Knowl Inf Syst (2008) 17:1–15
DOI 10.1007/s10115-007-0116-0

REGULAR PAPER

Clustering based on matrix approximation: a unifying view

Tao Li

Received: 13 December 2005 / Revised: 19 March 2007 / Accepted: 8 October 2007 /
Published online: 8 January 2008

Abstract Clustering is the problem of identifying the distribution of patterns and intrinsic correlations in large data sets by partitioning the data points into similarity classes. Recently, a number of methods have been proposed and demonstrated good performance based on matrix approximation. Despite significant research on these methods, few attempts have been made to establish the connections between them while highlighting their differences. In this paper, we present a unified view of these methods within a general clustering framework where the problem of clustering is formulated as matrix approximations and the clustering objective is minimizing the approximation error between the original data matrix and the reconstructed matrix based on the cluster structures. The general framework provides an elegant base to compare and understand various clustering methods. We provide characterizations of different clustering methods within the general framework including traditional one-side clustering, subspace clustering and two-side clustering. We also establish the connections between our general clustering framework with existing frameworks.

Keywords Clustering · Matrix approximation · Alternating optimization · Subspace

1 Introduction

Clustering is the problem of partitioning a finite set of points in a multi-dimensional space into classes (called clusters) so that (i) the points belonging to the same class are *similar* and (ii) the points belonging to different classes are *dissimilar*. Generally clustering problems are determined by four basic components: (a) the (physical) representation of the given data set; (b) the distance/dissimilarity measures between data points; (c) The criterion/objective function which the clustering solutions should aim to optimize; (d) The optimization procedure. For a given data clustering problem, the four components are tightly coupled. Clustering has been extensively studied in machine learning, databases, and statistics from various perspectives.

T. Li (✉)
School of Computer Science, Florida International University, Miami, FL 33199, USA
e-mail: taoli@cs.fiu.edu

Many applications of clustering have been discussed and many clustering techniques have been developed.

Recently, a number of authors [2,7,8,17,19,20,24–27] have suggested clustering methods based on matrix computations and have demonstrated good performance on various datasets. These methods are attractive as they utilize many existing numerical algorithms in matrix computations. Nevertheless, the use of matrix computations in the context of clustering needs more studies. In this paper, we present a generalized clustering framework[1] where the problem of clustering is formulated as matrix approximations. The goal of clustering is then transformed to minimizing the approximation error between the original data matrix and the reconstructed matrix based on the cluster structures. In the framework, the data are usually represented as matrices and the distance measures between data points are Euclidean distances. Hence our discussion in this paper focuses on the criterion/objective function and the optimization procedure. This framework encompasses many previously known clustering algorithms including traditional one-side clustering, co-clustering, and subspace clustering and provides an elegant base to compare and understand various clustering methods. While seemingly quite different, these different algorithms are closely related, and in fact, different variations derived from the general framework with different constraints and relaxations. In other words, the general framework provides a basis to establish the connections between various methods while highlighting their differences.

We address the following two questions in the paper: (*a*) *What are the different possible clustering methodologies* (*or, matrix reconstruction schemes*) *can be derived from the general model?* (*b*) *What are the relations between the general model with other existing models?* To address the first question, we provide characterizations of different clustering methods within the general framework. We show the close connections between various clustering methods and also explain their distinguishing features. To address the second question, we explore the relationships between our general framework with other existing models. In particular, we show the connections between our general model with the information-theoretic co-clustering framework.

The rest of the paper is organized as follows: Sect. 2 introduces the notations and describes the general clustering model; Sect. 3 provides characterizations of different clustering methods within the general framework; Sect. 4 explores the connections between our general model with other models, and finally, our conclusions are presented in Sect. 5.

2 Clustering model

The notations used in the paper are introduced in Table 1. We first present a general model for clustering problem. The model is formally specified as follows:

$$W = AXB^{\mathrm{T}} + E \tag{1}$$

where matrix E denotes the error component. The first term AXB^{T} characterizes the information of W that can be described by the cluster structures. A and B designate the cluster memberships for data points and features, respectively. X specifies cluster representation. Let $\hat{W}$ denote the approximation AXB^{T} and the goal of clustering is to minimize the approximation

[1] In this paper, we use model and framework interchangeably.

Table 1 Notations used throughout the paper

$W = (w_{ij})_{n\times m}$	The data set
$D = (d_1, d_2, \ldots, d_n)$	Set of data points
$F = (f_1, f_2, \ldots, f_m)$	Set of features
K	Number of clusters for data points
C	Number of clusters for features
$P = \{P_1, P_2, \ldots, P_K\}$	Partition of D into K clusters
$i \in P_k, 1 \le k \le K$	i-th data point in cluster P_k
$p_1, p_2, \ldots, p_K$	Sizes for the K data clusters
$Q = \{Q_1, Q_2, \ldots, Q_C\}$	Partition of F into C clusters
$q_1, q_2, \ldots, q_C$	Sizes for the C feature clusters
$j \in Q_c, 1 \le c \le C$	jth feature in cluster Q_c
$A = (a_{ik})_{n\times K}$	Matrix designating the data membership
$B = (b_{jc})_{m\times C}$	Matrix designating the feature membership
$X = (x_{kc})_{K\times C}$	Matrix specifies/indicates the association between data and features or the cluster representation
Trace(M)	Trace of the matrix M

