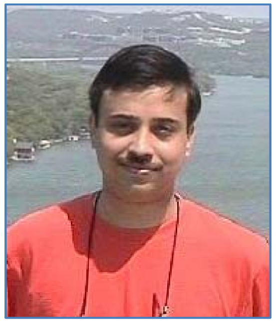
A Probabilistic Framework for Semi-Supervised Clustering

Sugato Basu, Mikhail Bilenko, and Raymond J. Mooney@SIGKDD 2004



Presenter: Jun-Cheng Chen

Date: 11/01/2011

Outline

- Introduction
  - Problem Description
- Hidden Markov Random Field (HMRF) Framework
- Adaptive Distortion(Distance) Measures
- HMRF-KMeans (EM algorithm)
- Experiments
- Summary

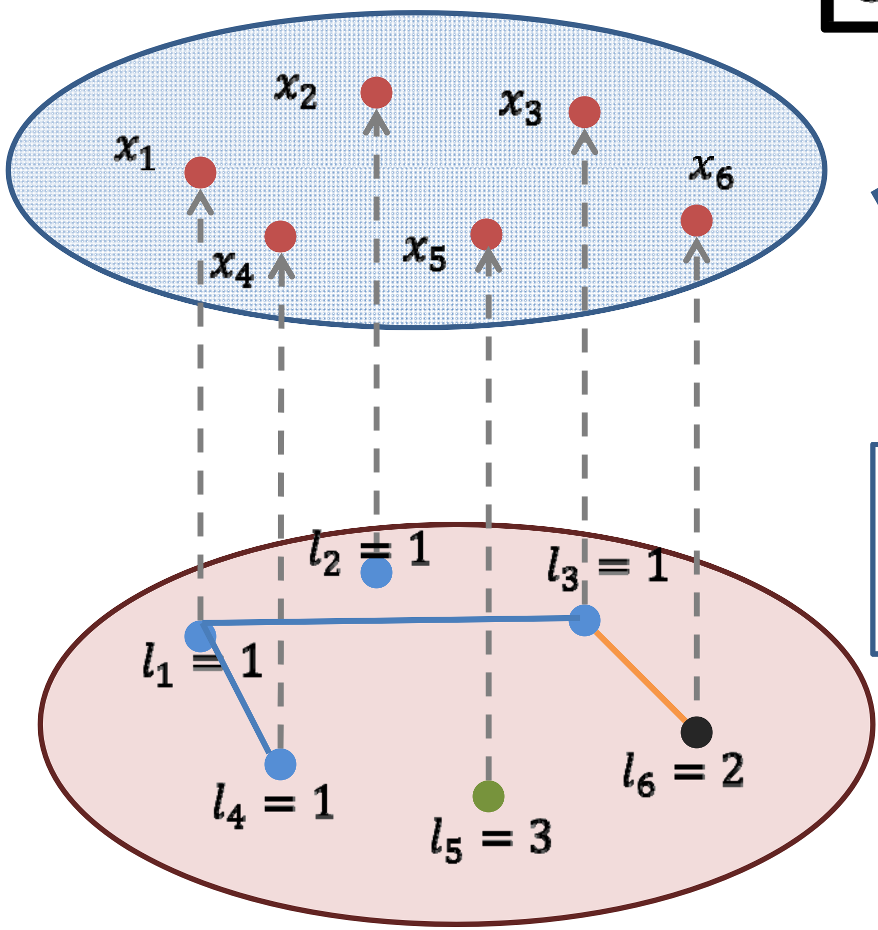
Introduction

- The paper focused on partition-based clustering
  - The number of clusters is given.
  - e.g. K-Means
- Semi-supervised learning
  - It makes use of both labeled and unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data. -- Wikpedia

The problem!

Clustering problem → Labeling problem

Observation



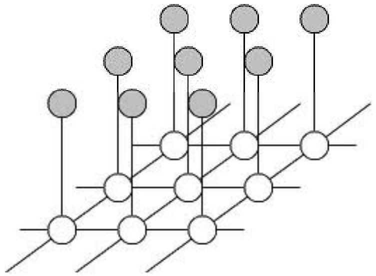
A Hidden Markov Random Field

Goal: Given x_i Infer its label ℓ_i

Must link constraints $\mathcal{M}$

Cannot link constraints $\mathcal{C}$

Hidden Markov Random Field (HMRF)



1. Hidden field (labels): $\mathcal{L} = \{l_i\}_{i=1}^{N}$
2. Observations (data): $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N}$
3. Markov property: (Conditional Independence)

$$\forall i.\,\mathbf{Pr}(l_i | \mathcal{L} - \{l_i\}) = \mathbf{Pr}(l_i | \{l_j : l_j \in \mathcal{N}_i\})$$

