SUM 2008

2nd International Conference on Scalable Uncertainty Management

Napoli, Italy, October 1-3, 2008

Clustering Uncertain Data via K-Medoids

F. Gullo, G. Ponti, A. Tagarelli

DEIS – University of Calabria - Italy

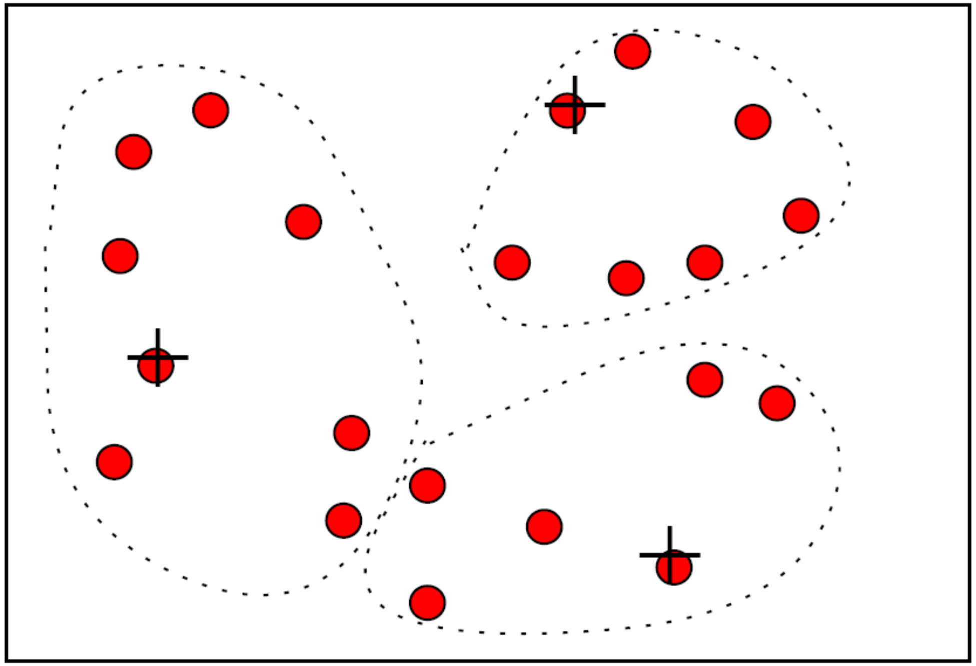
UNIVERSITÀ DELLA CALABRIA

Dipartimento di ELETTRONICA, INFORMATICA E SISTEMISTICA

Introduction

Clustering or *unsupervised classification:*

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



Introduction

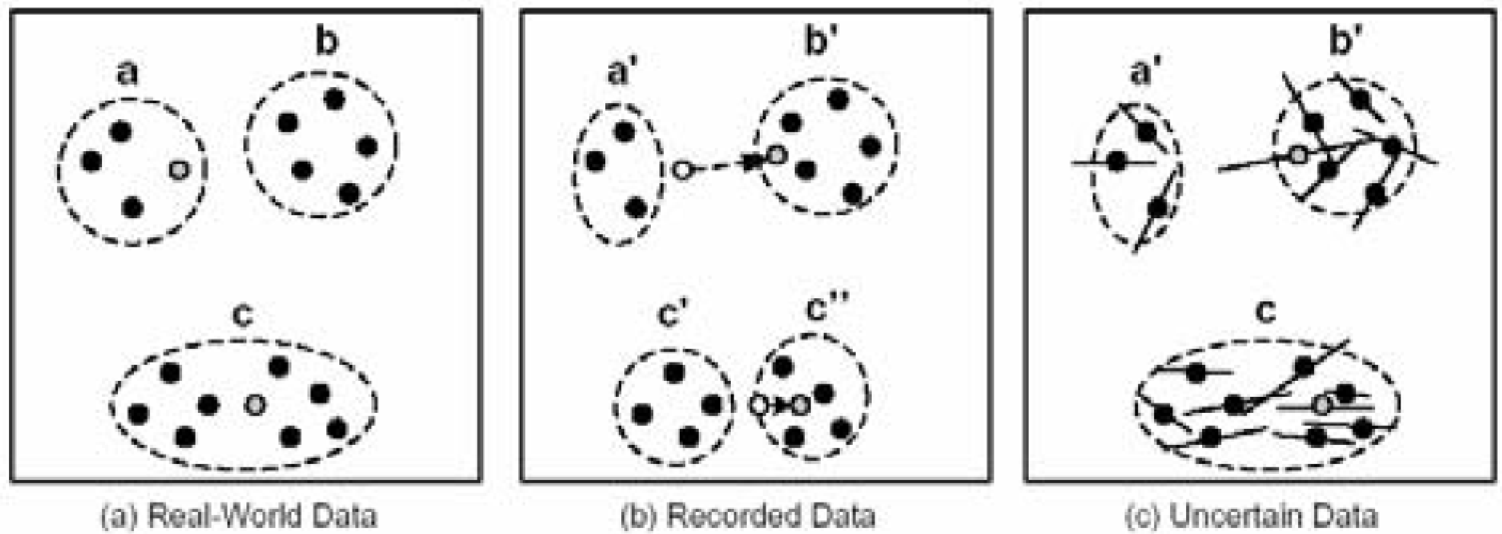
Data uncertainty is inherent in many applications due to, e.g.,

- *randomness in data generation/acquisition*
- *imprecision in physical measurements*
- *data staling*

Applications*: data cleaning, data integration, information extraction, sensor networks, market surveillance, moving object management, …*

Introduction

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



Outline

- Introduction
- Modeling uncertainty
- Our proposal: *a K-medoids-based algorithm for clustering uncertain data*
- Experimental results
- Conclusions

Modeling uncertainty

- **Granularity**: *e.g., table-level, tuple-level, attribute-level*
- **Modeling**: *e.g., % error, intervals, multi-represented objects, probabilistic models*

