Be *certain* of how-to before mining *uncertain* data

F. Gullo * G. Ponti † A. Tagarelli ‡

* Yahoo Labs
Barcelona, Spain

† ENEA Research Center
Portici (NA), Italy

‡ University of Calabria
Cosenza, Italy

7th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases
(ECML PKDD 2014)
September 15-19, 2014, Nancy (France)

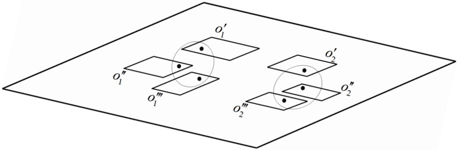
Uncertainty

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

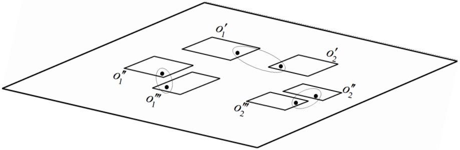
- sensor data
- location-based services (e.g., moving objects data)
- biomedical and biometric data (e.g., gene expression data)
- distributed applications
- RFID data

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

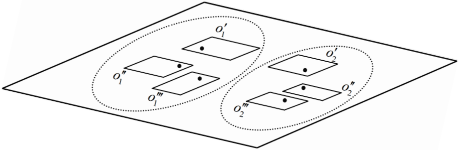
Uncertainty



(a)



(b)



(c)

Uncertainty representation

- Different granularities:
  - table
  - tuple
  - attribute
- Different models:
  - fuzzy
  - evidence-oriented
  - probabilistic

Attribute-level uncertainty modeled according to a probabilistic model (i.e., a probability distribution) $\Rightarrow$ **uncertain object**

Uncertain object

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

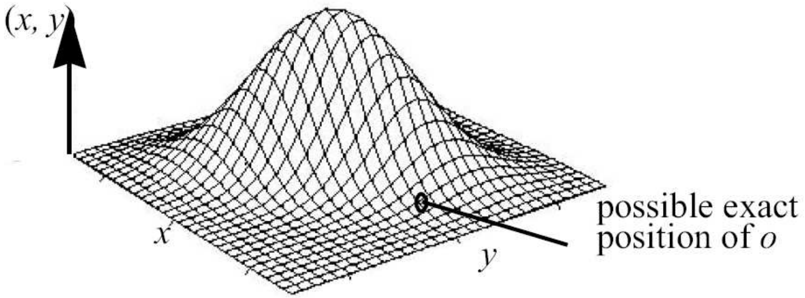


Figure borrowed from [Kriegel and Pfeifle, ICDM 2005]

Uncertain object

- 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
- $f : \mathbb{R}^m \rightarrow \mathbb{R}_0^+$ is the probability density function of o at each point $\mathbf{x} \in \mathbb{R}^m$ such that:

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

Dealing with uncertainty

Two main general tasks:

1. Defining a proximity measure between uncertain objects
   - needed in almost all major data-management and data-mining tasks (e.g., visualization, classification, clustering)
2. Defining a model to summarize a set of uncertain objects
   - required for tasks like data compression or clustering, and to speed-up complex data-analysis/management tasks

Similarity detection in uncertain data

Distance between uncertain objects

Traditional approaches:

1. Difference between expected values
2. Expected Distance (ED)

$$ED(o_1, o_2) = \int_{\mathbf{x} \in R_1} \int_{\mathbf{y} \in R_2} \|\mathbf{x} - \mathbf{y}\|_2^2 \; f_1(\mathbf{x}) \; f_2(\mathbf{y}) \; \mathrm{d}\mathbf{x} \; \mathrm{d}\mathbf{y}$$

Main drawbacks:

1. Difference between expected values is inaccurate: it considers only very little information stored in the pdfs:
2. Expected distance is slow: it has quadratic complexity in the number of statistical samples used to represent/approximate pdfs

