Information-theoretic Co-clustering

by Inderjit S. Dhillon et. al.

Presenter: Yiwei Zhang[1] Supervisor: Dr. Ben Kao[1]

[1]Department of Computer Science
The University of Hong Kong

May 30, 2008

Outline

1. Introduction and Motivation
   - Clustering
   - Co-clustering
   - Information Theory
2. Problem Formulation
3. Algorithm
   - Matrix Approximation
   - Demonstration
   - Finding the Optimal Arrangement
   - Pseudo-code
   - Convergence
4. Experimental Results
5. Remarks

Clustering

- Clustering: an important task in data mining.
  - Maximize the inter-cluster distances
  - Minimize the intra-cluster distances
- Obstacles:
  - High Dimensionality
  - Sparsity
  - Objective Function
- Need for robust and scalable clustering algorithms.

Co-clustering: Definition

Co-clustering

also known as *Bi-clustering*, refers to **simultaneous** clustering along multiple dimensions.

Two-dimensional(or two-side) Co-clustering

1. Represent the data by a contingency table or a matrix
2. Do clustering of rows and columns simultaneously.

Example: document-word co-clustering

- cluster the documents according to the words that they contain
- simultaneously, cluster the words according to the documents that contain them.

Co-clustering: Motivation

1. Simultaneously cluster row data and column data.
   - For example, we may be interested in finding similar documents and their interplay with word clusters
2. *Row clustering* and *column clustering* can bootstrap each other.
   - Actually a kind of subspace clustering.
   - Implicitly performs an adaptive dimensionality reduction at each iteration.
3. Robust to sparsity.

Information Theory: Motivation

Clustering Obstacles:

1. High Dimensionality: co-clustering
2. Sparsity: co-clustering
3. Objective Function: ?

Main Contribution

Use *information-theoretic* technique to construct the objective function

Information Theory: Basic Concepts

Entropy

Entropy of a random variable X with probability distribution $p(x)$ is defined by:

$$\mathcal{H}(X) = -\sum_{x \in X} p(x) \log p(x), \tag{1}$$

which measures the *uncertainty* of the random variable X.

Information Theory: Basic Concepts

KL Distance or Relative Entropy

Relative Entropy between two probability distributions p and q is defined by:

$$\mathcal{KL}(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)}, \tag{2}$$

which measures the *distance* between two distributions.

Information Theory: Basic Concepts

Mutual Information

Mutual Information between two random variables X and Y is:

$$\mathcal{I}(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x) \cdot p(y)}, \tag{3}$$

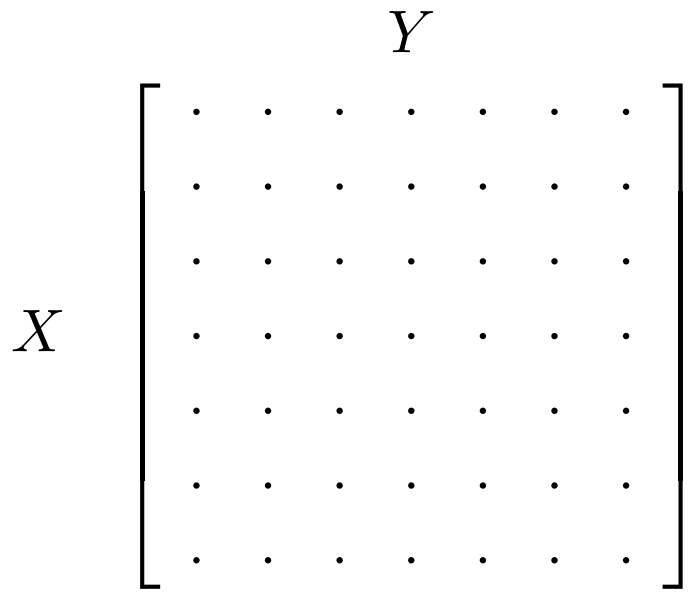
which measures the amount of information that one random variable contains about another random variable.

