A Hierarchical Algorithm for Clustering Uncertain Data via an Information-Theoretic Approach

Francesco Gullo Giovanni Ponti
Andrea Tagarelli Sergio Greco

Dept. Electronics, Computer and Systems Science
University of Calabria, Italy

IEEE International Conference on Data Mining (ICDM) 2008

Outline

Motivations

Mining in Uncertain Data

Wide range of emerging database applications

- location-based services, sensor networks, RFID systems, biomedical and biometric data, etc.
- inherently associated with uncertainty: measurement imperfection, sampling error, network latency, etc.

Uncertain Data Objects

Traditional modelling by *probability density functions (pdfs)*

Uncertain Data Clustering

Relatively recent in data mining [ICDM, KDD, PAKDD, etc.]

Major approaches:

- partitional clustering methods: uncertain versions of k-Means
- density-based clustering methods: uncertain versions of DBSCAN, OPTICS

A major issue: Computing distance between uncertain objects

- *either* based on aggregated values (e.g., expected value) from the pdfs $\Longrightarrow$ accuracy issue
- *or* according to the whole pdfs $\Longrightarrow$ efficiency issue

Information-theoretic distances for pdfs

- use the original information of the pdfs (efficiently)
- but require common event space — usually unsatisfied for uncertain objects

Our approach in a nutshell

Agglomerative hierarchical clustering
with **centroid-based** linkage criterion and
information-theoretic cluster distance

- **Cluster prototypes**: mixture densities summarizing the pdfs of the objects within the cluster
- **Cluster merging**: based on an information-theoretic distance between cluster prototypes

Our approach in a nutshell (2)

Highlights

- no need for computing EDs between objects and (deterministic) cluster centroids — our cluster prototypes are still uncertain objects
- no need for a notion of distance between the objects being clustered
  - using a dense summary (mixture densities)
  - exploiting the larger overlaps between the cluster prototypes' domain regions

Definitions

1. Modelling uncertain data objects
   - multivariate / univariate definitions
2. Modelling uncertain cluster prototypes
   - multivariate / univariate definitions
3. Computing distance between uncertain cluster prototypes
   - multivariate / univariate definitions

Multivariate uncertainty model

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

Definition (multivariate uncertain object)

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

- $R = [l_1, u_1] \times \cdots \times [l_m, u_m]$ is the m-dimensional region in which o is defined
- $f : \Re^m \rightarrow \Re_0^+$ is the pdf of o at each point $\vec{x} \in R$, such that:

$$\int_{\vec{x} \in R} f(\vec{x}) \mathrm{d}\vec{x} = 1 \quad \text{and} \quad \int_{\vec{x} \in \Re^m \setminus R} f(\vec{x}) \mathrm{d}\vec{x} = 0$$

Univariate uncertainty model

- an m-dimensional tuple of pairs:
  - interval of definition
  - univariate pdf

Definition (univariate uncertain object)

A univariate uncertain object o is a tuple $(a^{(1)}, \ldots, a^{(m)})$.
Each attribute $a^{(h)}$ is a pair $(I^{(h)}, f^{(h)})$, for each $h \in [1..m]$, where:

- $I^{(h)} = [l^{(h)}, u^{(h)}]$ is the interval of definition of $a^{(h)}$
- $f^{(h)} : \Re \rightarrow \Re_0^+$ is the probability density function that assigns a probability value to each $x \in I^{(h)}$, such that:

$$\int\limits_{x \in I^{(h)}} f^{(h)}(x)\mathrm{d}x = 1 \qquad \text{and} \qquad \int\limits_{x \in \Re \setminus I^{(h)}} f^{(h)}(x)\mathrm{d}x = 0$$

Multivariate Uncertain Prototype

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

Definition (multivariate uncertain prototype)

Let $\mathcal{C} = \{o_1, ..., o_n\}$ 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 multivariate uncertain object $\mathcal{P}_\mathcal{C} = (R_\mathcal{C}, f_\mathcal{C})$, where:

- $R_\mathcal{C}=\left[\min_{i\in[1..n]} l_{i_1}, \max_{i\in[1..n]} u_{i_1}\right]\times\cdots\times\left[\min_{i\in[1..n]} l_{i_m}, \max_{i\in[1..n]} u_{i_m}\right]$
- $f_\mathcal{C}(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\vec{x})$

Univariate Uncertain Prototype

For each dimension:

interval the interval with lower (upper) bound set equal to the minimum lower (maximum upper) bound over the objects

pdf the average over the univariate dimension-pdfs of the objects

Definition (univariate uncertain prototype)

Let $\mathcal{C} = \{o_1, ..., o_n\}$ be a set of univariate uncertain objects, where $o_i = ((I_i^{(1)}, f_i^{(1)}), \ldots, (I_i^{(m)}, f_i^{(m)}))$, $I_i^{(h)} = [l_i^{(h)}, u_i^{(h)}]$, for each $h \in [1..m]$, $i \in [1..n]$.

The univariate uncertain prototype of $\mathcal{C}$ is a univariate uncertain object $\mathcal{P}_{\mathcal{C}} = ((I_{\mathcal{C}}^{(1)}, f_{\mathcal{C}}^{(1)}), \ldots, (I_{\mathcal{C}}^{(m)}, f_{\mathcal{C}}^{(m)}))$ such that, for each $h \in [1..m]$:

- $I_{\mathcal{C}}^{(h)} = \left[\min_{i\in[1..n]} l_i^{(h)}, \max_{i\in[1..n]} u_i^{(h)}\right]$
- $f_{\mathcal{C}}^{(h)}(x) = \frac{1}{n}\sum_{i=1}^{n} f_i^{(h)}(x)$

Information-Theoretic distance for pdfs

Goal: to define a distance measure between uncertain prototypes based on the whole information stored in the pdfs

Bhattacharyya distance:

$$\mathrm{B}(p(\vec{x}), q(\vec{x})) = \sqrt{1 - \rho(p(\vec{x}), q(\vec{x}))}$$

where $\rho(p(\vec{x}), q(\vec{x})) = \int\limits_{\vec{x} \in \Re^m} \sqrt{p(\vec{x})\ q(\vec{x})}\ \mathrm{d}\vec{x}$ is the *Bhattacharyya coefficient* defined over the pdfs p and q

(Some) advantages

- easier to compute than, e.g., Chernoff distance
- satisfies the triangle inequality, unlike, e.g., Chernoff and Kullback-Leibler
- ranges within [0, 1]
- satisfies the additive property even if the random variables are not identically distributed

Multivariate Uncertain Prototype distance

Definition (multivariate uncertain prototype distance)

Given:

- a set $\mathcal{D}$ of multivariate uncertain objects,
- any two sets $\mathcal{C}_i, \mathcal{C}_j \subseteq \mathcal{D}$, with prototypes $\mathcal{P}_{\mathcal{C}_i} = (R_{\mathcal{C}_i}, f_{\mathcal{C}_i})$ and $\mathcal{P}_{\mathcal{C}_j} = (R_{\mathcal{C}_j}, f_{\mathcal{C}_j})$.

The multivariate uncertain prototype distance between $\mathcal{P}_{\mathcal{C}_i}$ and $\mathcal{P}_{\mathcal{C}_j}$ is:

$$\Delta(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j}) = \gamma \, \Delta'(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j}) + (1-\gamma) \, \Delta''(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j})$$

where

$$\Delta'(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j}) = \mathrm{B}(f_{\mathcal{C}_i}, f_{\mathcal{C}_j}), \qquad \Delta''(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j}) = \frac{d(E[f_{\mathcal{C}_i}], E[f_{\mathcal{C}_j}])}{E_{max}(\mathcal{D})}$$

$$\gamma = \mathcal{V}(R_{\mathcal{C}_i} \cap R_{\mathcal{C}_j}) \;/\; \min\{\mathcal{V}(R_{\mathcal{C}_i}), \mathcal{V}(R_{\mathcal{C}_j})\}$$

- d is a distance over a m-dimensional real-valued space,
- $E[f]$ is the expected value of the pdf f,
- $\mathcal{V}(R)$ is the hyper-volume of the m-dimensional region R,
- $E_{max}(\mathcal{D}) = \max_{o_u, o_v \in \mathcal{D}} d(E[f_u], E[f_v])$