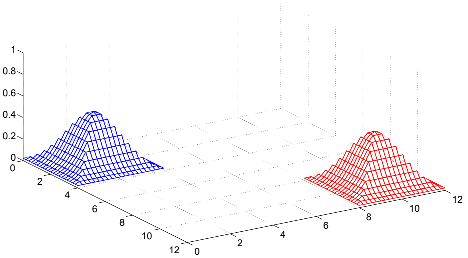
Distance between uncertain objects

- Need for a novel distance measure that trades off between accuracy and efficiency
- Idea: resort to *Information Theory*
- Information Theory alone is not enough

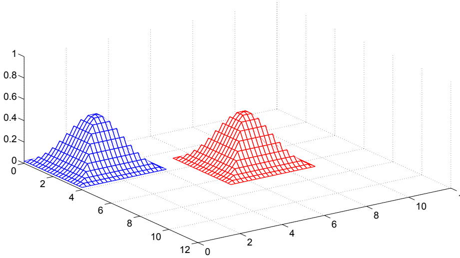
Distance measures for pdfs

Distance measures for pdfs: *information-theoretic* (IT) measures: *Kullback-Leibler* (KL), *Chernoff*, *Hellinger*, ...

IT measures are accurate, but they work out for pdfs that share a reasonably large overlapping probability values area



(a)



(b)

Compound distance for uncertain objects

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

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

Two critical choices for defining Δ:

1. IT-measure used for Δ_{IT} $\Rightarrow$ *Hellinger distance* ($\mathcal{H}$)

$$\rho(f, f') = \int_{\mathbf{x} \in \Re^m} \sqrt{f(\mathbf{x})\ f'(\mathbf{x})}\ \mathrm{d}\mathbf{x} \qquad \mathcal{H}(f, f') = \sqrt{1 - \rho(f, f')}$$

2. way of combining Δ_{IT} and Δ_{EV} $\Rightarrow$ Δ_{IT} should prevail on Δ_{EV} as long as discriminating among different cases by means of IT-measures is possible

Compound distance for uncertain objects

Definition (uncertain distance)

The *uncertain distance* between two uncertain objects $o = (\mathcal{R}, f)$ and $o' = (\mathcal{R}', f')$ is defined as

$$\Delta(o, o') = \underbrace{\mathcal{H}(f, f')}_{\Delta_{IT} \text{ term}} - \underbrace{\left(1 - \sqrt{\rho(f, f')}\right)}_{\substack{\text{combination} \\ \text{between } \Delta_{IT} \text{ and } \Delta_{EV}}} \times \underbrace{e^{-ED_2(\tilde{f}, \tilde{f}')}}_{\Delta_{EV} \text{ term}}$$

- $ED_2(\tilde{f}, \tilde{f}')$ is the expected distance between the uniform-approximation of f and f'

Centroid-based agglomerative hierarchical clustering

F. Gullo, G. Ponti, A. Tagarelli, S. Greco [ICDM'08]

- **Application**: hierarchical clustering of uncertain objects

The U-AHC Algorithm

Input: a set of uncertain objects $\mathcal{D} = \{o_1, \ldots, o_n\}$
Output: 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** $|\mathbf{C}| = 1$

Motivations:

- Hierarchical clustering is computationally expensive: need for a fast (yet accurate) proximity measure
- The way of combining Δ_{IT} and Δ_{EV} theoretically guarantees high accuracy in an agglomerative hierarchical clustering scheme

Uncertain data summarization

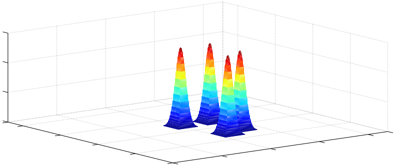
Summarization of a set of uncertain objects

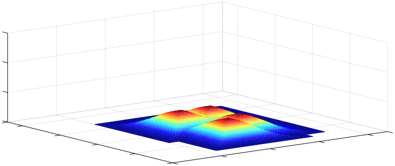
- Traditional approaches (e.g., Chau et al., UK-means, PAKDD'06) $\Rightarrow$ uncertain prototype defined as the average of the expected values of the objects to be summarized

Main drawbacks:

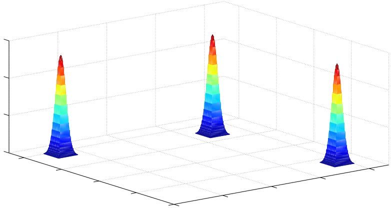
- Deterministic representation $\Rightarrow$ a lot of information is discarded
- Only central tendency is expressed $\Rightarrow$ *variance* is completely ignored

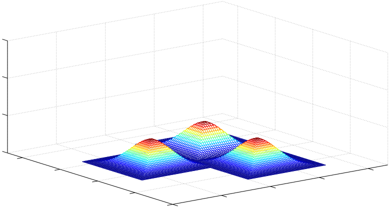
Summarization of a set of uncertain objects





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)

Summarization of a set of uncertain objects

Solutions:

1. Mixture-model-based uncertain data summarization
2. Random-variable-based uncertain data summarization

Mixture-model-based uncertain data summarization

Idea

Compute a prototype of a set of uncertain objects as *mixture model* :

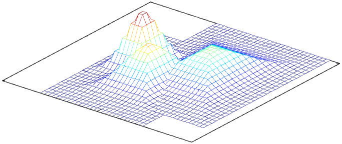
- set of uncertain objects $S = \{o_i\}_{i=1}^{k}$
- uncertain prototype $\mathcal{P}_S = (\mathcal{R}_S, f_S)$, where

$$\mathcal{R}_S = \bigcup_{o=(\mathcal{R},f)\in S} \mathcal{R},$$
$$f_S(\mathbf{x}) = (|S|)^{-1} \sum_{o=(\mathcal{R},f)\in S} f(\mathbf{x})$$

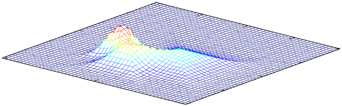
Mixture-model-based uncertain data summarization

Despite its simplicity, the mixture-model-based prototype plays a key role in a task of clustering uncertain objects: capability of employing a novel clustering criterion that *does not require any distance measure between uncertain objects*

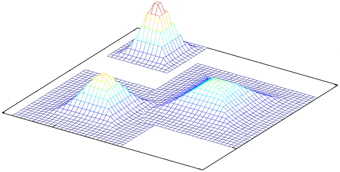
⇒ **minimizing the variance of cluster prototypes**



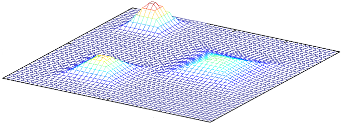
(a)



(b)



(c)



(d)

(a)–(c): *Sets of uncertain objects*

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

Minimizing the variance of cluster mixture models for clustering uncertain objects

F. Gullo, G. Ponti, A. Tagarelli [ICDM'10, SAM'13]

A novel criterion for clustering uncertain objects: 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:
$$J(\mathcal{C}') = J(\mathcal{C}) - (\sigma^2(\mathcal{P}_C) + \sigma^2(\mathcal{P}_{\widehat{C}})) + (\sigma^2(\mathcal{P}_{C'}) + \sigma^2(\mathcal{P}_{\widehat{C}'}))$$

The MMVar algorithm

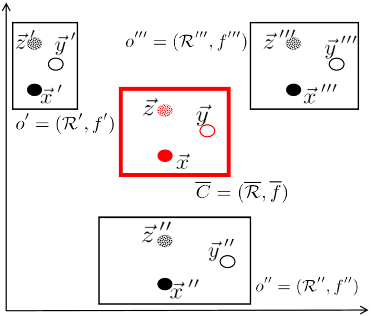
Input: A set $\mathcal{D}$ of UO; the number k of output clusters
Output: A partition $\mathcal{C}$ of $\mathcal{D}$