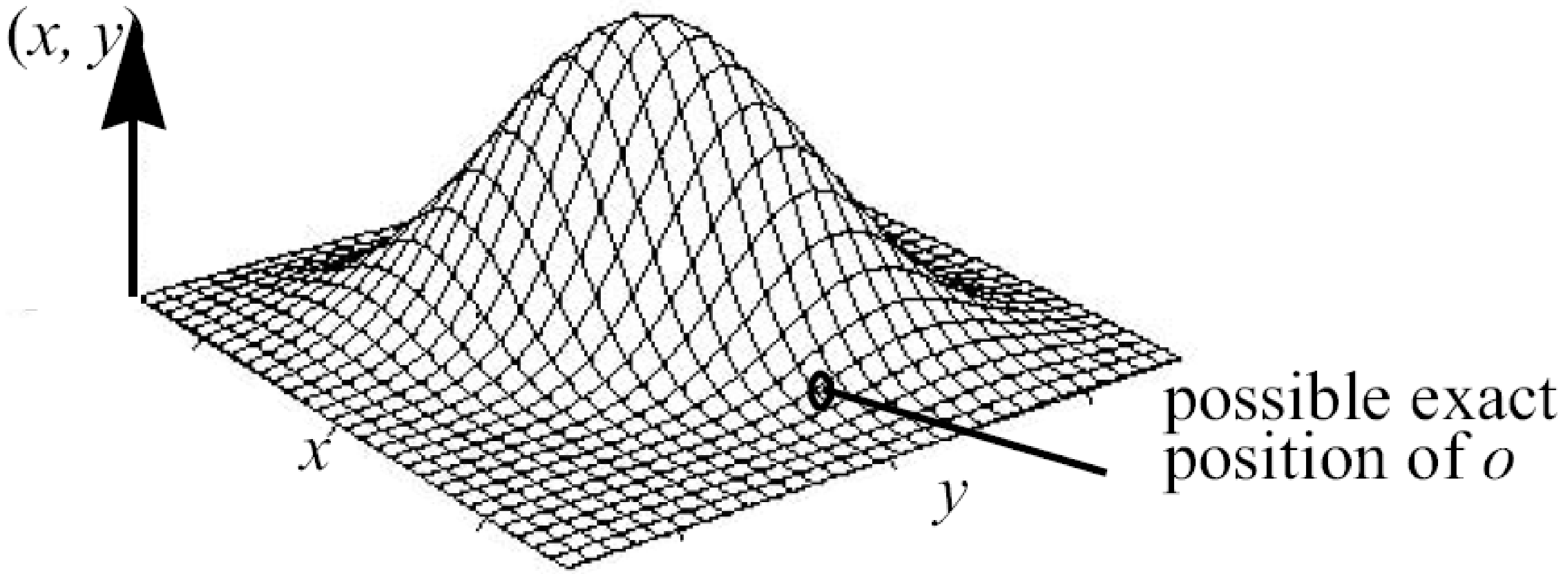
Modeling uncertainty: *uncertain objects*

An uncertain object is a data object represented by means of *probability density functions* (*pdfs*) that describe the probability that the object appears in a multidimensional space

➠ *pdfs can be either continuous or discrete*

Modeling uncertainty:
multivariate uncertain objects

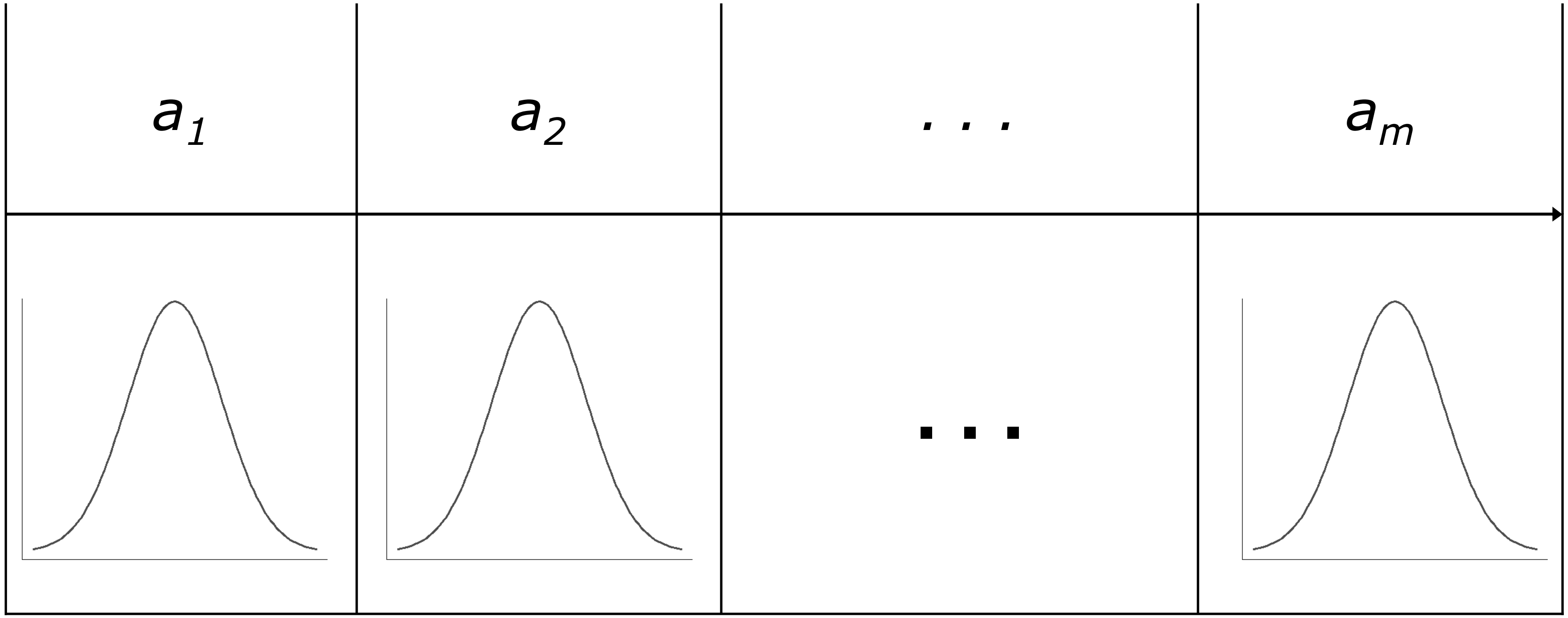
$$o = (R, f), R \subset \Re^m, f : \Re^m \to \Re_0^+$$



