Hierarchical Clustering of Uncertain Data

F. Gullo G. Ponti
Advisor: Prof. Sergio Greco

Department of Electronics, Computer and Systems Science (DEIS)
University of Calabria, Italy

GII Doctoral School on Advances in Databases - 2009
Doctoral Symposium

Outline

Motivations

Uncertain Data

Uncertainty is inherently present in a wide range of emerging application domains:

- sensor data
- location-based services (e.g., moving objects data)
- biomedical and biometric data (e.g., gene expression data)
- distributed applications
- RFID data
- ...

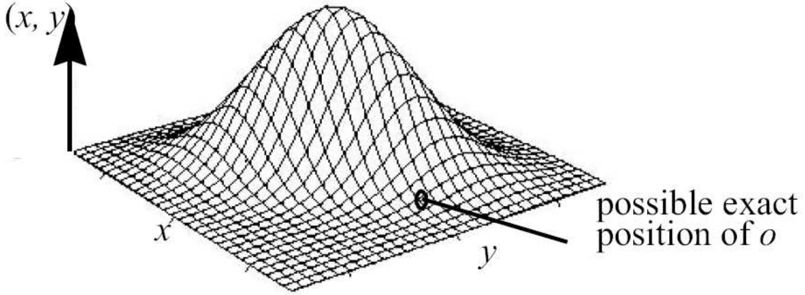
Motivations (2)

Mining of uncertain Data

- For application domains producing data inherently imprecise, uncertainty should be carefully taken into account to avoid wrong results
- But traditional data mining techniques are designed to work only on "deterministic" data
- Two possible solutions:
  - deterministic representation of uncertain data: traditional methods are still working without any change
  - uncertain data modeled in a more accurate way: traditional methods must be redesigned to deal with the new kind of representation

Uncertain Data Objects

Modeling by *regions (domains) of definition* and *probability density functions (pdfs)*



Multivariate Uncertain Objects

- m-dimensional region
- multivariate pdf defined over the region

Definition (multivariate uncertain object)

A multivariate uncertain object o is a pair (R, f):

- $R = [l^{(1)}, u^{(1)}] \times \cdots \times [l^{(m)}, u^{(m)}]$
- $f : \Re^m \rightarrow \Re_0^+$ is a multivariate pdf such that:

$$\int_{\vec{x} \in \Re^m \setminus R} f(\vec{x}) \mathrm{d}\vec{x} = 0 \qquad \text{and} \qquad f(\vec{x}) > 0, \ \forall \vec{x} \in R$$

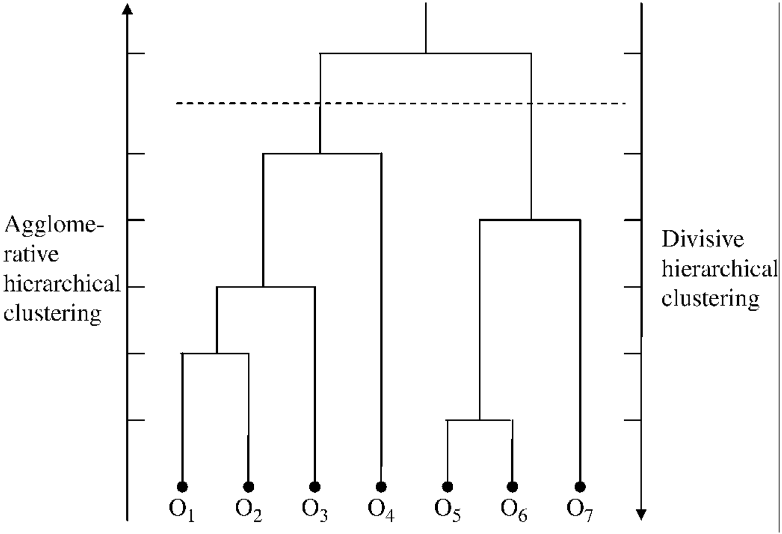
Clustering of Uncertain Objects

Major approaches:

- partitional clustering methods:
  - uncertain version of k-Means [Chau et Al., PAKDD'06] and its relative optimizations [Ngai et Al., ICDM'06, Lee et Al., ICDM Work.'07, Chui et Al., ICDM'08]
  - uncertain version of k-Medoids [Gullo et Al., SUM'08]
- density-based clustering methods:
  - uncertain version of DBSCAN [Kriegel and Pfeifle, KDD'05]
  - uncertain version of OPTICS [Kriegel and Pfeifle, ICDM'05]

Poor research on hierarchical approaches...