error (or *sum-of-squared-error*)

$$\begin{aligned} O(A, X, B) &= \|W - \hat{W}\|_{\mathrm{F}}^2 \\ &= \mathrm{Trace}[(W - \hat{W})(W - \hat{W})^{\mathrm{T}}] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m}(w_{ij} - \hat{w}_{ij})^2 \qquad (2) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m}\left(w_{ij} - \sum_{k=1}^{K}\sum_{c=1}^{C} a_{ik}b_{jc}x_{kc}\right)^2 \qquad (3) \end{aligned}$$

Note that the Frobenius norm, $\|M\|_{\mathrm{F}}$, of a matrix $M = (M_{ij})$ is given by $\|M\|_{\mathrm{F}} = \sqrt{\sum_{i,j} M_{ij}^2}$.

The general model provides a good basis for characterizing various matrix-based clustering approaches and it encompasses many previously known clustering algorithms including traditional one-side clustering, co-clustering, and subspace clustering.

3 Different clustering algorithms

Based on different constraints on the matrices A, B and X, this general model encompasses different clustering algorithms. In this section, we provide characterizations of different clustering methods within the general framework. A summary of the derivations is listed in Table 2.

Table 2 Summary on the clustering methods derived from the general model. Each row lists a clustering method and its associated constraints

Methods	B	A	X	Optimization procedure
One-side K-means	$B = I$	$a_{ik} \in \{0, 1\}, \sum_{i=k}^{K} a_{ik} = 1$	$X = (A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}} W$	Alternating least square
One-side low dimensional clustering	$B = I$	$a_{ik} \in \{0, 1\}, \sum_{i=k}^{K} a_{ik} = 1$	$\mathrm{Rank}(X) = t, t \leq \min(K - 1, m)$	Low-rank approximation
Spectral relaxation	$B = I$	Orthonormal	$X = (A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}} W$	Trace maximization
Concept/non-negative	$B = I$ Matrix factorization	non-negative	X = RW (Linear combination of points in the cluster)	Constrained optimization
Double K-means	$b_{jc} \in \{0, 1\}$ $\sum_{j=c}^{C} b_{jc} = 1$	$a_{ik} \in \{0, 1\}, \sum_{i=k}^{K} a_{ik} = 1$	$x_{kc} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ik} b_{jc} w_{ij}}{\sum_{i \in n} \sum_{j \in m} a_{ik} b_{jc}}$	Alternating least square (two-side)
Iterative feature data clustering	Arbitrary	Arbitrary	$X = I$	Mutually reinforcing optimization
Generalized spectral relaxation	Orthonormal	Orthonormal	$X = A^{\mathrm{T}} W B$	Two-side trace maximization
Subspace clustering	$B \in R^{m \times K}$	$a_{ik} \in \{0, 1\}, \sum_{i=k}^{K} a_{ik} = 1$	$X = (A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}} W$	Explicit subspace identification

3.1 One-side clustering

Consider the case when $C = m$, then each feature is a cluster by itself and $B = I_{m\times m}$. The model thus reduces to popular one-side clustering, i.e., grouping the data points into clusters.[2]

3.1.1 One-side K-means clustering

Suppose $A = (a_{ik}), a_{ik} \in \{0, 1\}, \sum_{k=1}^{K} a_{ik} = 1$ (i.e., A denotes the data membership), then the model reduces to

$$\begin{aligned}
O(A, X) &= \|W - AX\|_{\mathrm{F}}^2 \\
&= \mathrm{Trace}[(W - AX)(W - AX)^{\mathrm{T}}] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}(w_{ij} - \sum_{k=1}^{K} a_{ik}x_{kj})^2 \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} a_{ik}\sum_{j=1}^{m}(w_{ij} - x_{kj})^2 \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} a_{ik}\sum_{j=1}^{m}(w_{ij} - y_{kj})^2 + \sum_{k=1}^{K} p_k \sum_{j=1}^{m}(y_{kj} - x_{kj})^2 \qquad (4) \\
&\quad \text{where } p_k = \sum_{i=1}^{n} a_{ik} \text{ and } y_{kj} = \frac{1}{p_k}\sum_{i=1}^{n} a_{ik}w_{ij}
\end{aligned}$$

Given A, the objective criterion O is minimized by setting $x_{kj} = y_{kj} = \frac{1}{p_k}\sum_{i=1}^{n} a_{ik}w_{ij}$. Without loss of generality, we assume that the rows belong to a particular cluster are contiguous, so that all data points belonging to the first cluster appear first and the second cluster next, etc.[3] Then A can be represented as

$$A = \begin{bmatrix}
1 & 0 & \cdots & 0 \\
1 & 0 & \cdots & 0 \\
\vdots & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
\vdots & \vdots & \cdots & \vdots \\
0 & 0 & \cdots & 1 \\
\vdots & \vdots & \cdots & \vdots \\
0 & 0 & \cdots & 1
\end{bmatrix}.$$

[2] Here we only discuss the one-side clustering for data points. It should be note that, similarly, we can derive one-side feature clustering when $K = n$, $A = I$.