4. $\mathbf{Pr}(\mathcal{L})$ can be expressed as a Gibbs distribution

$$\mathbf{Pr}(\mathcal{L}) = \frac{1}{Z_1}\exp(-V(\mathcal{L})) = \frac{1}{Z_1}\exp(-\sum_{\mathcal{N}_i \in \mathcal{N}} V_{\mathcal{N}_i}(\mathcal{L}))$$

$$V(i,j) = \begin{cases} f_M(\mathbf{x}_i, \mathbf{x}_j) & \text{if } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{M} \\ f_C(\mathbf{x}_i, \mathbf{x}_j) & \text{if } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{C} \\ 0 & \text{otherwise} \end{cases}$$

Clique Potential function

M: Must link set, *C*: Cannot link set

MAP Estimation in HMRFs

- Posterior probability: $\mathbf{Pr}(\mathcal{L}|\mathcal{X}) \propto \mathbf{Pr}(\mathcal{L})\mathbf{Pr}(\mathcal{X}|\mathcal{L})$

Cluster Representatives

Distortion/Distance

– $\mathbf{Pr}(\mathcal{X}|\mathcal{L}) = p(\mathcal{X}, \{\boldsymbol{\mu}_h\}_{h=1}^{K}) \boxed{= \frac{1}{Z_3} \exp(-\sum_{x_i \in \mathcal{X}} D(\mathbf{x}_i, \boldsymbol{\mu}_{l_i}))}$

– $\mathbf{Pr}(\mathcal{L}|\mathcal{X}) = p(\mathcal{X}, \{\boldsymbol{\mu}_h\}_{h=1}^{K}) \cdot \left(\frac{1}{Z_2} \exp(-\sum_i \sum_j V(i,j))\right)$

Take Logarithm,

$$\boxed{\begin{aligned} \mathcal{J}_{\text{obj}} = & \textstyle\sum_{x_i \in \mathcal{X}} D(\mathbf{x}_i, \boldsymbol{\mu}_{l_i}) + \sum_{(x_i,x_j) \in \mathcal{M}} w_{ij} \varphi_D(\mathbf{x}_i, \mathbf{x}_j) \mathbb{1}[l_i \neq l_j] \\ & + \textstyle\sum_{(x_i,x_j) \in \mathcal{C}} \overline{w}_{ij} \left(\varphi_{D\max} - \varphi_D(\mathbf{x}_i, \mathbf{x}_j)\right) \mathbb{1}[l_i = l_j] + \log Z \end{aligned}}$$

where

$$\boxed{V(i,j) = \begin{cases} f_M(\mathbf{x}_i, \mathbf{x}_j) & \text{if } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{M} \\ f_C(\mathbf{x}_i, \mathbf{x}_j) & \text{if } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{C} \\ 0 & \text{otherwise} \end{cases}}$$

$$f_M(\mathbf{x}_i, \mathbf{x}_j) = w_{ij} \varphi_D(\mathbf{x}_i, \mathbf{x}_j) \mathbb{1}[l_i \neq l_j]$$

$$f_C(\mathbf{x}_i, \mathbf{x}_j) = \overline{w}_{ij} \left(\varphi_{D\max} - \varphi_D(\mathbf{x}_i, \mathbf{x}_j)\right) \mathbb{1}[l_i = l_j]$$

Adaptive Distortion Measures

- Different distortion measures assume different cluster condition probability distribution
  - e.g A Normal distribution if we use L_2 norm:
  - $p(X, \{\boldsymbol{\mu}_h\}_{h=1}^{K}) = \frac{1}{Z_3} \exp(-\sum_{x_i \in X} D(\mathbf{x}_i, \boldsymbol{\mu}_{l_i}))$

- I-divergence (aka. Generalized KL divergence)
  - Bregman divergence (i.e. proved convergence rate on expected error) Stephen Wright's NIPS tutorial on optimization
  - $D_I(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{d} x_{im} \log \frac{x_{im}}{x_{jm}} - \sum_{m=1}^{d} (x_{im} - x_{jm})$, x_i and x_j can be vectors here.
  - Multinomial distribution

- Directional similarity
  - Cosine distance
  - Von Mises-Fisher distribution

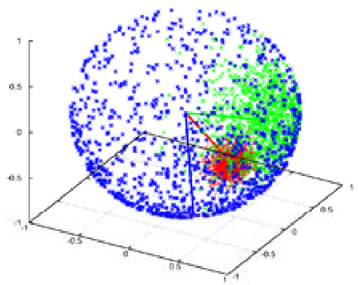


Image credit to wikipedia

Parameterized Distortion functions$^{(1/2)}$

- Parameterized Cosine Similarity
  - $D_{\cos_{\mathbf{A}}}(\mathbf{x}_i, \mathbf{x}_j) = 1 - \dfrac{\mathbf{x}_i^T \mathbf{A} \mathbf{x}_j}{\|\mathbf{x}_i\|_{\mathbf{A}} \|\mathbf{x}_j\|_{\mathbf{A}}}$, where $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^T \mathbf{A} \mathbf{x}}$, $\mathbf{A}^{1/2}\colon \mathbf{x} \rightarrow \mathbf{A}^{1/2}\mathbf{x}$
  - For simplicity, consider $\mathbf{A}$ as a diagonal matrix (i.e. $\mathbf{a} = \mathrm{diag}(\mathbf{A})$)

