}^b\mathbf{W}}{\mathbf{W}^T\mathbf{S}^w\mathbf{W}}\right), \tag{1}$$

where $\mathbf{S}^w = \sum_{i,j=1}^n (N_{ij} \times l_i \times l_j)(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$ denotes the biased within-class scatter, $\mathbf{S}^b = \sum_{i,j=1}^n (N_{ij} \times |l_i - l_j|)(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$ denotes the biased between-class scatter, and $N_{ij} = exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma)$ measures the closeness between two data samples $\mathbf{x}_i$ and $\mathbf{x}_j$. The optimization problem could be solved by generalized eigen-decomposition.

2.2 Supervised Manifold Regression

Supervised manifold regression (SMR) [4] aims to find the latent subspace, where two data points should be close to each other if they possess similar interestingness levels. The objective function of SMR is given as follows:

$$\mathbf{W} = \arg\min_{\mathbf{W}} \sum_{i,j=1}^{n} \|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|^2 \cdot \left(\alpha S_{ij}^l + (1-\alpha)N_{ij}\right), \tag{2}$$

where $S_{ij}^l = |l_i - l_j|$ measures the similarity between the interestingness level of $\mathbf{x}_i$ and that of $\mathbf{x}_j$.

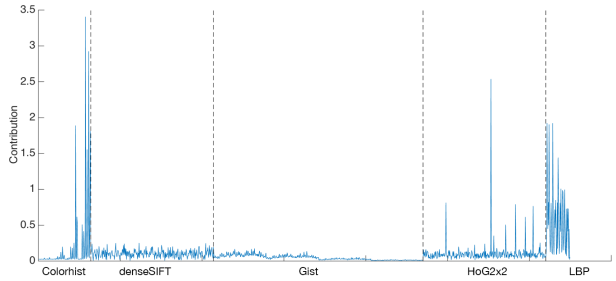
For each high-dimensional data point $\mathbf{x}_i$, we can obtain its low-dimensional representation by $\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i$. Then we apply SVM to $\mathbf{y}_i$ for interestingness prediction.

3 EXPERIMENTS

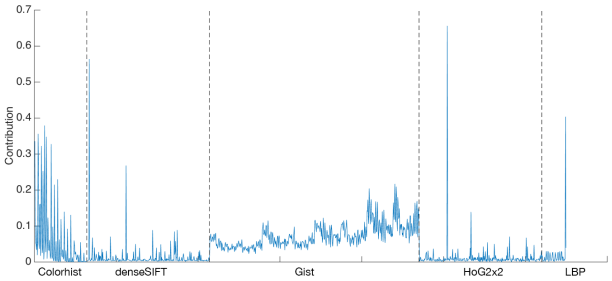
For each image data sample, we construct a 1299-D feature vector by selecting features from the feature set provided by the task organizers, including 128-D color histogram features, 300-D denseSIFT features, 512-D gist features, 300-D hog2×2, and 59-D LBP features. For the video data, we treat each frame as a separate image, and calculate the average and standard deviation over all frames in this shot, and thus we have a 2598-D feature set for each video. We normalize each dimension of the training data to the range $[0, 1]$ by $\hat{x} = \frac{x - x_{min}}{x_{max} - x_{min}}$ before dimensionality reduction, where x_{min} and x_{max} denote the minimum and maximum values in the corresponding dimension, respectively. Details about the dataset description can be found in [3].

For Run 1 of image data, we use the normalized 1299-D feature vector as the input of SVM. For Runs 2-5 of image data, we reduce the original data to the 23-D, 25-D, 26-D, 27-D subspaces via BDE (for discrete labels) and SMR (for continuous labels), respectively. For Run 1 of video data, we

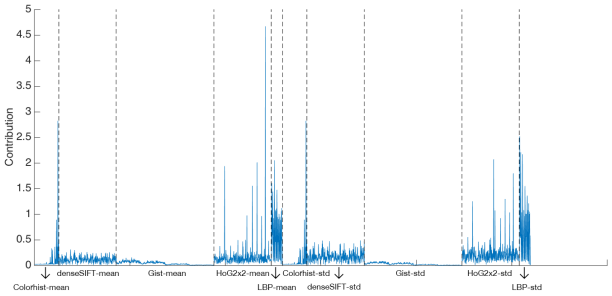
MediaEval'17, 13-15 September 2017, Dublin, Ireland