Modeling uncertainty:
univariate uncertain objects

$$o = ((I^{(1)}, f^{(1)}), \ldots, (I^{(m)}, f^{(m)}))$$

$$I^{(h)} = [l^{(h)}, u^{(h)}], f^{(h)} : \Re \rightarrow \Re_0^+, h \in [1..m]$$



Clustering of uncertain objects: *UK-means*

[Chau et Al., PAKDD'06]

- UK-means is an adapted version of K-means which handles uncertain objects
- It works on **multivariate** uncertain objects
- It provides the notion of centroid of a cluster of uncertain objects
- It defines the Expected Distance (ED) between centroids and uncertain objects

UK-means

[Chau et Al., PAKDD'06]

select n objects as initial centroids

REPEAT

assign each object to the closest cluster based on its distance to centroids

recompute centroids

UNTIL ***centroids do not change***

UK-means: *computing centroids*

a_1	a_2	. . .	a_m
$\mathbf{x_{11}}$	$\mathbf{x_{12}}$	. . .	$\mathbf{x_{1m}}$
$\mathbf{x_{21}}$	$\mathbf{x_{22}}$	. . .	$\mathbf{x_{2m}}$
⋮			
$\mathbf{x_{n1}}$	$\mathbf{x_{n2}}$	. . .	$\mathbf{x_{nm}}$

centroid:

$\frac{1}{n}\sum_{i=1}^{n} x_{i1}$	$\frac{1}{n}\sum_{i=1}^{n} x_{i2}$	. . .	$\frac{1}{n}\sum_{i=1}^{n} x_{im}$

UK-means: *computing centroids*

[Chau et Al., PAKDD'06]

Centroid $\vec{c} \in \Re^m$ of cluster C :

$$\vec{c} = E\left[\frac{1}{|C|}\sum_{o \in C} f\right] =$$

$$= \frac{1}{|C|}\sum_{o \in C}\int_{\vec{x} \in \Re^m} \vec{x}\ f(\vec{x})\ \mathrm{d}\vec{x}$$

UK-means: *computing EDs*

[Chau et Al., PAKDD'06]

ED between a centroid $\vec{c}$ and an uncertain object $o = (R, f)$:

$$ED(\vec{c}, o) = E\left[\|\vec{c} - f\|^2\right] =$$

$$= \int_{\vec{x} \in \Re^m} \|\vec{c} - \vec{x}\|^2 \; f(\vec{x}) \; \mathrm{d}\vec{x}$$

UK-means

[Chau et Al., PAKDD'06]

Two major weaknesses:

- representing centroids (*accuracy issue*)
- computing *Expected Distance (ED)* between centroids and uncertain objects (*efficiency issue*)

Our proposal

medoids instead of centroids…

Accuracy improvement:

cluster representatives are not computed as a trivial mean of expected values

Efficiency improvement:

the bottleneck of computing EDs at each iteration can be reduced by computing offline the pair-wise distances for each pair of objects

Our proposal: *UK-medoids*

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

compute distances $\delta(o_i, o_j), \forall o_i, o_j \in D$
compute the set $S = \{m_1, \ldots, m_k\}$ of initial medoids
repeat
  $S' \leftarrow S$
  $S \leftarrow \emptyset$
  $\mathcal{C} = \{C_1, \ldots, C_k\} \leftarrow \{\emptyset, \ldots, \emptyset\}$
  for all $o \in D$ **do**
    {*assign each object to the closest cluster, based on its uncertain distance to cluster medoids*}
    $m_j \leftarrow \arg\min_{o' \in S'} \delta(o, o')$
    $C_j \leftarrow C_j \cup \{o\}$
  end for
  for all $C \in \mathcal{C}$ **do**
    {*recompute the medoid of each cluster*}
    $m \leftarrow \arg\min_{o \in C} \sum_{o' \in C} \delta(o, o')$
    $S \leftarrow S \cup \{m\}$
  end for
until $S \neq S'$
return $\mathcal{C}$

UK-medoids

What about the distance between uncertain objects ?

Uncertain distance: *multivariate objects*

$\delta(o_i, o_j)$ is computed by taking into account the distances between all the possible deterministic locations $\vec{x}$, $\vec{y}$, for o_i and o_j, respectively, and their corresponding probabilities $f_i(\vec{x})$, $f_j(\vec{y})$

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

Uncertain distance: *univariate objects*

$$\delta(o_i, o_j) = f_{dist}(\psi^{(1)}(o_i, o_j), \ldots, \psi^{(m)}(o_i, o_j))$$

$$\psi^{(h)}(o_i, o_j) = \int_{x \in I_i^{(h)}} \int_{y \in I_j^{(h)}} |x - y| \; f_i^{(h)}(x) \; f_j^{(h)}(y) \; \mathrm{d}x \; \mathrm{d}y$$

Experiments

- Comparison between UK-means and UK-medoids
  - Accuracy evaluation
  - Efficiency evaluation

Experiments: *datasets*

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

On each of the selected datasets, the uncertainty for any object was synthetically generated according to both the univariate and multivariate models

Pdfs used: *Uniform*, *Normal*, *Binomial*

Experiments: *accuracy evaluation*

To assess the accuracy of clustering solutions, the availability of reference classifications for the datasets

$$\Gamma = \{\Gamma_1, \ldots, \Gamma_k\}$$

reference classification

$$C = \{C_1, \ldots, C_k\}$$

output classification

$$P = \frac{1}{k} \sum_{i=1}^{k} \frac{|C_i \cap \Gamma_i|}{|C_i|}$$ ***precision***

$$R = \frac{1}{k} \sum_{i=1}^{k} \frac{|C_i \cap \Gamma_i|}{|\Gamma_i|}$$ ***recall***