$$\boxed{\begin{aligned}\mathcal{J}_{\text{obj}} = {} & \textstyle\sum_{x_i \in \mathcal{X}} D(\mathbf{x}_i, \boldsymbol{\mu}_{l_i}) \\ & + \textstyle\sum_{(x_i,x_j) \in \mathcal{M}} w_{ij} \varphi_D(\mathbf{x}_i, \mathbf{x}_j) \mathbb{1}[l_i \neq l_j] \\ & + \textstyle\sum_{(x_i,x_j) \in \mathcal{C}} \overline{w}_{ij} \left(\varphi_{D\max} - \varphi_D(\mathbf{x}_i, \mathbf{x}_j)\right) \mathbb{1}[l_i = l_j] + \log Z\end{aligned}}$$

$$\varphi_D(\mathbf{x}_i, \mathbf{x}_j) = D_{\cos_{\mathbf{a}}}(\mathbf{x}_i, \mathbf{x}_j)$$

$$\boxed{\begin{aligned}\mathcal{J}_{\cos_{\mathbf{a}}} = {} & \sum_{x_i \in \mathcal{X}} D_{\cos_{\mathbf{a}}}(\mathbf{x}_i, \boldsymbol{\mu}_{l_i}) \\ & + \sum_{(x_i,x_j) \in \mathcal{M}} w_{ij} D_{\cos_{\mathbf{a}}}(\mathbf{x}_i, \mathbf{x}_j) \mathbb{1}[l_i \neq l_j] \\ & + \sum_{(x_i,x_j) \in \mathcal{C}} \overline{w}_{ij} \left(D_{\cos_{\mathbf{a}}\max} - D_{\cos_{\mathbf{a}}}(\mathbf{x}_i, \mathbf{x}_j)\right) \mathbb{1}[l_i = l_j] + \log Z\end{aligned}}$$

Parameterized Distortion functions$^{(2/2)}$

- Parameterized I-Divergence

$$D_{I_{\mathbf{a}}}(\mathbf{x}_i,\mathbf{x}_j)=\sum_{m=1}^{d}a_m x_{im}\log\frac{x_{im}}{x_{jm}}-\sum_{m=1}^{d}a_m(x_{im}-x_{jm})$$

$$\begin{aligned}\mathcal{J}_{\text{obj}} &= \textstyle\sum_{x_i\in\mathcal{X}} D(\mathbf{x}_i,\boldsymbol{\mu}_{l_i})\\ &\quad+\textstyle\sum_{(x_i,x_j)\in\mathcal{M}} w_{ij}\varphi_D(\mathbf{x}_i,\mathbf{x}_j)\mathbb{1}[l_i\neq l_j]\\ &\quad+\textstyle\sum_{(x_i,x_j)\in\mathcal{C}} \overline{w}_{ij}\left(\varphi_{D\max}-\varphi_D(\mathbf{x}_i,\mathbf{x}_j)\right)\mathbb{1}[l_i=l_j]+\log Z\end{aligned}$$

$$\varphi_D(\mathbf{x}_i,\mathbf{x}_j)=D_{IM_{\mathbf{a}}}(\mathbf{x}_i,\mathbf{x}_j)$$

$$\begin{aligned}\mathcal{J}_{I_{\mathbf{a}}} &= \sum_{x_i\in\mathcal{X}} D_{I_{\mathbf{a}}}(\mathbf{x}_i,\boldsymbol{\mu}_{l_i})\\ &\quad+\sum_{(x_i,x_j)\in\mathcal{M}} w_{ij}D_{IM_{\mathbf{a}}}(\mathbf{x}_i,\mathbf{x}_j)\mathbb{1}[l_i\neq l_j]\\ &\quad+\sum_{(x_i,x_j)\in\mathcal{C}} \overline{w}_{ij}\left(D_{IM_{\mathbf{a}}\max}-D_{IM_{\mathbf{a}}}(\mathbf{x}_i,\mathbf{x}_j)\right)\mathbb{1}[l_i=l_j]+\log Z\end{aligned}$$

Overview of HMRF-Kmeans

Algorithm: HMRF-KMEANS

Input: Set of data points $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$, number of clusters K,
set of *must-link* constraints $\mathcal{M} = \{(\mathbf{x}_i, \mathbf{x}_j)\}$,
set of *cannot-link* constraints $\mathcal{C} = \{(\mathbf{x}_i, \mathbf{x}_j)\}$,
distance measure D, constraint violation costs $\mathcal{W}$ and $\overline{\mathcal{W}}$.