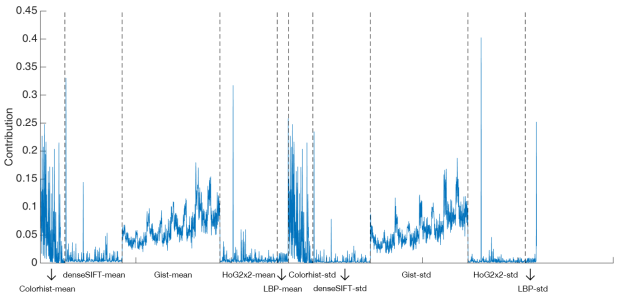
(a) BDE on image data



(b) SMR on image data



(c) BDE on video data



(d) SMR on video data

Figure 1: Contribution of each individual feature in image/video discrete/continuous prediction tasks.

Table 1: MAP@10 and MAP of the proposed model.

	Images		Videos	
	MAP@10	MAP	MAP@10	MAP
Run 1	0.1184	0.2812	0.0556	0.1813
Run 2	0.132	0.2916	0.0468	0.1761
Run 3	0.1332	0.2898	0.0468	0.1761
Run 4	0.1315	0.2884	0.0463	0.1742
Run 5	0.1369	0.291	0.0445	0.1746

use the normalized 2598-D feature vector as the input of SVM. For Runs 2-5 of video data, we reduce the original data to the 23-D, 25-D, 26-D, 27-D subspaces via BDE (for discrete labels) and SMR (for continuous labels), respectively. To predict the binary interestingness labels, we use ν-SVC [5] with an RBF kernel. We set $\nu = 0.1$ and $gamma = 100$ (for image data)/64 (for video data). To predict the continuous interestingness level, we use ϵ-SVR [2] with an RBF kernel. We set $cost = 1$, $\epsilon = 0.01$, and $\gamma = 1/D$. Table 1 reports the evaluation results of the proposed model provided by the task organizers. For image data, the reduced features perform better than the original ones, which indicates that the subspaces learned by BDE and SMR capture important information in terms of media interestingness. For video data, the performance of reduced features is slightly worse than that of the original ones. The reason might be that video data are more complex than image data so that such a low-dimensional representation cannot fully capture the key discriminant information embedded in the original space.

We further analyze the contribution of each dimension in the original feature space. The contribution of the i-th dimension is defined as $Contribution_i = \sum_j \lambda_j |w_{ij}|$, where λ_j denotes the j-th eigenvalue, w_{ij} denotes the (i,j)-th element of $\mathbf{W}$, and $|\cdot|$ denotes the absolute value operator. From Figures 1(a) and 1(c), we can observe that color histogram and LBP features contribute more than the others while the GIST features contribute the least in the discrete prediction task. In continuous prediction (Figures 1(b) and 1(d)), the color histogram and GIST features contribute the most among the five feature sets.

4 DISCUSSION AND OUTLOOK

This paper introduces our model designed for media interestingness prediction. For the future work, we aim to improve the performance of video interestingness prediction by incorporating the video temporal information. Moreover, as the ground truth (labels) of interestingness are provided by human beings, they generally vary with each individual and are somewhat subjective. We are therefore particularly interested in refining the human labeled ground truth (especially for continuous case) via machine learning technologies.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 61503317, and in part by the Faculty Research Grant of Hong Kong Baptist University (HKBU) under Project FRG2/16-17/032.

REFERENCES

[1] Daniel E. Berlyne. 1960. *Conflict, arousal and curiosity*. McGraw-Hill.

[2] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2 (2011), 27:1–27:27. Issue 3.

[3] C.-H. Demarty, M. Sjoberg, B. Ionescu, T.-T. Do, M. Gygli, and N. Q. K Duong. MediaEval 2017 Predicting Media Interestingness Task. In *Proc. of the MediaEval 2017 Workshop*. Dublin, Ireland, Sept. 13–15, 2017.

[4] Y. Liu, Z. Gu, and Y.-M. Cheung. Supervised Manifold Learning for Media Interestingness Prediction. In *Proc. of the MediaEval 2016 Workshop*. Hilversum, Netherlands, Oct. 20–21, 2016.

[5] Bernhard Schölkopf, Alex J. Smola, Robert C. Williamson, and Peter L. Bartlett. 2000. New Support Vector Algorithms. *Neural Comput.* 12, 5 (2000), 1207–1245.

[6] Paul J. Silvia. 2006. *Exploring the psychology of interest*. Oxford University Press.

[7] Craig Smith and Phoebe Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology* 48, 4 (1985), 813–838.