TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 12, No. 9, September 2014, pp. 6993 ~ 6998
DOI: 10.11591/telkomnika.v12i9.4845

A Robust *k*-Means Type Algorithm for Soft Subspace Clustering and its Application to Text Clustering

Tiantian Yang, Jun Wang*
School of Digital Media, Jiangnan University
*Corresponding author, e-mail: wangjun_sytu@hotmail.com

Abstract

Soft subspace clustering are effective clustering techniques for high dimensional datasets. In this work, a novel soft subspace clustering algorithm RSSKM are proposed. It is based on the incorporation of the alternative distance metric into the framework of k-means type algorithm for soft subspace clustering and can automatically calculates the feature weights of each cluster in the clustering process. The properties of RSSKM are also investigated. Experiments on real world text datasets are conducted and the results show that RSSKM outperformed some popular clustering algorithms for text mining, while still maintaining efficiency of the k-means clustering process.

Keywords*: k-means, soft subspace clustering, text clustering*

1. Introduction

Recently, subspace clustering has become an effective data mining tool for high dimensional text data. It pursues two tasks, locating the subspaces in which clusters can be found and discovering the clusters from different subspaces. According to the ways in which the subspaces are identified, subspace clustering can be classified into two categories. The first category, which is called hard subspace clustering, is to locate the exact subspaces of different clusters [1-4]. In the algorithms of this category, the membership of a feature belonging to one cluster is identified by a binary value. The second category is to cluster data objects in the entire data space but assign different weights to different features of clusters in the clustering process, based on the importance of the features in identifying the corresponding clusters [5, 6]. We call these methods soft subspace clustering.

Subspace clustering techniques need to compute the cluster memberships of data objects and the subspace of each cluster simultaneously [14], which throws a key challenge to researchers. Up to now, many subspace clustering algorithms have been presented and they have become effective and powerful methods in clustering high dimensional text data. However, most of them are still sensitive to noisy data. In this study, we will develop a novel robust *k*-means type clustering algorithm under the framework of soft subspace clustering. By incorporating the alternative distance metric [7], the robust statistics is incorporated into the soft subspace clustering algorithms, which makes the algorithm suitable for the high dimensional sparse data and insensitive to the noise in the dataset.

The rest of the paper is organized as follows. In Section II, we review some popular soft subspace clustering algorithms. In section III, we propose our algorithm RSSKM and study its convergence property and robustness. In section IV, we show the experimental results and verify its priority over some representative clustering algorithms.

2. Related Works

Recently, many soft subspace clustering algorithms have been proposed. Generally speaking, most of them can be unified as the problem of finding the local minimum of the objective function.

$$J(\mathbf{U}, \mathbf{W}, \mathbf{V}) = f\left(\sum_{h=1}^{s} w_{ih}{}^{\sim} \left(x_{kh} - v_{ih}\right)^2\right) + \mathrm{H} \quad (1)$$

Received October 18, 2013; Revised Apr 26, 2014; Accepted May 10, 2014

Under the constraints $\sum_{i=1}^{c} u_{ik}$ =1 and $\sum_{h=1}^{d} w_{ih}$ =1. For most existing algorithms, the first term $f\left(\sum_{h=1}^{s} w_{ih}^{\sim}\left(x_{kh}-v_{ih}\right)^2\right)$ is interpreted as the total weighted distance of each data object to cluster centers and often computed as $\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}\sum_{h=1}^{s} w_{ih}^{\sim}\left(x_{kh}-v_{ih}\right)^2$. The second term H is a penalty term which is used to enhance the performance of the clustering algorithm. According to different forms of H, many soft subspace clustering algorithms are proposed in literatures, typical representatives of them are AWA [5], FWKM [8], FSC [9][10], EWKM[11], and COSA [12].