compute $\mu(o)$, $\mu_2(o)$, $\forall o \in \mathcal{D}$
$\mathcal{C} \leftarrow \mathit{randomPartition}(\mathcal{D}, k)$
compute $\mu(\mathcal{P}_C)$, $\mu_2(\mathcal{P}_C)$, $\forall C \in \mathcal{C}$
$v \leftarrow J(\mathcal{C})$
repeat
    for all $o \in \mathcal{D}$ **do**
        let $C \in \mathcal{C}$ be the cluster s.t. $o \in C$
        $C^* \leftarrow \arg\min_{\widehat{C}} J_{\mathcal{C}}(C, o, \widehat{C})$
        if $C^* \neq C$ **then**
            $v = J_{\mathcal{C}}(C, o, \widehat{C})$
            recompute $\mathcal{C}$ by moving o from C to C^*
            recompute $\mu(\mathcal{P}_C)$, $\mu_2(\mathcal{P}_C)$, $\mu(\mathcal{P}_{C^*})$, $\mu_2(\mathcal{P}_{C^*})$
until no object in $\mathcal{D}$ is relocated

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

One step further from mixture model: *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 still addressed
- Clear stochastic meaning (unlike mixture-model-based prototypes)

U-centroid: main advantages

F. Gullo, A. Tagarelli [VLDB'12]

The notion of U-centroid can be coupled with a cluster criterion that aims at minimizing the expected distance between uncertain objects and U-centroid

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

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

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

An efficient local-search method can be employed to optimize J:

1. Start with a random partition
2. At each step, perform the object move that leads to the best increment of J (if any)
3. Stop when J cannot be improved anymore (warranty to end up with a local optimum of J)

Conclusions

- Similarity detection and summarization are critical tasks that are commonly encountered when dealing with uncertain data
- We show how traditional measures for similarity detection in uncertain data can be empowered by combining notions from Information Theory and central-tendency-based comparison methods
- We discuss how to improve existing uncertain data summarization techniques by incorporating the variance of the uncertain objects to be summarized
- We provide evidence on how the tasks of similarity detection and summarization in uncertain data find natural application in data mining/machine learning

Thanks!

Backup: experiments about U-AHC

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
  - 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, univariate models)

dataset	*pdf*	UK-means	CK-means	UK-medoids	$\mathcal{F}$DBSCAN	$\mathcal{F}$OPTICS	**U-AHC**
Iris	Uniform	0.841	*0.963*	0.886	0.919	0.886	**0.993**
	Normal	0.849	0.849	0.855	0.871	*0.907*	0.905
	Gamma	0.622	0.501	0.848	0.893	*0.905*	0.628
Wine	Uniform	0.500	0.724	*0.810*	0.664	0.695	**0.984**
	Normal	0.500	0.704	0.578	0.653	*0.713*	0.954
	Gamma	0.500	0.581	0.581	0.692	*0.713*	0.595
Glass	Uniform	0.639	0.670	0.697	*0.768*	0.718	**0.828**
	Normal	*0.577*	0.552	0.513	0.514	0.438	0.822
	Gamma	0.379	0.314	*0.644*	0.468	0.438	0.550
Ecoli	Uniform	0.653	*0.795*	0.696	0.436	0.477	**0.915**
	Normal	0.609	*0.741*	0.528	0.544	0.477	0.726
	Gamma	0.533	0.412	*0.693*	0.401	0.477	0.450
Yeast	Uniform	0.497	0.562	*0.618*	0.515	0.543	**0.719**
	Normal	*0.471*	0.458	0.288	0.291	0.316	0.577
	Gamma	0.403	0.306	*0.469*	0.331	0.316	0.406
ImageSegmentation	Uniform	*0.810*	0.798	0.769	0.426	0.419	0.552
	Normal	0.623	*0.655*	0.451	0.416	0.419	**0.836**
	Gamma	0.545	0.353	*0.656*	0.339	0.419	0.503
Abalone	Uniform	0.331	0.294	*0.590*	0.447	0.439	**0.719**
	Normal	*0.288*	0.217	0.265	0.136	0.209	0.577
	Gamma	0.360	0.200	0.313	0.565	*0.607*	0.406
LetterRecognition	Uniform	0.529	0.629	*0.776*	0.344	0.318	**0.792**
	Normal	0.449	0.451	*0.490*	0.247	0.318	0.531
	Gamma	0.432	0.215	*0.584*	0.265	0.318	0.603
	avg. score	0.539	0.539	0.608	0.506	0.521	0.690
	avg. gain	15.1%	15.1%	8.2%	18.4%	16.9%	—