[3] This can be achieved by multiplying W with a permutation matrix if necessary.

Note that

$$A^{\mathrm{T}}A = \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & p_K \end{bmatrix}$$

is a diagonal matrix with the cluster size on the diagonal. The inverse of $A^{\mathrm{T}}A$ serves as a weight matrix to compute the centroids. Thus we have the following equation for representing centroids:

$$X = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}W. \tag{5}$$

On the other hand, given X, $O(A, X)$ is minimized by

$$\hat{a}_{ik} = \begin{cases} 1 & \text{if } \sum_{j=1}^{m}(w_{ij} - y_{kj})^2 < \sum_{j=1}^{m}(w_{ij} - y_{lj})^2 \\ & \text{for } l = 1, \ldots, K, l \neq k \\ 0 & \text{Otherwise} \end{cases} \tag{6}$$

The alternative minimization leads to traditional the K-means clustering procedure [14].

3.1.2 One-side low dimensional clustering

When data are column-centered, the K cluster centroids always define a $(K-1)$ dimensional subspace [24]. Sometimes, a low-dimensional representation of the cluster structure is very useful and each cluster is represented by a centroid in a low dimensional space. To achieve dimensional reduction, we can restrict that the K centroids lie in an t-dimensional subspace by restricting $\mathrm{Rank}(X) = t, t <= \min(K-1, m)$.

Based on Eq. (4), if we treat A as a constant, then minimizing $O(A, X)$ reduces to minimizing the following optimization criterion:

$$\begin{aligned} O(X) &= \sum_{k=1}^{K} p_k \sum_{j=1}^{m} (y_{kj} - x_{kj})^2 \\ &= \|E(Y - X)\|_{\mathrm{F}}^2 \\ \text{where } E &= \mathrm{diag}(\sqrt{p_1}, \ldots, \sqrt{p_K}) \end{aligned} \tag{7}$$

with the rank constraint on X. This can be solved using low-rank approximation. Mathematically, the optimal rank r approximation of a matrix W, under the Frobenius norm can be formulated as follows: *Find a matrix $\hat{X}$ with $rank(\hat{X}) = r$ such that $\hat{X} = argmin_{\mathrm{rank}(\hat{X}=r)}$* $\|X - \hat{X}\|_{\mathrm{F}}^2$. The matrix $\hat{X}$ can be readily obtained by computing the Singular Value Decomposition (SVD) of X, as stated in the following theorem [11].

Theorem 1 *Let the Singular Value Decomposition of $X \in R^{n\times m}$ be $X = USV^{\mathrm{T}}$, where U and V are orthogonal, $S = diag(\sigma_1, \ldots, \sigma_t, 0, \ldots, 0)$, $\sigma_1 \geq \cdots \geq \sigma_t > 0$ and $t = rank(X)$. Then for $1 \leq r \leq t$, $\sum_{i=r+1}^{t} \sigma_i^2 = min\{\|X - \hat{X}\|_{\mathrm{F}}^2 | rank(\hat{X}) = r\}$. The minimum is achieved with $\hat{X} = X_r$, where $X_r = U_r diag(\sigma_1, \ldots, \sigma_r)V_r^{\mathrm{T}}$, and U_r and V_r are the matrices formed by the first r columns of U and V, respectively.*

Back to Eq. (7), let $U_t S_t V_t^{\mathrm{T}}$ be the rank t truncated singular value decomposition of EY, it can be then be shown that $X = E^{-1}U_t S_t V_t^{\mathrm{T}}$ gives the rank t matrix minimizing $O(X)$.

3.1.3 Spectral relaxation

Based on Eq. (5), we have

$$\begin{aligned} O(A, X) &= \|W - AX\|_{\mathrm{F}}^2 \\ &= \|W - A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}W\|_{\mathrm{F}}^2 \end{aligned}$$

If we denote

$$R = \begin{bmatrix} \frac{1}{\sqrt{p_1}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{p_1}} & 0 & \cdots & 0 \\ \vdots & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{p_2}} & \cdots & 0 \\ 0 & \frac{1}{\sqrt{p_2}} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{p_K}} \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{p_K}} \end{bmatrix}$$

then $RR^{\mathrm{T}} = A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$. Hence

$$\begin{aligned} O(A, X) &= \|W - A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}W\|_{\mathrm{F}}^2 \\ &= \|(I - RR^{\mathrm{T}})W\|_{\mathrm{F}}^2 \\ &= \mathrm{Trace}(W^{\mathrm{T}}(I - RR^{\mathrm{T}})(I - RR^{\mathrm{T}})^{\mathrm{T}}W) \\ &= \mathrm{Trace}(W^{\mathrm{T}}(I - RR^{\mathrm{T}})W) \\ &= \mathrm{Trace}(WW^{\mathrm{T}}) - \mathrm{Trace}(RWW^{\mathrm{T}}R^{\mathrm{T}}) \end{aligned}$$

Since $I - RR^{\mathrm{T}}$ is a projection matrix, so $(I - RR^{\mathrm{T}})(I - RR^{\mathrm{T}})^{\mathrm{T}} = I - RR^{\mathrm{T}}$.