Multivariate Uncertain Prototype distance (2)

Remarks:

- the Bhattacharyya distance (Δ') of two pdfs considers their portions defined over the common domain region
  - $\Delta' = 1$, if there is no common event space
- Δ'' considers the distance between the expected values of the prototype pdfs
- γ is proportional to the width of the common region
- $\Delta', \Delta'', \gamma \in [0, 1] \quad \Rightarrow \quad \Delta \in [0, 1]$

Univariate Uncertain Prototype distance

Definition (univariate uncertain prototype distance)

Given:

- a set $\mathcal{D}$ of univariate uncertain objects,
- any two sets $\mathcal{C}_i, \mathcal{C}_j \subseteq \mathcal{D}$, with prototypes $\mathcal{P}_{\mathcal{C}_i} = ((I_{\mathcal{C}_i}^{(1)}, f_{\mathcal{C}_i}^{(1)}), \ldots, (I_{\mathcal{C}_i}^{(m)}, f_{\mathcal{C}_i}^{(m)}))$ and $\mathcal{P}_{\mathcal{C}_j} = ((I_{\mathcal{C}_j}^{(1)}, f_{\mathcal{C}_j}^{(1)}), \ldots, (I_{\mathcal{C}_j}^{(m)}, f_{\mathcal{C}_j}^{(m)}))$.

The univariate uncertain prototype distance between $\mathcal{P}_{\mathcal{C}_i}$ and $\mathcal{P}_{\mathcal{C}_j}$ is:

$$\Delta(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j}) = f_{dist}(\delta^{(1)}, \ldots, \delta^{(m)})$$

where $$\delta^{(h)} = \gamma^{(h)} \, \mathrm{B}(f_{\mathcal{C}_i}^{(h)}, f_{\mathcal{C}_j}^{(h)}) + (1 - \gamma^{(h)}) \left(\frac{1}{E_{max}^{(h)}(\mathcal{D})} \left| E\left[f_{\mathcal{C}_i}^{(h)}\right] - E\left[f_{\mathcal{C}_j}^{(h)}\right] \right| \right)$$

$$\gamma^{(h)} = \mathcal{V}(I_{\mathcal{C}_i}^{(h)} \cap I_{\mathcal{C}_j}^{(h)}) / \min\{\mathcal{V}(I_{\mathcal{C}_i}^{(h)}), \mathcal{V}(I_{\mathcal{C}_j}^{(h)})\}$$

- f_{dist} is a distance over a m-dimensional real-valued space, e.g., $\sqrt{(1/m)\sum_{h=1}^{m} (\delta^{(h)})^2}$
- $E_{max}^{(h)}(\mathcal{D}) = \max_{o_u, o_v \in \mathcal{D}} |E[f_u^{(h)}] - E[f_v^{(h)}]|$

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 $\frac{1}{2}(\Delta(\mathcal{P}_{\mathcal{C}_i \cup \mathcal{C}_j}, \mathcal{P}_{\mathcal{C}_i}) + \Delta(\mathcal{P}_{\mathcal{C}_i \cup \mathcal{C}_j}, \mathcal{P}_{\mathcal{C}_j}))$ is minimum

5: $\mathbf{C} \leftarrow \{\mathcal{C} \in \mathbf{C} \ : \ \mathcal{C} \neq \mathcal{C}_i, \mathcal{C} \neq \mathcal{C}_j\} \cup \{\mathcal{C}_i \cup \mathcal{C}_j\}$

6: $\mathbf{D} \leftarrow \mathbf{D} \cup \{\mathbf{C}\}$

7: **until** $|\mathbf{C}| = 1$

8: **return** $\mathbf{D}$

- U-AHC follows the classic AHC scheme
- The selection of clusters to be merged (Line 4) employs the multivariate/univariate notion of distance between uncertain prototypes

