Uncertain Centroid based Partitional Clustering of Uncertain Data

Francesco Gullo * Andrea Tagarelli †

* Yahoo! Research
Barcelona, Spain

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

38th International Conference on Very Large Data Bases (VLDB)
August 27-31, 2012
Istanbul, Turkey

Uncertainty

Uncertainty inherently affects data from 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

Generally due to noisy factors, such as signal noise, instrumental errors, wireless transmission

Uncertain Objects (UO) (1)

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

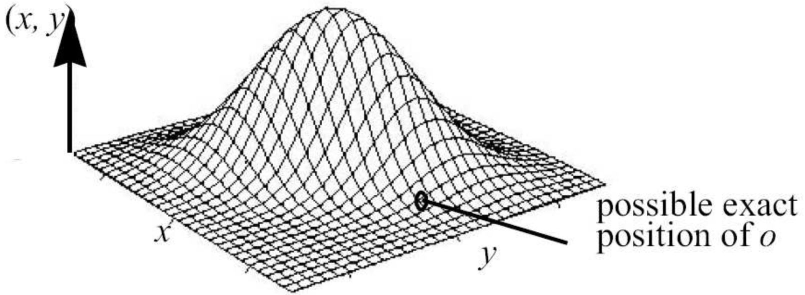


Figure borrowed from [Kriegel and Pfeifle, ICDM 2005]

Uncertain Objects (UO) (2)

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

Definition (uncertain object)

An uncertain object o is a pair $(\mathcal{R}, f)$:

- $\mathcal{R} \subseteq \mathbb{R}^m$ is the m-dimensional domain region in which o is defined
- $f : \mathbb{R}^m \to \mathbb{R}_0^+$ is the probability density function of o at each point $\vec{x} \in \mathbb{R}^m$ such that:

$$f(\vec{x}) > 0, \ \forall \vec{x} \in R \qquad \text{and} \qquad f(\vec{x}) = 0, \ \forall \vec{x} \in \mathbb{R}^m \setminus \mathcal{R}$$

Clustering Uncertain Objects

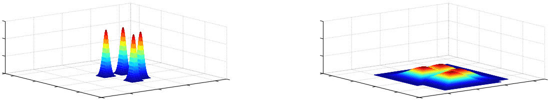
Major approaches:

- partitional approaches:
  - uncertain version of k-Means [Chau et al., PAKDD 2006] and its relative optimizations [Lee et al., ICDM Work. 2007, Kao et al., TKDE 2010, Ngai et al., Information Systems 2011]
  - uncertain version of k-Medoids [Gullo et al., SUM 2008]
- density-based approaches:
  - uncertain version of DBSCAN [Kriegel and Pfeifle, KDD 2005]
  - uncertain version of OPTICS [Kriegel and Pfeifle, ICDM 2005]
- hierarchical approaches [Gullo et al., ICDM 2008]

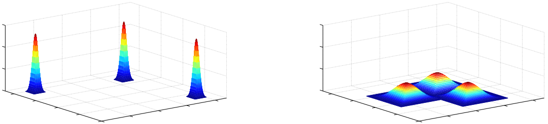
Partitional approaches include the fastest methods so far defined

Intuition

Approaches to partitional clustering of uncertain objects should take into account both central tendency and variance of the input uncertain objects



Uncertain objects with the same central tendency: lower-variance, more-compact cluster (left) and higher-variance, less-compact cluster (right)



Uncertain objects with different central tendency: lower-variance, less-compact cluster (left) and higher-variance, more-compact cluster (right)

Contributions

- We formally show that existing formulations of partitional clustering of uncertain objects do not comply with the intuition about central tendency and variance
- We propose a novel formulation to the problem of clustering uncertain objects based on the notion of *U-centroid*
- Given that the expression of the U-centroid is not analytically computable, we derive some theoretical properties to be efficiently exploited as *closed-form update rules* for the proposed objective function
- We define an efficient *local-search procedure* based on these rules

