

Handling Uncertainty in Clustering Art-exhibition Visiting Styles

Giovanni Ponti¹

joint work with

Francesco Gullo² and Andrea Tagarelli³

Salvatore Cuomo⁴, Pasquale De Michele⁴, Francesco Piccialli⁴

¹DTE-ICT-HPC ENEA Research Center, Portici (NA)

²DIMES, University of Calabria, Rende (CS)

³UniCredit, R&D Department, Rome

⁴University of Naples, Federico II

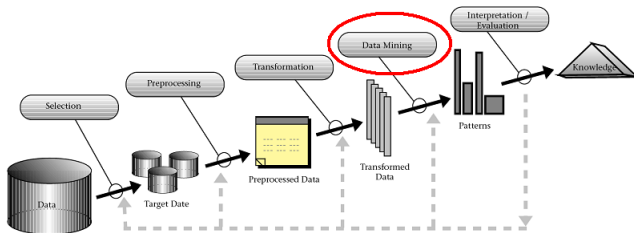
7th EAI International Conference on Big Data
Technologies and Applications

November 17th, 2016 - Seoul, South Korea

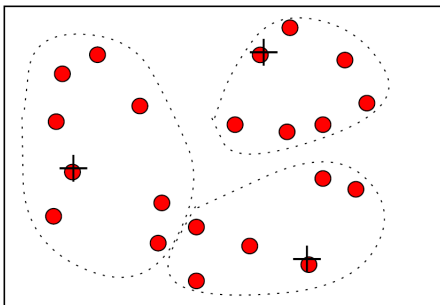
- 1 Data Mining & Clustering
- 2 Uncertain Data
 - Applications
 - Issues in Data Analysis
 - Uncertainty modeling
- 3 Proposal – UK-medoids algorithm
- 4 Real Case application – Cultural Heritage
 - Naples Maschio Angioino art exhibition
 - Visitor behaviours & animal styles
- 5 Experiments
 - Uncertainty generation
 - Accuracy results
 - Efficiency results

Knowledge Discovery in Databases

“Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying novel, valid, potentially useful, and ultimately understandable patterns in data” [Fayyad et Al., '96]



KDD $\Rightarrow$ Data Mining $\Rightarrow$ **Clustering**



- Low intra-cluster distance
- High inter-cluster distance

Data uncertainty naturally arises from:

- implicit randomness in a process of data generation/acquisition
- imprecision in physical measurements
- application of approximation methods
- ...

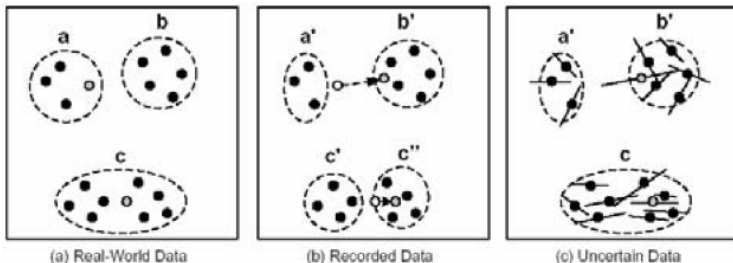
Application domains

sensor data, moving objects databases, biomedical data (e.g., probe-level uncertainty in microarray data), forecasting, RFID applications, ...

Uncertainty in Data

Issues in Clustering

Clustering of uncertain data may lead to wrong results if uncertainty is not taken into account

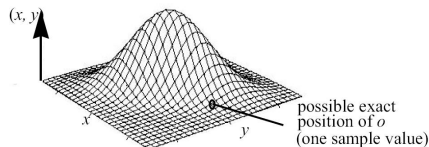


Uncertainty modeling

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)

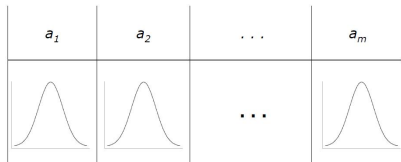
A **multivariate uncertain object** o is a pair (R, f) , where:

- $R = [l_1, u_1] \times \cdots \times [l_m, u_m]$ 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:

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

Univariate Uncertain Objects

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



Definition (univariate uncertain object)

A **univariate uncertain object** o is a tuple $(a^{(1)}, \dots, a^{(m)})$.