A Word about Notations

Notations	Examples	Meanings
upper-letters	X, Y	random variable
lower-letters	x, y	elements of the set
hatted letters	$\hat{X}$, $\hat{x}$	variables associated with clusters

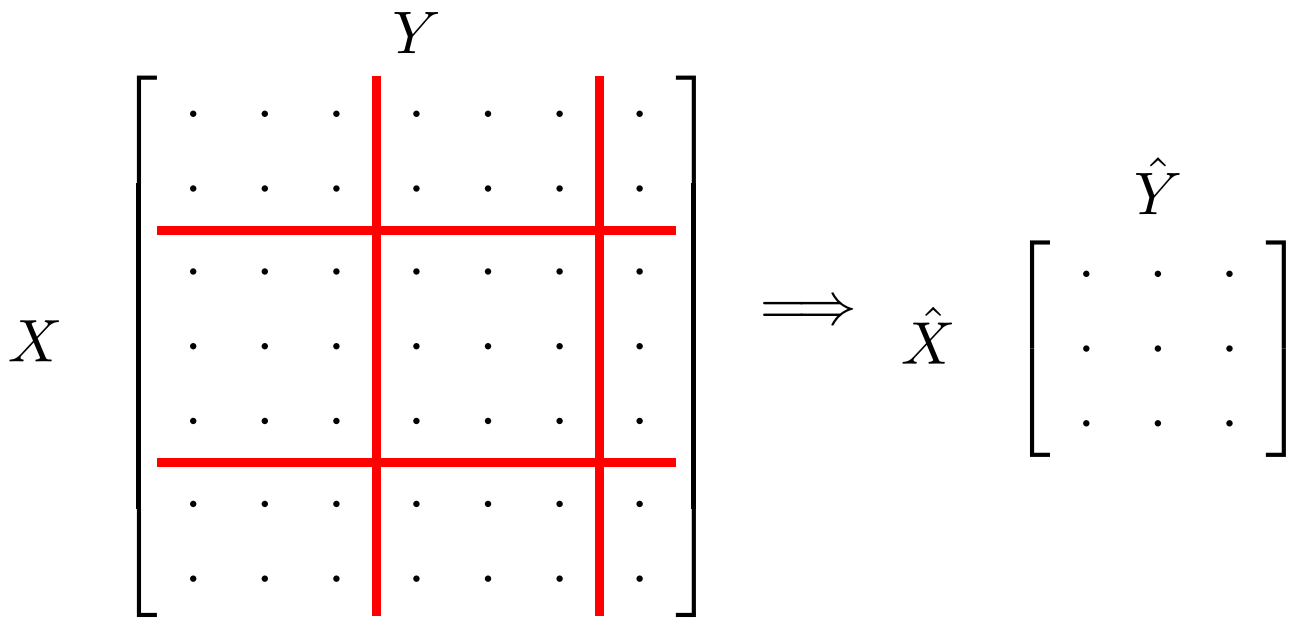
Problem Formulation

1. represent the data in a contingency table
2. treat the contingency table as a joint probability distribution between row and column random variables.



Problem Formulation

3. our object is:



given the *fixed number* of row and column clusters desired, seek a *hard*-clustering of both dimensions such that minimizes *loss in mutual information*

$$\mathcal{I}(X;Y) - \mathcal{I}(\hat{X};\hat{Y})$$

Lemma 1

Lemma

$$\mathcal{I}(X,Y) - \mathcal{I}(\hat{X},\hat{Y}) = \mathcal{KL}(p(x,y) \parallel q(x,y)), \tag{4}$$

where

$$q(x,y) = p(\hat{x},\hat{y}) \cdot p(x|\hat{x}) \cdot p(y|\hat{y}) \ \textit{where}\ x \in \hat{x} \ \textit{and}\ y \in \hat{y}.$$

Object Conversion

Minimize the loss of mutual information

↓

Find an approximation $q(X,Y)$ to $p(X,Y)$ with minimal *KL Distance*

Lemma 1

Proof.

