Scienxt Journal of Computer Science & Information Technology

Year-2023|| Volume-1|| Issue-3|| pp. 11-20

Optimum variable weighting with clustering and mind map technique

[*1]A. Sasikala

Assistant Professor, Department of Computer Science,

PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India

[2]Dr. Ashok Kumar Singh

Assistant professor, (Computer Science)

Shiksha Snatak Mahavidyalaya, Mandhar, Raipur Chattisgarh

**Corresponding Author: A. Sasikala*

Email: shashisekar@gmail.com

Abstract:

Two additional steps are added to the iterative k-means clustering process to automatically compute the variable weights and view weights. We used two real-life data sets to investigate the properties of two types of weights in TW-k-means and investigated the difference between the weights of TW-k-means and the weights of the individual variable weighting method. The research has discovered the convergence property of the view weights in TW-k-means. We evaluate TW-k-means with five clustering algorithms on three real-life data sets and the results have shown that the TW-k-means algorithm significantly outperformed the other five clustering algorithms in four evaluation indices. In this proposed work we have done modification work with the two types of weights, compact views and significant variables can be identified and effect of low-quality views and noise variables can be reduced. Therefore, TW-k-means can obtain better clustering results than individual variable weighting clustering algorithms from multi view data. We discussed the difference of the weights between TW-k-means and EW-k means algorithms. The experiments also discovered the convergence property of the view weights in TW-k-means. We compared TW-k-means with five clustering algorithms on three real-life data sets and the results have shown that the TW-k-means algorithm significantly outperformed and also mind mapping technique also introduced for the multi view data with this we can easily maintain the user search data.

Keywords:

Tw K Means, Mapping, Variable Weighting, Ew K means.

1. Introduction:

Variable weighting clustering has been important research topic in cluster analysis. It automatically computes a weight for individual variable, and identifies significant variables and insignificant variables through variable weights. The multi view data could be considered as have two levels of variables. In clustering the Multi view data, the difference of views and the importance of individual variables in each view should be taken into account. The conventional variable weighting clustering methods only compute weights for individual variables and ignore the differences in views in the multi view data. Therefore, they are not suitable for multi view data. To our knowledge, SYNCLUS is the first variable weighting multi view clustering algorithm which uses weights for both views and individual variables in the clustering process.

But it only computes variable weights automatically and the view weights are given by users. Recently, Tzortzis and Likas proposed a weighted combination of exemplar-based mixture models (WCMM) that assigns different weights to the views and learns those weights automatically, but their method does not reflect on variable weights. The two algorithms have big limitations that they are not scalable to large data sets. Multi-view algorithms train two independent hypotheses which bootstrap by providing each other with labels for the unlabelled data. The training algorithms tend to maximize the agreement between the two independent hypotheses. Dasgupta have shown that the disagreement between two independent hypotheses is an upper bound on the error rate of one hypothesis, this observation explains at least some of the often remarkable success of multi-view learning. It also enhances to the question whether the multi-view approach can be used to improve clustering algorithms. Partitioning methods - such as k-Means, k-Medoids, and EM- and hierarchical, agglomerative methods are among the clustering approaches most frequently used in data mining. We study multi-view versions of these families of algorithms for document clustering.

2. Related Works:

2.1. Data Model:

Suppose we are given a data set with both numerical and categorical features. Standard k-means is designed to work with numerical data, and does not work well with categorical data. Hence, in our setting, at the very least, we would like to have two feature spaces. It is possible to further break-up numerical and categorical features into smaller feature spaces. However, we linearly scale each numerical feature (that is, we subtract the mean and divide by the square-root of the

variance) and use a 1-in-q representation for each q-ary definite feature. This makes all numerical features approximately homogeneous to each other and all categorical features roughly homogeneous to each other, thus obviating any need for auxiliary division. For the numerical aspect space, we use the squared-Euclidean distance. Assuming no missing values, all the categorical feature vectors have the same norm. We only retain the "direction" of the categorical feature vectors, that is, we normalize each categorical feature vector to have a unit Euclidean norm, and apply the cosine distance. Clustering algorithms can be divided into two categories: generative (or model-based) approaches and discriminative approaches.