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 $a^{(h)}$
- $f^{(h)} : \mathbb{R} \rightarrow \mathbb{R}_0^+$ is the probability density function that assigns a probability value to each $x \in I^{(h)}$, such that:

$$\int_{x \in I^{(h)}} f^{(h)}(x) dx = 1 \quad \text{and} \quad \int_{x \in \mathbb{R} \setminus I^{(h)}} f^{(h)}(x) dx = 0$$

A major issue: Computing distance between uncertain objects

Traditional approaches:

- aggregated values (e.g., expected value) from the pdfs $\Rightarrow$ **accuracy issue**
- Expected Distance (ED) $\Rightarrow$ **efficiency issue**

Traditional approaches do not compare the whole information stored in the pdfs

Uncertain Data Clustering

Approaches

Major approaches:

- **partitional** clustering methods:
 - **UK-Means** and its optimizations
- **density-based** clustering methods:
 - **FDBSCAN**
 - **FOPTICS**

Idea

Using **medoids** instead of centroids

Advantages:

- **More versatile.** It can be applied both to multivariate and univariate uncertainty modeling
- **More accurate**
 - cluster representatives are not computed as a trivial mean of expected values
 - we defined a proper distance function for uncertain objects (while UK-Means uses the Expected Distance between a deterministic object, i.e., cluster centroid, and an uncertain object)
- **More efficient.** The bottleneck of computing EDs at each iteration can be reduced by computing offline the pair-wise distances for each pair of objects

The UK-medoids Algorithm

Input: a set of uncertain objects $D = \{o_1, \dots, o_n\}$; the number of output clusters k

Output: a set of clusters $\mathcal{C}$

```

1: compute distances  $\delta(o_i, o_j), \forall o_i, o_j \in D$ 
2: compute the set  $S = \{m_1, \dots, m_k\}$  of initial medoids
3: repeat
4:    $S' \leftarrow S$ 
5:    $S \leftarrow \emptyset$ 
6:    $\mathcal{C} = \{C_1, \dots, C_k\} \leftarrow \{\emptyset, \dots, \emptyset\}$ 
7:   for all  $o \in D$  do
8:     {assign each object to the closest cluster, based on its
       uncertain distance to cluster medoids}
9:      $m_j \leftarrow \arg \min_{o' \in S'} \delta(o, o')$ 
10:     $C_j \leftarrow C_j \cup \{o\}$ 
11:  end for
12:  for all  $C \in \mathcal{C}$  do
13:    {recompute the medoid of each cluster}
14:     $m \leftarrow \arg \min_{o \in C} \sum_{o' \in C} \delta(o, o')$ 
15:     $S \leftarrow S \cup \{m\}$ 
16:  end for
17: until  $S = S'$ 
    
```

- UK-medoids follows the classic partitional-relocation scheme
- The pairwise uncertain distances are computed as

$$\delta(o_i, o_j) = \int_{\vec{x} \in R_i} \int_{\vec{y} \in R_j} \text{dist}(\vec{x}, \vec{y}) f_i(\vec{x}) f_j(\vec{y}) d\vec{x} d\vec{y}$$

The UK-medoids Algorithm

Input: a set of uncertain objects $D = \{o_1, \dots, o_n\}$; the number of output clusters k

Output: a set of clusters $\mathcal{C}$

```

1: compute distances  $\delta(o_i, o_j), \forall o_i, o_j \in D$ 
2: compute the set  $S = \{m_1, \dots, m_k\}$  of initial medoids
3: repeat
4:    $S' \leftarrow S$ 
5:    $S \leftarrow \emptyset$ 
6:    $\mathcal{C} = \{C_1, \dots, C_k\} \leftarrow \{\emptyset, \dots, \emptyset\}$ 
7:   for all  $o \in D$  do
8:     {assign each object to the closest cluster, based on its
       uncertain distance to cluster medoids}
9:      $m_j \leftarrow \arg \min_{o' \in S'} \delta(o, o')$ 
10:     $C_j \leftarrow C_j \cup \{o\}$ 
11:  end for
12:  for all  $C \in \mathcal{C}$  do
13:    {recompute the medoid of each cluster}
14:     $m \leftarrow \arg \min_{o \in C} \sum_{o' \in C} \delta(o, o')$ 
15:     $S \leftarrow S \cup \{m\}$ 
16:  end for
17: until  $S = S'$ 
    
```

- UK-medoids follows the classic partitional-relocation scheme
- The pairwise uncertain distances are computed as

$$\delta(o_i, o_j) = \int_{\vec{x} \in R_i} \int_{\vec{y} \in R_j} \text{dist}(\vec{x}, \vec{y}) f_i(\vec{x}) f_j(\vec{y}) d\vec{x} d\vec{y}$$

A real case application — visiting styles in an art exhibition

Case Study

Real Data from Art exhibition within Maschio Angioino Castle, in Naples (Italy), of sculptures by Francesco Jerace promoted by DATABENC

To classify the **visitor behaviours** w.r.t. the **usage of mobile devices** in an exhibit. In Zancaro et al. (2007), visitor behaviours are compared to that of four typical animals:

- ANT (A);
- FISH (F);
- BUTTERFLY (B);
- GRASSHOPPER (G).



We adapt this classification to find **how visitors interact with the ICT** technology and **how a lot they are interested in the exhibition**.

