Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 122 (2017) 276–283

Procedia Computer Science

www.elsevier.com/locate/procedia

Information Technology and Quantitative Management (ITQM 2017)

An Entropy based Method for Overlapping Subspace Clustering

Charu Puri[1], Naveen Kumar[1]

[a]*Dept. of Computer Science, University of Delhi, New Delhi-10065, INDIA*

Abstract

We present subspace clustering method based on entropy and modification of Gustafson-Kessel clustering. The proposed method associates with each cluster its center, variance co-variance matrix and a weight matrix. It represents clusters and their features through gradation in membership, and hence reflects a realistic representation of clusters. Evaluation of experiments on data from UCI ae well as text with the comparative methods shows better results of proposed method.

Peer-review under responsibility of the scientific committee of the 5th International Conference on Information Technology and Quantitative Management, ITQM 2017.

Keywords: Entropy; subspace; document clustering.

1. Introduction

Large document collection often contains vast amount of important scientific data which requires an automatic method for organizing documents. A collection of documents may contain multiple topics and hence, may belong to multiple categories. Also, text data has a large number of dimensions and is highly correlated. Thus, the primary issues for document categorization are hard clustering, curse of dimensionality and correlation in data. Class of fuzzy co-clustering algorithms generate descriptive clusters in high dimension and discover clusters with overlap i.e. it enables a document to be in multiple clusters. Subspace clustering finds cluster with fewer dimensions. Soft subspace clustering finds cluster in the entire space with continuously dimension grading. In high dimensional text data, only a subset of dimensions is relevant for each cluster [14] as varying key words categorize the feature space of each document cluster. Also, a document may belong to multiple categories and multiple specific set of key words may belong to a document cluster. Some of the techniques that have been proposed for variable weighting subspace clustering for document categorization are LAC [6], FWKM [15], and EWKM [14]. These algorithms discover soft subspaces for soft partitioning of document categories.

We propose an algorithm for entropy based overlapping subspace clustering and demonstrate the performance on document categorization. The proposed algorithm associates a center, variance co-variance matrix and a weight matrix with each of the clusters. It represents cluster and their features through a gradation in memberships, and hence reflects a realistic representation of document clusters.

It takes into account the correlation between the features at two stages: (i) during the clustering (ii) while finding the subspaces. The proposed algorithm has the feature of local adaptation of Mahalanobis distance metric

*Charu Puri Tel.: +91-971-755-6021;
E-mail address: cpuri.cs.du@gmail.com, nk.cs.du@gmail.com.

1877-0509
Peer-review under responsibility of the scientific committee of the 5th International Conference on Information Technology and Quantitative Management, ITQM 2017.
10.1016/j.procs.2017.11.370

used as the similarity measure which takes into account the co-variance of the data for clustering. For finding subspaces locally for each cluster the weight matrix assigns the degree of relevance of features and is inversely proportional to ratio of dispersion. In order to effectively capture the correlation in the data, we incorporated mahalanobis distance metric in our algorithm. The primary contributions of this paper are:

We present an entropy based gustafon kessel subspace clustering algorithm with locally adaptive distance metric. Entropy based subspace clustering algorithm finds meaningful subspaces and clusters with fewer dimensions from text. The contributions of EGKS clustering algorithm are:

1. A correlation-based subspace clustering objective function is proposed which minimize cluster error.
2. Finds multiple categories with multiple set of keywords.
3. Captures the correlation in data at level of clustering as well to find subspaces.
4. Extracts arbitrarily positioned overlapping subspace clusters simultaneously.
5. Has feature of local adaptation of distance metric and weight matrix.
6. Ability to identify subspace clusters to categorize text data.

The paper is structured in the following manner. Section 2 presents the existing research literature close to the proposed method on subspace clustering. In Section 3 the proposed Entropy based Gustafson Kessel Subspace algorithm is presented. Section 4 presents the results of experiments on both UCI as well as text datasets along with the experimental comparisons with other comparative methods. Finally conclusion is presented in Section 5.

