A Hierarchical Information Theoretic Technique for the Discovery of Non Linear Alternative Clusterings

James Bailey and Xuan-Hong Dang

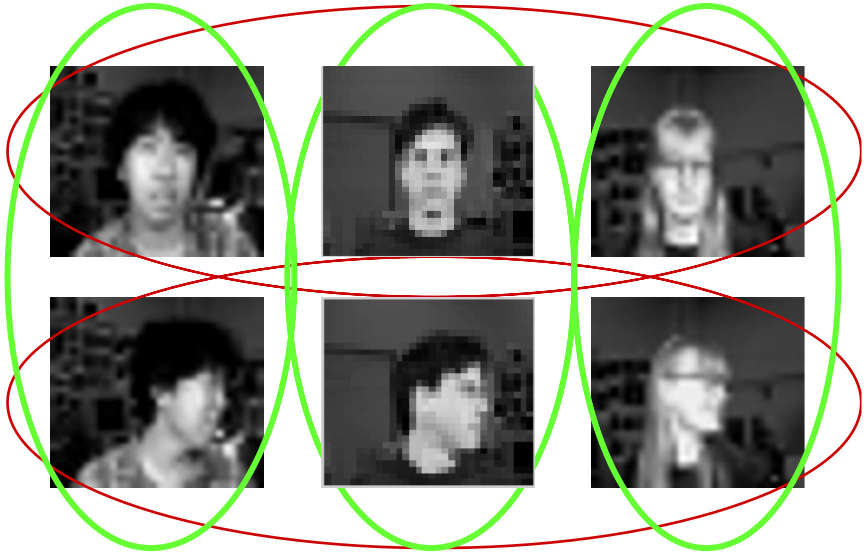
Department of Computer Science and Software Engineering
The University of Melbourne, Australia

KDD 2010

Introduction

- Cluster analysis: group "similar" objects into clusters
- No single solution
- Examples:
  - Documents
  - Genes
  - Images

=> Equally important, different views regarding the data



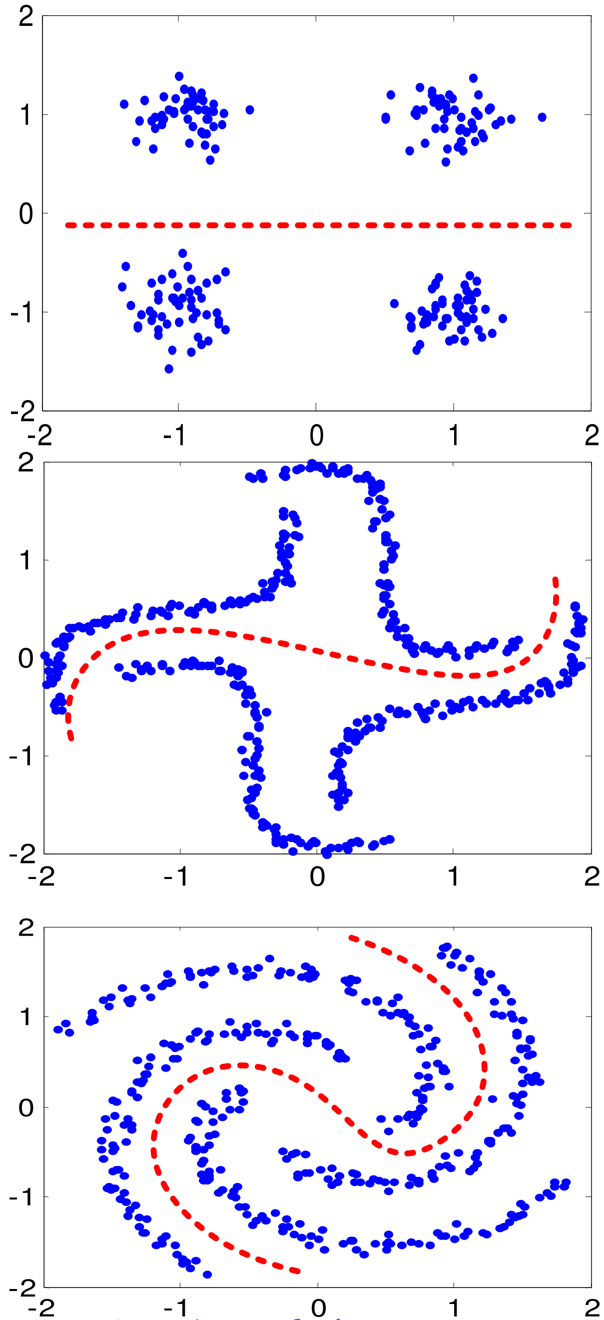
Cluster by pose or individual (CMU data)?