Hierarchical Clustering



Standard scheme:

- starting point: each object forms a cluster
- at each iteration, the pair of closest clusters are merged
- criteria to determine the closest clusters:
  - *Single-Linkage* (*SL*):

$$\langle \mathcal{C}_i, \mathcal{C}_j \rangle = \arg\min_{\langle \mathcal{C}', \mathcal{C}'' \rangle \in \mathbf{C} \times \mathbf{C}} \min_{\substack{o' \in \mathcal{C}', \\ o'' \in \mathcal{C}''}} d(o', o'')$$

  - *Complete-Linkage* (*CL*):

$$\langle \mathcal{C}_i, \mathcal{C}_j \rangle = \arg\min_{\langle \mathcal{C}', \mathcal{C}'' \rangle \in \mathbf{C} \times \mathbf{C}} \max_{\substack{o' \in \mathcal{C}', \\ o'' \in \mathcal{C}''}} d(o', o'')$$

  - *Average-Linkage* (*AL*):

$$\langle \mathcal{C}_i, \mathcal{C}_j \rangle = \arg\min_{\langle \mathcal{C}', \mathcal{C}'' \rangle \in \mathbf{C} \times \mathbf{C}} \frac{1}{|\mathcal{C}'||\mathcal{C}''|} \sum_{\substack{o' \in \mathcal{C}', \\ o'' \in \mathcal{C}''}} d(o', o'')$$

Agglomerative Hierarchical Clustering of Uncertain Data

Naïve approach:
Defining a proper distance measure d between uncertain object and exploit the standard agglomerative hierarchical clustering scheme

Major issue:
Defining d is particularly critical; traditional approaches:

- difference between expected values (*drawback: accuracy*)
- Expected Distance (ED) (*drawback: efficiency*—$\mathcal{O}(s^2)$)

Agglomerative Hierarchical Clustering of Uncertain Data

Idea:
resorting to a *centroid-linkage* cluster merging criterion:

$$\langle C_i, C_j \rangle = \arg\min_{\langle \mathcal{C}', \mathcal{C}'' \rangle \in \mathbf{C} \times \mathbf{C}} \Delta(\mathcal{P}_{\mathcal{C}'}, \mathcal{P}_{\mathcal{C}''})$$

Requirements:

1. notion of *centroid* (or *prototype*) of a cluster of uncertain objects defined by somehow exploiting the entire information coming from the pdfs of the objects to be summarized
2. effective and efficient criterion Δ for comparing prototypes

How Requirement 1 can be efficiently satisfied?

Multivariate Uncertain Prototype

$\mathcal{P}_{\mathcal{C}}$ is computed as a mixture model of the pdfs of the objects in $\mathcal{C}$

In particular, $\mathcal{P}_{\mathcal{C}}$ is defined as a "new" uncertain object:

region: the product of the "stretched" dimension intervals of definition

- for each of these intervals, the lower (upper) bound is the minimum lower (maximum upper) bound over the objects

pdf: the average over the pdfs of the objects

Multivariate Uncertain Prototype (2)

Definition (multivariate uncertain prototype)

Let $\mathcal{C} = \{o_1, \ldots, o_n\}$ be a set of multivariate uncertain objects, where $o_i = (R_i, f_i)$, $R_i = [l_i^{(1)}, u_i^{(1)}] \times \ldots \times [l_i^{(m)}, u_i^{(m)}]$, for each $i \in [1..n]$. The *multivariate uncertain prototype* of $\mathcal{C}$ is a pair $\mathcal{P}_\mathcal{C} = (R_\mathcal{C}, f_\mathcal{C})$, where

$$R_\mathcal{C} = \left[\min_{i \in [1..n]} l_i^{(1)}, \max_{i \in [1..n]} u_i^{(1)}\right] \times \cdots \times \left[\min_{i \in [1..n]} l_i^{(m)}, \max_{i \in [1..n]} u_i^{(m)}\right]$$

$$f_\mathcal{C}(\vec{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\vec{x})$$

How Requirement 2 can be satisfied?