Here minimizing $O(A, X)$ is reduced to maximizing $\mathrm{Trace}(RWW^{\mathrm{T}}R^{\mathrm{T}})$. If we ignore the special structure of R and let it be an arbitrary orthonormal matrix, the clustering problem then reduced to the trace maximization problem which can be solved by eigenvalue decomposition of the symmetric matrix WW^{T} [27].

3.1.4 Concept factorization

In K-means clustering described in Sect. 3.1.1, X represents the centroid (i.e., the average mean) of the data points in the cluster. In general, the cluster centroid can be thought as a linear combination of the data points in the cluster [7]. In other words, $X = SW$ where S is a $K \times n$ coefficient matrix. Then

$$\begin{aligned} O(A, X) &= \|W - AX\|_{\mathrm{F}}^2 \;=\; \|W - ASW\|_{\mathrm{F}}^2 \\ &= \|(I - AS)W\|_{\mathrm{F}}^2 \\ &= \mathrm{Trace}((I - AS)WW^{\mathrm{T}}(I - AS)^{\mathrm{T}}) \\ &= \mathrm{Trace}(WW^{\mathrm{T}} - 2S^{\mathrm{T}}A^{\mathrm{T}}WW^{\mathrm{T}} + ASWW^{\mathrm{T}}S^{\mathrm{T}}A^{\mathrm{T}}) \\ &= \mathrm{Trace}(WW^{\mathrm{T}}) - 2\mathrm{Trace}(S^{\mathrm{T}}A^{\mathrm{T}}WW^{\mathrm{T}}) \\ &\quad + \mathrm{Trace}(ASWW^{\mathrm{T}}S^{\mathrm{T}}A^{\mathrm{T}}) \end{aligned}$$

If we also treat A as a non-negative coefficient matrix, which denotes the associated degrees of each data point to the clusters, we can use the multiplicative update algorithm described in [22,25] to perform the optimization. If we require the entries in both A and X to be non-negative, the one-side clustering problem is then related to non-negative matrix factorization [16]. The minimization problem is then a constrained optimization problem which can be solved use the Lagrange multiplier methods [26].

3.2 Subspace clustering

The general model can also be reduced to subspace clustering. Many of the existing clustering algorithms do not work efficiently in high dimensional spaces (*curse of dimensionality*). As demonstrated in [1], the correlations among the dimensions are often specific to data locality, in the sense that some data points are correlated with a given set of features and others are correlated with respect to different features. In other words, in high dimensional space, each cluster usually has its own subspace structure.

To explicitly model the subspace structure for each cluster, let B be a $m \times K$ matrix, whose entries denote the coefficients of each feature associated with each cluster. Note that WB is the projection of W into the subspace defined by B. Since AX is an approximation of W, hence AXB gives the approximation of WB. To perform the subspace clustering, we want the approximation loss in the projected space to be minimized. This can be thought as a special case of the model described in Eq. (1) where the approximation error in the original space is minimized.

The approximation error in the projected space is

$$\begin{aligned} O(A, X, B) &= \|WB - AXB\|_{\mathrm{F}}^2 \\ &= \|WB - A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}WB\|_{\mathrm{F}}^2 \text{ (based on Eq. (5))} \\ &= \|[W - A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}W]B\|_{\mathrm{F}}^2 \end{aligned}$$

The columns of B are the coefficients of the features associated with different clusters. They are usually orthogonal. So, the objective criterion is minimized by taking the smallest K eigenvectors of $W^{\mathrm{T}}(I_n - A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}})W$, or equivalently, the first K eigenvectors of $W^{\mathrm{T}}(A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}} - I_n)W$. Given B, matrix A can be obtained using the least square minimization [19].

3.3 Two-side clustering

Now suppose B is not an identity matrix, then the model leads to many formulations of two-side clustering, i.e., the problem of simultaneously clustering both data points (rows) and features (columns) of a data matrix [5,13].