F-measure results (benchmark datasets, multivariate models)

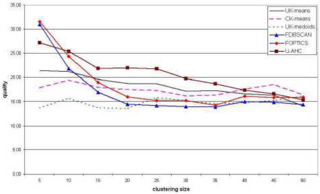
dataset	*pdf*	UK-means	CK-means	UK-medoids	$\mathcal{F}$DBSCAN	$\mathcal{F}$OPTICS	**U-AHC**
Iris	Uniform	0.948	*0.962*	0.907	0.929	0.907	**1**
	Normal	0.859	0.897	0.888	*0.929*	0.907	0.962
Wine	Uniform	0.735	0.747	0.761	*0.767*	0.713	**0.826**
	Normal	0.707	0.705	*0.749*	0.691	0.713	0.795
Glass	Uniform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

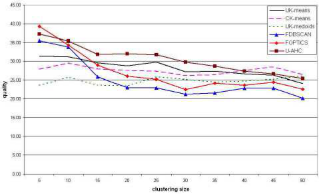
Remarks:

- U-AHC achieved the highest accuracy on all datasets
- average gains (univariate): from 8.2%(vs. UK-medoids) to 18.4%(vs $\mathcal{F}$DBSCAN)
- average gains (multivariate): from 10%(vs. UK-medoids) to 18%(vs $\mathcal{F}$OPTICS)
- results on univariate and multivariate cases were quite similar each other

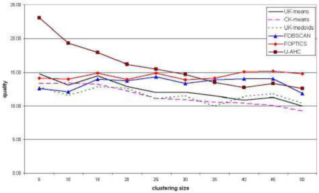
Quality results (microarray datasets)



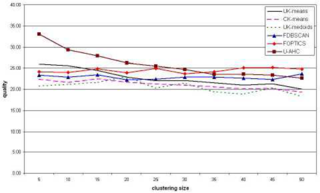
(a) Leukaemia — Normal pdf



(b) Leukaemia — Percentiles-based pdf



(c) Neuroblastoma — Normal pdf



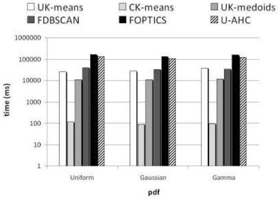
(d) Neuroblastoma — Percentiles-based pdf

Quality results (microarray datasets) (2)

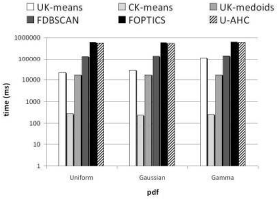
Remarks:

- 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

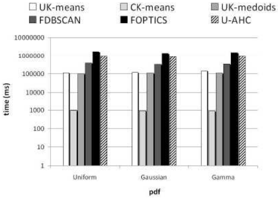
Efficiency results



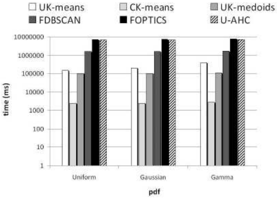
(a) Yeast



(b) ImageSegmentation



(c) Abalone



(d) LetterRecognition

Efficiency results (2)

Remarks:

- performances followed the (on-line) computational complexities of the corresponding algorithms:
  - $\mathcal{O}(t\ n)$, for CK-means
  - $\mathcal{O}(t\ n^2)$, for UK-medoids
  - $\mathcal{O}(t\ s\ n)$, for UK-means
  - $\mathcal{O}(n^2)$, for $\mathcal{F}$DBSCAN
  - $\mathcal{O}(s\ n^2)$, for U-AHC and $\mathcal{F}$OPTICS
- U-AHC performed closely to the density-based algorithms $\mathcal{F}$DBSCAN and $\mathcal{F}$OPTICS