Since we are considering hard clustering,

$$p(\hat{x}, \hat{y}) = \sum_{x \in \hat{x}} \sum_{y \in \hat{y}} p(x, y) \tag{5}$$

$$p(\hat{x}) = \sum_{x \in \hat{x}} p(x) \tag{6}$$

$$p(\hat{y}) = \sum_{y \in \hat{y}} p(y) \tag{7}$$

Lemma 1

Proof.(Cont.)

$$
\begin{aligned}
\mathcal{I}(X;Y) - \mathcal{I}(\hat{X};\hat{Y}) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} - \sum_{\hat{x}} \sum_{\hat{y}} p(\hat{x},\hat{y}) \log \frac{p(\hat{x},\hat{y})}{p(\hat{x})p(\hat{y})} \\
&= \sum_{\hat{x}} \sum_{\hat{y}} \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{\hat{x}} \sum_{\hat{y}} \sum_x \sum_y \left(p(x,y) \left(\log \frac{p(x,y)}{p(x)p(y)} - \log \frac{p(\hat{x},\hat{y})}{p(\hat{x})p(\hat{y})} \right) \right) \\
&= \sum_{\hat{x}} \sum_{\hat{y}} \sum_x \sum_y \left(p(x,y) \cdot \log \frac{p(x,y)}{p(\hat{x},\hat{y})p(x|\hat{x})p(y|\hat{y})} \right) \\
&= \sum_{\hat{x}} \sum_{\hat{y}} \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)} \\
&= \mathcal{KL}(p \parallel q)
\end{aligned}
$$

Demonstration: Rearrange

$$p(x,y) = \begin{bmatrix} 0 & .05 & 0 & .05 & .05 & 0 \\ .05 & 0 & .05 & 0 & 0 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .05 & 0 & .05 & 0 & 0 & .05 \\ 0 & .05 & 0 & .05 & .05 & 0 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix} \longrightarrow \begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix}$$

Demonstration: Co-cluster

$p(x, y) =$

$$\begin{bmatrix} 0 & .05 & 0 & .05 & .05 & 0 \\ .05 & 0 & .05 & 0 & 0 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .05 & 0 & .05 & 0 & 0 & .05 \\ 0 & .05 & 0 & .05 & .05 & 0 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix} \longrightarrow \left[\begin{array}{ccc|ccc} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ \hline .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{array}\right]$$

Demonstration: Co-cluster

$$\left[\begin{array}{ccc|ccc} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ \hline .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{array}\right]$$

$$\left[\begin{array}{ccc} .5 & 0 & 0 \\ .5 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .5 \\ 0 & 0 & .5 \end{array}\right] \quad \left[\begin{array}{cc} .3 & 0 \\ 0 & .3 \\ .2 & .2 \end{array}\right] \quad \left[\begin{array}{cccccc} .36 & .36 & .28 & 0 & 0 & 0 \\ 0 & 0 & 0 & .28 & .36 & .36 \end{array}\right]$$

$p(x|\hat{x})$ $\quad$ $p(\hat{x}, \hat{y})$ $\quad$ $p(y|\hat{y})$

Demonstration: Calculate KL Divergence

$$q(x, y) = p(x|\hat{x}) \cdot p(\hat{x}, \hat{y}) \cdot p(y|\hat{y})$$

$$\Longrightarrow \begin{bmatrix} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ .036 & .036 & 0.28 & .028 & .036 & .0.36 \\ .036 & .036 & 0.28 & .028 & .036 & .0.36 \end{bmatrix} = q(x, y)$$

KL Distance

$$\mathcal{KL}(p \parallel q) = 0.0957,$$

which is the smallest loss in mutual information.

Marginal Probability Preservation

$$p(x,y)\begin{bmatrix} .05 & .05 & .05 & 0 & 0 & 0 \\ .05 & .05 & .05 & 0 & 0 & 0 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ 0 & 0 & 0 & .05 & .05 & .05 \\ .04 & .04 & 0 & .04 & .04 & .04 \\ .04 & .04 & .04 & 0 & .04 & .04 \end{bmatrix} \quad q(x,y)\begin{bmatrix} .054 & .054 & .042 & 0 & 0 & 0 \\ .054 & .054 & .042 & 0 & 0 & 0 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ 0 & 0 & 0 & .042 & .054 & .054 \\ .036 & .036 & 0.28 & .028 & .036 & .0.36 \\ .036 & .036 & 0.28 & .028 & .036 & .0.36 \end{bmatrix}$$

Proposition

$$p(x) = q(x) \text{ and } p(y) = q(y)$$

Proof.