UK-means and MMVar

Partitional clustering of uncertain objects relies on two main notions: *cluster centroid* ($\overline{C}$), and *cluster compactness* (J)

Most prominent existing formulations:

- *UK-means* [Chau et al., PAKDD'06] $\longrightarrow$ cluster centroid is a *deterministic* object

$$\overline{C}_{UK} = \frac{1}{|C|} \sum_{o \in C} \vec{\mu}(o)$$

$$J_{UK}(C) = \sum_{o \in C} ED(o, \overline{C}_{UK}), \text{ where } ED(o, \overline{C}_{UK}) = \int_{\vec{x} \in \mathcal{R}} \|\vec{x} - \overline{C}_{UK}\|^2 \, f(\vec{x}) \, \mathrm{d}\vec{x}$$

- *MMvar* [Gullo et al., ICDM'10] $\longrightarrow$ cluster centroid is an *uncertain* object

$$\overline{C}_{MM} = (\overline{\mathcal{R}}_{MM}, \overline{f}_{MM}), \text{ where } \overline{\mathcal{R}}_{MM} = \bigcup_{o \in C} \mathcal{R} \text{ and } \overline{f}_{MM}(\vec{x}) = \frac{1}{|C|} \sum_{o \in C} f(\vec{x})$$

$$J_{MM}(C) = \sigma^2(\overline{C}_{MM})$$

Issues of UK-means and MMVar formulations

- The deterministic centroid representation in UK-means is not able to discriminate among different variances
- The MMvar formulation does not overcome this issue, although its centroid representation involves uncertainty

Proposition

Given a cluster C of m-dimensional uncertain objects, where $o = (\mathcal{R}, f)$, $\forall o \in C$, it holds that $J_{MM}(C) = |C|^{-1} J_{UK}(C)$.

A straightforward (inappropriate) solution

- Idea: combine the notions of MMVar centroid with the UK-means cluster compactness criterion
$$\widehat{J}(C) = \sum_{o \in C} \widehat{ED}(o, \overline{C}_{MM}),$$
where $\widehat{ED}(o, \overline{C}_{MM}) = \int_{\vec{x} \in \mathcal{R}} \int_{\vec{y} \in \overline{\mathcal{R}}_{MM}} \|\vec{x} - \vec{y}\|^2 f(\vec{x}) \overline{f}_{MM}(\vec{y}) \, \mathrm{d}\vec{x} \, \mathrm{d}\vec{y}$
- Unfortunately, such an objective function $\widehat{J}$ is not appropriate as it is equivalent to functions J_{UK} and J_{MM}

Proposition

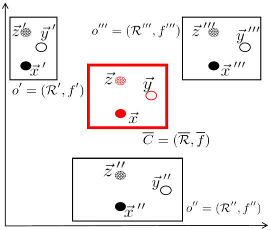
Given a cluster C of m-dimensional uncertain objects, where $o = (\mathcal{R}, f)$, $\forall o \in C$, it holds that $\widehat{J}(C) = 2\,|C|\, J_{MM}(C) = 2\, J_{UK}(C)$.

Our proposal

1. Introducing a novel notion of cluster centroid
2. Defining a cluster compactness criterion based on this novel cluster centroid definition which meets the requirements about central tendency and variance

U-centroid

Cluster centroid as *random variable* summarizing all possible deterministic representations of the objects in the cluster



Two key advantages:

- Shortcomings of a deterministic centroid notion are addressed
- Clear stochastic meaning

U-centroid: analytical expression (1)