2. Related Work

$Kmeans$ algorithm is a widely used partitional clustering algorithm which gives equal weights to each of the variables in the clustering process. However, in many real-world applications where the true clusters tend to appear only in a subset of dimensions it is more desirable to generalize the algorithm to incorporate feature weights. Huang et al. [12] presented $kmeans$ type method $Wkmeans$ which automatically assigns weights to the variables on the basis of their relevance to the clustering solution. $Wkmeans$ modifies the basic k-means by including the weights to the features as well an update mechanism for the weight variables on the basis of current data partitioning.

Wang et al. have presented fuzzy c-means algorithm with varying weights called Fuzzy W-k means. It creates partition of soft clusters [20]. They have modified K-means to create soft partitions of the clusters and introduced the parameters α and β as fuzzification parameters. They computes memberships of data points and dimensions simultaneously.

Jing et al. have presented clustering method FW-KMeans [15] for finding subspaces. It assigns weight to dimensions specific to clusters. Thus a dimension may be assigned a high weight for a cluster (implying high relevance) and a low weight for another cluster (implying low relevance).

Subsequently Jing et al. have modified the FW-KMeans for clustering textual data. In text clustering, large weight identifies a key. In text data many words may not be a part of documents of certain classes. This adds sparsity to the text data. However, cases where a word may not appear in any document of the corresponding cluster or a word may appear in every document with equal frequency results in zero dispersion which makes ω_{ir} infinite. For text data sets FW-Kmeans algorithm becomes unsolvable as weights for these terms become infinite.

Jing et al. have proposed an extension of k-means i.e. Entropy Weighting K-Means. EWKM is a soft subspace finding approach [14]. They have extended k-means to compute a membership of every dimension in each cluster. These weight values identify the subset of relevant dimensions which categorize different every cluster. The weight ω_{ir} of a feature measures the contribution of that dimension in forming the cluster where as the entropy of the feature measures the certainty of a feature in the cluster identification.

In this paper, we present an entropy and Gustafson Kessel based subspace finding method for document categorization. It discovers multiple partitions of document in variable weighting subspaces of words. It finds overlapping clusters of documents in the overlapping subspaces i.e. it clusters documents and words simultaneously where word membership helps in generating more useful descriptions of document clusters. The proposed algorithm takes into account the correlation in the data calculated using the inverse of variance co-variance matrix.

Gradation in degree of membership of feature weights is computed using weight matrix which is inversely proportional to the ratio of dispersion. Detailed formulation of entropy based Gustafson Kessel subspace clustering (EGKS) algorithm. We demonstrated its effectiveness with the help of experimental study through performance comparison with other document clustering algorithms like FCCM, GK and EWKM on the benchmark document datasets. Results show the efficacy of EGKS algorithm over other standard fuzzy clustering and co-clustering algorithms in the web environment.

3. Entropy based Gustafson-Kessel Subspace Clustering

We discuss proposed Entropy based Gustafson-Kessel Subspace clustering algorithm in this section. We have extended Gustafson-Kessel clustering objective function as follows:

$$J_{\lambda,\xi}(U,W,Z) = \sum_{j=1}^{n}\sum_{i=1}^{k}\sum_{r=1}^{d}\mu_{ij}\omega_{ir}d_{ijr}^2 + \lambda^{-1}\sum_{j=1}^{n}\sum_{i=1}^{k}\mu_{ij}log(\mu_{ij}) + \xi^{-1}\sum_{r=1}^{d}\sum_{i=1}^{k}\omega_{ir}log(\omega_{ir}) \tag{1}$$

$J_{\lambda,\xi}$ the entropy based subspace Gustafson-Kessel objective function. where A_i is the norm inducing matrix for cluster i. Distance with respect to the r^{th} dimension of point x_j and the i^{th} class center [8] is defined as:

$$d_{ijr}^2 = \sum_{s=1}^{d}(x_{jr} - z_{ir})a_{rs}(x_{js} - z_{is}), \tag{2}$$

$A_i = [a_{rs}]_{d\times d}$

is a symmetric and positive definite matrix. It induces for a norm for each cluster [10]. As$J_{\lambda,\xi}$ is linear in A_i, it leads to singularity problem. Thus it is difficult to minimizing $J_{\lambda,\xi}$. Thus, A_i is constrained to obtain feasible solution in the following way

$$det(A_i) = \rho_i > 0, \tag{3}$$

ρ_i is fixed for each i. It permits varying cluster sizes [11].