$$\begin{aligned} q(x) &= \sum_{y} q(x,y) = \sum_{\hat{y}} \sum_{y \in \hat{y}} p(\hat{x},\hat{y}) p(x|\hat{x}) p(y|\hat{y}) \\ &= p(x|\hat{x}) \sum_{\hat{y}} p(\hat{x},\hat{y}) \sum_{y \in \hat{y}} p(y|\hat{y}) = p(x|\hat{x}) p(\hat{x}) = p(x,\hat{x}) = p(x) \end{aligned}$$

□

General Idea

Iterative Method

- inspired by the idea of *k-mean* clustering, at each iteration,
- cluster rows according to the *"row cluster centroids"*;
- simultaneously, cluster columns according to the *"column cluster centroids"*;
- until convergence.

Problem

How to define the "centroid", more generally, the *cluster prototype*?

Lemma 2

Lemma

$$\begin{aligned}\mathcal{KL}(p(x,y) \parallel q(x,y)) &= \sum_{\hat{x}} \sum_{x \in \hat{x}} p(x) \mathcal{KL}(p(y|x) \parallel q(y|\hat{x})) \\ &= \sum_{\hat{y}} \sum_{y \in \hat{y}} p(y) \mathcal{KL}(p(x|y) \parallel q(x|\hat{y}))\end{aligned}$$

Lemma 2

Proof.

$$
\begin{aligned}
\mathcal{KL}(p(x,y) \parallel q(x,y)) &= \sum_{\hat{x}} \sum_{x \in \hat{x}} \sum_{\hat{y}} \sum_{y \in \hat{y}} p(x,y) \log \frac{p(x,y)}{q(x,y)} \\
&= \sum_{\hat{x}} \sum_{x \in \hat{x}} \sum_{\hat{y}} \sum_{y \in \hat{y}} (p(x)p(y|x)) \log \frac{p(x)p(y|x)}{p(x)q(y|\hat{x})} \\
&= \sum_{\hat{x}} \sum_{x \in \hat{x}} p(x) \sum_{\hat{y}} \sum_{y \in \hat{y}} p(y|x) \log \frac{p(y|x)}{q(y|\hat{x})} \\
&= \sum_{\hat{x}} \sum_{x \in \hat{x}} p(x) \mathcal{KL}(p(y|x) \parallel q(y|\hat{x}))
\end{aligned}
$$

□

Lemma 2

Proof.

$$
\begin{aligned}
\mathcal{KL}(p(x,y) \parallel q(x,y)) &= \sum_{\hat{x}} \sum_{x \in \hat{x}} \sum_{\hat{y}} \sum_{y \in \hat{y}} p(x,y) \log \frac{p(x,y)}{q(x,y)} \\
&= \sum_{\hat{x}} \sum_{x \in \hat{x}} \sum_{\hat{y}} \sum_{y \in \hat{y}} (p(x)p(y|x)) \log \frac{p(x)p(y|x)}{p(x)q(y|\hat{x})} \\
&= \sum_{\hat{x}} \sum_{x \in \hat{x}} p(x) \sum_{\hat{y}} \sum_{y \in \hat{y}} p(y|x) \log \frac{p(y|x)}{q(y|\hat{x})} \\
&= \sum_{\hat{x}} \sum_{x \in \hat{x}} p(x) \mathcal{KL}(p(y|x) \parallel q(y|\hat{x}))
\end{aligned}
$$

$$
q(y|\hat{x}) = q(y, \hat{y}|\hat{x}) = \frac{q(y, \hat{y}, \hat{x})}{q(\hat{x})} = q(y|\hat{y}, \hat{x}) q(\hat{y}|\hat{x}) = q(y|\hat{y}) q(\hat{y}|\hat{x})
$$

Lemma 2

Lemma

$$\begin{aligned}\mathcal{KL}(p(x,y) \parallel q(x,y)) &= \sum_{\hat{x}} \sum_{x \in \hat{x}} p(x) \mathcal{KL}(p(y|x) \parallel q(y|\hat{x})) \\ &= \sum_{\hat{y}} \sum_{y \in \hat{y}} p(y) \mathcal{KL}(p(x|y) \parallel q(x|\hat{y}))\end{aligned}$$

Lemma 2

Lemma

$$\begin{aligned}\mathcal{KL}(p(x,y) \parallel q(x,y)) &= \sum_{\hat{x}} \sum_{x \in \hat{x}} p(x) \mathcal{KL}(\boxed{p(y|x)} \parallel q(y|\hat{x})) \\ &= \sum_{\hat{y}} \sum_{y \in \hat{y}} p(y) \mathcal{KL}(p(x|y) \parallel q(x|\hat{y}))\end{aligned}$$

- $p(Y|x)$: a piece of row data x.