Theorem

Given a cluster $C = \{o_1, \ldots, o_{|C|}\}$ of m-dimensional uncertain objects, where $o_i = (\mathcal{R}_i, f_i)$ and $\mathcal{R}_i = \left[\ell_i^{(1)}, u_i^{(1)}\right] \times \cdots \times \left[\ell_i^{(m)}, u_i^{(m)}\right]$, $\forall i \in [1..|C|]$, let $\overline{C} = (\overline{\mathcal{R}}, \overline{f})$ be the U-centroid of C defined by employing the squared Euclidean norm as distance to be minimized. It holds that:

$$\overline{f}(\vec{x}) = \int\limits_{\vec{x}_1 \in \mathcal{R}_1} \cdots \int\limits_{\vec{x}_{|C|} \in \mathcal{R}_{|C|}} \mathrm{I}\left[\vec{x} = \frac{1}{|C|} \sum_{i=1}^{|C|} \vec{x}_i\right] \prod_{i=1}^{|C|} f_i(\vec{x}_i) \mathrm{d}\vec{x}_1 \cdots \mathrm{d}\vec{x}_{|C|}$$

$$\overline{\mathcal{R}} = \left[\frac{1}{|C|}\sum_{i=1}^{|C|}\ell_i^{(1)}, \frac{1}{|C|}\sum_{i=1}^{|C|}u_i^{(1)}\right] \times \cdots \times \left[\frac{1}{|C|}\sum_{i=1}^{|C|}\ell_i^{(m)}, \frac{1}{|C|}\sum_{i=1}^{|C|}u_i^{(m)}\right]$$

where $\mathrm{I}[A]$ is the indicator function, which is 1 when the event A occurs, 0 otherwise.

U-centroid based cluster compactness criterion

Two main requirements for the proposed cluster compactness criterion J:

- It should rely on the U-centroid notion so to meet the requirements about central tendency and variance
- The expression of the pdf $\overline{f}$ in the proposed U-centroid is not analytically computable $\Rightarrow$ J should be such that it can be optimized without requiring to explicitly compute $\overline{f}$

A first solution: minimizing the U-centroid variance

Minimizing the variance of the U-centroid (similarly to MMVar) does not work, as it is equivalent to minimizing the average variance of the individual uncertain objects in the cluster:

Theorem

Given a cluster $C = \{o_1, \ldots, o_{|C|}\}$ of m-dimensional uncertain objects, where $o_i = (\mathcal{R}_i, f_i)$, $\forall i \in [1..|C|]$, let $\overline{C} = (\overline{\mathcal{R}}, \overline{f})$ be the U-centroid of C. It holds that $\sigma^2(\overline{C}) = |C|^{-2} \sum_{i=1}^{|C|} \sigma^2(o_i)$.

Minimizing the expected distance between uncertain objects and U-centroid (1)

$$J(C) = \sum_{o \in C} \widehat{ED}(o, \overline{C})$$

Observation 1: J takes into account both central tendency and variance

Theorem

Let $C = \{o_1, \ldots, o_{|C|}\}$ be a cluster of uncertain objects, where $o_i = (\mathcal{R}_i, f_i)$, and $\overline{C} = (\overline{\mathcal{R}}, \overline{f})$ be the U-centroid of C. It holds that:

$$J(C) = \sum_{j=1}^{m} \left(\frac{\Psi_C^{(j)}}{|C|} + \Phi_C^{(j)} - \frac{\Upsilon_C^{(j)}}{|C|} \right) = \frac{1}{|C|} \sum_{i=1}^{|C|} \sigma^2(o_i) + \sum_{o \in C} ED\left(o, \frac{1}{|C|} \sum_{o \in C} \vec{\mu}(o)\right)$$

where

$$\Psi_C^{(j)} = \sum_{i=1}^{|C|} (\sigma^2)_j(o_i) \qquad \Phi_C^{(j)} = \sum_{i=1}^{|C|} (\mu_2)_j(o_i) \qquad \Upsilon_C^{(j)} = \left(\sum_{i=1}^{|C|} \mu_j(o_i) \right)^2$$

Minimizing the expected distance between uncertain objects and U-centroid (2)

Observation 2: Given a cluster C, the value of J of any other cluster resulting from adding/removing an object to/from C can be computed according to an efficient closed-form expression