3.3.1 Double K-means approach

Suppose $A = (a_{ik}), a_{ik} \in \{0, 1\}, \sum_{k=1}^{K} a_{ik} = 1, B = (b_{jc}), b_{jc} \in \{0, 1\}, \sum_{c=1}^{C} b_{jc} = 1$ (i.e., A and B denote the data and feature memberships, respectively). Thus, based on Eq. (3), we obtain

$$
\begin{aligned}
O(A, X, B) &= \|W - \hat{W}\|_{\mathrm{F}}^2 \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} \left(w_{ij} - \sum_{k=1}^{K} \sum_{c=1}^{C} a_{ik} b_{jc} x_{kc} \right)^2 \\
&= \sum_{k=1}^{K} \sum_{c=1}^{C} \sum_{i \in P_k} \sum_{j \in Q_c} (w_{ij} - x_{kc})^2
\end{aligned}
\tag{8}
$$

For fixed P_k and Q_c, it is easy to check that the optimum X is obtained by

$$
x_{kc} = \frac{1}{p_k q_c} \sum_{i \in P_k} \sum_{j \in Q_c} w_{ij} \tag{9}
$$

In other words, X can be thought as the matrix of centroids for the two-side clustering problem and it represents the associations between the data clusters and the feature clusters [4]. $O(A, X, B)$ can then be minimized via an iterative procedure of the following steps:

1. Given X and B, then the feature partition Q is fixed, $O(A, X, B)$ is minimized by

$$
\hat{a}_{ik} = \begin{cases} 1 & \text{if } \sum_{c=1}^{C} \sum_{j \in Q_c} (w_{ij} - x_{kj})^2 < \sum_{c=1}^{C} \sum_{j \in Q_c} (w_{ij} - x_{lj})^2 \\ & \text{for } l = 1, \ldots, K, l \neq k \\ 0 & \text{otherwise} \end{cases} \tag{10}
$$

2. Similarly, Given X and A, then the data partition P is fixed, $O(A, X, B)$ is minimized by

$$
\hat{b}_{jc} = \begin{cases} 1 & \text{if } \sum_{k=1}^{K} \sum_{i \in P_k} (w_{ij} - x_{ic})^2 < \sum_{k=1}^{K} \sum_{i \in P_k} (w_{ij} - x_{il})^2 \\ & \text{for } l = 1, \ldots, C, l \neq c \\ 0 & \text{otherwise} \end{cases} \tag{11}
$$

3. Given A and B, X can be computed using Eq. (9).

This leads to natural extensions of the K-means type algorithm for two-side case [3,4,21]. In general, if we do not require $a_{ik} \in \{0, 1\}$ and $b_{jc} \in \{0, 1\}$, then

$$
O(A, X, B) = \sum_{k=1}^{K} \sum_{c=1}^{C} \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ik} b_{jc} (w_{ij} - x_{kc})^2
$$

For fixed A and B, the optimum X is obtained by

$$
x_{kc} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} a_{ik} b_{jc} w_{ij}}{\sum_{i \in n} \sum_{j \in m} a_{ik} b_{jc}} \tag{12}
$$

The optimization of A and B can be performed via a penalty clustering which considers both the objective function's partial derivatives and the constraints [10].

3.3.2 Iterative feature and data clustering

Consider the case when X is a diagonal matrix. Then in the general model, we have $C = K$, i.e., both data points and features have the same number of clusters. The assumption also implies that, after appropriate permutation of the rows and columns, the approximation data take the form of a block diagonal matrix [12].

When W is binary data matrix and X is identity matrix, this leads to the cluster model described in [18]. The objective function can be rewritten as

$$\begin{aligned} O(A, X, B) &= \|W - AB^{\mathrm{T}})\|_{\mathrm{F}}^2 \\ &= \mathrm{Trace}((W - AB^{\mathrm{T}})(W - AB^{\mathrm{T}})^{\mathrm{T}}) \\ &= \mathrm{Trace}(WW^{\mathrm{T}}) - 2\mathrm{Trace}(WAB^{\mathrm{T}}) \\ &\quad + \mathrm{Trace}(AB^{\mathrm{T}}AB^{\mathrm{T}})) \end{aligned}$$

Note that if we relax A and B and let them be arbitrary matrices, then based on

$$\frac{\partial O}{\partial A} = -WB + AB^{\mathrm{T}}B \tag{13}$$

$$\frac{\partial O}{\partial B} = -W^{\mathrm{T}}A + BA^{\mathrm{T}}A \tag{14}$$

we would get the optimization rules $A = WB(B^{\mathrm{T}}B)^{-1}$ and $B = W^{\mathrm{T}}A(A^{\mathrm{T}}A)^{-1}$. By imposing orthogonal requirements, we could obtain two simplified updating rules which has a natural interpretation analogous to the HITS ranking algorithm [15].

$$B = W^{\mathrm{T}}A \tag{15}$$

$$A = WB \tag{16}$$

In fact, Eqs. (15) and (16) can be thought of as the use of power iteration method for computing the singular vectors of WW^{T} [11]. Basically, the optimizing rules show a mutually reinforcing relationship between the data and the features for binary dataset which can be naturally expressed as follows: if a feature f (or, data point d) is shared by many points (or, features) that have high weights associated with a cluster c, then feature f (or, data point d) has a high weight associated with c.

