Ogata, H. et al. (Eds.) (2015). Proceedings of the 23rd International Conference on Computers in Education. China: Asia-Pacific Society for Computers in Education

Weighted Multi-view Clustering for Handwritten Numerals

Haidong GUO , Bo JIANG *, Feiyue QIU
College of Educational Science and Technology, Zhejiang University of Technology, China
*bjiang@zjut.edu.cn

Abstract: Many problems in educational data mining involve datasets that come from multiple different views or sources, which make the data mining task more challenging. However, most existing methods rely equally on every view, something lead to performance degradation in the case of incompatible views. In this work, we focus on a typical multi-view problem, the handwritten numerals clustering. In the proposed algorithm, each view is assigned a weight to express its importance and a simple yet efficient dynamical weight updating strategy is given.

Keywords: educational data mining, clustering, multi-view clustering, handwritten numerals clustering

1. Introduction

In educational data mining, data is often collected from multiple different sources, each of which reflects the hidden pattern of data from a specifically view. For example, in traditional educational data mining, students' academic performance can be predicted using their GPA, skill in programming, learning experiences and family financial status(Ibrahim & Rusli, 2007). In massive open online courses(MOOCs), various features are used for modeling learner engagement, such as behavioral features, linguistic features, temporal features and structural features(Ramesh, Goldwasser, Huang, & Daume III, 2014). These kinds of data are called *multi-view* data in data mining area. Multi-view data are instances that have multiple views from different feature spaces(Xiaojun Chen, 2013).

Handwritten numerals recognition plays important role in many real-world applications, such as automatic mail sorting and test paper management(Liu, Nakashima, Sako, & Fujisawa, 2003). In test paper management problem, the management system needs to recognize the student number and score automatically, both of which are handwritten in most cases. To improve recognition accuracy, multiple features of the handwritten numerals were extracted, such as the shape features, pixel features and morphological features(Liu et al., 2003). Observing that multiple views often provide complementary information, it is natural to utilize them to obtain better recognition performance rather than relying on single view. In this work, we propose a simple yet efficient multi-view clustering method. The proposed fuzzy weighting strategy is integrated into classical k-means algorithm, which has often been applied to solve large-scale data clustering problems. To solve the induced loss function, an alternative optimization strategy is developed to compute weight vectors and cluster centroids automatically.

2. The Proposed Method

Consider a dataset $\mathbf{X}$ with N unlabeled objects, D features and H views: $\mathbf{X} = \{X_i\}_{i=1}^{N}$ and $X_i = \{x_i^{(h)}\}_{h=1}^{H}$, where $x_i^{(h)} \in R^{d^{(h)}}$ and $\sum_{h=1}^{H} d^{(h)} = D$. The D features are divided into H views $\{T_h\}_{h=1}^{H}$. The *k*-means algorithm optimized the following Euclidean-based loss function:

$$J(Z,U) = \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{D} u_{ki}(z_{kj} - x_{ij})^2 \quad (1)$$

where $U = [u_{ki}]_{K\times N}$ is the hard partition matrix, $Z = [z_{kj}]_{K\times D}$ is the cluster centroids. To discriminate feature and view, a feature weight w and view weight v are assigned to each feature and view of the entire objects, respectively. The optimization model of this work is defined as follow.

$$\min J_{\alpha,\beta}(Z,W,V,U) = \sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{h=1}^{H}\sum_{j\in T_h} u_{ki} w_j^{\alpha} v_h^{\beta} (z_{kj} - x_{ij})^2 \quad (2)$$

$$s.t.\begin{cases} \sum_{k=1}^{K} u_{ki} = 1 & u_{ki} \in \{0,1\}, i = N_{1,K} \\ \sum_{j \in T_h} w_j = 1 & w_j \in [0,1], j = N_{1,D}, h = N_{1,H} \\ \sum_{h=1}^{H} v_h = 1 & v_h \in [0,1] \end{cases} \tag{3}$$

where $W = [w_j]_{1\times D}$ is the feature weight and $V = [v_h]_{1\times H}$ is the view weight. α and β, used to control the weight distributions, are two fuzzy exponents that need to be provided by user. To solve this optimization problem, an iterative algorithm that alternates between updating the clusters and computing the two weight vectors is proposed. The presented algorithm is called Simultaneous Weighting on View and Weight (SWVF), whose major steps of the proposed algorithm are as follows.

1) Updating Z for given as follow.

$$z_{kj} = \frac{\sum_{i=1}^{N} u_{ki} x_{ij}}{\sum_{i=1}^{N} u_{ki}} \tag{4}$$

2) Updating W for given V^*, U^* and Z^* as Theorem 1.

Theorem 1. Let $V = V^*, U = U^*$, $Z = Z^*$, $\alpha > 1$ and $\beta > 1$ are fixed, $J_{\alpha,\beta}(U^*, Z^*, W, V^*)$ is minimized if and only if

$$w_j = 1 / \sum_{j' \in T_h} \left(\frac{E_j}{E_{j'}} \right)^{\frac{1}{\alpha - 1}} \text{ if } \alpha > 1 \tag{5}$$

where $E_j = \sum_{k=1}^{K} \sum_{i=1}^{N} u_{ki} v_h^{\beta} (z_{kj} - x_{i,j})^2$.

3) Updating V for given U^*, Z^* and W^* as Theorem 2.