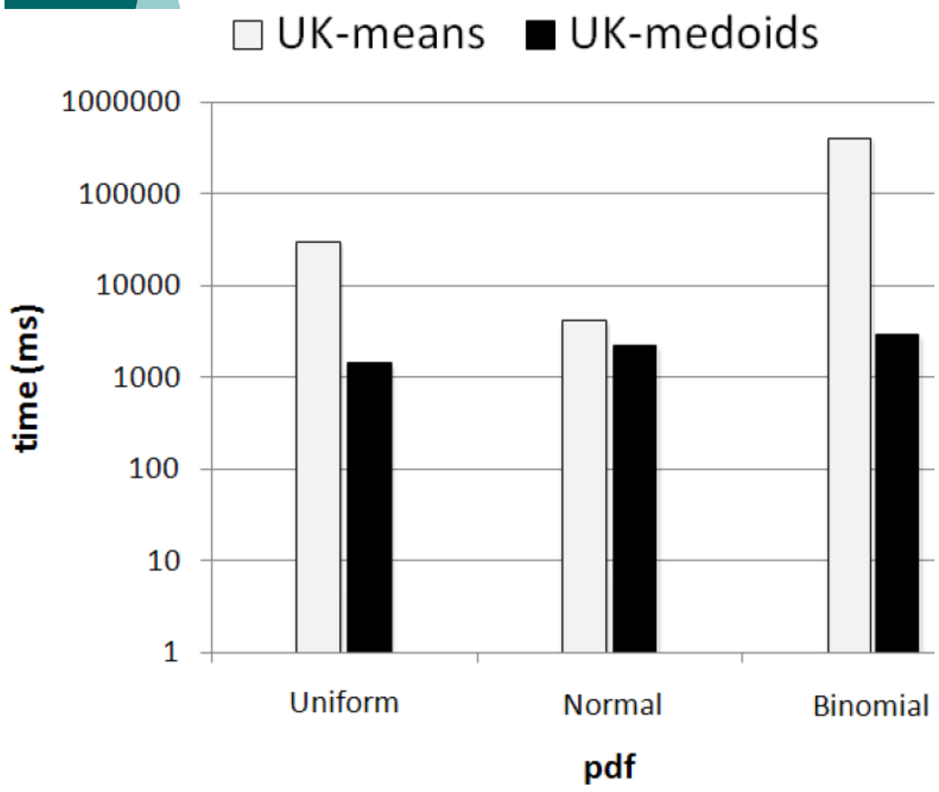
$$F = \frac{2PR}{P + R}$$ ***f-measure***

Experiments: *accuracy results*

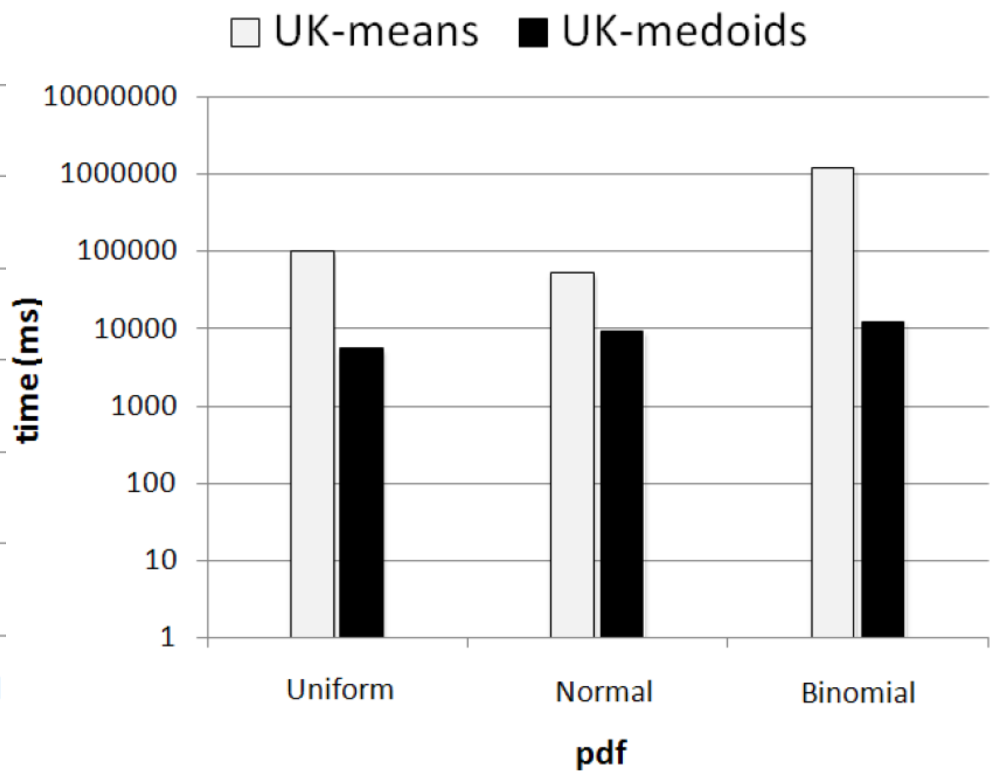
dataset	*pdf*	UK-means	**UK-medoids**
Iris	Uniform	0.45	**0.84**
	Normal	0.84	**0.88**
	Binomial	*0.62*	**0.87**
Wine	Uniform	0.46	**0.80**
	Normal	0.69	**0.70**
	Binomial	*0.63*	**0.73**
Glass	Uniform	0.26	**0.71**
	Normal	*0.63*	**0.68**
	Binomial	0.27	**0.67**
Ecoli	Uniform	0.30	**0.73**
	Normal	0.73	**0.77**
	Binomial	*0.50*	**0.72**