Distance measures for pdfs

Distance measures for pdfs: *information-theoretic* (IT) measures, such as those falling into the *Ali-Silvey* class of distance functions.

As an example, the *Kullback-Leibler* (KL) divergence:

$$\mathrm{KL}(g_1, g_2) = \int_{\vec{x} \in \Re^m} g_1(\vec{x}) \, \log \frac{g_2(\vec{x})}{g_1(\vec{x})} \, \mathrm{d}\vec{x}$$

Further examples: *Chernoff* distance, *Bhattacharyya* distance, ...

IT-adequacy

IT measures are accurate and efficient ($\mathcal{O}(s)$), but they are defined for pdfs defined over a common event space; therefore, they work out for pdfs that share a reasonably large overlapping probability values area

Notion of IT-adequacy (Υ):

$$\Upsilon_{R_1,R_2}(g_1,g_2) = \frac{1}{2}\left(\int\limits_{\vec{x}\in R_1\cap R_2} g_1(\vec{x})\ \mathrm{d}\vec{x}\ + \int\limits_{\vec{x}\in R_1\cap R_2} g_2(\vec{x})\ \mathrm{d}\vec{x} \right)$$

where $\int\limits_{\vec{x}\in\Re^m\setminus R_i} g_i(\vec{x})\mathrm{d}\vec{x} = 0$ and $g_i(\vec{x}) > 0$, $\forall\vec{x}\in R_i,\ i\in\{1,2\}$

Compound distance for uncertain objects

$$\Delta(o_i, o_j) = \mathrm{f}(\Delta_{IT}(o_i, o_j), \Delta_{ED}(o_i, o_j))$$

- Δ_{IT} involves a comparison by means of a certain IT measure
- Δ_{ED} measures the distance proportionally to the difference of the expected values

Compound distance for uncertain objects (2)

Two critical choices for defining Δ:

1. the IT-measure used for computing Δ_{IT}
2. the way of combining Δ_{IT} and Δ_{ED}

Compound distance for uncertain objects (3)

Choice 1: $\Delta_{IT} = $ Bhattacharyya distance B:

Bhattacharyya coefficient:

$$\rho(g_1, g_2) = \int_{\vec{x} \in \Re^m} \sqrt{g_1(\vec{x})\ g_2(\vec{x})}\ \mathrm{d}\vec{x}$$

Bhattacharyya distance:

$$\mathrm{B}(g_1, g_2) = \sqrt{1 - \rho(g_1, g_2)}$$

Compound distance for uncertain objects (4)

Main motivations for choosing $\Delta_{IT} = \mathrm{B}$:

- $\mathrm{B} \in [0, 1]$, which makes B easily comparable and combinable with other measures
- theoretical result stated in the following Proposition

Proposition

Let g_1 and g_2 be two m-dimensional pdfs ($m \geq 1$), and $R_1 \subseteq \Re^m$, $R_2 \subseteq \Re^m$ be two m-dimensional regions such that (for $i \in \{1, 2\}$):

$$\int_{\vec{x} \in \Re^m \setminus R_i} g_i(\vec{x}) \mathrm{d}\vec{x} \approx 0, \quad \textit{and} \quad g_i(\vec{x}) > 0, \ \forall \vec{x} \in R_i$$

It holds that:

$$\rho(g_1, g_2) \leq \Upsilon_{R_1, R_2}(g_1, g_2)$$

Compound distance for uncertain objects (4)

Choice 2:

- making Δ_{ED} comparable with Δ_{IT} ($\in [0,1]$):

$$\Delta_{ED} = 1 - e^{-EDdist}$$

- controlling the combination of Δ_{IT} and Δ_{ED} by some proper factor α:

$$\alpha = \mathrm{f}(\Upsilon)$$

- α is reasonably based on the degree of overlap of the pdf areas
- α gives a more robust way to combine Δ_{IT} and Δ_{ED} than the width of the shared domain region

Compound distance for uncertain objects (5)

$$\Delta_{IT} = \mathrm{B} = \sqrt{1-\rho} \qquad \Delta_{ED} = 1 - e^{-distED}$$

$$\Delta = 1 - [(1 - \Delta_{IT}) + \alpha\,(1 - \Delta_{ED})]$$

According to the previous Proposition ($\rho \leq \Upsilon$) it holds that $1 - \Delta_{IT} \leq 1 - \sqrt{1-\Upsilon}$; then α can be defined as equal to $1 - (1 - \sqrt{1-\Upsilon})$

$$\Rightarrow \quad \Delta = 1 - [(1 - \Delta_{IT}) + (1 - (1 - \sqrt{1-\Upsilon}))(1 - \Delta_{ED})]$$

$$\boldsymbol{\Delta = \Delta_{IT} - \sqrt{1-\Upsilon}\,(1 - \Delta_{ED})}$$

Multivariate uncertain distance

Definition (multivariate uncertain distance)

The *multivariate uncertain distance* between two multivariate uncertain objects $o_i = (R_i, f_i)$ and $o_j = (R_j, f_j)$ is defined as

$$\Delta(o_i, o_j) = \mathrm{B}(f_i, f_j) - \sqrt{1 - \Upsilon(o_i, o_j)}\ e^{-dist(E[f_i], E[f_j])}$$

Why does using Δ as a criterion for comparing uncertain prototypes (defined as mixture densities) in an AHC algorithm make sense?

Impact of Δ on the behavior of U-AHC

- $\mathbf{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_{n-k+1}\}$ is a set of prototypes of the form $\mathcal{P}_q = (R_q, f_q)$, for $q \in [1..n-k+1]$, which correspond to the clustering solution $\mathbf{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{n-k+1}\}$ obtained by U-AHC (at the k-th iteration);
- $i, j \in [1..n-k+1]$ are two indices such that $\mathcal{C}_i, \mathcal{C}_j \in \mathbf{C}$ is the pair of clusters to be merged, and
- $\overline{\mathcal{C}} = \mathcal{C}_i \cup \mathcal{C}_j$ is the new cluster formed;
- $\overline{\mathcal{P}} = (\overline{R}, \overline{f})$ is the prototype of $\overline{\mathcal{C}}$.

Impact of Δ on the behavior of U-AHC (2)

Theorem

- $\alpha \in [0, 1]$ *is a constant*
- $\Psi_u(\mathbf{C}, \alpha) = \Upsilon(\mathcal{P}_u, \overline{\mathcal{P}}) - (\alpha\Upsilon(\mathcal{P}_u, \mathcal{P}_i) + (1 - \alpha)\Upsilon(\mathcal{P}_u, \mathcal{P}_j))$
- $\hat{\alpha} = |\mathcal{C}_i|/(|\mathcal{C}_i| + |\mathcal{C}_j|)$

For each $u \in [1..n - k + 1]$, $u \neq i$, $u \neq j$, *it holds that:*

$$\Psi_u(\mathbf{C}, \hat{\alpha}) = \frac{1}{2}\left((1 - \hat{\alpha}) \int_{R_i} f_u(\vec{x})\mathrm{d}\vec{x} + \hat{\alpha} \int_{R_j} f_u(\vec{x})\mathrm{d}\vec{x}\right)$$

Impact of Δ on the behavior of U-AHC (3)

Corollary

For each $u \in [1..n-k+1]$, $u \neq i$, $u \neq j$, *if* $\Upsilon(\mathcal{P}_u, \mathcal{P}_i) \neq 0$ *or* $\Upsilon(\mathcal{P}_u, \mathcal{P}_j) \neq 0$, *then* $\Psi(\mathbf{C}, \hat{\alpha}) > 0$; *otherwise* $\Psi(\mathbf{C}, \hat{\alpha}) = 0$.