Presentation Outline

- Introduction
- Clustering Objectives
- Information Theoretic Approach
- Experiments
- Conclusions
- Q&A

Clustering Objectives

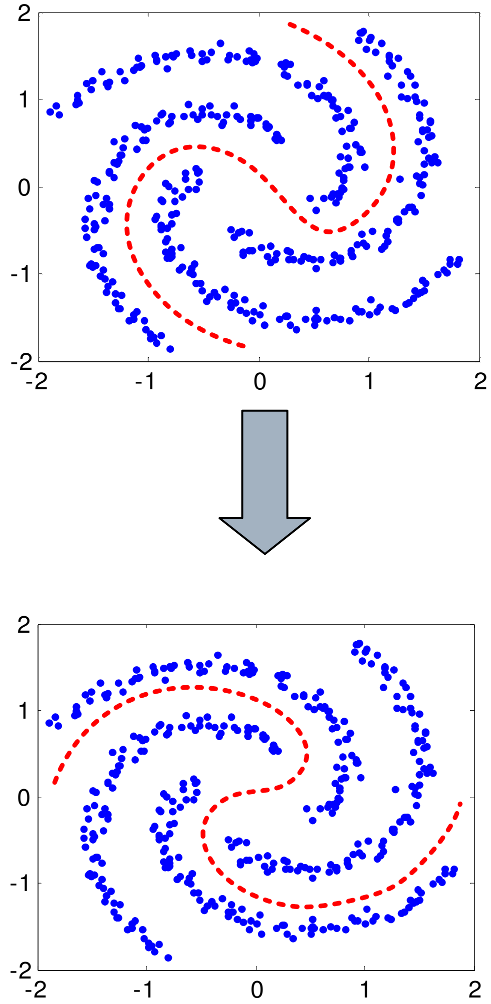
- Many algorithms have been developed!
  - Assumptions about data distributions (implicitly/explicitly) made.
- We address different aspect:
  - No assumptions imposed regarding data distributions
  - Clusters' boundary functions can be non-linear!



Number of clusters: K=2

Clustering Objectives

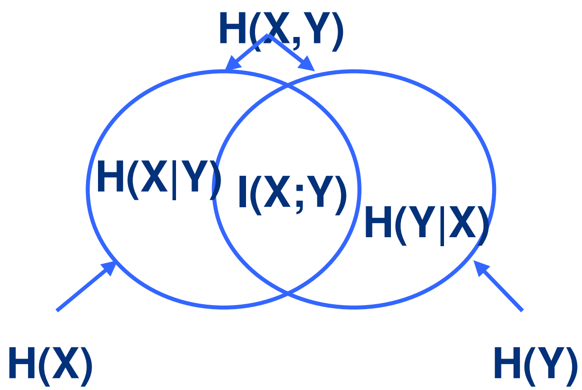
- Given a dataset X = {x1,…,xn} and a reference clustering C^-
- Find C^+ from X s.t.
  - High dissimilarity (from C^-)
  - High quality (strong prob. relationship with X)
- Purely relying on Information Theory; fully exploit information embedded in data



Information Theoretic Approach

- Lower bound for probability of error (Fano's theorem):

$$Pr(c^+ \neq \widehat{c^+}) \geq \frac{H(C^+|X) - 1}{\log(|C^+|)} = \frac{H(C^+) - I(C^+;X)}{\log(|C^+|)}$$



X, Y are random variables
H(X): Entropy of X
H(X|Y): Cond. entropy of X given Y
I(X;Y): mutual info. btw X and Y

- C^+ has little uncertainty given observation X

- X contains much information of C^+.

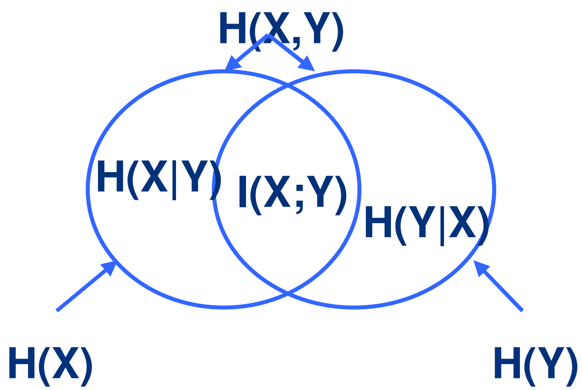
- Thus, a good clustering if C^+ and X has strong probabilistic relationship.