- *Uniform* : + 34-45%
- *Normal* : + 1-5%
- *Binomial* : + 10-40%

Experiments: *efficiency results*

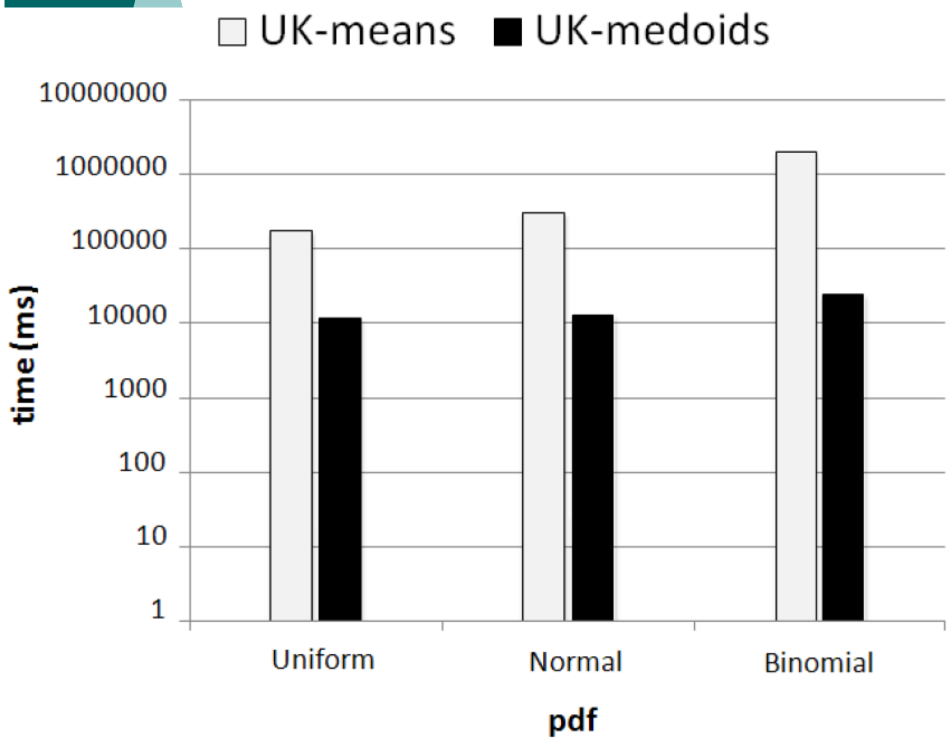


Iris

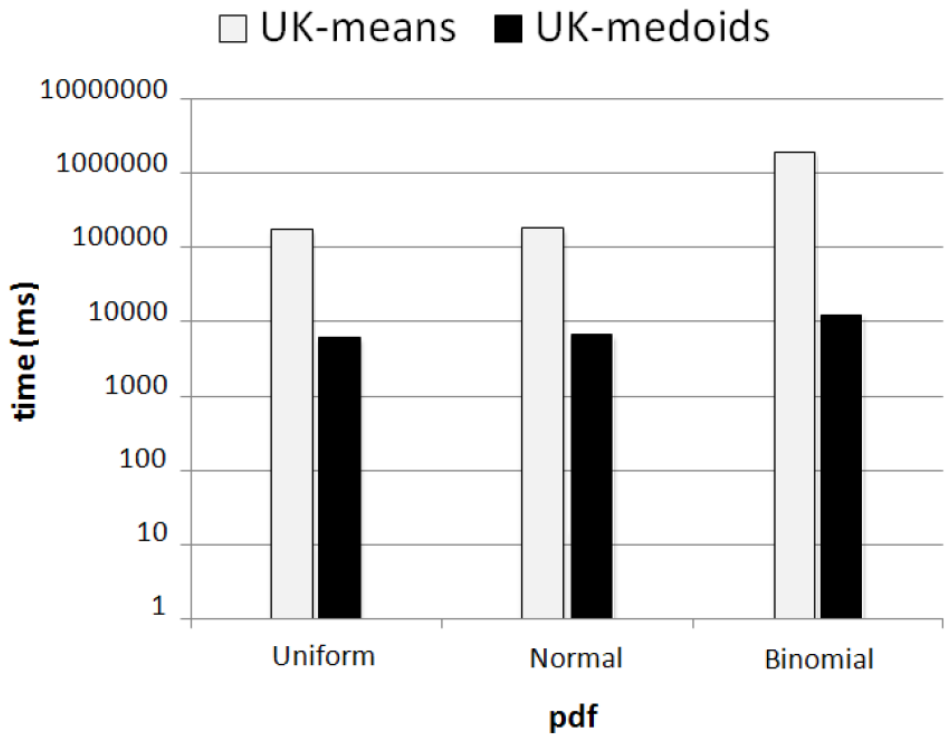


Wine

Experiments: *efficiency results*



Ecoli



Glass

Experiments: *efficiency results*

- Efficiency results:
  - *UK-medoids is 1-2 orders of magnitude faster than UK-means*

Conclusions

- UK-medoids: a K-medoids-based algorithm for clustering uncertain objects
  - Notion of *medoid*
  - Notions of uncertain distance between multivariate and univariate uncertain objects
- High accuracy, good efficiency

Thanks

Traditional (numerical) data objects

A data object represented by a vector of deterministic values

a_1	a_2	. . .	a_m
$\mathbf{x_1}$	$\mathbf{x_2}$	**. . .**	$\mathbf{x_m}$

UK-medoids: *uncertain distance function*

$\Delta(o_i, o_j, z)$ returns the probability that the distance between o_i and o_j is equal to z

$$\int_{z \in \Re} \Delta(o_i, o_j, z)\ \mathrm{d}z = 1, \quad \forall o_i, o_j \in D,$$

$$\Delta(o_i, o_j, z) = \begin{cases} 1, & \textit{if} \;\; i = j, z = 0 \\ 0, & \textit{if} \;\; i = j, z \neq 0 \end{cases}$$

Uncertain distance function: *multivariate objects*

$\Delta(o_i, o_j, z)$ is computed by taking into account all the possible values $\vec{x}$, $\vec{y}$, for o_i and o_j, respectively, such that the distance between $\vec{x}$ and $\vec{y}$ is equal to z

$$\Delta(o_i, o_j, z) = \int_{\vec{x} \in R_i} \int_{\vec{y} \in R_j} I[dist(\vec{x}, \vec{y}) = z] \; f_i(\vec{x}) \; f_j(\vec{y}) \; \mathrm{d}\vec{x} \; \mathrm{d}\vec{y}$$

Uncertain distance function: *univariate objects*

$$\Delta(o_i, o_j, z) = \int_{x_1 \in \Re} \cdots \int_{x_m \in \Re} I[f_{dist}(x_1, \ldots, x_m) = z] \prod_{h=1}^{m} \Psi^{(h)}(o_i, o_j, x_h)\ \mathrm{d}x_1 \cdots \mathrm{d}x_m$$