Theorem 2. Let $W = W^*, U = U^*$, $Z = Z^*$, $\alpha > 1$ and $\beta > 1$ are fixed, $J_{\alpha,\beta}(U^*, Z^*, W^*, V)$ is minimized if and only if

$$v_h = 1 / \sum_{h'}^{H} \left(\frac{D_h}{D_{h'}} \right)^{\frac{1}{\beta - 1}} \text{ if } \beta > 1 \tag{6}$$

where $D_h = \sum_{k=1}^{K} \sum_{i=1}^{N} \sum_{j \in V_h} u_{ki} w_j^{\alpha} (z_{kj} - x_{ij})^2$.

4) Updating U for given Z^*, W^* and V^*: When the cluster centroids, feature and view weights are known, the partitional matrix that minimize the intra-cluster distance is given by

$$u_{ki}^* = \begin{cases} 1 & \text{if } k = \arg\min_{k \in N_{1,K}} \left\{ \sum_{h=1}^{H} \sum_{j \in T_h} w_j^{\alpha} v_h^{\beta} (z_{kj} - x_{ij})^2 \right\} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

The above four steps are iteratively executed in SWVF to minimize the objective function (2). The Theorem 1 and 2 is easy to be proved using first-order gradient-descent theorem.

3. Application in Handwritten Numerals Clustering

3.1 Handwritten Numeral Dataset

This dataset consists of features of handwritten numerals (`0'--`9'). 200 patterns per class (for a total of 2,000 patterns) have been digitized in binary images. These digits are represented in terms of the following six feature subsets (views): 1. mfeat-fou: 76 Fourier coefficients of the character shapes; 2. mfeat-fac: 216 profile correlations; 3. mfeat-kar: 64 Karhunen-Love coefficients; 4. mfeat-pix: 240 pixel averages in 2 x 3 windows; 5. mfeat-zer: 47 Zernike moments; 6. mfeat-mor: 6 morphological features. In each file the 2000 patterns are stored in ASCI on 2000 lines. The first 200 patterns are of

class `0', followed by sets of 200 patterns for each of the classes `1' - `9'. Corresponding patterns in different feature sets (files) correspond to the same original character.

Figure 1 Example of the handwritten numerals

4. Experimental Results

Table 1 shows the clustering quality of the compared algorithms on handwritten numerals dataset. Table 1 also shows the results of t-test at the confidence level of 5% between the SWVF and each of the other algorithms. "+"and "−" indicate that the SWVF is significantly better and worse than the compared algorithm, respectively. "≈" indicates that the difference is not statistically significant. As it can be seen, SWVF outperforms most of the baselines in accuracy and Rand Index on the three datasets.

From Table 1, we can see that, quite a few approaches including LAC, WkMeans and TWKM are close to our fuzzy subspace clustering approach on both measures. Both TWKM and SWVF demonstrate encouraging clustering quality. SWVF performs slightly better than TWKM on accuracy measure. In addition, the difference is especially noticeable between EWKM and SWVF, where SWVF achieves about 32% and 15% quality gains in the two indicators.

Table 1 Clustering quality of the six algorithms on the handwritten numerals dataset

	LAC	EWKM	WkMeans	TWKM	MWKKM	SWVF
Rand Index	93.29(+)	79.79(+)	94.68(+)	95.29(≈)	89.85(+)	**95.29**
Accuracy	72.72(+)	48.01(+)	78.16(+)	79.93(≈)	61.75(+)	**80.10**

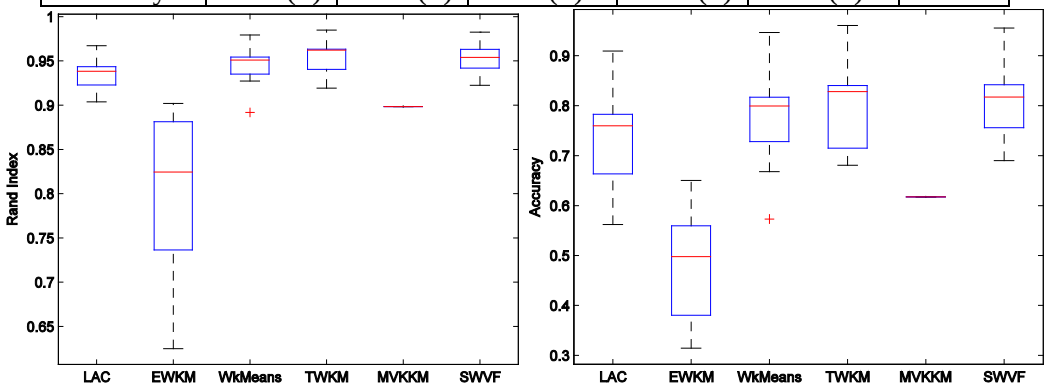

Figure 2 Box plots of the six algorithms on the handwritten numerals dataset

Acknowledgements

This work is partly supported by the National Natural Science Foundation of China under No. 61503340 and Scientific Research Fund of Zhejiang Educational Department under No.Y201432071.

References

Ibrahim, Z., & Rusli, D. (2007). *Academic Performance: Comparing Artificial Neural Network, Decision Tree and Linear Regression.* Paper presented at the Proceedings of the 21º Annual SAS Malaysia Forum, Kuala Lumpur, Malaysia.

Liu, C.-L., Nakashima, K., Sako, H., & Fujisawa, H. (2003). Handwritten digit recognition: benchmarking of state-of-the-art techniques. *Pattern Recognition, 36*(10), 2271-2285.

Ramesh, A., Goldwasser, D., Huang, B., & Daume III, H. G., Lise. (2014). *Learning Latent Engagement Patterns of Students in Online Courses.* Paper presented at the Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence.

Xiaojun Chen, X. X., J. Z. Huang, Yunming Ye. (2013). TW-k-means: Automated two-level variable weighting clustering algorithm for multiview dat. *IEEE Transactions on Knowledge & Data Engineering, 25*(4), 932-944.