$J_{\lambda,\xi}$ is the sum of squared-errors for each i^{th}cluster additionally weighted by weight memberships and functions E_1 and E_2 multiplied by weight factors $\lambda^{-1} > 0$ and $\xi^{-1} > 0$ respectively. It defines the total scatter within each group. The $\lambda^{-1} > 0$ and $\xi^{-1} > 0$ are predefined and called the degree of fuzzy entropy. In limiting cases of weight factors $\lambda^{-1} > 0$ and $\xi^{-1} > 0$, as λ^{-1} and $\xi^{-1} \to 0$ EGKS clustering tends to become hard clustering. Also, if λ^{-1} and $\xi^{-1} \to \infty$, it results in single cluster [19].

$\mathrm{E}_1(U) = -\sum_{i=1}^{k}\sum_{j=1}^{n}\mu_{ij}\log(\mu_{ij})$

$\mathrm{E}_2(W) = -\sum_{r=1}^{d}\sum_{i=1}^{k}\omega_{ir}\log(\omega_{ir})$

$E_1(U)$ and $E_2(W)$ express respectively the uncertainty in determining to what extent j^{th} object belongs to a given cluster and to what extent r^{th} dimension belongs to subspace of a cluster. $E_1(U)$ and $E_2(W)$ signify the average degree of non-membership of members in a cluster [19]. $E_1(U)$ and $E_2(W)$ are maximum if $\mu_{ij} = 1/k$ and $\omega_{ir} = 1/k \quad \forall i$ respectively and are minimum if μ_{ij} and $\omega_{ir} = 1 \quad or \quad 0$ respectively. In the context of document categorizations the terms $-\sum_{i=1}^{k}\sum_{j=1}^{n}\mu_{ij}\log(\mu_{ij})$ and $-\sum_{r=1}^{d}\sum_{i=1}^{k}\omega_{ir}\log(\omega_{ir})$ are the document and word fuzzifiers respectively [18]. These fuzzifiers have dual objective i.e. first is to act as a regularization terms. If the fuzzifiers are not introduced in the objective function, maximization of objective function becomes ill posed problem i.e. no unique solution for μ_{ij} and ω_{ir} [18] exists. Secondly regularization terms help in adjusting the fuzziness of the resulting document and word clusters so that document and word memberships are forced not to take crisp values [18]. Different clusters have different minimization of objective function of EGKS algorithm leads to minimizations of its first term and maximization of entropy functions E_1 and E_2 respectively with continuous feature weighting. The objective function $J_{\lambda,\xi}$ is minimized over U, W and Z to obtain partitioning for X and the set of attributes respectively. The matrices A_i also act as optimization variables in the objective function $J_{\lambda,\xi}$.

Equations below are the necessary conditions for minimizing objective function:

$$A_i = ((det(F_i)\rho_i))^{1/d}F_i^{-1}, \ 1 \le i \le k \tag{4}$$

$$\mu_{ij} = \frac{e^{-\lambda \sum_{r=1}^{d} \omega_{ir} d_{ijr}^2}}{\sum_{i'=1}^{k} e^{-\lambda \sum_{r=1}^{d} \omega_{i'r} d_{i'jr}^2}}, \quad 1 \le i \le k,\ 1 \le j \le n, \tag{5}$$

$$\omega_{ir} = \frac{e^{-\xi \sum_{j=1}^{n} \mu_{ij} d_{ijr}^2}}{\sum_{r'=1}^{d} e^{-\xi \sum_{j=1}^{n} \mu_{ij} d_{ijr'}^2}}, \quad 1 \le i \le k,\ 1 \le r \le d, \tag{6}$$

$$z_{ir} = \frac{\sum_{j=1}^{n} \mu_{ij} x_{jr}}{\sum_{j=1}^{n} \mu_{ij}}, \quad 1 \le i \le k, \quad 1 \le r \le d \tag{7}$$

3.1. The Proposed EGKS Clustering Algorithm

Using the update equations described earlier, we will now describe below the Entropy based Gustafson-Kessel Subspace Clustering (EGKS) algorithm. The algorithm works on high dimensional data set X of size n. Parameters λ and ξ are chosen in suitable range. Initially prototypes are chosen randomly. The process of updating the cluster prototypes and partition matrices is repeated until the membership matrices in successive iteration fall within a predefined threshold range of ϵ or the limit of the maximum number of iteration is reached.