Corollary

Let C be a cluster of uncertain objects, and $C^+ = C \cup \{o^+\}$, $C^- = C \setminus \{o^-\}$ be two clusters defined by adding an object $o^+ \notin C$ to C and removing an object $o^- \in C$ from C, respectively. It holds that:

$$J(C^+)=\sum_{j=1}^{m}\left(\frac{\Psi_{C^+}^{(j)}}{|C|+1}+\Phi_{C^+}^{(j)}-\frac{\Upsilon_{C^+}^{(j)}}{|C|+1}\right) \quad J(C^-)=\sum_{j=1}^{m}\left(\frac{\Psi_{C^-}^{(j)}}{|C|-1}+\Phi_{C^-}^{(j)}-\frac{\Upsilon_{C^-}^{(j)}}{|C|-1}\right)$$

The UCPC local-search algorithm

Input: A set $\mathcal{D}$ of UO; the number k of output clusters
Output: A partition $\mathcal{C}$ of $\mathcal{D}$, where $|\mathcal{C}| = k$

compute $\vec{\mu}(o)$, $\vec{\mu}_2(o)$, $\vec{\sigma}^2(o)$, $\forall o \in \mathcal{D}$
$\mathcal{C} \leftarrow initialPartition(\mathcal{D}, k)$, compute $\Psi_C^{(j)}$, $\Phi_C^{(j)}$, $\Upsilon_C^{(j)}$, $J(C)$
repeat
    $V \leftarrow \sum_{C \in \mathcal{C}} J(\mathcal{C})$
    for all $o \in \mathcal{D}$ **do**
        $C^* \leftarrow \mathrm{argmin}_{C \in \mathcal{C}} V - [J(C^o) + J(C)] + [J(C^o \setminus \{o\}) + J(C \cup \{o\})]$
        if $C^* \neq C^o$ **then**
            $\mathcal{C} \leftarrow \mathcal{C} \setminus \{C^*, C^o\} \cup \{C^+, C^-\}$
            replace $\Psi_{C^*}^{(j)}$, $\Phi_{C^*}^{(j)}$, $\Upsilon_{C^*}^{(j)}$, $J(C^*)$ with $\Psi_{C^+}^{(j)}$, $\Phi_{C^+}^{(j)}$, $\Upsilon_{C^+}^{(j)}$, $J(C^+)$, $\forall j \in [1..m]$
            replace $\Psi_{C^o}^{(j)}$, $\Phi_{C^o}^{(j)}$, $\Upsilon_{C^o}^{(j)}$, $J(C^o)$ with $\Psi_{C^-}^{(j)}$, $\Phi_{C^-}^{(j)}$, $\Upsilon_{C^-}^{(j)}$, $J(C^-)$, $\forall j \in [1..m]$
until no object in $\mathcal{D}$ is relocated

- UCPC converges to a local optimum of function J in a finite number I of iterations
- UCPC works in $\mathcal{O}(I\ k\ |\mathcal{D}|\ m)$

Evaluation methodology (1)

- Benchmark datasets from UCI (Iris, Wine, Glass, Ecoli, Yeast, Image, Abalone, Letter) where uncertainty is generated **synthetically** and modeled according to *Uniform* (U), *Normal* (N), and *Exponential* (E) pdfs
- Real (gene expression) datasets where uncertainty is inherently present

(a) Benchmark datasets

dataset	*obj.*	*attr.*	*classes*
Iris	150	4	3
Wine	178	13	3
Glass	214	10	6
Ecoli	327	7	5
Yeast	1,484	8	10
Image	2,310	19	7
Abalone	4,124	7	17
Letter	7,648	16	10

(b) Real datasets

dataset	*obj.*	*attr.*
Neuroblastoma	22,282	14
Leukaemia	22,690	21

Evaluation methodology (2)

- Evaluation in terms of:
  - **accuracy** (external and internal clustering evaluation)
  - **efficiency**
- Competitors: MMVar (MMV), UK-means (UKM), UK-medoids (UKmed), UAHC, $\mathcal{F}$DBSCAN ($\mathcal{F}$DB), $\mathcal{F}$OPTICS ($\mathcal{F}$OPT)