Methodology

Goals

- Assessment of effectiveness and efficiency of the U-AHC algorithm in clustering uncertain data
- Comparison of U-AHC with state-of-the-art algorithms
  - k-means based algorithms: UK-means [Chau et al., PAKDD'06] and CK-means [Lee et al., ICDM'07 Workshops]
  - density-based algorithms: $\mathcal{F}$DBSCAN [Kriegel & Pfeifle, KDD'05] and $\mathcal{F}$OPTICS [Kriegel & Pfeifle, ICDM'05]

Data: benchmark datasets from the UCI Machine Learning Repository

Assessment criteria: (external) F-measure, precision, recall

Method setups: param. tuning, integral estimation (Monte Carlo)

dataset	*objects*	*attributes*	*classes*
Iris	150	4	3
Wine	178	13	3
Glass	214	10	6
Ecoli	327	7	5

Generating uncertainty

Univariate objects

For each attribute $a^{(h)}$ of object o

- $I^{(h)}$: subinterval within $[min_{o_h}, max_{o_h}]$, where min_{o_h} (resp. max_{o_h}) is the minimum (resp. maximum) deterministic value of the h-th attribute, over all the objects belonging to the same ideal class of o
- $f^{(h)}$: *Uniform*, *Normal* and *Gamma* pdfs.
  The parameters of Normal and Gamma pdfs in such a way that their mode corresponded to the deterministic value of the h-th attribute of o

Multivariate objects

- R: the product of the intervals randomly generated for each attribute of o
- f: *Uniform* and *Normal* pdfs.
- setting the pdf parameters and the region intervals: analogous to the univariate case

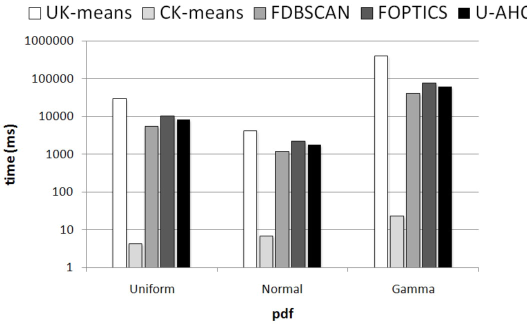
F-measure results (univariate models)

dataset	*pdf*	UK-means	CK-means	$\mathcal{F}$DBSCAN	$\mathcal{F}$OPTICS	**U-AHC**
Iris	Uniform	*0.93*	0.92	0.92	0.92	**0.93**
	Normal	0.84	0.85	*0.90*	*0.90*	**0.92**
	Gamma	0.60	0.50	*0.79*	0.77	**0.87**
Wine	Uniform	0.75	*0.76*	0.65	0.68	**1**
	Normal	0.70	0.71	*0.77*	0.76	**0.89**
	Gamma	*0.67*	0.58	0.64	0.64	**0.73**
Glass	Uniform	0.55	*0.69*	0.43	0.47	**0.81**
	Normal	0.58	0.55	0.60	*0.61*	**0.83**
	Gamma	0.46	0.51	0.62	*0.64*	**0.92**
Ecoli	Uniform	0.39	0.40	0.48	*0.51*	**0.79**
	Normal	0.73	*0.74*	0.68	0.68	**0.83**
	Gamma	*0.48*	0.41	0.47	0.47	**0.83**
	avg. score	0.64	0.635	0.663	0.67	0.863
	avg. gain	22.25%	22.75%	20%	19.17%	–

Remarks:

- U-AHC outperforms the other methods with average improvements from 19% ($\mathcal{F}$OPTICS) to about 23% (CK-means)
- Density-based algorithms perform better than k-means-based algorithms (around 3%)

Time performances



Remarks:

- U-AHC is one order of magnitude faster than UK-means on average
- CK-means outperforms the other methods
  - due to the optimization of the EDs calculation
- U-AHC, $\mathcal{F}$OPTICS and $\mathcal{F}$DBSCAN performances are comparable each other
  - but U-AHC is more accurate

Conclusion and further work

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

- Uncertain cluster prototype for univariate and multivariate uncertainty models
- Information-theoretic distance between uncertain prototypes

Experimental results:

accuracy U-AHC outperforms existing methods

efficiency U-AHC performs comparably to density-based clustering algorithms

Further work:

- improve the notions of prototype and prototype distance
- do a lot of new experiments with more real data, non-convex data, different distribution functions, etc.