3.3.3 Two-side spectral relaxation

In general, if A and B denote the cluster membership, then $A^{\mathrm{T}}A = \mathrm{diag}(p_1, \ldots, p_K)$ and $B^{\mathrm{T}}B = \mathrm{diag}(q_1, \ldots, q_C)$ are two diagonal matrices. If we relax the conditions on A and B, requiring $A^{\mathrm{T}}A = I_K$ and $B^{\mathrm{T}}B = I_C$, we would obtain a new variation of two-side clustering algorithm. Note that

$$\begin{aligned} O(A, X, B) &= \|W - AXB^{\mathrm{T}}\|_{\mathrm{F}}^2 \\ &= \mathrm{Trace}((W - AXB^{\mathrm{T}})(W - AXB^{\mathrm{T}})^{\mathrm{T}}) \\ &= \mathrm{Trace}(WW^{\mathrm{T}}) + \mathrm{Trace}(XX^{\mathrm{T}}) - 2\mathrm{Trace}(AXB^{\mathrm{T}}W^{\mathrm{T}}) \end{aligned}$$

Since $\mathrm{Trace}(WW^{\mathrm{T}})$ is constant, hence minimizing $O(A, X, B)$ is equivalent to minimizing

$$O'(A, X, B) = \mathrm{Trace}(XX^{\mathrm{T}}) - 2\mathrm{Trace}(AXB^{\mathrm{T}}W^{\mathrm{T}}). \tag{17}$$

The minimum of Eq. (17) is achieved where $X = A^{\mathrm{T}}WB$ as $\frac{\partial O'}{\partial X} = X - A^{\mathrm{T}}WB$.

Plugging $X = A^{\mathrm{T}}WB$ into Eq. (17), we have

$$\begin{aligned} O'(A, X, B) &= \mathrm{Trace}(XX^{\mathrm{T}}) - 2\mathrm{Trace}(AXB^{\mathrm{T}}W^{\mathrm{T}}) \\ &= \mathrm{Trace}(A^{\mathrm{T}}WBB^{\mathrm{T}}W^{\mathrm{T}}A) - 2\mathrm{Trace}(AA^{\mathrm{T}}WBB^{\mathrm{T}}W^{\mathrm{T}}) \\ &= \mathrm{Trace}(WW^{\mathrm{T}}) - 2\mathrm{Trace}(A^{\mathrm{T}}WBB^{\mathrm{T}}W^{\mathrm{T}}A) \end{aligned}$$

Since the first term Trace(WW^{T}) is constant, minimizing $O'(A, X, B)$ is thus equivalent to maximizing Trace$(A^{\mathrm{T}}WBB^{\mathrm{T}}W^{\mathrm{T}}A)$.

To maximize Trace$(A^{\mathrm{T}}WBB^{\mathrm{T}}W^{\mathrm{T}}A)$, we perform the following alternating optimization procedure. Let $G = WB$. Given B, A should maximize Trace$(A^{\mathrm{T}}GG^{\mathrm{T}}A)$. This can be easily obtained by constructing A with the eigenvectors of GG^{T} corresponding to the K largest eigenvalues [11]. Note that Trace$(A^{\mathrm{T}}WBB^{\mathrm{T}}W^{\mathrm{T}}A) =$ Trace$(B^{\mathrm{T}}W^{\mathrm{T}}AA^{\mathrm{T}}WB)$. Denote $H = W^{\mathrm{T}}A$. So, given A, B should maximize Trace$(B^{\mathrm{T}}HH^{\mathrm{T}}B)$. This can be easily obtained by constructing B with the eigenvectors of HH^{T} corresponding to the C largest eigenvalues [11]. The above alternative optimization procedure can be thought as a two-side generalization of spectral relaxation. After obtaining the relaxed A and B, the final cluster assignments of the data points and features are obtained by applying ordinary K-means clustering in the reduced spaces. A short description of the clustering procedure is presented as Algorithm 1.

Algorithm 1 Two-side spectral relaxation

```
Input: ($W_{n\times m}$, $K$ and $C$)
Output: $P$, $Q$: set of clusters;
begin
1       Initialize $A$;
2.      Iteration: Do while the stop criterion is not met
        begin
2.1         Update $B$ to maximize $Trace(A^T WBW^T B^T A)$
2.2         Compute $X = A^T WB$
2.3         Update $A$ to maximize $Trace(B^T W^T AA^T WB)$
        end
3.      Get the final clusterings $P$ and $Q$
end
```

It should be noted that there are some connections between the cluster solutions to iterative feature and data clustering and the two-side spectral relaxation. If we compute the QR decomposition of A and B for iterative feature and data clustering, we could obtain the cluster solutions to the two-side spectral relaxation. On the other hand, for two-side spectral relaxation, if we compute Singular Value Decomposition(SVD) of $X = USV$ and set $A = AUS$ and $B = VB$, we could derive the cluster solutions to iterative feature and data clustering.

4 Relations with other models

In this section, we show the relations between our general models with the information-theoretic clustering framework and the error-variance approach.

4.1 Information-theoretic clustering

Recently, an information-theoretic clustering framework applicable to empirical joint probability distributions was developed for two-dimensional contingency table or co-occurrence matrix [6]. In this framework, the (scaled) data matrix W is viewed as a joint probability distribution between row and column random variables taking values over the rows and columns. The clustering objective is to seek a hard-clustering of both dimensions such that loss in *mutual information* $I(W) - I(\bar{W})$, where $\bar{W}$ denotes the reduced data matrix, is minimized [23].