Model-based approaches attempt to learn generative models from the documents, with each model representing one cluster. Usually generative clustering approaches are based on the Expectation Maximization algorithm. The EM algorithm is an iterative statistical technique for maximum likelihood estimation in settings with incomplete data. Given a model of data generation, and data with some omitted values, EM will locally exploit the probability of the model parameters and give estimates for the missing values. Similarity-based clustering approaches optimize an objective function that involve the pair wise document similarities, aiming at maximizing the average similarities within clusters and minimize the average similarities between clusters. Most of the similarity based clustering algorithms follow the hierarchical agglomerative approach, where a dendrogram is build up by iteratively merging closest examples/ clusters.

2.2. Multi-View EM Clustering:

In this section we want to analyze whether we can extend EM based cluster algorithms, so that they incorporate the multi-view setting with independent views. Different EM applications differ in specific models. We focus on models that are suitable for document clustering. Gaussian models could be used for multi-view EM as well, but are not applicable for document clustering. We firstly describe the general EM algorithm extended for two views, and then we describe two instances of this algorithm and present and analyze empirical results.

2.3. General Multi View EM Algorithm:

In the field of semi-supervised learning, co-EM based methods Positive results on the co-EM algorithm for the problem of semi-supervised learning lead to the question whether co-EM can improve on EM for unsupervised learning setting as well. The co-EM algorithm, in each iteration i, each view v finds the model parameters £ (v) i which maximize the likelihood given the expected values for the hidden variables of the other view. In turns M, E steps in view one

and M, E steps in view two are executed. The single expectation and maximization steps are equivalent to the E and M steps of the original EM algorithm. The algorithm is not guaranteed to converge. Our experiments show that the algorithm often does not converge. As displayed in Table. 1, we do not run the algorithm until convergence but until a special stopping criterion is met.

Table 1. Multi-View EM.

Input: Unlabeled data $D = \{(x_1^{(1)}, x_1^{(2)}), \ldots, (x_n^{(1)}, x_n^{(2)})\}$.

1. Initialize $\Theta_0^{(2)}$, T, $t = 0$.
2. E step view 2: compute expectation for hidden variables given the model parameters $\Theta_0^{(2)}$
3. Do until stopping criterion is met:
   (a) For $v = 1 \ldots 2$:
      i. $t = t + 1$
      ii. M step view v: Find model parameters $\Theta_t^{(v)}$ that maximize the likelihood for the data given the expected values for the hidden variables of view $\bar{v}$ of iteration $t - 1$
      iii. E step view v: compute expectation for hidden variables given the model parameters $\Theta_t^{(v)}$
   (b) End For v.
4. return combined $\hat{\Theta} = \Theta_{t-1}^{(1)} \cup \Theta_t^{(2)}$

3. Results and Discussion:

As an example, Fig.1 draws the average clustering accuracies of six clustering algorithms on the Multiple Features data set in the first experiment. From these results, we can observe that TW-k-means produced better results with large value of than the other five algorithms. When _ was large, it produced relatively stable results with the change of. WCMM produced the worst results, which indicates that WCMM failed to recover the clusters from this high-dimensional multiview data. EWKM produced unstable and worse results than W-k-means, LAC and TW-k-means. EW-k-means produced similar results as W-k-means, which indicates that the regularization term affects the result not too much. In the second experiment, we set the parameter values of six clustering algorithms as shown in Table 2.

Table. 2: Parameter values of six clustering algorithms in the experiments on the three real-life data sets

Algorithms	***MF***	***IA***	***IS***
W-*k*-means (β)	8	10	30
EW-k-means (ŋ)	20	1	30
LAC (*h*)	1	15	30
EWKM(λ)	5	40	30
WCMM (*a*)	1.5	4	1
TW-*k*-means (λ, ŋ)	(30,7)	(80,25)	(70,40) ββ

MF: the multiple features data set, IA: the internet advertisement data set, IS: the image segmentation data set

Table.3 summarizes the total 1,503 clustering results. From these results, we can see that TW-k-means significantly outperformed the other five algorithms in almost all results, especially on the Multiple Features and Internet Advertisement data sets. Although TW-k-means is an extension to EW-k means, the introduction of weights on views improved its results. WCMM produced the worst results on all three data sets. To sum up, TW-k-means is superior to the other five clustering algorithms in clustering multiview data.

