Minimizing the Variance of Cluster Mixture Models for Clustering Uncertain Objects

F. Gullo G. Ponti A. Tagarelli

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

10th IEEE International Conference on Data Mining (ICDM)
December 14-17, 2010
Sydney, Australia

Uncertainty

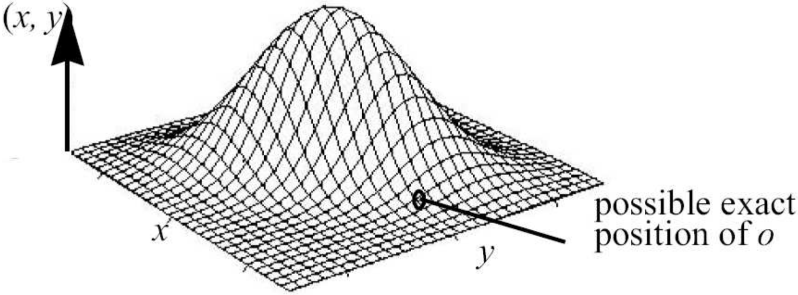
Uncertainty inherently affects data from a wide range of emerging application domains:

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

It is 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)*



Uncertain Objects (UO) (2)

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

Definition (uncertain object)

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

- $\mathcal{R} \subseteq \Re^m$ is the m-dimensional in which o is defined
- $f : \Re^m \longrightarrow \Re_0^+$ is the probability density function of o at each point $\vec{x} \in \Re^m$ such that:

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

Clustering Uncertain Objects (1)

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]
- hierarchical clustering methods [Gullo et Al., ICDM'08]

Clustering Uncertain Objects (2)

Issues of existing algorithms:

1. they require some notion of distance between uncertain objects
   - hard task as existing notions are either inaccurate or inefficient
2. they generally suffer from efficiency issues
   - intrinsically due to the adopted formulations, which require to continuously execute critical operations

Minimizing Mixture Model Variances for Clustering Uncertain Objects

Goal: to solve both the issues arising from existing algorithms for clustering uncertain objects

Proposal

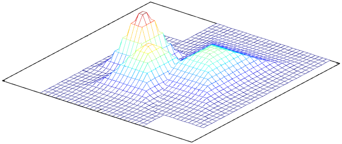
Novel formulation to the problem of clustering uncertain objects whose major features are:

1. exploiting *mixture models* for representing the clusters to be identified
2. employing the minimization of the *variance* of the mixture models as optimization criterion

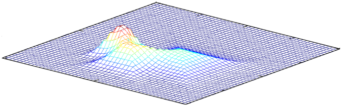
Cluster Mixture Models (Uncertain Prototypes)

Mixture model (uncertain prototype) of cluster C: $\mathcal{P}_C = (\mathcal{R}_C, f_C)$

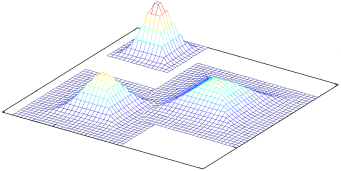
$$\mathcal{R}_C = \bigcup_{o=(\mathcal{R},f)\in C} \mathcal{R} \qquad f_C(\vec{x}) = (|C|)^{-1} \sum_{o=(\mathcal{R},f)\in C} f(\vec{x})$$



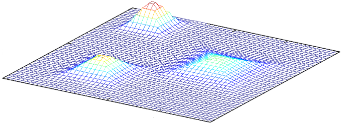
(a)



(b)



(c)



(d)

(a)–(c): *Sets of uncertain objects*

(b)–(d): *The corresponding mixture models*

Proposed Formulation

Idea: minimizing variance of cluster mixture models

$$J(\mathcal{C}) = \sum_{C \in \mathcal{C}} \sigma^2(\mathcal{P}_C)$$

- **accuracy**: the lower the variance, the higher the cluster compactness
- **efficiency**: capability of exploiting interesting analytical properties

Computing objective function J

- Moving object o from $C \in \mathcal{C}$ to $\widehat{C} \in \mathcal{C}$ leads to a new $\mathcal{C}' = \mathcal{C} \setminus (C \cup \widehat{C}) \cup (C' \cup \widehat{C}')$, where $C' = C \setminus \{o\}$, $\widehat{C}' = \widehat{C} \cup \{o\}$
- $J(\mathcal{C}')$ can be efficiently computed in $\mathcal{O}(m)$ as:

$$J(\mathcal{C}') = J(\mathcal{C}) - (\sigma^2(\mathcal{P}_C) + \sigma^2(\mathcal{P}_{\widehat{C}})) + (\sigma^2(\mathcal{P}_{C'}) + \sigma^2(\mathcal{P}_{\widehat{C}'}))$$