Output: Disjoint K-partitioning $\{\mathcal{X}_h\}_{h=1}^K$ of $\mathcal{X}$ such that
objective function $\mathcal{J}_{\mathrm{obj}}$ in Eqn.(9) is (locally) minimized.

Method:

1. Initialize the K clusters centroids $\{\boldsymbol{\mu}_h^{(0)}\}_{h=1}^K$, set t $\leftarrow$ 0
2. Repeat until *convergence*

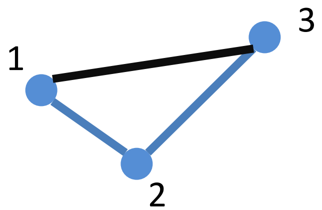
2a. `E-step`: Given $\{\boldsymbol{\mu}_h^{(t)}\}_{h=1}^K$, re-assign cluster labels $\{l_i^{(t+1)}\}_{i=1}^N$ on the points $\{\mathbf{x}_i\}_{i=1}^N$ to minimize $\mathcal{J}_{\mathrm{obj}}$.

2b. `M-step(A)`: Given cluster labels $\{l_i^{(t+1)}\}_{i=1}^N$, re-calculate cluster centroids $\{\boldsymbol{\mu}_h^{(t+1)}\}_{h=1}^K$ to minimize $\mathcal{J}_{\mathrm{obj}}$.

2c. `M-step(B)`: Re-estimate distance measure D to reduce $\mathcal{J}_{\mathrm{obj}}$.

2d. t $\leftarrow$ t+1

HMRF-Kmeans$^{(1/3)}$



- Initialization: (Find good representatives)
  - Neighborhood inference
    - Find all possible connected components by transitive closure on $\mathcal{M}$ and $\mathcal{C}$ (i.e. $\{\mathcal{N}_p\}_{p=1}^{\lambda}$ and $\{\mathcal{N}_{p'}\}_{p=1}^{\lambda}$)
    - Augment the pairwise constraints (must-link and cannot-link)
  - Cluster Selection
    - Choose K representatives from the cluster centers of $\{\mathcal{N}_p\}_{p=1}^{\lambda}$
      - Weighted farthest first (to select centroids that are relatively far apart and large in size)
    - If $K > \lambda$, choose $K - \lambda$ rep. as the global centroid with random pertubation.

HMRF-KMeans$^{(2/3)}$

- E-Step:
  - Assign each x_i to a cluster using ICM (Iterated conditional mode)

$$\mathcal{J}_{\text{obj}}(\mathbf{x}_i, \boldsymbol{\mu}_h) = D(\mathbf{x}_i, \boldsymbol{\mu}_h) + \sum_{(x_i, x_j) \in \mathcal{M}} w_{ij} \varphi_D(\mathbf{x}_i, \mathbf{x}_j) \mathbb{1}[h \neq l_j] + \sum_{(x_i, x_j) \in \mathcal{C}} \overline{w}_{ij} \left(\varphi_{D\max} - \varphi_D(\mathbf{x}_i, \mathbf{x}_j)\right) \mathbb{1}[h = l_j]$$

- M-Step:
  - Re-estimate each cluster representatives

$$D_{I_{\mathbf{a}}} \ : \ \boldsymbol{\mu}_h^{(I_{\mathbf{a}})} = \frac{1}{1+\alpha} \left(\frac{\sum_{\mathbf{x}_i \in \mathcal{X}_h} \mathbf{x}_i}{|\mathcal{X}_h|} + \alpha \frac{1}{n} \right) \quad [1]$$

$$D_{\cos_{\mathbf{a}}} \ : \ \boldsymbol{\mu}_h^{(\cos_{\mathbf{a}})} = \frac{\sum_{\mathbf{x}_i \in \mathcal{X}_h} \mathbf{x}_i}{\|\sum_{\mathbf{x}_i \in \mathcal{X}_h} \mathbf{x}_i\|_{\mathbf{A}}} \quad [2]$$

[1] I.S. Dhillon et al. ,Information theoretic clustering of sparse co-occurrence data, ICDM03, 2003
[2] A. Banerjee et al. ,Generative model-based clustering of directional data, SIGKDD03, 2003

HMRF-Kmeans$^{(3/3)}$

(M-step Contd.: Update for distance metric)