In this section, we explore the relations between our general framework and the information-theoretic framework. If we view entries of W as values of a joint probability distribution between row and column random variables, then $I(W) = \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} \log \frac{w_{ij}}{w_{i.}w_{.j}}$ where $w_{i.} = \sum_{j=1}^{m} w_{ij}$ and $w_{.j} = \sum_{i=1}^{n} w_{ij}$.

Once we have a simplified $K \times C$ matrix $\bar{W}$, we can construct an $n \times m$ matrix $\hat{W}$ as the approximation of original matrix W by

$$\hat{W}_{ij} = \bar{w}_{kc}\left(\frac{w_{i.}}{\bar{w}_{k.}}\right)\left(\frac{w_{.j}}{\bar{w}_{.c}}\right) \tag{18}$$

where $i \in P_k$, $j \in Q_c$ and $\bar{w}_{k.} = \sum_{c=1}^{C} \bar{w}_{kc}$ and $\bar{w}_{.c} = \sum_{k=1}^{K} \bar{w}_{kc}$. As the approximation preserves marginal probability [6], it can easily check that

$$\bar{w}_{kc} = \sum_{i\in P_k}\sum_{j\in Q_c} \hat{w}_{ij} = \sum_{i\in P_k}\sum_{j\in Q_c} w_{ij} \tag{19}$$

$$\hat{w}_{i.} = w_{i.} \tag{20}$$

$$\hat{w}_{.j} = w_{.j} \tag{21}$$

Hence we have

$$\begin{aligned}
I(\hat{W}_{ij}) &= \sum_{i=1}^{n}\sum_{j=1}^{m} \hat{w}_{ij} \log \frac{\hat{w}_{ij}}{w_{i.}w_{.j}} \quad \text{(based on Eqs. (20) and (21))}\\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} \hat{w}_{ij} \log \frac{\hat{w}_{ij}\frac{\bar{w}_{kc}}{\bar{w}_{kc}}}{\bar{w}_{k.}\left(\frac{w_{i.}}{\bar{w}_{k.}}\right)\bar{w}_{.c}\left(\frac{w_{.j}}{\bar{w}_{.c}}\right)}\\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} \hat{w}_{ij} \log \frac{\bar{w}_{kc}}{\bar{w}_{k.}\bar{w}_{.c}} \quad \text{(based on Eq. (18))}\\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} log \frac{\bar{w}_{kc}}{\bar{w}_{k.}\bar{w}_{.c}} \quad \text{(based on Eq. (19))} \qquad (22)\\
&= \sum_{k=1}^{K}\sum_{c=1}^{C} \bar{w}_{kc} log \frac{\bar{w}_{kc}}{\bar{w}_{k.}\bar{w}_{.c}} \quad \text{(based on Eq. (19))} \qquad (23)\\
&= I(\bar{W}) \qquad (24)
\end{aligned}$$

So

$$\begin{aligned}
I(W) - I(\bar{W}) &= I(W) - I(\hat{W}) \quad \text{(based on Eq. (24))}\\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} \log \frac{w_{ij}}{w_{i.}w_{.j}} - \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} \frac{\bar{w}_{kc}}{\bar{w}_{k.}\bar{w}_{.c}}\\
&\quad \text{(based on Eq. (22))}\\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} \log \frac{w_{ij}}{w_{i.}w_{.j}} - \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} \frac{\hat{w}_{ij}}{w_{i.}w_{.j}}\\
&\quad \text{(based on Eq. (18))}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} \log \frac{w_{ij}}{\hat{w}_{ij}} \\
&\approx \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{(w_{ij} - \hat{w}_{ij})^2}{w_{ij}}
\end{aligned}
\tag{25}
$$

The last step from the above derivation is based on power series approximation of logarithm. The approximation is valid if the absolute difference $|w_{ij} - \hat{w}_{ij}|$ are not large as compared with w_{ij}. The right side of Eq. (25) can be thought as a weighted version of the right side of Eq. (2). Thus minimizing the criterion $O(A, X, B)$ is conceptually consistent with the loss of *mutual information*, i.e., $I(W) - I(\bar{W})$.

4.2 Error-variance approach

It should also note that the criterion in Eq. (2) is related to the fraction of variance defined in [9]

$$V = 1 - \frac{\sum_{i,j}(w_{ij} - \hat{w_{ij}})^2}{\sum_{i,j}(w_{ij} - \bar{w_{ij}})^2}$$

where $\bar{w_{ij}} = \frac{1}{nm}\sum_{i,j} w_{ij}$. Minimizing the criterion $O(A, X, B)$ is equivalent to maximizing the variance V defined above.

5 Conclusion

In this paper, we present a generalized clustering framework by formulating the problem as matrix approximations. The clustering procedure then aims at minimizing the approximation error between the original data matrix and the reconstructed matrix induced by the cluster structures. We also provide characterizations of different clustering methods within the general framework including traditional one-side clustering, subspace clustering and two-side clustering and establish the connections between our general clustering framework with existing frameworks.