By inspecting these algorithms, it is clear that all the cluster centers along with each feature are computed as:

$$\mathbf{v}_{ih}^{*}=\frac{\sum_{k=1}^{n} u_{ik}\mathbf{x}_{kh}}{\sum_{k=1}^{n} u_{ik}},\ I=1,2,\ldots,c,\ ,\ h=1,2,\ldots,s. \tag{2}$$

Equation (2) implies that each data point belonging to cluster *i* has equal weight 1 even though a data point is far away from other data points, which makes the cluster center heavily affected by the noisy data. In order to make the cluster centers more robust, we should give a smaller weight to those noisy data and a large weight to those compact data in the dataset. In order to achieve this goal, we use the following Equation (3) as the distance function:

$$d=\sqrt{1-\exp\left(-\ \left\|\mathbf{x}_1-\mathbf{x}_2\right\|^2\right)} \tag{3}$$

Obviously, it satisfies the following conditions [13]:

(1) $d(x,y)>0,\ \forall\ x\neq y,\ d(x,x)=0$;

(2) $d(x, y) = d(y, x)$;

(3) $d(x, y) \leq d(x, z) + d(z, y),\ \forall\ z$.

After incorporating Equation (3) into the framework of soft subspace clustering, Equation (2) can be modified as:

$$v_{ih}^{*}=\frac{\sum_{k=1}^{n} u_{ik}x_{kh}\exp\left(-\ _{h}\left(x_{kh}-v_{ih}\right)^2\right)}{\sum_{k=1}^{n} u_{ik}\exp\left(-\ _{h}\left(x_{kh}-v_{ih}\right)^2\right)}\ I=1,2,\ldots,c,\ h=1,2,\ldots,s \tag{4}$$

As we expect, it assigns larger weights to the data objects which are closer to cluster centers $\mathbf{v}_i$ and smaller weights to those far away from **v**. Thus algorithm will be more robust if the distance function Equation (3) is utilized.

3. Algorithm RSSKM

By incorporating Equation (2) into the framework of soft subspace clustering, we consider a novel algorithm named RSSKM with the following objective function:

$$J(\mathbf{U},\mathbf{W},\mathbf{V})=\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}\sum_{h=1}^{s} w_{ih}^{\sim}\left(1-\exp\left(-\ _{h}\left(x_{kh}-v_{ih}\right)^2\right)\right) \tag{5}$$

Under the following constraints:

$$u_{ik} \in \{0,1\}, i = 1, 2, \ldots, c, k = 1, 2, \ldots, n \quad (6a)$$

$$\sum_{i=1}^{c} u_{ik} = 1, k = 1, 2, \ldots, n \quad (6b)$$

$$0 < \sum_{k=1}^{n} u_{ik} < n, I = 1, 2, \ldots, c \quad (6c)$$

And,

$$0 \le w_{ih} \le 1, i = 1, 2, \ldots, c, h = 0, 1, \ldots, s \quad (7a)$$

$$\sum_{h=1}^{s} w_{ih} = 1, I = 1, 2, \ldots, c. \quad (7b)$$

In which $c \ge 2$ is the cluster number, n is the number of data points. In Equation (5), $\sqrt{1-\exp\left(-\ _h\left(x_{kh}-v_{ih}\right)^2\right)}$ is used to compute the distance between data point $\mathbf{x}_k$ and cluster center $\mathbf{v}_i$ along with the hth feature.

Similarly, the objective function of RSSKM can be minimized by iteratively solving the following three minimization problems:

(a) Problem P1: Fix $\mathbf{W}=\mathbf{W}^*$, $\mathbf{V}=\mathbf{V}^*$, solve the reduced problem that minimizes $J_m(\mathbf{W}^*, \mathbf{V}^*, \mathbf{U})$ under the constraint Equation (6).

(b) Problem P2: Fix $\mathbf{V}=\mathbf{V}^*$, $\mathbf{U}=\mathbf{U}^*$, solve the reduced problem that minimizes $J_m(\mathbf{W}, \mathbf{V}^*, \mathbf{U}^*)$ under the constraint Equation (7);

(c) Problem P3: Fix $\mathbf{W}=\mathbf{W}^*$, $\mathbf{U}=\mathbf{U}^*$, solve the reduced problem that minimizes $J_m(\mathbf{W}^*, \mathbf{V}, \mathbf{U}^*)$;

By using Lagrange mulipliers, Problem P1 is solved by:

$$v_{ih}^* = \frac{\sum_{k=1}^{n} u_{ik} x_{kh} \exp\left(-\ _h\left(x_{kh}-v_{ih}\right)^2\right)}{\sum_{k=1}^{n} u_{ik} \exp\left(-\ _h\left(x_{kh}-v_{ih}\right)^2\right)} \quad I = 1, 2, \ldots, c, h = 1, 2, \ldots, s \quad (8)$$

In order to compute the cluster centers $\mathbf{V}$ with Equation (8), the fix-point iteration should be employed. It is an inefficient and time-consuming process. However, according to our experiment, we observe that it is enough to estimate the cluster centers $\mathbf{V}$ with one step, which can also achieve ideal results. This modification makes RSSKM convergent to its local minimum value faster.