- **Gradient Descent:**

$$a_m = a_m + \eta \frac{\partial \mathcal{J}_{\text{obj}}}{\partial a_m} \Rightarrow \frac{\partial \mathcal{J}_{\text{obj}}}{\partial a_m} = \sum_{x_i \in X} \frac{\partial D(\mathbf{x}_i, \boldsymbol{\mu}_{l_i})}{\partial a_m} + \sum_{(x_i, x_j) \in \mathcal{M}} w_{ij} \frac{\partial D(\mathbf{x}_i, \mathbf{x}_j)}{\partial a_m} \mathbb{1}[l_i \neq l_j] + \sum_{(x_i, x_j) \in \mathcal{C}} \overline{w}_{ij} \left[\frac{\partial D_{\max}}{\partial a_m} - \frac{\partial D(\mathbf{x}_i, \mathbf{x}_j)}{\partial a_m} \right] \mathbb{1}[l_i = l_j]$$

$$\frac{\partial D_{\cos_{\mathrm{A}}}(\mathbf{x}_i, \mathbf{x}_j)}{\partial a_m} = \frac{x_{im} x_{jm} \|\mathbf{x}_i\|_{\mathbf{A}} \|\mathbf{x}_j\|_{\mathbf{A}} - \mathbf{x}_i^T \mathbf{A} \mathbf{x}_j \frac{x_{im}^2 \|\mathbf{x}_j\|_{\mathbf{A}}^2 + x_{jm}^2 \|\mathbf{x}_i\|_{\mathbf{A}}^2}{2 \|\mathbf{x}_i\|_{\mathbf{A}} \|\mathbf{x}_j\|_{\mathbf{A}}}}{\|\mathbf{x}_i\|_{\mathbf{A}}^2 \|\mathbf{x}_j\|_{\mathbf{A}}^2}$$

$$\frac{\partial D_{I_{\mathrm{A}}}(\mathbf{x}_i, \mathbf{x}_j)}{\partial a_m} = x_{im} \log \frac{x_{im}}{x_{jm}} - (x_{im} - x_{jm})$$

Hope: similar points are brought closer, while dissimilar points are pulled apart through this modification

Experiments(1/6)

- 3 Text Datasets (from 20-Newsgroups collection [4])
  - Each dataset consists of 3 newsgroups

    (1) News-Similar-3 : 300 points in 1864 dims

    (2) News-Related-3 : 300 points in 3225 dims

    (3) News-Different-3 : 300 points in 3251 dims
  - Preprocessed by
    - (1) Stop-word removal, (2) TF-IDF weighting,
      (3) Removal of very high and low-frequency words, etc.

Experiments(2/6)

- Evaluation criterion
  - Normalized mutual Information

$$NMI = \frac{I(C;K)}{(H(C)+H(K))/2}$$

where

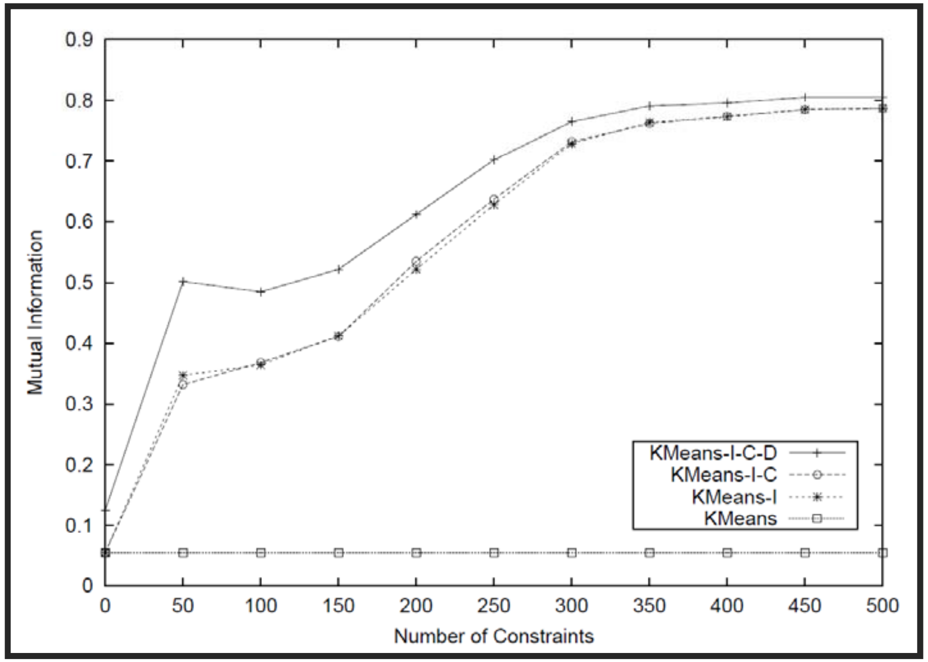
- *Cluster Assignment*: C
- *Class Label*: K
- *Sha* $H(X)$
- *Mutual Information*: $I(X;Y) = H(X) - H(X|Y)$

