Supervised Manifold Learning for Media Interestingness Prediction

Yang Liu[1,2], Zhonglei Gu[3], Yiu-ming Cheung[1,2,4]
[1]Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong SAR, China
[2]Institute of Research and Continuing Education, Hong Kong Baptist University, Shenzhen, China
[3]AAOO Tech Limited, Shatin, Hong Kong SAR, China
[4]United International College, Beijing Normal University-Hong Kong Baptist University, Zhuhai, China
{csygliu,ymc}@comp.hkbu.edu.hk, allen.koo@aaoo-tech.com

ABSTRACT

In this paper, we describe the models designed for automatically selecting multimedia data, e.g., image and video segments, which are considered to be interesting for a common viewer. Specifically, we utilize an existing dimensionality reduction method called Neighborhood MinMax Projections (NMMP) to extract the low-dimensional features for predicting the discrete interestingness labels. Meanwhile, we introduce a new dimensionality reduction method dubbed Supervised Manifold Regression (SMR) to learn the compact representations for predicting the continuous interestingness levels. Finally, we use the nearest neighbor classifier and support vector regressor for classification and regression, respectively. Experimental results demonstrate the effectiveness of the low-dimensional features learned by NMMP and SMR.

1. INTRODUCTION

Effective prediction of media interestingness plays an important role in many real-world applications such as image/video search, retrieval, and recommendation [5–9, 12]. The *MediaEval 2016 Predicting Media Interestingness Task* requires participants to automatically select images and/or video segments which are considered to be the most interesting for a common viewer. The data used in this task are extracted from ca 75 movie trailers of Hollywood-like movies. More details about the task requirements as well as the dataset description can be found in [3].

Supervised manifold learning, which aims to discover the data-label mapping relation while capturing the manifold structure of the dataset, plays an important role in many multimedia content analysis tasks such as face recognition [4] and video classification [10]. In this paper, we aim to solve both image and video interestingness prediction via supervised manifold learning. There are two kinds of interestingness labels in the given task, i.e., discrete and continuous. For the case of discrete labels, we utilize an existing competitive dimensionality reduction method called Neighborhood MinMax Projections (NMMP) to extract the low-dimensional features from the original high-dimensional space. For the case of continuous labels, we propose a new dimensionality reduction method dubbed Supervised Manifold Regression (SMR) to learn the compact representations of the original data. Finally, we use nearest neighbor classifier and support vector regressor to predict the discrete and continuous labels of the given images/videos, respectively.

MediaEval 2016 Workshop, October 20-21, 2016, Hilversum, Netherlands

2. METHOD

2.1 Feature Extraction via NMMP and SMR

2.1.1 Neighborhood MinMax Projections

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 the label vector $\mathbf{l} = [l_1, l_2, ..., l_n]$, where $l_i \in \{0, 1\}$ denotes the corresponding label of $\mathbf{x}_i$, 1 for interesting and 0 for non-interesting, Neighborhood MinMax Projections (NMMP) aims to find a linear transformation, after which the nearby points within the same class are as close as possible, while those between different classes are as far as possible [11]. The objective function of NMMP is given as follows:

$$\mathbf{W} = \arg\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}} \frac{tr(\mathbf{W}^T\tilde{\mathbf{S}}_b\mathbf{W})}{tr(\mathbf{W}^T\tilde{\mathbf{S}}_w\mathbf{W})}, \tag{1}$$

where $tr(\cdot)$ denotes the matrix trace operator, $\mathbf{W}$ denotes the transformation matrix to be learned, $\tilde{\mathbf{S}}_b$ denotes the between-class scatter matrix defined on nearby data points, and $\tilde{\mathbf{S}}_w$ denotes the within-class scatter matrix defined on nearby data points. The optimization problem in Eq. (1) can be effectively solved by eigen-decomposition. More details of NMMP can be found in [11].