Similar with Problem P1, Problem P2 is solved by:

$$w_{ih}^* = \frac{1}{\sum_{l=1}^{s}\left(\frac{D_{ih}}{D_{il}}\right)^{\frac{1}{\sim -1}}}, I = 1, 2, \ldots, c, h = 1, 2, \ldots, s, \quad (9)$$

where $D_{ih} = \sum_{k=1}^{n} u_{ik}\left(1-\exp\left(-\ _h\left(x_{kh}-v_{ih}\right)^2\right)\right)$,

And Problem P3 is solved by:

$$u_{ik}^* = \begin{cases} 1 & \sum_{h=1}^{s} w_{ih}^{\sim}\left(1-\exp\left(-\ _h\left(x_{kh}-v_{ih}\right)^2\right)\right) \\ & \le \sum_{h=1}^{s} w_{lh}^{\sim}\left(1-\exp\left(-\ _h\left(x_{kh}-v_{lh}\right)^2\right)\right), l = 1, 2, \ldots, c \\ 0 & otherwise \end{cases} \quad (10)$$

Now, the proposed algorithm RSSKM is given as follows:

Algorithm RSSKM
Input: The number of clusters *c*, parameter *y* Randomly select *c* cluster centers and set all initial weights to 1/*s*; REPEAT Update the partition matrix **U** with Eq.(10)); Update the feature weights matrix **W** with Eq.(9); Update the cluster centers **V** with Eq.(8) using the fixed-point iteration; UNTIL (the objective function obtains its local minimum value); Output: The partition matrix **U** and feature weights matrix **W**.

After a finite number of iterations, RSSKM algorithm converges to the local minimal of the objective function. Using Equation (8), Equation (9) and Equation (10), we can show that the sequence $J^{(t)}(\mathbf{U},\mathbf{W},\mathbf{V})$ generated by Eq.(5) decreases strictly. Meanwhile, we can also observe that each possible partition **U** only occurs once in the clustering process. Thus, RSSKM algorithm converges in a finite number of iterations [10].

Assuming *s* is the number of features, *n* is the number of data objects and *c* is the number of clusters, the computational complexity of RSSKM per iteration is O(*snc*). On the other hand, we need O(*ns*) space to store *n* data points, O(*cs*) space to store *c* cluster centers **V**, O(*cs*) space to store the feature weight matrix **W** and O(*cn*) space to store the partition matrix **U**. That is to say, both the computational complexity and the storage complexity of RSSKM are linearly dependent on the number of data objects when the number of features is fixed. Thus, the proposed RSSKM algorithm is well suitable for high dimensional datasets and to large scaled datasets.

The proposed RSSKM is different from some previous work on soft subspace clustering. In this work, we incorporate the alternative distance into the framework of *k*-mean type algorithms for soft subspace clustering, which makes the data points far away from the cluster centers have smaller weights. Thus RSSKM will be more robust and the performance of RSSKM on noisy dataset is improved.

4. Experiments

In this section, we present the clustering results obtained by RSSKM on the well-known text datasets *20 Newsgroups*, which was a publicly available dataset from website *http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html*. In our experiment, the original text data was first preprocessed to strip the news messages from the e-mail headers and special tags and eliminate the stop words and stem words to their root forms. Then, 1000 features were selected according to their IDF (inverse document frequency) values. In order to improve the performance of the tested algorithms, the dataset was further processed using the following *tf-idf* formulas:

$$\begin{aligned} tf_{ij} &= \frac{n_{ij}}{\sum_k n_{kj}} \\ idf_i &= \log \frac{|D|}{\left|\{d : d \ni t_i\}\right|} \\ tfidf_{ij} &= tf_{ij} \bullet idf_i \end{aligned} \tag{11}$$

Where n_{ij} denoted the term frequency of term t_i in document d_j, $|D|$ denoted the total number of the documents in dataset, $\left|\{d : d \ni t_i\}\right|$ denoted the number of documents in which the term t_i appeared.