- $\Psi^{(h)} : D \times D \times \Re \longrightarrow \Re$,
- $\Psi^{(h)}(o_i, o_j, x_h) = \int_{u \in I_i^{(h)}} \int_{v \in I_j^{(h)}} I[|u - v| = x_h]\ f_i^{(h)}(u)\ f_j^{(h)}(v)\ \mathrm{d}u\ \mathrm{d}v, \quad h \in [1..m]$,
- $f_{dist} : \Re^m \longrightarrow \Re$ is a function that computes a scalar value from the components of a vector $(x_1, \ldots, x_m)$

UK-medoids: *uncertain distance*

Given an uncertain distance function Δ, the *uncertain distance* δ is defined by extracting a single, well-representative value from Δ

$$\delta(o_i, o_j) = \int_{z \in \Re} z \Delta(o_i, o_j, z) \ \mathrm{d}z$$

Clustering: *partitional clustering*

Partitional *(or* ***partitioning****) clustering iteratively assigns objects to the clusters according to the intra- and inter-cluster distance*

Precision and Recall

$$\Gamma = \{\Gamma_1, \ldots, \Gamma_k\}$$

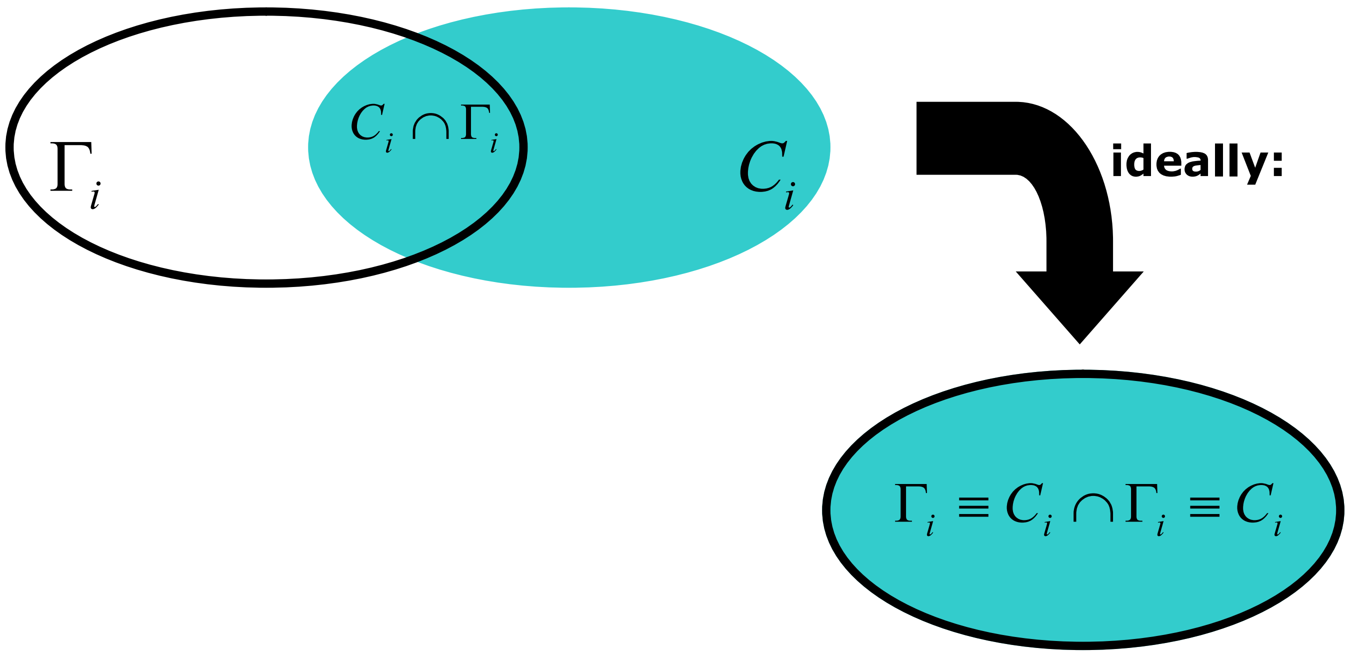
desired classification

$$C = \{C_1, \ldots, C_k\}$$

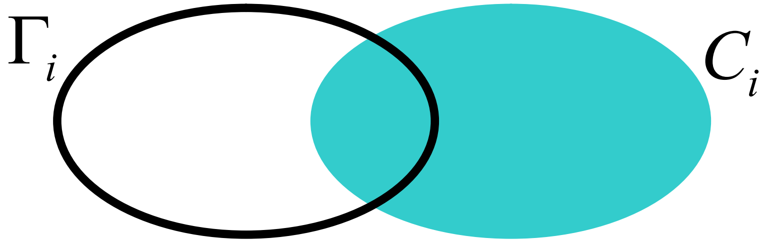
output classification

$$P = \frac{1}{k}\sum_{i=1}^{k}\frac{|C_i \cap \Gamma_i|}{|C_i|}$$ ***precision***

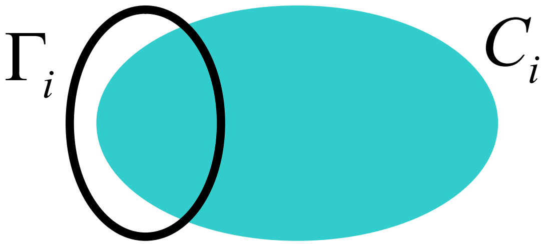
$$R = \frac{1}{k}\sum_{i=1}^{k}\frac{|C_i \cap \Gamma_i|}{|\Gamma_i|}$$ ***recall***