2.1.2 Supervised Manifold Regression

Different from the binary form in discrete case, the continuous interestingness label is a real number, i.e., $l_i \in [0, 1]$. The idea behind Supervised Manifold Regression (SMR) is simple: the more similar the interestingness levels of two media data, the closer the two feature vectors should be in the learned subspace. Meanwhile, we aim to preserve the manifold structure of the dataset in the original feature space. The objective function of SMR is formulated as follows:

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

where $S_{ij}^l = |l_i - l_j|$ measures the similarity between the interestingness level of $\mathbf{x}_i$ and that of $\mathbf{x}_j$ $(i, j = 1, ..., n)$, $S_{ij}^m = exp(-\frac{||\mathbf{x}_i - \mathbf{x}_j||_2^2}{2\sigma})$ denotes the similarity between $\mathbf{x}_i$

Table 1: Performance of proposed system (provided by the organizers)

	MAP	Accuracy	Precision	Recall	F-score
Run 1: Original image features ($D = 1299$)	0.1835	0.838	$[0.900, 0.139]$	$[0.922, 0.110]$	$[0.911, 0.123]$
Run 2: Reduced image features ($d = 100$)	0.1806	0.802	$[0.902, 0.134]$	$[0.874, 0.169]$	$[0.888, 0.149]$
Run 3: Original video features ($D = 2598$)	0.1552	0.828	$[0.901, 0.084]$	$[0.910, 0.076]$	$[0.905, 0.080]$
Run 4: Reduced video features ($d = 100$)	0.1733	0.834	$[0.902, 0.098]$	$[0.916, 0.084]$	$[0.909, 0.091]$

and $\mathbf{x}_j$ in the original space, and $\alpha \in [0, 1]$ denotes the balancing parameter, which is empirically set to be 0.5 in our experiments. Following some standard operations in linear algebra, the above optimization problem could be reduced to the following one:

$$\mathbf{W} = \arg\min_{\mathbf{W}} tr(\mathbf{W}^T \mathbf{X}\mathbf{L}\mathbf{X}^T \mathbf{W}), \tag{3}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \mathbb{R}^{D \times n}$ is the data matrix, $\mathbf{L} = \mathbf{D} - \mathbf{S}$ is the $n \times n$ Laplacian matrix [1], and $\mathbf{D}$ is a diagonal matrix defined as $D_{ii} = \sum_{j=1}^{n} S_{ij}$ $(i = 1, ..., n)$, where $S_{ij} = \alpha S_{ij}^l + (1 - \alpha) S_{ij}^m$. By transforming (2) to (3), the optimal $\mathbf{W}$ can be easily obtained by employing the standard eigen-decomposition.

2.2 Prediction via NN and SVR

2.2.1 Nearest Neighbor Classifier

Given the feature matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$ and the label vector $\mathbf{l} = [l_1, l_2, ..., l_n]$, for a new test data sample $\mathbf{x}$, its label l is decided by $l = l_{i*}$, where

$$i* = \arg\min_i ||\mathbf{x} - \mathbf{x}_i||_2 \tag{4}$$

2.2.2 Support Vector Regressor

To predict the continuous interestingness level, we use the ϵ-SVR [2]. The final optimization problem, i.e., the dual problem that ϵ-SVR aims to solve is:

$$\begin{aligned} \min_{\alpha, \alpha^*} \quad & \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T \mathbf{K}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) + \epsilon \mathbf{e}^T(\boldsymbol{\alpha} + \boldsymbol{\alpha}^*) + \mathbf{l}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \\ s.t. \quad & \mathbf{e}^T(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) = 0, \quad 0 \le \alpha_i, \alpha_i^* \le C, \quad i = 1, ..., n, \end{aligned} \tag{5}$$

where α_i, α_i^* are the Lagrange multipliers, $\mathbf{K}$ is a positive semidefinite matrix, in which $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ is the kernel function, $\mathbf{e} = [1, ..., 1]^T$ is the n-dimensional vector of all ones, and $C > 0$ is the regularization parameter. The level of a new sample $\mathbf{x}$ is predicted by:

$$l = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) K(\mathbf{x}_i, \mathbf{x}) + b. \tag{6}$$

3. EVALUATION RESULTS

In this section, we report the experimental settings and the evaluation results. For the image data, we construct a 1299-D feature set, including 128-D color hist 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.