Algorithm of EGKS

Input:

n: data set size X
k: cluster number, where $1 < k < n$
λ: matrix U weight exponent , $\lambda^{-1} > 0$
ξ: matrix W weight exponent , $\xi^{-1} > 0$
ϵ: closeness, $\epsilon > 0$
A: norm-inducing matrix,
$maxiter$: max iteration.

Output:

U: membership matrix of objects
W: membership matrix of dimensions
Z: center of cluster

Procedure:

The partition matrices U, W are randomized initially
$\left\|U^t - U^{t-1}\right\| > \in$ or $t \le maxiter$
Step 1. Updation of prototypes of cluster
$z_{ir} = \sum_{j=1}^{n} \mu_{ij}^t x_{jr} / \sum_{j=1}^{n} \mu_{ij}^t$

Step 2. Update the cluster covariance matrices
$F_i^t = \sum_{j=1}^{n} (\omega_{ir}^{t-1})(\mu_{ij}^{t-1}) (x_{jr} - z_{ir}^t) (x_{js} - z_{is}^t)\, 1 < r, s \le d.$

Step 3. Update the distances:
$d_{ijr}^2 = [(x_{j1} - z_{i1}^t) \dots (x_{jd} - z_{id}^t)] A_i [(x_{j1} - z_{i1}^t) \dots (x_{jd} - z_{id}^t)]^T$

Step 4. Update the partition matrices:
$\mu_{ij}^t = \frac{e^{-\lambda \sum_{r=1}^{d} \omega_{ir}^{t-1} d_{ijr}^2}}{\sum_{i'=1}^{k} e^{-\lambda \sum_{r=1}^{d} \omega_{i'r}^{t-1} d_{i'jr}^2}}$

$$\omega_{ir}^{t} = \frac{e^{-\xi \sum_{j=1}^{n} \mu_{ij}^{t-1} d_{ijr}^{2}}}{\sum_{r'=1}^{d} e^{-\xi \sum_{j=1}^{n} \mu_{ij}^{t-1} d_{ijr'}^{2}}}$$

Step 5. $t = t + 1$

U, W, and Z are updated iteratively till there is scope for improvement for partitions. Computational complexity of EGKS algorithm in updating Z, U, W and F is $O(nkd)$, $O(nk^2d)$, $O(nkd^2)$ and $O(nkd^2)$ respectively.

4. Experimental Results

We will now present the experimental results of the EGKS clustering algorithm using the standard UCI machine learning repository data sets [1] and several other document datasets. We will also compare the performance of EGKS with other state-of-the-art algorithms. Finally, we will also present some of the qualitative results obtained by these algorithms on text datasets.

4.1. Experimental Setup

Table 1. UCI and text datasets description.

Datasets	Instances	Attributes	Classes
Diabetes	768	9	2
Lung Cancer	32	57	3
Ionosphere	351	35	2
Statlog	2310	19	7
Email-1431	1431	3490	3
SW	341	859	4
Lingspam	1156	2259	4
Spambase	4601	57	3

Dataset Description: We used four UCI datasets [1] for the experimental evaluation. These data sets are different with respect to size, clusters count, and class distribution. In addition to demonstrate the performance of the algorithm on real high-dimensional scenarios, we have also used the following four popular text datasets in our evaluation.

1. Email-1431 dataset[1] contains 1431 documents. It has three classes: conference (370 documents), jobs (272 documents) and spam (789 documents).
2. Syskill & Webert (SW) dataset [1] consists of 341 web documents distributed into 4 classes with unequal distribution of documents. The 4 different classes are sheep (70 documents), goats (74 documents), band (61 documents), and biomedical (136 documents).
3. Ling-Spam corpus[2] has four versions. They are with or without stemming and with or without stop word removal. Each data set consists of 289 mails. We use Lemmatized, stop-list enabled and Lemmatize and stop-list enabled versions from the Ling-Spam corpus.
4. Spambase dataset [1] contains 4601 documents. These documents belongs to one of the two classes: spam and nonspam.