Lemma 2

Lemma

$$\mathcal{KL}(p(x,y) \parallel q(x,y)) = \sum_{\hat{x}} \sum_{x \in \hat{x}} p(x) \mathcal{KL}(p(y|x) \parallel \boxed{q(y|\hat{x})})$$
$$= \sum_{\hat{y}} \sum_{y \in \hat{y}} p(y) \mathcal{KL}(p(x|y) \parallel q(x|\hat{y}))$$

- $p(Y|x)$: a piece of row data x.
- $q(Y|\hat{x})$: the prototype of row cluster $\hat{x}$.

Lemma 2

Lemma

$$\begin{aligned}\mathcal{KL}(p(x,y) \parallel q(x,y)) &= \sum_{\hat{x}} \sum_{x \in \hat{x}} p(x) \boxed{\mathcal{KL}(p(y|x) \parallel q(y|\hat{x}))} \\ &= \sum_{\hat{y}} \sum_{y \in \hat{y}} p(y) \mathcal{KL}(p(x|y) \parallel q(x|\hat{y}))\end{aligned}$$

- $p(Y|x)$: a piece of row data x.
- $q(Y|\hat{x})$: the prototype of row cluster $\hat{x}$.
- $\mathcal{KL}(p(y|x) \parallel q(y|\hat{x}))$: the distance between x and row cluster prototype $\hat{x}$.

Method Skeleton

minimize the loss in mutual information: $\mathcal{I}(X;Y) - \mathcal{I}(\hat{X};\hat{Y})$

↓ Lemma 1

find approximiated $q(X,Y)$ to $p(X,Y)$ with minimal $\mathcal{KL}(p \parallel q)$

↓ Lemma 2

iteratively find *row prototypes* and *column prototypes* until convergence

Algorithm 1: Information-theoretic Co-clustering Algorithm

Input: $p(X,Y), k, l$
Output: C_X, C_Y

1 Initialization: Set $t = 0$. Start with some initial partition functions C_X^0, C_Y^0. Compute

$$q^0(\hat{X},\hat{Y}), q^0(X|\hat{X})), q^0(Y|\hat{Y})$$

and the distributions $q^0(Y|\hat{x})1 \leq \hat{x} \leq k$;

2 Compute row clusters: For each row x, find its new cluster index as

$$C_X^{t+1} = \text{argmin}_{\hat{x}} \mathcal{KL}(p(Y|x)||q^t(Y|\hat{x})),$$

resolving ties arbitrarily. Let $C_Y^{t+1} = C_Y^t$;

3 Compute distributions

$$q^{t+1}(\hat{X},\hat{Y}), q^{t+1}(X|\hat{X})), q^{t+1}(Y|\hat{Y})$$

and $q^{t+1}(X|\hat{y})1 \leq \hat{y} \leq l$;

4 Compute column clusters: For each column y, find its new cluster index as

$$C_Y^{t+2} = \text{argmin}_{\hat{y}} \mathcal{KL}(p(X|y)||q^{t+1}(X|\hat{y})),$$

resolving ties arbitrarily. Let $C_X^{t+2} = C_X^{t+1}$;

5 Compute

$$q^{t+1}(\hat{X},\hat{Y}), q^{t+2}(X|\hat{X})), q^{t+2}(Y|\hat{Y})$$

and the distribution $q^{t+2}(Y|\hat{x})1 \leq \hat{x} \leq k$;

6 **if** $\mathcal{KL}(p(X,Y) \parallel q^t(X,Y)) - \mathcal{KL}(p(X,Y) \parallel q^{t+2}(X,Y)) \leq \epsilon$ **then**

7 **return** $C_X = C_X^{t+2}, C_Y = C_Y^{t+2}$;

8 **else**

9 Set $t = t + 2$;

10 Goto step 2;

Convergence of Algorithm

Theorem

Algorithm 1 monotonically decreases the objective function

$$\mathcal{KL}(p(x,y) \parallel q(x,y)).$$

Convergence of Algorithm

Theorem

Algorithm 1 monotonically decreases the objective function

$$\mathcal{KL}(p(x,y) \parallel q(x,y)).$$

Proof.