Results

- the IT-adequacy of the prototype of the (new) merged cluster to any $\mathcal{P}_u$ is not lower than (the linear combination of) the individual IT-adequacy of $\mathcal{P}_i$ and $\mathcal{P}_j$ with respect to $\mathcal{P}_u$
- if there is an overlap between the region of $\mathcal{P}_i$ (or $\mathcal{P}_j$) and $\mathcal{P}_u$, the IT-adequacy resulting from the merging step will increase

Impact of Δ on the behavior of U-AHC (4)

Results stated in the previous Theorem and Corollary intuitively prove that a strong relationship holds among the main ingredients that entail our proposal (i.e., the centroid-linkage-based AHC scheme, the prototype definition as a mixture model, the proposed criterion for comparing prototypes)

In other words, the distance measure Δ, which is in principle not generally applicable for comparing uncertain objects, is well-founded if used as a criterion for comparing uncertain prototypes defined as mixture models in a centroid-linkage-based agglomerative hierarchical algorithm for clustering uncertain objects

Centroid-based Aggl. Hierarch. Clustering

The U-AHC Algorithm

Require: a set of uncertain objects $\mathcal{D} = \{o_1, \ldots, o_n\}$

Ensure: a set of partitions $\mathbf{D}$

1: $\mathbf{C} \leftarrow \{\{o_1\}, \ldots, \{o_n\}\}$
2: $\mathbf{D} \leftarrow \{\mathbf{C}\}$
3: **repeat**
4: let $\mathcal{C}_i, \mathcal{C}_j$ be the pair of clusters in $\mathbf{C}$ such that $\Delta(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j})$ is minimum
5: $\mathbf{C} \leftarrow \mathbf{C} \setminus \{\mathcal{C}_i, \mathcal{C}_j\} \cup \{\mathcal{C}_i \cup \mathcal{C}_j\}$
6: $\mathbf{D} \leftarrow \mathbf{D} \cup \{\mathbf{C}\}$
7: **until** $|\mathbf{C}| = 1$

Methodology

Goals

- Assessment of effectiveness of the U-AHC algorithm in clustering uncertain data
- Comparison of U-AHC with state-of-the-art algorithms
  - UK-means, CK-means, UK-medoids, $\mathcal{F}$DBSCAN, $\mathcal{F}$OPTICS

Datasets

Table: Benchmark datasets used in the experiments

dataset	*# of objects*	*# of attributes*	*# of classes*
Iris	150	4	3
Wine	178	13	3
Glass	214	10	6
Ecoli	327	7	5
Yeast	1,484	8	10
ImageSegmentation	2,310	19	7
Abalone	4,124	7	17
LetterRecognition	7,648	16	10

Table: Non-benchmark datasets used in the experiments

dataset	*# of objects (genes)*	*# of attributes*
Leukaemia	22,690	21
Neuroblastoma	22,282	14

Clustering validity criteria

- External criteria (benchmark datasets): *F-measure*, *Precision*, *Recall*
- Internal criteria (non-benchmark datasets): *intra-cluster distance*, *inter-cluster distance*

F-measure results (benchmark datasets, multivariate models)