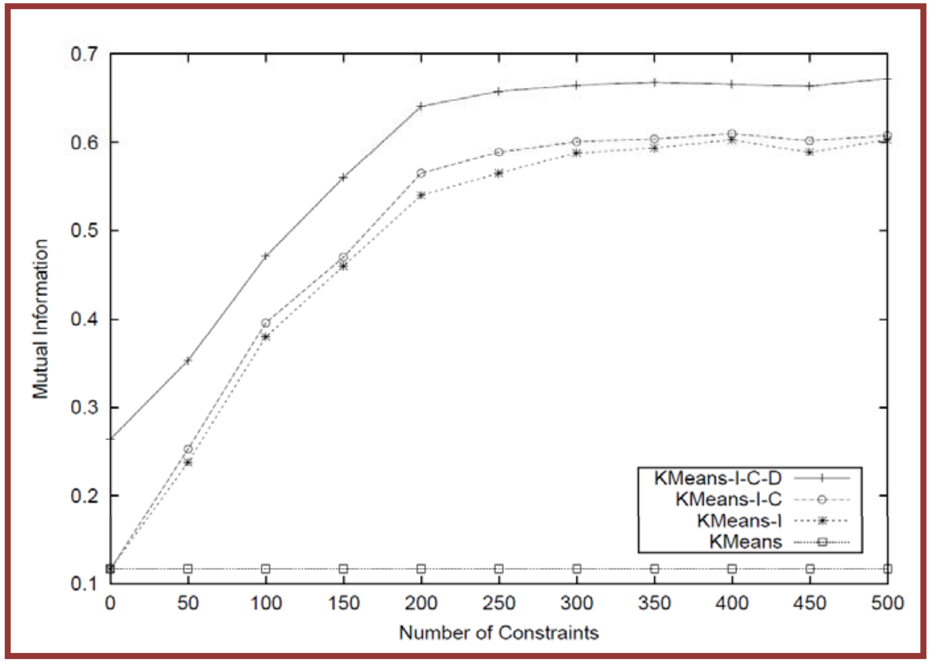
Experiments(3/6)

- I: use of supervised data in initialization
- C: incorporate constraints in cluster assignment
- D: perform distance learning

Experiments(4/6)

- Clustering on News-Different-3 dataset
  - (1) alt.atheism, (2) rec.sport.baseball, and (3) sci.space