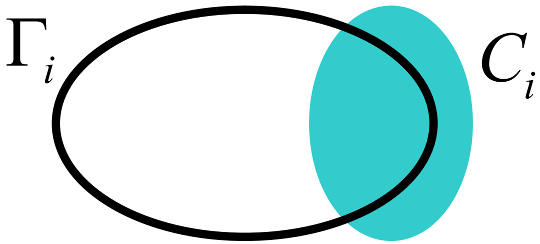
Precision and Recall



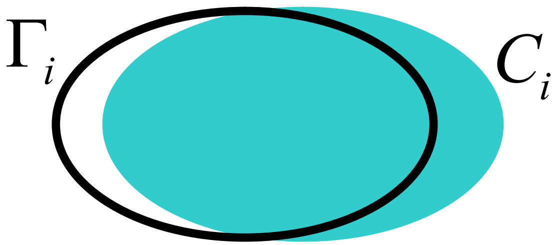
LOW Precision
LOW Recall



LOW Precision
HIGH Recall

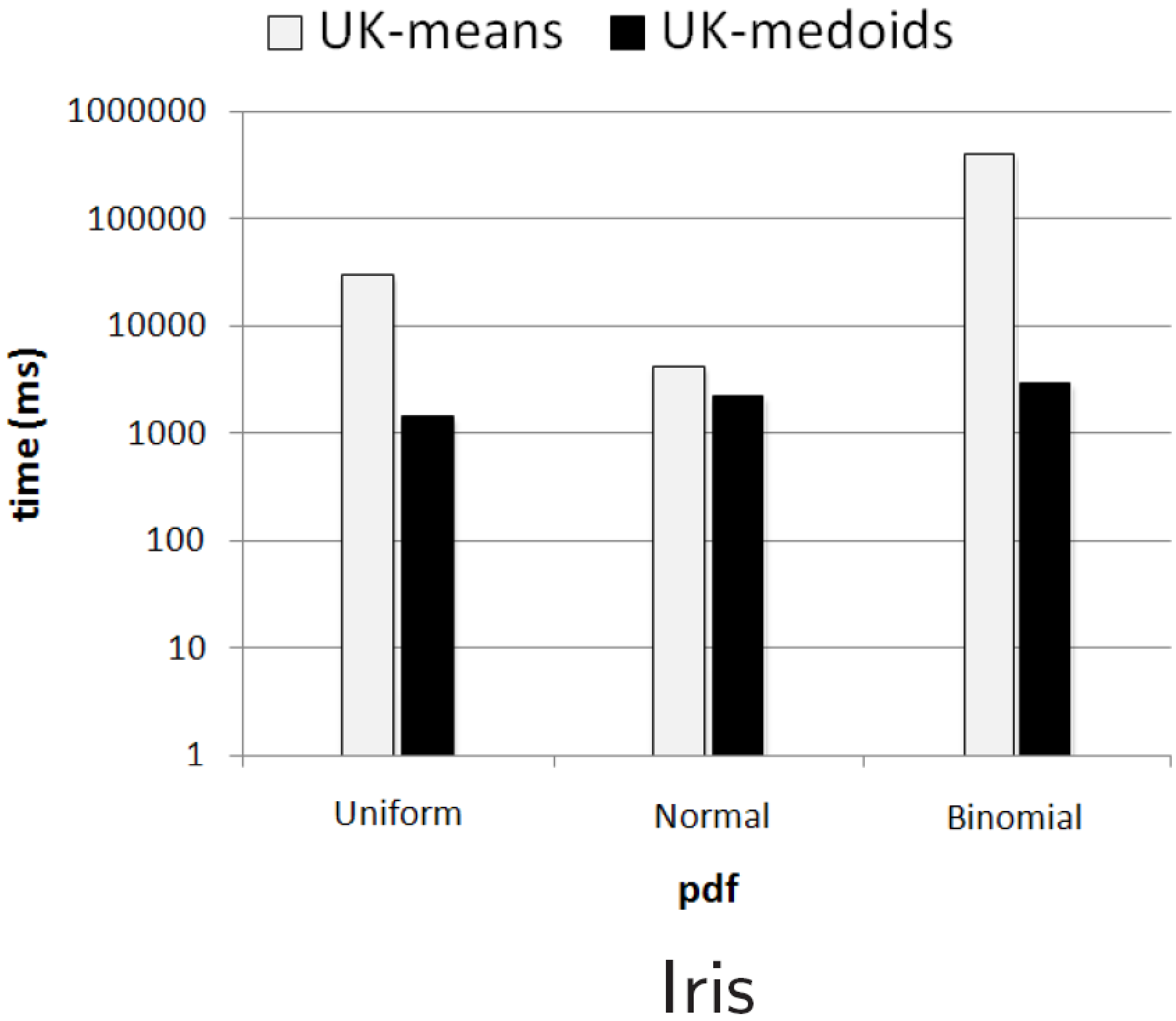