Text datasets form a good testbed for subspace algorithms due to their high-dimensional nature. Table 1 summerizes the data sets used in experiment.

Comparison Algorithms: We compare the EGKS algorithm with the following methods: EWKM [14], FWKM [15], and LAC [6] algorithms that are described in the related work section. While EWKM and LAC are subspace

[1] http://www2.imm.dtu.dk/rem/data/

[2] http://www.aueb.gr/users/ion/

clustering algorithms, FWKM compute a common subset of dimensions for all clusters. All three algorithms are implemented in WEKA software[3]. The comparisons are made using some of the standard evaluation metrics. The implementation of EGKS has been done through MATLAB. Systematic grid search was made in the range 0.05-5 in steps of 0.1 independently for setting the parameters for α and β.

4.2. Performance Comparison Results

We will show the comparison of different algorithms using several different evaluation metrics. Table 2 presents result of accuracy for all competing algorithms and data sets. Since all of the datasets have associated class labels information, we matched the membership of the cluster result with that of the class labels to determine the values of various performance metrics.

Table 2. Comparison of Accuracy values for real-world datasets.

Data Sets	EGKS	EWKM	FWKM	LAC
Diabetes	0.6640	0.6614	0.5924	**0.6666**
Lung Cancer	**0.6250**	**0.6250**	0.4687	0.5625
Ionosphere	**0.7179**	0.3809	0.3875	0.7094
Statlog	0.6398	0.3964	0.3823	**0.6831**
Email-1431	**0.9364**	0.5521	0.5538	0.5552
SW	**0.5982**	0.5416	0.5446	0.5148
Lingspam	**0.6314**	0.2502	0.2502	0.2502
Spambase	**0.6568**	0.6034	0.5934	0.5089

In order to make our evaluation more comprehensive, other alternative measures of evaluation such as F1-measure, recall rate (or sensitivity) and specificity, are also used in the experiments.

Tables 3, 4, and 5 show the comparison results in terms of F1-measure, recall, and specificity, respectively. These tables show EGKS generally performs superior compared to the other methods. In general, the improvements of the EGKS algorithm are much higher on high-dimensional text datasets. On low-dimensional UCI datasets, LAC performs better in terms of F1-measure. More specifically, the performance of EGKS on text datasets in terms of F1-measure and recall rates (widely used metrics in text classification) is significantly better compared to all the other methods.

4.3. Qualitative Results

Table 6 shows the results on text datasets generated by EGKS, EWKM, and LAC algorithms. It gives the top 10 words with the highest membership values in the respective clusters for the Email-1431, SW, and Spambase datasets. Since the FWKM algorithm obtains a common set of features for all the clusters, we could not it in

[3]http://www.cs.waikato.ac.nz/ml/weka/

Table 3. Comparison of F1-measure values on real-world datasets.

Data Sets	EGKS	EWKM	FWKM	LAC
Diabetes	0.6005	0.6005	0.6005	**0.6666**
Lung Cancer	**0.6387**	0.6227	0.4785	0.5673
Ionosphere	0.4270	0.4270	0.4270	**0.7149**
Statlog	0.6383	0.3267	0.3309	**0.6827**
Email-1431	**0.9377**	0.4025	0.3970	0.4385
SW	**0.5766**	0.4801	0.4865	0.4857
Lingspam	**0.5167**	0.1274	0.1125	0.1454
Spambase	**0.6310**	0.4760	0.4518	0.4114

Table 4. Comparison of Sensitivity (or Recall) values on real-world datasets.

Data Sets	EGKS	EWKM	FWKM	LAC
Diabetes	0.5861	0.5861	0.5861	**0.6666**
Lung Cancer	**0.6387**	0.6250	0.4687	0.5625
Ionosphere	0.4279	0.4279	0.4279	**0.7094**
Statlog	0.6398	0.3964	0.3823	**0.6831**
Email-1431	**0.9457**	0.5521	0.5538	0.5552
SW	**0.5498**	0.5416	0.5446	0.5148
Lingspam	**0.5130**	0.2502	0.2502	0.2502
Spambase	0.5759	**0.6034**	0.5934	0.5089

Table 5. Comparison of Specificity values on real-world datasets.