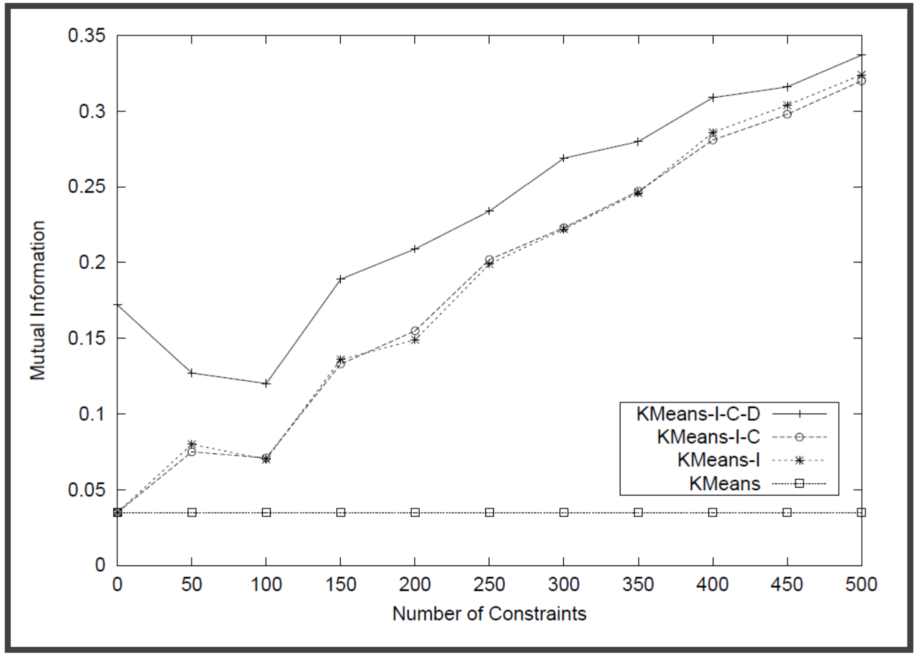




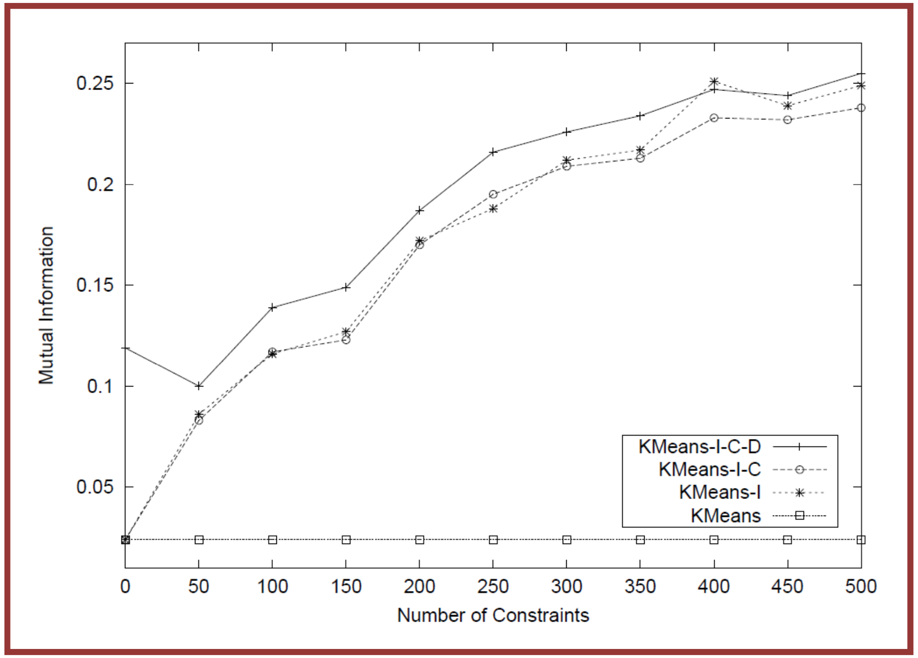
$D_{\cos_a}$ D_{I_a}

Experiments$^{(5/6)}$

- Clustering on News-Related-3 dataset
  - (1) talk.politics.misc, (2) talk.politics.guns, and (3) talk.politics.mideast