In our experiment, 6 datasets were established from the *20 Newsgroups*. As can be seen from Table 1, these datasets are divided into two series: series A and series B. The categories in series A are more semantically different than that in series B. The datasets of each series are generated incrementally by adding two more categories to the former dataset, resulting to name them NG20-A2 (B2), NG20-A4 (B4) and NG20-A6 (B6) accordingly. The

number after the series code shows the number of categories in this dataset. For example, A2 denotes a dataset in series A with two categories in it. Each category contains n_{doc} documents which were chosen randomly from the original *20 Newsgroups* dataset.

In our experiment, we evaluate them by the performance index *RandIndex*, which was computed as follows:

$$RI = \frac{a+d}{a+b+c+d}. \tag{12}$$

In which the value of $a+b$ can be interpreted as the total pairs predicted in the same cluster and the value of $a+d$ can be interpreted as the total pairs in the in the same class.

Table 1. Two Series of Newsgroup Datasets

NG20-A2	n_{doc}	NG20-B2	n_{doc}
comp.sys.ibm.pc.hardware (4)	500	comp.sys.ibm.pc.hardware (4)	500
talk.politics.guns (17)	500	comp.sys.mac.hardware (5)	500
NG20-A4	n_{doc}	NG20-B4	n_{doc}
comp.sys.ibm.pc.hardware (4)	500	comp.os.ms-windows.misc (3)	500
rec.autos (8)	500	comp.sys.ibm.pc.hardware (4)	500
sci.electronics (13)	500	comp.sys.mac.hardware (5)	500
talk.politics.guns (17)	500	comp.windows.x (6)	500
NG20-A6	n_{doc}	NG20-B6	n_{doc}
comp.sys.ibm.pc.hardware (4)	500	comp.os.ms-windows.misc (3)	500
rec.autos (8)	500	comp.sys.ibm.pc.hardware (4)	500
rec.sport.baseball (10)	500	comp.sys.mac.hardware (5)	500
sci.electronics (13)	500	comp.windows.x (6)	500
soc.religion.christian (16)	500	talk.politics.guns(17)	500
talk.politics.guns (17)	500	talk.politics.mideast(18)	500

In the experiment, the performance of RSSKM was compared with four *k*-means type clustering algorithms, namely *k*-means, W-*k*-means [14], AWA [5] and FWKM [8]. The parameters used in these algorithms are tabulated in Table 2. Since the clustering results can be easily affected by the initial cluster centers, the random selection method for initial centers was used. In our experiments, each algorithm was repeated ten times. The clustering results were tabulated in Table 3, from which it can be easily observed that RSSKM outperforms its rivals in most cases. This indicates that the utilization of Equation (3) as the distance function can make RSSKM more robust and improve its performance greatly.

Table 2. Parameter Settings of the Algorithms

Algorithms	Parameter settings
RSSKM	α=1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3
W-*k*-means	β = 2; 5; 10; 50; 100; 1000; 10^4; 10^5
AWA	α=1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3
FWKM	$\sigma = \frac{\sum_{k=1}^{\hat{n}}\sum_{h=1}^{s}\left(x_{kh}-v_{oh}\right)^2}{\hat{n}\cdot s}$

Table 3. *RI* Performance for Different Algorithms

Datasets	**RSSKM**	***k*-means**	**W-*k*-means**	**FWKM**	**AWA**
NG20-A2	0.7136±0.0034	0.5099±0.0215	0.4993±0.0001	0.6201±0.0317	0.6037±0.0097
NG20-A4	0.7297±0.0005	0.5000±0.0010	0.4993±0.0002	0.6304±0.0480	0.6002±0.0020
NG20-A6	0.5367±0.0376	0.3491±0.0684	0.3546±0.0101	0.4511±0.0201	0.4437±0.0738
NG20-B2	0.5856±0.0293	0.4121±0.0372	0.2608±0.0040	0.5201±0.1102	0.4994±0.0976
NG20-B4	0.5620±0.0957	0.3685±0.0694	0.3391±0.1130	0.5104±0.0976	0.4896±0.1279
NG20-B6	0.5802±0.0377	0.4489±0.0475	0.2877±0.0038	0.5211±0.1201	0.4990±0.1050