Data Sets	EGKS	EWKM	FWKM	LAC
Diabetes	0.5861	0.6247	0.6171	**0.6332**
Lung Cancer	**0.8051**	0.5768	0.6102	0.6102
Ionosphere	**0.7179**	0.4395	0.2821	0.2843
Statlog	**0.93997**	0.8578	0.8719	0.8782
Email-1431	**0.9457**	0.6673	0.6682	0.6811
SW	**0.8757**	0.8000	0.7965	0.7933
Lingspam	0.5130	0.6667	**0.7500**	**0.7500**
Spambase	0.5759	0.4969	0.5103	**0.5793**

our qualitative comparisons. From these qualitative results, we can clearly observe that the EGKS describes the word clusters with a better set of words for the clusters compared to the EWKM and LAC algorithms. In other words, the features extracted by the EGKS are more representative of the clusters and can qualitatively describe the clusters in a better way compared to the other algorithms available in the literature.

5. Conclusion

We developed a novel entropy based algorithm for extracting overlapping subspace clusters from high-dimensional data. We have discussed in detail the formulation of the algorithm and the derivation of the new update rules along with proof of its convergence. The proposed EGKS algorithm generates soft membership results which in turn led to more qualitatively interpretable results for the high-dimensional document categorization tasks in many real-world datasets. In addition, the EGKS algorithm generates clusters with fewer dimensions than the original dataset compared to the ones produced by its competitors.

References

[1] Website. http://www.ics.uci.edu/ mlearn/MLRepository.html.
[2] J. Abonyi and B. Feil. *Cluster Analysis for Data Mining and System Identification*. Birkhauser, 2006.
[3] J. Bezdek, R. Hathaway, M. Sobin, and W. Tucker. Convergence theory for fuzzy c-means: counterexamples and repairs. *IEEE Trans. Syst. Man Cybern.*, 17:873–877, 1987.
[4] James C. Bezdek. An convergence theorem for the fuzzy isodata clustering algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1:1–8, 1980.
[5] James C. Bezdek, Chris Coray, Robert Gunderson, and James Watson. Detection and characterization of cluster substructure. *SIAM Journal on Applied Mathematics*, 40:358–372, 1981.
[6] Carlotta Domeniconi, Dimitrios Gunopulos, Sheng Ma, Bojun Yan, Muna Al-Razgan, and Dimitris Papadopoulos. Locally adaptive metrics for clustering high dimensional data. *Data Min. Knowl. Discov.*, 14:63–97, 2007.
[7] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3:32–57, 1973.
[8] H. Frigui and O. Nasraoui. Unsupervised learning of prototypes and attribute weights. *Pattern Recognition*, 37, 2004.
[9] G. Gan and J. Wu. A convergence theorem for the fuzzy subspace clustering (fsc) algorithm. *Pattern Recogn.*, 41:1939–1947, 2008.

Table 6. Top 10 Words obtained from text datasets using different algorithms.