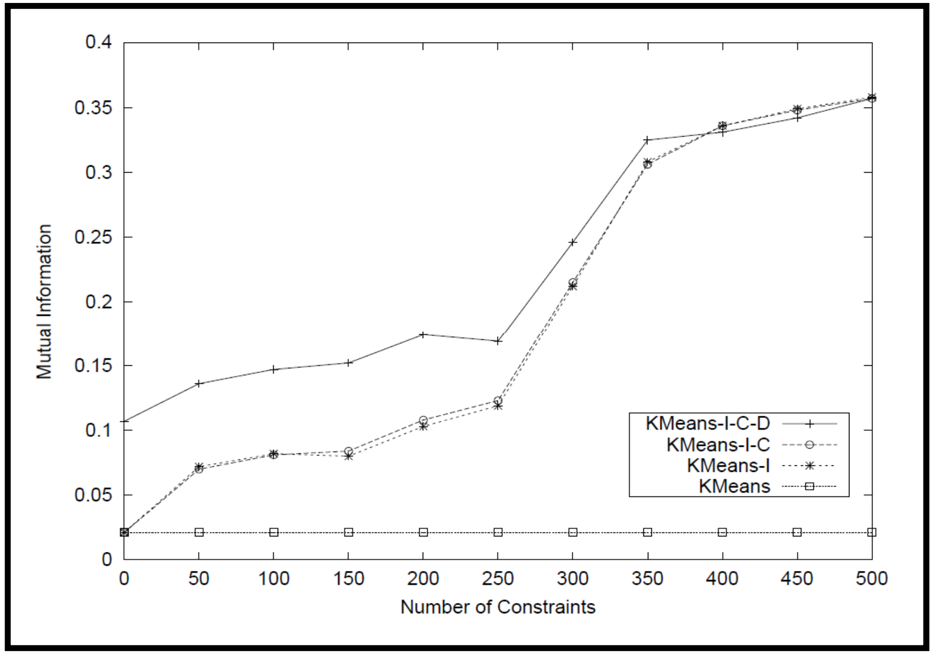
$D_{\cos_a}$



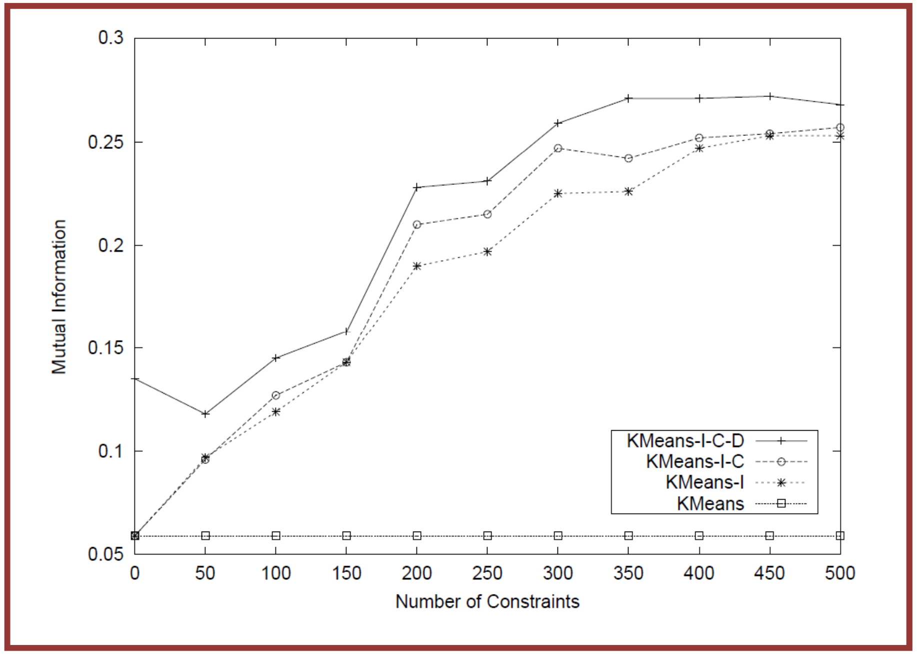
D_{I_a}

Experiments$^{(6/6)}$

- Clustering on News-Similar-3 dataset
  - (1) comp.graphics, (2) comp.os.ms-windows, and (3) comp.windows.x



$D_{\cos_a}$



D_{I_a}

Summary

- The framework combines distance and constraint clustering together.
- It uses labeled data to gain better initialization (i.e. # of clusters have been pre-specified).
- Cannot handle non-partitional clustering problem.
- Do not address how to choose appropriate distance function.

Reference

1. Basu, S. and Bilenko, M. and Mooney, R.J. "A probabilistic framework for semi-supervised clustering", ACM SIGKDD, 2004
2. Online video: Basu, S. "Semi-supervised Clustering: Probabilistic Models, Algorithms and Experiments"
   – http://research.microsoft.com/apps/video/default.aspx?id=104674
3. The presentation slide from the author:
   – http://209.128.81.248/view/1812e2-NmQ4M/A_Probabilistic_Framework_for_SemiSupervised_Clustering_flash_ppt_presentation
4. 20-Newsgroups collection
   – http://www.ai.mit.edu/people/jrennie/20Newsgroups

Q&A

Thank you!!!

Backup Slides

- Parameterized I-Divergence

$$D_{I_{\mathbf{a}}}(\mathbf{x}_i,\mathbf{x}_j) = \sum_{m=1}^{d} a_m x_{im} \log \frac{x_{im}}{x_{jm}} - \sum_{m=1}^{d} a_m (x_{im} - x_{jm})$$

$$\boxed{\begin{aligned}\mathcal{J}_{\text{obj}} = {} & \textstyle\sum_{x_i \in \mathcal{X}} D(\mathbf{x}_i, \boldsymbol{\mu}_{l_i}) \\ & + \textstyle\sum_{(x_i,x_j)\in\mathcal{M}} w_{ij}\varphi_D(\mathbf{x}_i,\mathbf{x}_j)\mathbb{1}[l_i \neq l_j] \\ & + \textstyle\sum_{(x_i,x_j)\in\mathcal{C}} \overline{w}_{ij}\left(\varphi_{D\max} - \varphi_D(\mathbf{x}_i,\mathbf{x}_j)\right)\mathbb{1}[l_i = l_j] + \log Z\end{aligned}}$$

$$\begin{aligned}\varphi_D(\mathbf{x}_i,\mathbf{x}_j) &= D_{IM_{\mathbf{a}}}(\mathbf{x}_i,\mathbf{x}_j) \\ &= \sum_{m=1}^{d} a_m \Big(x_{im} \log \frac{2x_{im}}{x_{im}+x_{jm}} \\ &\quad + x_{jm} \log \frac{2x_{jm}}{x_{im}+x_{jm}}\Big)\end{aligned}$$

$$\boxed{\begin{aligned}\mathcal{J}_{I_{\mathbf{a}}} = {} & \sum_{x_i \in \mathcal{X}} D_{I_{\mathbf{a}}}(\mathbf{x}_i, \boldsymbol{\mu}_{l_i}) \\ & + \sum_{(x_i,x_j)\in\mathcal{M}} w_{ij} D_{IM_{\mathbf{a}}}(\mathbf{x}_i,\mathbf{x}_j)\mathbb{1}[l_i \neq l_j] \\ & + \sum_{(x_i,x_j)\in\mathcal{C}} \overline{w}_{ij}\left(D_{IM_{\mathbf{a}}\max} - D_{IM_{\mathbf{a}}}(\mathbf{x}_i,\mathbf{x}_j)\right)\mathbb{1}[l_i = l_j] + \log Z\end{aligned}}$$