...

□

Convergence of Algorithm

Theorem

Algorithm 1 monotonically decreases the objective function

$$\mathcal{KL}(p(x,y) \parallel q(x,y)).$$

Proof.

...

□

Conclusion

Algorithm 1 is guaranteed to convergent and returns a *local optimal* co-clustering.

Datasets

1. 20 Newsgroups data
   - 20 classes, 20000 documents
2. Classic3 data set
   - 3 classes, 3893 documents.

Result: Classic3

Co-clustering			1D-clustering		
992	4	8	944	9	98
40	1452	7	71	1431	5
1	4	1387	18	20	1297

Table: CLASSIC3: Co-clustering recovers original clusters more accurately.

Result: Convergence

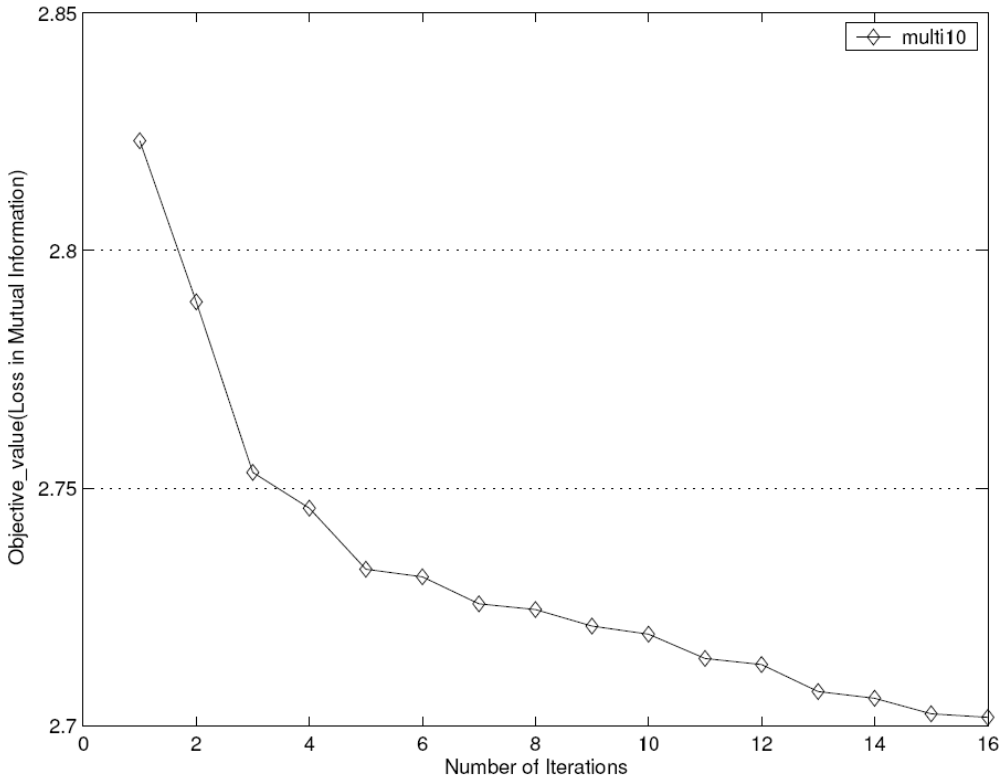


Figure: Loss in mutual information decreases monotonically with the number of iterations.

Remarks

- Theoretically solid paper.
- The method has a flavor of *k-means*. But it uses different formula to compute cluster prototype (centroid in k-means).
- Needs to specify the number of clusters of row and column in advance.
- Return a local optimal solution, therefore the initial values are important to the algorithm.

Thank You