6. Conclusion

In this study, a novel robust soft subspace clustering algorithm RSSKM is proposed by introducing a novel distance metric into the learning criterion of the algorithm. This work involves the following aspects: (1) A novel objective function integrating the alternative distance metric and soft subspace clustering is proposed based on the *k*-means type clustering algorithms; (2) A novel robust soft subspace clustering RSSKM is developed and its properties are investigated; (3) Experiments on high dimensional text datasets are carried out to verify the performance of the RSSKM algorithm on high dimensional datasets.

It is necessary to find a suitable value for parameter α in RSSKM. In this study, we only set these parameters empirically. Our future work involves further theoretical study on the parameters, which will be of great importance in providing useful and convenient guidelines for the real-world applications of the soft subspace clustering algorithms.

This study will be further extended to improve its performance by extending the *k*-means type clustering algorithms into the fuzzy clustering algorithms. In addition, other ideas such as entropy weighting as well as multi-view learning can also be integrated. In this way, the performance of the robust soft subspace clustering will be further improved.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 61170122, and by the Natural Science Foundation of Jiangsu Province under Grant BK2009067 and BK2011417, the Fundamental Research Funds for the Central Universities (Grant JUSRP111A38).

References

[1] R Agrawal, J Gehrke, D Gunopulos, P Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. *Proc. ACM SIGMOD Int'l Conf. Management of Data*. 1998; 94-105.
[2] C Aggarwal, C Procopiuc, JL Wolf, PS Yu, JS Park. Fast algorithms for projected clustering. *Proc. ACM SIGMOD Int'l Conf. Management of Data*. 1999; 61-72.
[3] CC Aggarwal, PS Yu. Finding generalized projected clusters in high dimensional spaces. *Proc. ACM SIGMOD Int'l Conf. Management of Data*. 2000; 70-81.
[4] KY Yip, DW Cheung, MK Ng. A practical projected clustering algorithm. *IEEE Trans. Knowledge and Data Eng*., 2004; 16(11): 1387-1397.
[5] Y Chan, W Ching, MK Ng, JZ Huang. An optimization algorithm for clustering using weighted dissimilarity measures. *Pattern Recognition*. 2004; 37(5): 943-952.
[6] Z Deng, K Choi, F Chung, S Wang. Enhanced soft subspace clustering integrating within-cluster and between-cluster information. *Pattern Recognition*. 2010; 43(3): 767–781.
[7] KL Wu, MS Yang. Alternative c-means clustering algorithms. *Pattern Recognition.* 2002; 35(10): 2267-2278.
[8] L Jing, MK Ng, J Xu, JZ Huang. Subspace Clustering of Text Documents with Feature Weighting K-Means Algorithm," *Proceedings of the Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 2005; 802–812.
[9] G Gan, J Wu, Z Yang. *A fuzzy subspace algorithm for clustering high dimensional data.* X. Li, O. Zaiane, Z. Li (Eds.). Lecture Notes in Artificial Intelligence. Springer, Berlin. 2006; 4093: 271–278.
[10] G Gan, J Wu. A convergence theorem for the fuzzy subspace clustering (FSC) algorithm. *Pattern Recognition*. 2008; 41: 1939-1947.
[11] JZ Huang, MK Ng, H Rong, Z Li. Automated Variable Weighting in *k*-Means Type Clustering. *IEEE Trans. Pattern Analysis and Machine Intelligence*. 2005; 27(5): 1-12.
[12] H Friedman, JJ Meulman. Clustering objects on subsets of attributes, *J. R. Statist. Soc. B*. 2004; 66(4): 815-849.
[13] W Rudin. Principles of Mathematical Analysis, McGraw-Hill Book Company, New York. 1976.
[14] Liping Jing, Michael K. Ng, Joshua Zhexue Huang. An Entropy Weighting *k*-Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data*. IEEE Transactions on Knowledge and Data Engineering*. 2007; 19(8): 1026-1041.