Information Theoretic Approach

- Our dual-objective clustering function:

$$C^+ = \arg\max_{C^+}\{I(C^+;X) - \eta I(C^+;C^-)\}$$

  - C^+ and X are statistically *dependent*
  - C^+ and C^- are statistically *independent*



X, Y are random variables
H(X): Entropy of X
H(X|Y): Cond. entropy of X given Y
I(X;Y): mutual info. btw X and Y

- Unfortunately, estimating I(X;Y) in Shannon's definition is practically hard

$$I(X;Y) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy$$
$$= D_{KL}(p(x,y)\,||\,p(x)p(y))$$

  - Require availability of all variables' distributions
  - Numerical integration

Information Theoretic Approach

- Our task is to optimize MI, rather than computing it exactly.
- In such cases, a more general divergence can be used:

$$D(p||q) = \frac{1}{\alpha(\alpha-1)} \sum_{i=1}^{n} \left(p^{\alpha}(x_i) - \alpha \frac{p(x_i)}{q^{1-\alpha}(x_i)} + (\alpha - 1) q^{\alpha}(x_i)\right)$$

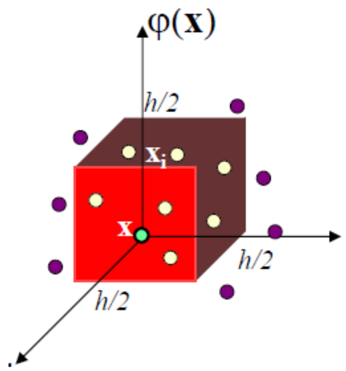
where $\alpha \neq 0, 1$.

- Selecting $\alpha=2$ results in Quadratic Mutual Information (with Renyi entropy):

$$I_{R_2}(X;Y) = \iint (p(x,y) - p(x)p(y))^2 dxdy$$

  - In quadratic form, but practically computed from data!

Information Theoretic Approach



Hypercube kernel

- Why?
  - Non-parametric methods for pdfs estimation
    - no assumptions of the underlying densities' form
    - approx. for arbitrary distributions
- Parzen-windows:
  - Placing kernels at data samples

and density is sum of kernels

$$p(x) = \frac{1}{n}\sum_{i=1}^{n} G(x - x_i, \sigma^2)$$

(info. potential, local interaction between xi and xj)

  - Note for Gaussian kernel, convolution of 2 Gausses

$$\int G(x - x_i, \sigma^2) G(x - x_j, \sigma^2) dx = G(x_i - x_j, 2\sigma^2)$$

  - Computing quadratic MI is thus

computationally *INexpensive* when

combined with Parzen-windows.

With $p(x \mid c_i^+) = \frac{1}{n_i}\sum_{l=1}^{n_i} G(x - x_l, \sigma^2)$

$$I_{R_2}(C^+; C^-) = \sum_{c_i^+}\sum_{c_j^-} (p(c_i^+, c_j^-) - p(c_i^+)p(c_j^-))^2$$

$$I_{R_2}(C^+; X) = \sum_{c_i^+}\int_x (p(c_i^+, x) - p(c_i^+)p(x))^2 dx$$

Information Theoretic Approach

- Problem is simple with a hierarchical clustering technique
  - Start with n clusters and merging 2 at each iterative step.
  - Classical similarity matrix is replaced by two matrices:
    - D_{in}: account for variation in $I_{R2}(C^+;X)$
    - D_{btw}: account for variation in $I_{R2}(C^+;C^-)$
  - c_β^+ is merged to c_α^+ if

$$(\alpha, \beta) = \arg\max_{i,j}\{D_{in} - \eta D_{btw}\}$$

  - *Given matrix of info. potentials between any 2 samples,* D_{in} and D_{btw} are computed easily (see paper).