ANT (A)

tends to follow a specific path and spends a lot of time using the furnished technology

FISH (F)

moves around in the centre of the rooms and usually avoids looking at media content details

BUTTERFLY (B)

does not follow a specific path and stops frequently to look for more media content

GRASSHOPPER (G)

seems to have a specific preference for some preselected artworks and spends a lot of time observing the related media contents

The interpretation of data

We have tracked the visitor behaviours by using a suitable Extrapolation Algorithm (EA) that has the JSON file as input data.

```
IDUser : e7a5774700c1e88e1417618582735
# of artworks: 271
# of viewed artworks: 44
% of viewed artworks : 17.5%

...

-----
i-th viewed artwork : 2
ID artwork : 128
Available audio (sec.) : 32.922
Listen audio (sec.) : 32.922
Available images : 3
Viewed images : 0
Available text : True
Viewed text : False
Interaction time (sec.) : 58.259
Path is followed : True
-----

...

-----
i-th viewed artwork : 6
ID artwork : 17
Available audio (sec.) : 85.141
Listen audio (sec.) : 85.141
Available images : 4
Viewed images : 2
Available text : True
Viewed text : True
Interaction time (sec.) : 103.141
Path is followed : False
-----
```

- For each visitor, collected data are organized as **tuples** representing **fruition** percentages of the media contents of any artwork:
 - Audios
 - Images
 - Texts

- We devised an experimental evaluation aimed to assess the ability in clustering uncertain objects of the UK-medoids algorithm
- We consider the dataset derived from the analysis of how visitors of a museum interact with an IoT framework, according to animal behaviour model
- We model this data as a set of uncertain objects, and apply the UK-medoids algorithm to obtain clusters of similar visiting styles

Goal

To compare such a visiting-style grouping with a ground truth obtained by a well-established classification methodology and show that our method outperforms a baseline clustering method that does not take uncertainty into consideration.

Dataset and uncertainty generation

```
@RELATION ARTWORKS
@ATTRIBUTE a NUMERIC [0..1]
@ATTRIBUTE  $\tau$  NUMERIC [0..1]
@ATTRIBUTE v NUMERIC [0..1]
@ATTRIBUTE class {A,B,F,G}
@DATA
...
0.0836653386454,0.846588116217,?,G
0.0478087649402,0.317966675258,?,F
0.119521912351,?,0.714285714286,A
0.175298804781,?,0.470303571429,B
...
```

In case of univariate uncertain objects, we generated the uncertain interval $I^{(h)}$ and the pdf $f^{(h)}$ defined over $I^{(h)}$, for each attribute $a^{(h)}$, with $h \in [1..m]$ of the object o

Uncertainty generation

- The interval $I^{(h)}$ was randomly chosen as a 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 attribute h , over all the objects belonging to the same ideal class of o .
- As concerns $f^{(h)}$, we considered a continuous density function, namely *Uniform*, and a discrete mass function, namely *Binomial*. We set the parameters of Binomial pdf in such a way that their mode corresponded to the deterministic value of the h -th attribute of the object o .

To evaluate the impact of dealing with uncertainty in a clustering-based analysis

We are interesting in **comparing clustering results** achieved by our UK-medoids algorithm on the dataset with uncertainty w.r.t. the ones achieved by K-medoids algorithm on the dataset with deterministic values.

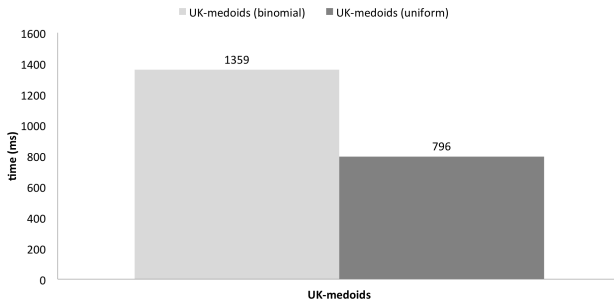
<i>pdf</i>	<i>UK-medoids gain</i>
Binomial	0.04298805%
Uniform	6.22712192%

Observations

- UK-medoids achieves higher accuracy results w.r.t. K-medoids
 - slight for binomial distribution (0.043%)
 - relevant for uniform distribution (6.227%)
- **introducing uncertainty in the dataset and handling it in the clustering task with our proper UK-medoids algorithm leads to improve the effectiveness of the results**

To evaluate time performances in clustering uncertain objects

We calculated the sum of the times obtained for the pre-computing phase (i.e., uncertain distances computation), together with the algorithm runtimes



Observations

- using a uniform pdf we obtain execution times about 2 times faster than those achieved with a binomial pdf
 - This is due to the fact that a binomial pdf requires to process a higher number of samples w.r.t. a uniform pdf

Thanks!