TABLE 3
Summary of Clustering Results on Three Real-Life Data Sets by Six Clustering Algorithms

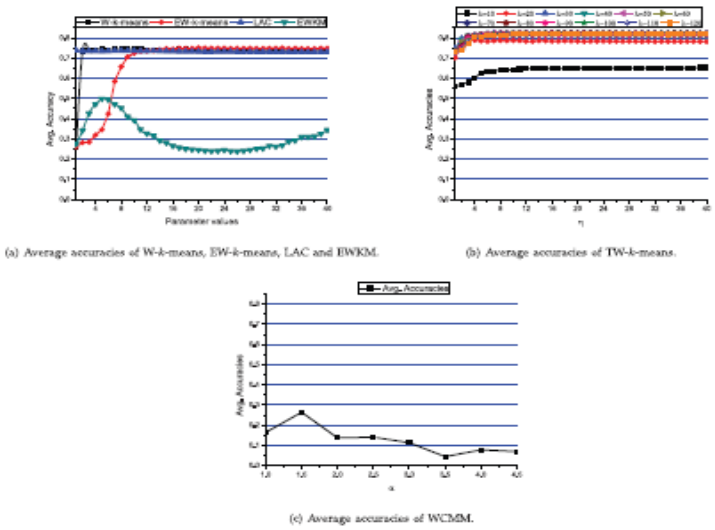
Data	Evaluation indices	W-*k*-means	EW-*k*-means	LAC	EWKM	WCMM	TW-*k*-means
MF	Precision	−0.06(.10)∗	−0.07(.10)∗	−0.07(.09)∗	−0.24(.08)∗	−0.59(.00)∗	0.79(.09)
	Recall	−0.09(.09)∗	−0.09(.09)∗	−0.09(.08)∗	−0.36(.12)∗	−0.56(.00)∗	0.82(.08)
	F-measure	−0.08(.10)∗	−0.08(.10)∗	−0.08(.08)∗	−0.41(.12)∗	−0.59(.00)∗	0.80(.09)
	Accuracy	−0.09(.09)∗	−0.09(.09)∗	−0.09(.08)∗	−0.36(.12)∗	−0.56(.00)∗	0.82(.08)
IA	Precision	−0.16(.19)∗	−0.16(.20)∗	−0.14(.20)∗	−0.22(.19)∗	−0.56(.00)∗	0.72(.12)
	Recall	−0.14(.04)∗	−0.10(.07)∗	−0.10(.08)∗	−0.13(.06)∗	−0.33(.00)∗	0.72(.07)
	F-measure	−0.23(.04)∗	−0.17(.12)∗	−0.17(.12)∗	−0.21(.09)∗	−0.47(.00)∗	0.69(.11)
	Accuracy	−0.14(.04)∗	−0.10(.07)∗	−0.10(.08)∗	−0.13(.06)∗	−0.33(.00)∗	0.72(.07)
IS	Precision	−0.03(.07)∗	−0.04(.08)∗	−0.03(.07)∗	−0.03(.09)∗	−0.37(.00)∗	0.62(.09)
	Recall	−0.03(.05)∗	−0.03(.03)∗	−0.03(.05)∗	−0.03(.05)∗	−0.41(.00)∗	0.64(.05)
	F-measure	−0.01(.07)∗	−0.02(.05)∗	−0.01(.07)∗	−0.02(.07)∗	−0.40(.00)∗	0.60(.07)
	Accuracy	−0.03(.05)∗	−0.03(.03)∗	−0.03(.05)∗	−0.03(.05)	−0.41(.00)∗	0.64(.05)

The value of the TW-k-means algorithm is the mean value of 100 results and the other values are the differences of the mean values between the corresponding algorithms and the TW-k-means algorithm. The value in brackets is the standard deviation of 100 results. "" indicates that the difference is significant.

3.1. Performance Metrics:

We set the parameter values of four clustering algorithms as 30 integers from 1 to 30. For TW-k-means, we set _ as 30 integers from 1 to 30 and _ as 12 values of f10; 20; 30; 40; 50; 60; 70;

80; 90; 100; 110; 120g. Since the clustering results of the five clustering algorithms excluding WCMM were affected by the initial cluster centers, we randomly generated 100 sets of initial cluster centers for each data set. For each parameter setting, we ran each of the five clustering algorithms to produce 100 clustering results on each of the three data sets. For WCMM, we set _ as eight values 1; 1:5; 2; 2:5; 3; 3:5; 4; 4:5g. Since WCMM can find global optima, we only ran WCMM once. In the second experiment, we first set the parameter values for six algorithms by selecting those with the best results in the first experiment. Similar to the first experiment, we produced 100 results for each of the five clustering algorithms excluding WCMM and 1 result for WCMM on each data set. In order to compare the classification performance, we used precision, recall, f-measure and accuracy to evaluate the results [35]. Precision is calculated as the fraction of correct objects among those that the algorithm believes belonging to the relevant class. Recall is the fraction of actual objects that were identified.