Acknowledgments The author is grateful to Dr. Shenghuo Zhu, Dr. Sheng Ma and Dr. Mitsunori Ogihara for their helpful discussions and suggestions. The work is partially supported by a NSF grant IIS-0546280, a 2005 IBM Faculty Award, and a 2005 IBM Shared University Research (SUR) Award.

References

1. Aggarwal CC, Wolf JL, Yu PS, Procopiuc C, Park JS (1999) Fast algorithms for projected clustering. In: Proceedings of the 1999 ACM SIGMOD international conference on Management of data (SIGMOD'99). ACM Press, pp 61–72
2. Ando RK, Lee L (2001) Iterative residual rescaling: an analysis and generalization of LSI. In: Proceedings of the 24th SIGIR, pp 154–162
3. Baier D, Gaul W, Schader M (1997) Two-mode overlapping clustering with applications to simultaneous benefit segmentation and market structuring. In: Klar R, Opitz O (eds) Classification and knowledge organization. Springer, Heidelberg, pp 577–566
4. Castillo W, Trejos J (2002) Two-mode partitioning: Review of methods and application and tabu search. In: Jajuga K, Sokolowski A, Bock H-H (eds) Classification, clustering and data analysis. Springer, Heidelberg, pp 43–51

5. Cho H, Dhillon IS, Guan Y, Sra S (2004) Minimum sum-squared residue co-clustering of gene experssion data. In: Proceedings of the SIAM data mining conference
6. Dhillon IS, Mallela S, Modha SS (2003) Information-theoretic co-clustering. In: Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining (SIGKDD 2003). ACM Press, pp 89–98
7. Dhillon IS, Modha DS (2001) Concept decompositions for large sparse text data using clustering. Mach Learn 42(1/2):143–175
8. Ding C, Li T, Peng W, Park H (2006) Orthogonal nonnegative matrix tri-factorizations for clustering. In: Proceedings of the twelfth ACM SIGKDD international conference on knowledge discovery and data mining, pp 126–135
9. Eckes T, Orlik P (1993) An error variance approach to two-mode hierarchical clustering. J Classif 10:52–74
10. Gaul W, Schader M (1996) A new algorithm for two-mode clustering. In: Bock H-H, Polasek W (eds) Data analysis and information systems. Springer, Heidelberg, pp 15–23
11. Golub GH, Van Loan CF (1996) Matrix computations. The Johns Hopkins University Press
12. Govaert G (1995) Simultaneous clustering of rows and columns. Control Cybernet 24(4):437–458
13. Hartigan JA (1975) Clustering algorithms. Wiley, New York
14. Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice Hall, Englewood Cliffs
15. Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632
16. Lee DD, Sebastian Seung H (2000) Algorithms for non-negative matrix factorization. In: NIPS, pp 556–562
17. Li T (2005) A general model for clustering binary data. In: KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, pp 188–197
18. Li T, Ma S (2004) IFD: iterative feature and data clustering. In: Proceedings of the 2004 SIAM international conference on data mining (SDM 2004). SIAM
19. Li T, Ma S, Ogihara M (2004) Document clustering via adaptive subspace iteration. In: Proceedings of twenty-seventh annual international ACM SIGIR conference on research and development in information retrieval (SIGIR 2004), pp 218–225
20. Long B, Zhang Z, Yu PS (2005) Co-clustering by block value decomposition. In: KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining, pp 635–640
21. Maurizio V (2001) Double k-means clustering for simultaneous classification of objects and variables. In: Borra S, Rocci R, Vichi M, Schader M (eds) Advances in classification and data analysis. Springer, Heidelberg, pp 43–52
22. Sha F, Saul LK, Lee DD (2002) Multiplicative updates for nonegative quadratic programming in support vector machines. In: Advances in neural information processing systems, pp 1065–1072
23. Slonim N, Tishby N (2000) Document clustering using word clusters via the information bottleneck method. In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'00). ACM Press, pp 208–215
24. Soete GD, douglas Carroll J (1994) K-means clustering in a low-dimensional euclidean space. In: New approaches in classification and data analysis. Springer, Heidelberg, pp 212–219
25. Xu W, Gong Y (2004) Document clustering by concept factorization. In: SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval. ACM Press, pp 202–209
26. Xu W, Liu X, Gong Y (2003) Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval(SIGIR'03). ACM Press, pp 267–273
27. Zha H, He X, Ding C, Simon H (2001) Spectral relaxation for k-means clustering. In: Proceedings of neural information processing systems

Author Biography

Dr. Tao Li is currently an assistant professor in the School of Computing and Information Sciences at Florida International University in Miami. He received his Ph.D. degree in Computer Science from University of Rochester in 2004. He is the recipient of a NSF CAREER Award (2006–2011) and IBM Faculty Research Awards (2005 & 2007). His primary research interests are: data mining, machine learning, music information retrieval, and bioinformatics.