Accuracy results: benchmark datasets

		F-measure ($\Theta \in [-1,1]$)						
	pdf	$\mathcal{F}$DB	$\mathcal{F}$OPT	UAHC	UKmed	UKM	MMV	**UCPC**
avg score	U	-.189	.055	.089	.210	.081	.193	.429
	N	-.081	-.046	.149	-.028	.019	.199	.287
	E	-.317	-.088	-.008	-.011	-.137	.200	223
overall avg. score		-.196	-.026	.077	.057	-.012	.198	.313
overall avg. gain		+.509	+.339	+.236	+.256	+.324	+.115	—

		Quality ($Q \in [-1,1]$)						
	pdf	$\mathcal{F}$DB	$\mathcal{F}$OPT	UAHC	UKmed	UKM	MMV	**UCPC**
avg score	U	.021	.089	.027	.084	.042	.345	.375
	N	.061	.115	.091	.089	.127	.139	.189
	E	-.001	.025	0	.011	.015	.199	.200
overall avg. score		.027	.076	.039	.061	.061	.228	.255
overall avg. gain		+.228	+.179	+.216	+.194	+.194	+.027	—

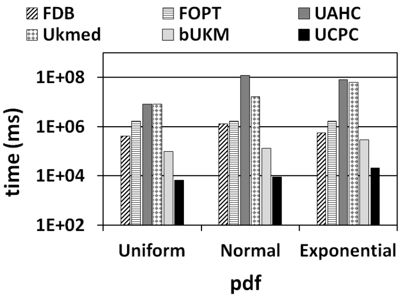
Accuracy results: real datasets

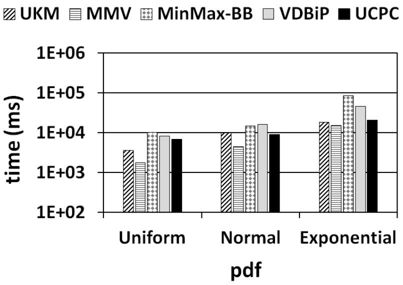
		Quality ($Q \in [-1, 1]$)						
data	*#clust.*	$\mathcal{F}$DB	$\mathcal{F}$OPT	UAHC	UKmed	UKM	MMV	**UCPC**
Neuro.	*avg score*	-.004	.010	.630	.045	.060	.544	.576
Leuk.	*avg score*	-.018	.190	.192	.231	.430	.433	.471
over.	*avg score*	-.011	.100	.411	.138	.245	.489	.523
over.	*avg gain*	+.534	+.423	+.112	+.385	+.278	+.034	—

Efficiency results: benchmark datasets

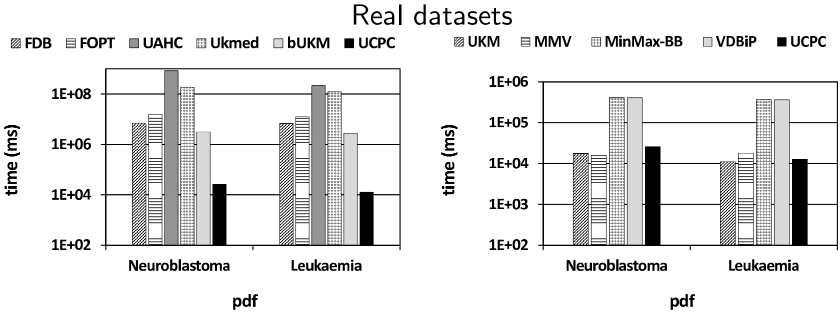
- Efficiency evaluation also involves optimized versions of UK-means, i.e., MinMax-BB and VDBiP

Letter





Efficiency results: real datasets



Conclusions

- Existing formulations of partitional clustering of uncertain objects miss some crucial requirements about central tendency and variance of the objects to be clustered
- Novel notion of cluster centroid, called U-centroid
- Effective and efficient U-centroid based cluster compactness criterion
- Efficient local-search heuristic to optimize the proposed objective function

Thanks!