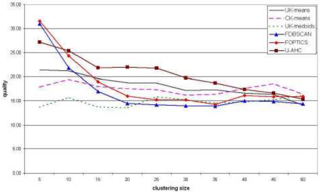
dataset	*pdf*	UK-means	CK-means	UK-medoids	$\mathcal{F}$DBSCAN	$\mathcal{F}$OPTICS	**U-AHC**
Iris	Uniform	0.948	<u>*0.962*</u>	0.907	0.929	0.907	**1**
	Normal	0.859	0.897	0.888	<u>*0.929*</u>	0.907	0.962
Wine	Uniform	0.735	0.747	0.761	<u>*0.767*</u>	0.713	**0.826**
	Normal	0.707	0.705	<u>*0.749*</u>	0.691	0.713	0.795
Glass	Uniform	0.677	<u>*0.703*</u>	0.653	0.575	0.636	0.779
	Normal	0.540	0.551	0.579	<u>*0.868*</u>	0.828	**0.891**
Ecoli	Uniform	0.787	<u>*0.790*</u>	0.728	0.443	0.477	0.743
	Normal	<u>*0.745*</u>	0.740	0.560	0.416	0.477	**0.795**
Yeast	Uniform	0.533	0.538	<u>*0.622*</u>	0.599	0.528	**0.684**
	Normal	0.455	<u>*0.457*</u>	0.318	0.374	0.420	0.486
ImageSegmentation	Uniform	0.780	<u>*0.801*</u>	0.765	0.482	0.419	**0.837**
	Normal	0.628	0.637	<u>*0.649*</u>	0.415	0.419	0.684
Abalone	Uniform	0.288	0.290	<u>*0.531*</u>	0.499	0.439	0.492
	Normal	0.215	0.217	0.288	0.497	<u>*0.558*</u>	**0.572**
LetterRecognition	Uniform	0.637	0.636	<u>*0.763*</u>	0.320	0.318	**0.798**
	Normal	0.442	0.435	<u>*0.595*</u>	0.353	0.318	0.613
	avg. score	0.624	0.632	0.647	0.571	0.567	0.747
	avg. gain	12.3%	11.5%	10.0%	17.6%	18.0%	—

F-measure results (benchmark datasets)

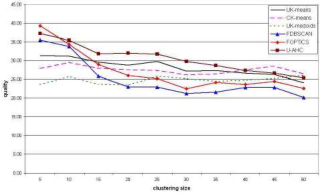
Remarks:

- U-AHC achieved the highest accuracy on all datasets
- average gains: from 10%(vs. UK-medoids) to 18%(vs $\mathcal{F}$OPTICS)

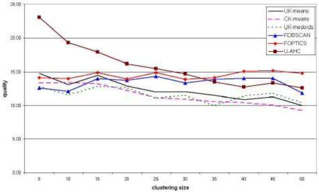
Quality results (microarray datasets)



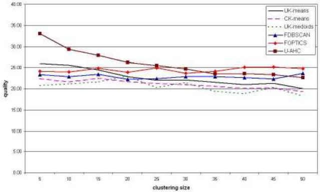
(a) Leukaemia — Normal pdf



(b) Leukaemia — Percentiles-based pdf



(c) Neuroblastoma — Normal pdf



(d) Neuroblastoma — Percentiles-based pdf

Quality results (microarray datasets) (2)

Remarks:

- U-AHC achieved the best results averaged over the cluster sizes
- highest quality on Leukaemia, whereas behaved on average better than the other methods on Neuroblastoma

Conclusions

U-AHC, the first (centroid-linkage-based) agglomerative hierarchical algorithm for uncertain data clustering

- Uncertain cluster prototype defined as a mixture model
- Information-theoretic-based compound distance for comparing uncertain prototypes

Experimental results:

- U-AHC outperforms existing methods in terms of accuracy

References

- F. Gullo, G. Ponti, A. Tagarelli, S. Greco, A Hierarchical Algorithm for Clustering Uncertain Data via an Information-Theoretic Approach. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM '08)*. Pisa, Italy, December 15-19, 2008
- F. Gullo, G. Ponti, A. Tagarelli, G. Tradigo, P. Veltri, Hierarchical Clustering of Microarray Data with Probe-level Uncertainty. In *Proceedings of the 22th IEEE International Symposium on Computer-Based Medical Systems (CBMS '09)*. Albuquerque, New Mexico (USA), August 3-4, 2009
- F. Gullo, G. Ponti, A. Tagarelli, S. Greco, Information-Theoretic Hierarchical Clustering of Uncertain Data. In *Proceedings of the 17th Italian Symposium on Advanced Database Systems (SEBD '09)*. Geneva, Italy, June 21-24, 2009

Thanks!