MMVar algorithm

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

```
compute μ(o), μ2(o), ∀o ∈ D
C ← randomPartition(D, k)
compute μ(P_C), μ2(P_C), ∀C ∈ C
v ← J(C)
repeat
    for all o ∈ D do
        let C ∈ C be the cluster s.t. o ∈ C
        C* ← arg min_Ĉ J_C(C, o, Ĉ)
        if C* ≠ C then
            v = J_C(C, o, Ĉ)
            recompute C by moving o from C to C*
            recompute μ(P_C), μ2(P_C), μ(P_C*), μ2(P_C*)
        end if
    end for
until no object in D is relocated
```

$$\begin{array}{l}
\text{compute } \vec{\mu}(o),\ \vec{\mu}_2(o),\ \forall o \in \mathcal{D} \\
\mathcal{C} \leftarrow \mathit{randomPartition}(\mathcal{D}, k) \\
\text{compute } \vec{\mu}(\mathcal{P}_C),\ \vec{\mu}_2(\mathcal{P}_C),\ \forall C \in \mathcal{C} \\
v \leftarrow J(\mathcal{C}) \\
\textbf{repeat} \\
\quad \textbf{for all } o \in \mathcal{D} \textbf{ do} \\
\quad\quad \text{let } C \in \mathcal{C} \text{ be the cluster s.t. } o \in C \\
\quad\quad C^* \leftarrow \arg\min_{\widehat{C}} J_{\mathcal{C}}(C, o, \widehat{C}) \\
\quad\quad \textbf{if } C^* \neq C \textbf{ then} \\
\quad\quad\quad v = J_{\mathcal{C}}(C, o, \widehat{C}) \\
\quad\quad\quad \text{recompute } \mathcal{C} \text{ by moving } o \text{ from } C \text{ to } C^* \\
\quad\quad\quad \text{recompute } \vec{\mu}(\mathcal{P}_C), \vec{\mu}_2(\mathcal{P}_C), \vec{\mu}(\mathcal{P}_{C^*}), \vec{\mu}_2(\mathcal{P}_{C^*}) \\
\quad\quad \textbf{end if} \\
\quad \textbf{end for} \\
\textbf{until} \text{ no object in } \mathcal{D} \text{ is relocated}
\end{array}$$

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

Evaluation Methodology

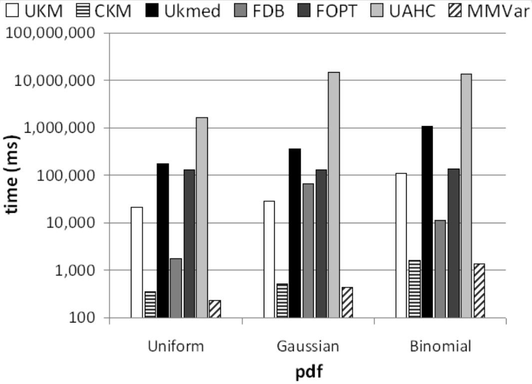
- Benchmark datasets from UCI (Iris, Wine, Glass, Ecoli, Yeast, Image, Abalone, Letter)
- Uncertainty generated **synthetically** and modeled according to *Uniform* (U), *Normal* (N), and *Binomial* (B) pdfs
- Evaluation in terms of:
  - **accuracy** (w.r.t. reference classifications according to *F-Measure*)
  - **efficiency**
- Competitors: UK-means (UKM), CK-means (CKM), UK-medoids (UKmed), $\mathcal{F}$DBSCAN ($\mathcal{F}$DB), $\mathcal{F}$OPTICS ($\mathcal{F}$OPT), U-AHC

Accuracy Results

		F-measure ($F \in [0,1]$)						
data	*pdf*	UKM	CKM	UKmed	$\mathcal{F}$DB	$\mathcal{F}$OPT	UAHC	**MMVar**
avg score	U	0.601	0.675	0.729	0.331	0.575	0.626	**0.731**
	N	0.54	0.582	0.493	0.441	0.475	0.606	**0.657**
	B	0.476	0.363	0.602	0.295	0.525	0.508	**0.716**
overall avg. score		0.539	0.54	0.608	0.356	0.525	0.58	**0.701**
overall avg. gain		0.162	0.161	0.093	0.345	0.176	0.121	—

- MMVar achieved the best overall scores, from +0.093 (w.r.t. UKmed) to +0.345 (w.r.t. $\mathcal{F}$DB)
- MMVar achieved the best avg scores on all the pdfs
  - maximum avg gain of 0.254 (Binomial)
  - minimum avg gain of 0.134 (Normal)

Efficiency Results



- MMVar performed faster than CKM
- MMVar drastically outperformed all other competitors but CKM (at least 1 order of magnitude, up to 5 orders)
- Slowest methods: UAHC and UKmed; fastest methods: CKM and $\mathcal{F}$DB

Conclusion

- Novel formulation to the problem of clustering uncertain objects
  - Cluster mixture models
  - Minimization of the variance of mixture models
- MMVar heuristic algorithm
- Significant advantages achieved by MMVar in terms of efficiency and accuracy w.r.t. existing algorithms

Thanks!

Datasets

dataset	*# objects*	*# attributes*	*# 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