- For Run 1, we use the 1299-D image feature vector as the input of each data sample.
- For Run 2, we first learn the 100-D subspaces of the original feature vector via NMMP (for discrete labels) and SMR (for continuous labels), respectively. After we obtain the transformation matrix $\mathbf{W} \in \mathbb{R}^{1299 \times 100}$, we define the contribution of the i-th dimension ($i = 1, ..., 1299$) of the original feature vector:

$$Contribution_i = \sum_j |w_{ij}|, \tag{7}$$

  where w_{ij} is the element in row i and column j of $\mathbf{W}$, and $|\cdot|$ denotes the absolute value operator. Then we select the features with $Contribution_i \ge 4$ to form the reduced feature space, the dimension of which is 117. We use this 117-D feature vector as the input of each data sample.
- For Run 3, we use the 2598-D video feature vector as the input of each data sample.
- For Run 4, we apply the same way used in Run 2 to select the most contributing features, the dimension of which is 140. We use this 140-D feature vector as the input of each data sample.

For each run, the NN classifier and ϵ-SVR are used to predict the discrete and continuous labels, respectively. For ϵ-SVR, we use RBF kernel with the default parameter settings from LIBSVM: $cost = 1$, $\epsilon = 0.1$, and $\gamma = 1/D$.

Table 1 reports the performance of the proposed system, which is provided by the organizers, on several standard evaluation criteria. For Precision, Recall, and F-score, the results follow the label order [non-interesting, interesting]. After dimensionality reduction, the performance of the reduced features is comparable to that of original features, which indicates that the reduced features capture most of the discriminant information of the dataset. Furthermore, we can observe that the performance on interesting data is not as good as that on non-interesting data. This might be caused by the imbalance between non-interesting (majority) and interesting (minority) data. Sampling techniques and cost-sensitive measures could therefore be utilized to further improve the performance.

4. CONCLUSIONS

In this paper, we have introduced our system for media interestingness prediction. The results shown that the features extracted by NMMP and SMR are informative. Our future work will focus on improving the system by considering the dynamic nature of the video data as well as exploring the technologies for learning imbalanced data.

Acknowledgments

The authors would like to thank the reviewer for the helpful comments. This work was supported in part by the National Natural Science Foundation of China under Grant 61503317.

5. REFERENCES

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[3] C.-H. Demarty, M. Sjoberg, B. Ionescu, T.-T. Do, H. Wang, N. Q. K. Duong, and F. Lefebvre. Mediaeval 2016 predicting media interestingness task. In *Working Notes Proceedings of the MediaEval 2016 Workshop*, Oct. 20-21, 2016, Hilversum, Netherlands.

[4] B. Ge, Y. Shao, and Y. Shu. Uncorrelated discriminant isometric projection for face recognition. In *Information Computing and Applications*, volume 307, pages 138–145. 2012.

[5] L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38(3), 2006.

[6] H. Grabner, F. Nater, M. Druey, and L. Van Gool. Visual interestingness in image sequences. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 1017–1026, 2013.

[7] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. V. Gool. The interestingness of images. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1633–1640, 2013.

[8] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1469–1482, 2014.

[9] Y.-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yang. Understanding and predicting interestingness of videos. In *Proceedings of The 27th AAAI Conference on Artificial Intelligence (AAAI)*, 2013.

[10] Y. Liu, Y. Liu, and K. C. Chan. Supervised manifold learning for image and video classification. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 859–862, 2010.

[11] F. Nie, S. Xiang, and C. Zhang. Neighborhood minmax projections. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 993–998, 2007.

[12] M. Soleymani. The quest for visual interest. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 919–922, 2015.