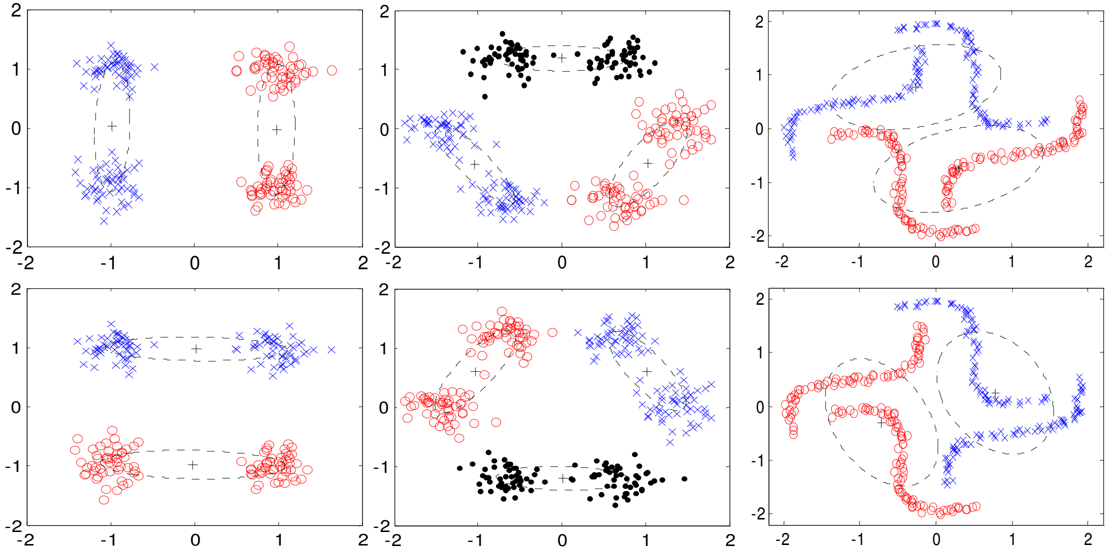
Information Theoretic Approach

- Clustering quality depends on kernel parameter sigma.
  - Work reasonably well for many datasets when sigma is selected s.t. mean squared error between estimator and true density p(x) is optimized.
- Algorithm complexity
  - Matrix of local interactions (info. potentials) between any 2 data samples: O(dn*n)
  - Calculation of MI's variation: O(n*n)
  - Search and delete element from matrix O(n*log(n))
  - Since n-1 steps of merging, overall complexity is O(n*nlog(n)+dn*n)
  - *Same time as that of a conventional tech. using group-avg similarity*

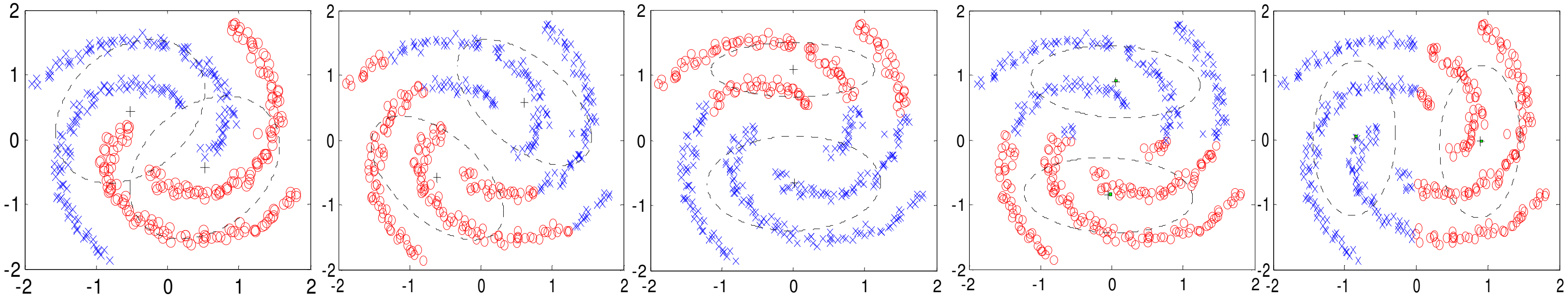
Experiments

- Compared against 8 other algorithms
- Use 4 syn. datasets and 4 real-world datasets
- Evaluation based on
  - Clustering quality (*higher -> better*)
    - F-measure if knowing true labels
    - Dunn Index if not
  - Clustering dissimilarity (*smaller -> better*)
    - Normalized Mutual Information
    - Jaccard Index

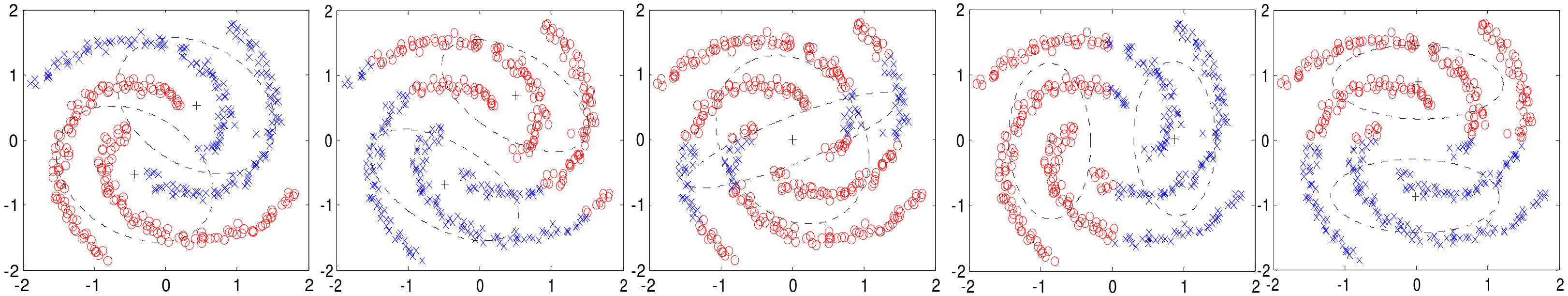
Experiments



(a) Syn1 dataset (b) Syn2 dataset (c) Syn3 dataset



(a) Pre-defined clustering (c) Algo1's alternative clustering (e) COALA's alternative clustering (g) Dec-kmeans's two alternative clusterings