Clusters	Algorithms	Top-10 Words obtained
Email-1431		
Job	EGKS	letter, cover, test, applicant, degree, refere, includ, researc, hospital, program
	EWKM	1st, 20spm, 2nd, 3rd, 4th, 7bit, 8bit, abil, abl, absolut
	LAC	1st, 20spm, 2nd, 3rd, 4th, 7bit, 8bit, abil, abl, absolut
Conference	EGKS	paper, submit, manuscript, student, registratio, confer, abstrac, program, talk
	EWKM	20spm, access, activ, actual, advantag, america, apolog, appoint, autonom
	LAC	advertis, approxim, bank, bonu, campu, color, daili, david, difficult, disciplin
Spam	EGKS	1st, 2nd, 3rd, 4th, 7bit, 8bit, abl, absolut, abstract, academ
	EWKM	1st, 2nd, 3rd, 4th, 7bit, 8bit, abl, absolut, abstract, academ
	LAC	1st, 20spm, 2nd, 3rd, 4th, 7bit, 8bit, abl, absolut, abstract
SW Text		
Bands	EGKS	music, guitar, song, band, rock, word, industri, acoust, audienc, cassett
	EWKM	across, audienc, band, bandag, began, believ, blue, cassett, combin, compos
	LAC	band, bandag, blue, cassett, compos, debut, drum, excerpt, excit, feel
Biomedical	EGKS	physician, medical, scienc, human, health, department, research, aid, bioscienc, across
	EWKM	academi, acoust, aid, arrang, bandag, began, bioinformat, biologi, biopag, bioscienc
	LAC	band bandag, blue, cassett, compos, debut, drum, excerpt, excit, feel
Goats	EGKS	goat, pig, milk, farm, diary, show, individu, unig, agricultur, account
	EWKM	abov, account, action, aid, album, alink, america, angel, artwork, audienc
	LAC	abstract, academ, account, agricultur, alink, analysi, answer, appli, applic, aspect
Sheep	EGKS	sheep, wool, wolves, thread, fiends, abov, connect, brain, diary, blue
	EWKM	abstract, academ, account, agricultur, alink, analysi, answer, appli, applic, aspect
	LAC	abstract, academ, academi, acoust, advanc, agenc, aid, angel, announc, arizona
Spambase		
NonSpam	EGKS	report, telnet, 857c, parts, original, project, table, 3d, font, pm
	EWKM	857c, 415d, telnet, 85e, capital_run_length_longest, #, labs, direct, technology, hpl
	LAC	857c, 415d, telnet, 85e, capital_run_length_longest, #, labs, direct, technology, hpl
3*Spam	EGKS	#, capital_run_length_average, capital_run_length_longest, [, !, (, 85e, parts, credit, conference
	EWKM	font, money, 3d, capital_run_length_average, capital_run_length_longest, credit, $, 000, !, remove
	LAC	font, money, 3d, capital_run_length_average, capital_run_length_longest, credit, $, 000, !, remove

[10] Donald E. Gustafson and William C. Kessel. Fuzzy clustering with a fuzzy covariance matrix. *IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, 17:761–766, 1978.

[11] F. Hoppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis, and Image Recognition*. John Wiley & Sons, 1999.

[12] Joshua Zhexue Huang, Michael K. Ng, Hongqiang Rong, and Zichen Li. Automated variable weighting in k-means type clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27:657–668, 2005.

[13] Jung-Yi Jiang, Ren-Jia Liou, and Shie-Jue Lee. A fuzzy self-constructing feature clustering algorithm for text classification. *IEEE Trans. on Knowl. and Data Eng.*, 23(3):335–349, 2011.

[14] Liping Jing, Michael K. Ng, and Joshua Zhexue Huang. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Trans. on Knowl. and Data Eng.*, 19:1026–1041, 2007.

[15] Liping Jing, Michael K. Ng, Jun Xu, and Joshua Zhexue Huang. On the performance of feature weighting k-means for text subspace clustering. *Proceedings of the 6th international conference on Advances in Web-Age Information Management*, pages 502–512, 2005.

[16] Argyris Kalogeratos and Aristidis Likas. Document clustering using synthetic cluster prototypes. *Data Knowl. Eng.*, 70:284–306, 2011.

[17] P-N. Tan, M Steinbach, and V. Kumar. Introduction to data mining. 2005.

[18] William-Chandra Tjhi and Lihui Chen. A partitioning based algorithm to fuzzy co-cluster documents and words. *Pattern Recogn. Lett.*, 27:151–159, 2006.

[19] D. Tran and M. Wagner. Fuzzy entropy clustering. *IEEE International Conference on Fuzzy Systems*, pages 152–157, 2000.

[20] Qiang Wang, Yunming Ye, and Joshua Zhexue Huang. Fuzzy k-means with variable weighting in high dimensional data analysis. *Proceedings of International Conference on Web-Age Information Management*, pages 365–372, 2008.

[21] W. Zangwill. *Non-Linear programming: A unified Approach-Englewood Cliffs, NJ*. prentice-Hall, 1969.