(a) Average accuracies of W-k-means, EW-k-means, LAC and EWKM. (b) Average accuracies of TW-k-means.

(c) Average accuracies of WCMM.

Figure. 1: The clustering results of six clustering algorithms versus their parameter values on the Multiple Features data set

F-measure is the harmonic mean of precision and recall and accuracy is the proportion of correctly classified objects. All four indices use the corresponding actual classification as the reference classification.

To statistically compare the clustering performance, the paired t-test comparing TW-k-means with the other five clustering methods was computed from each of the four evaluation indices. If the p-value was below the threshold of the statistical significance level (usually 0.05), then the null hypothesis was rejected in favor of an alternative hypothesis, which typically states that the comparing two distributions do differ. Thus, if the p-value of two approaches was less

than 0.05, the difference of the clustering results of the two approaches was considered to be significant, otherwise, insignificant.

4. Conclusion and Future Work:

TW k- means can compute weights for views and individual variables simultaneously in the clustering process. With the two types of weights, compact views and significant variables can be identified and effect of low-quality views and noise variables can be reduced. Therefore, TW-k-means can obtain better clustering results than individual variable weighting clustering algorithms from multi view data. We used two real-life data sets to investigate the properties of two types of weights in TW-k-means. We discussed the difference of the weights between TW-k-means and EW- k means algorithms. The research also discovered the convergence property of the view weights in TW-k-means. We compared TW-k-means with five clustering algorithms on three real-life data sets and the results have shown that the TW-k-means algorithm significantly outperformed the other five clustering algorithms in four evaluation indices. As such, it is a new variable weighting method for clustering of multi view data. In the future, we will combine the two-level variable weighting method with other techniques such as fuzzy techniques, subspace clustering techniques, semi-supervised techniques etc. so as to apply our method to more applications. Moreover, we will investigate approaches that can automatically group variables in the clustering process.

5. References:

(1) J. Mui and K. Fu, "Automated Classification of Nucleated Blood Cells Using a Binary Tree Classifier," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 2, no. 5, pp. 429-443, May 1980.

(2) J. Wang, H. Zeng, Z. Chen, H. Lu, L. Tao, and W. Ma, "ReCoM: Reinforcement Clustering of multiType Interrelated Data Objects," Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Informaion Retrieval, pp. 274-281, 2003.

(3) S. Bickel and T. Scheffer, "Multi-view Clustering," Proc. IEEE Fourth Int'l Conf. Data Mining, pp. 19-26, 2004.

(4) K. Kailing, H. Kriegel, A. Pryakhin, and M. Schubert, "Clustering Multi-Represented Objects with Noise," Proc. Eighth Pacific-Asia Conf. Knowledge Discovery and Data

Mining, H. Dai, R. Srikant, and C. Zhang, eds., vol. 3056, pp. 394-403, Springer Berlin/Heidelberg, 2004.

(5) V.R. de Sa, "Spectral Clustering with Two Views," Proc. IEEE 22nd Int'l Workshop Learning with Multiple Views (ICML), pp. 20-27, 2005.

(6) D. Zhou and C. Burges, "Spectral Clustering and Transductive Learning with Multiple Views," Proc. 24th Int'l Conf. Machine Learning, pp. 1159-1166, 2007.

(7) M.B. Blaschko and C.H. Lampert, "Correlational Spectral Clustering," Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR '08), pp. 1-8, 2008.

(8) K. Chaudhuri, S. Kakade, K. Livescu, and K. Sridharan, "Multiview Clustering via Canonical Correlation Analysis," Proc. 26th Ann. Int'l Conf. Machine Learning, pp. 129-136, 2009.

(9) G. Tzortzis and C. Likas, "Multiple View Clustering Using a Weighted Combination of Exemplar-Based Mixture Models," IEEE Trans. Neural Networks, vol. 21, no. 12, pp. 1925-1938, Dec. 2010.

(10) B. Long, P. Yu, and Z. Zhang, "A General Model for Multiple View Unsupervised Learning," Proc. Eighth SIAM Int'l Conf. Data Mining (SDM '08), 2008.