(b) NACI's alternative clustering (d) ADFT's alternative clustering (f) CIB's alternative clustering (h) CAMI's two alternative clusterings

Experiments

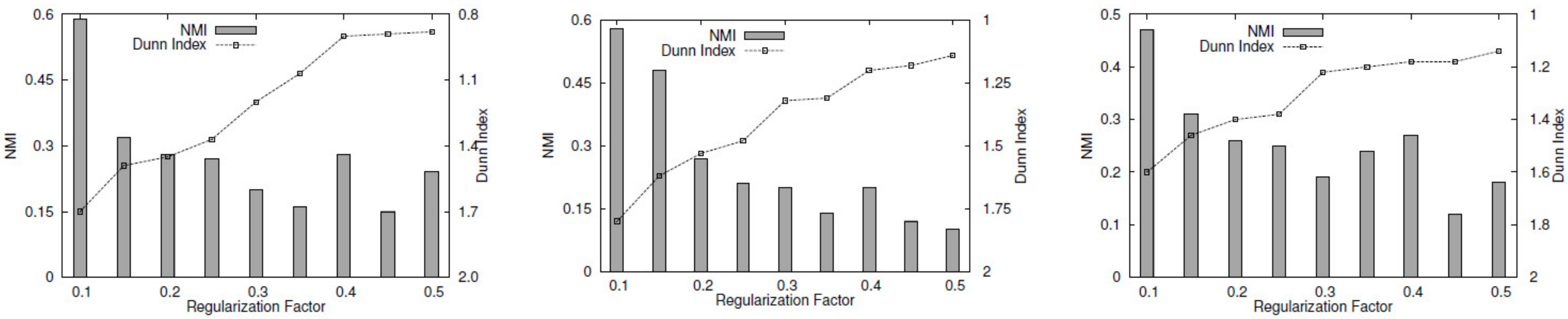
Results on CMU dataset

Methods	NMI	JI	F(pose)	F(person)
Algo1	0.31	0.34	0.68	0.87
Algo2	0.33	0.36	0.67	0.84
ADFT	0.29	0.33	0.69	0.89
COALA	0.27	0.32	0.71	0.87
CIB	0.28	0.34	0.69	0.86
Dec-kmeans	0.26	0.32	0.72	0.9
ConvEM	0.28	0.33	0.7	0.89
CAMI	0.24	0.31	0.74	0.89
NACI	**0.2**	**0.24**	**0.81**	**0.94**

Table 1: Results on CMU dataset

Methods	Segmentation			Vehicle			Vowel		
	NMI	JI	DI	NMI	JI	DI	NMI	JI	DI
Algo1	0.51	0.38	1.31	0.38	0.39	1.28	0.42	0.19	1.27
Algo2	0.44	0.3	1.27	0.39	0.44	1.46	0.43	0.21	1.3
ADFT	0.46	0.31	1.3	0.35	0.37	1.42	0.48	0.33	**1.41**
COALA	0.44	0.29	1.25	0.29	0.35	1.51	0.36	0.27	1.29
CIB	0.45	0.32	1.32	0.33	0.41	1.39	0.41	0.26	1.25
Deckm	0.39	0.29	1.26	0.26	0.36	1.4	0.27	0.17	1.26
ConvEM	0.41	0.3	1.27	0.25	0.34	1.41	0.31	0.19	1.23
CAMI	0.31	0.27	1.44	0.23	0.32	**1.53**	0.24	0.11	1.38
NACI	**0.26**	**0.25**	**1.46**	**0.21**	**0.28**	1.51	**0.22**	**0.11**	1.38

Table 2: Results on 3 real world datasets



(a) Segmentation dataset (b) Vehicle dataset (c) Vowel dataset

Impact of trade-off factor

Conclusions

- An unsupervised learning technique directly address non-linear boundary clustering function
- No assumptions made about data distributions
- Firmly rooted from information theory
- Well performing on various benchmark datasets
- Future work: convert to iterative approach to reduce computation time

Thank you
(Q&A)