Summary

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

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

Experimental results:

accuracy U-AHC outperforms existing methods

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

Backup: experiments about MMVar

Evaluation Methodology

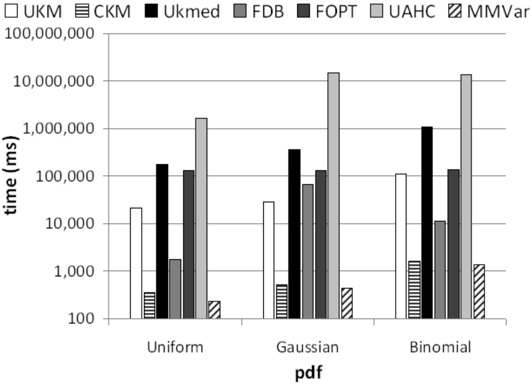
- Benchmark datasets from UCI (Iris, Wine, Glass, Ecoli, Yeast, Image, Abalone, Letter)
- Uncertainty generated **synthetically** and modeled according to *Uniform* (U), *Normal* (N), and *Binomial* (B) pdfs
- Evaluation in terms of:
  - **accuracy** (w.r.t. reference classifications according to *F-Measure*)
  - **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

data	pdf	F-measure ($F \in [0,1]$)						
		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

Backup: experiments about UCPC

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	—

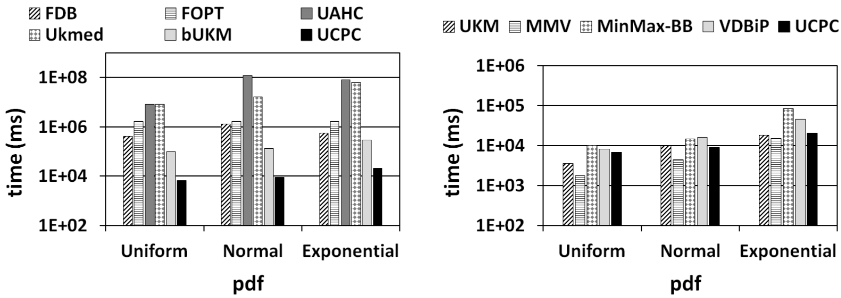
		Quality ($Q \in [-1, 1]$)						
	pdf	$\mathcal{F}$DB	$\mathcal{F}$OPT	UAHC	UKmed	UKM	MMV	**UCPC**
avg score	U	.021	.089	.027	.084	.042	.345	.375
	N	.061	.115	.091	.089	.127	.139	.189
	E	-.001	.025	0	.011	.015	.199	.200
overall avg. score		.027	.076	.039	.061	.061	.228	.255
overall avg. gain		+.228	+.179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

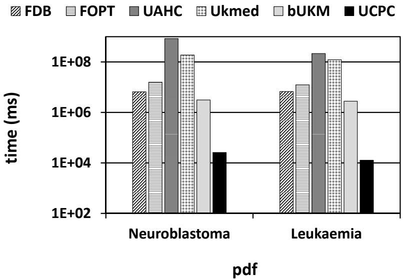
- Efficiency evaluation also involves optimized versions of UK-means, i.e., MinMax-BB and VDBiP

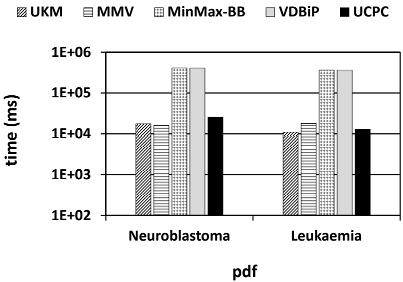
Letter



Efficiency results: real datasets

Real datasets





Backup: details about U-centroid

U-centroid: analytical expression

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:*

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

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

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

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:

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

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 $\mu(o)$, $\mu_2(o)$, $\sigma^2(o)$, $\forall o \in \mathcal{D}$
$\mathcal{C} \leftarrow \mathit{initialPartition}(\mathcal{D}, k)$, 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 $\mathcal{O}(I\ k\ |\mathcal{D}|\ m)$