HIGH Precision
LOW Recall

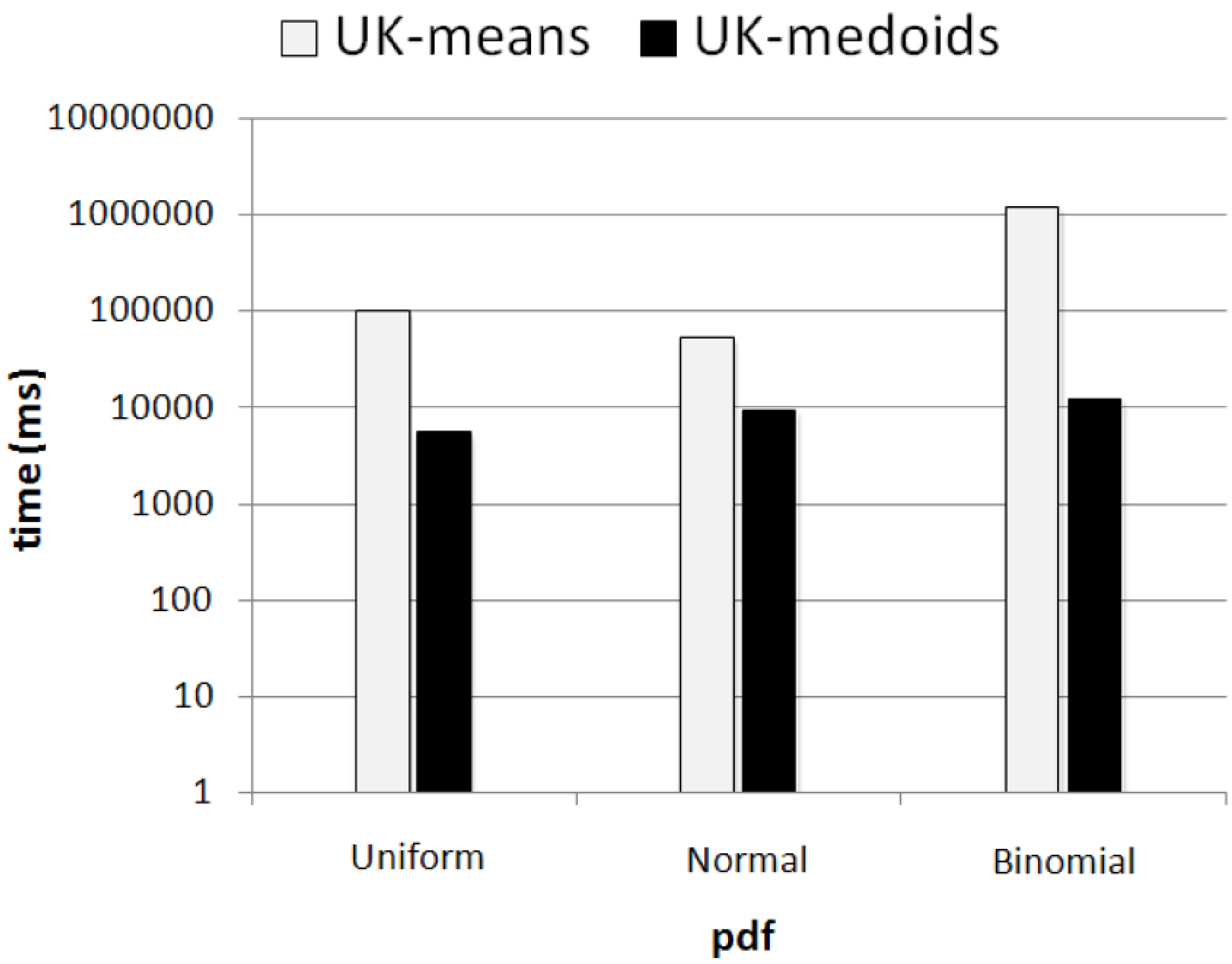


HIGH Precision
HIGH Recall

Experiments: *efficiency results*

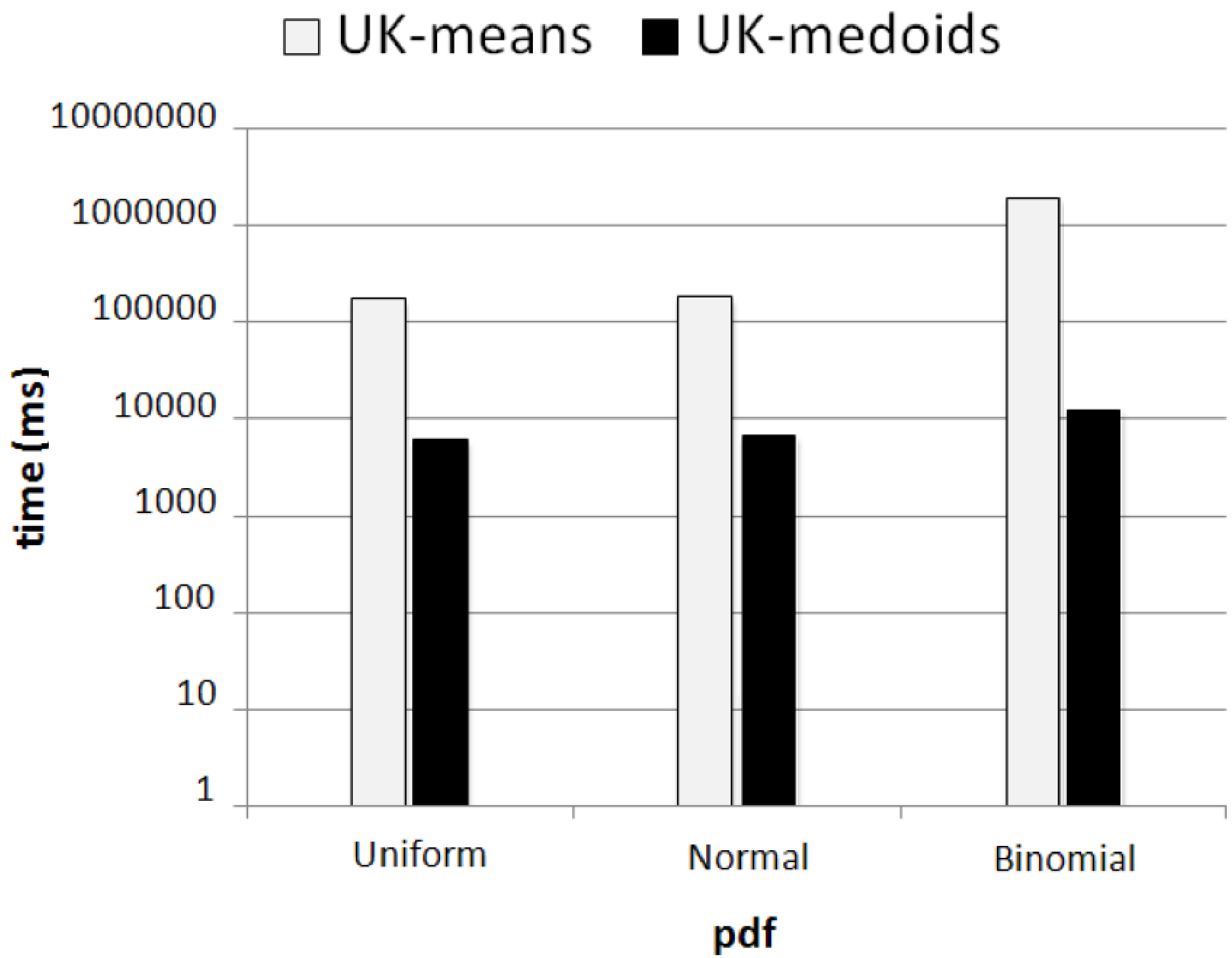


Experiments: *efficiency results*



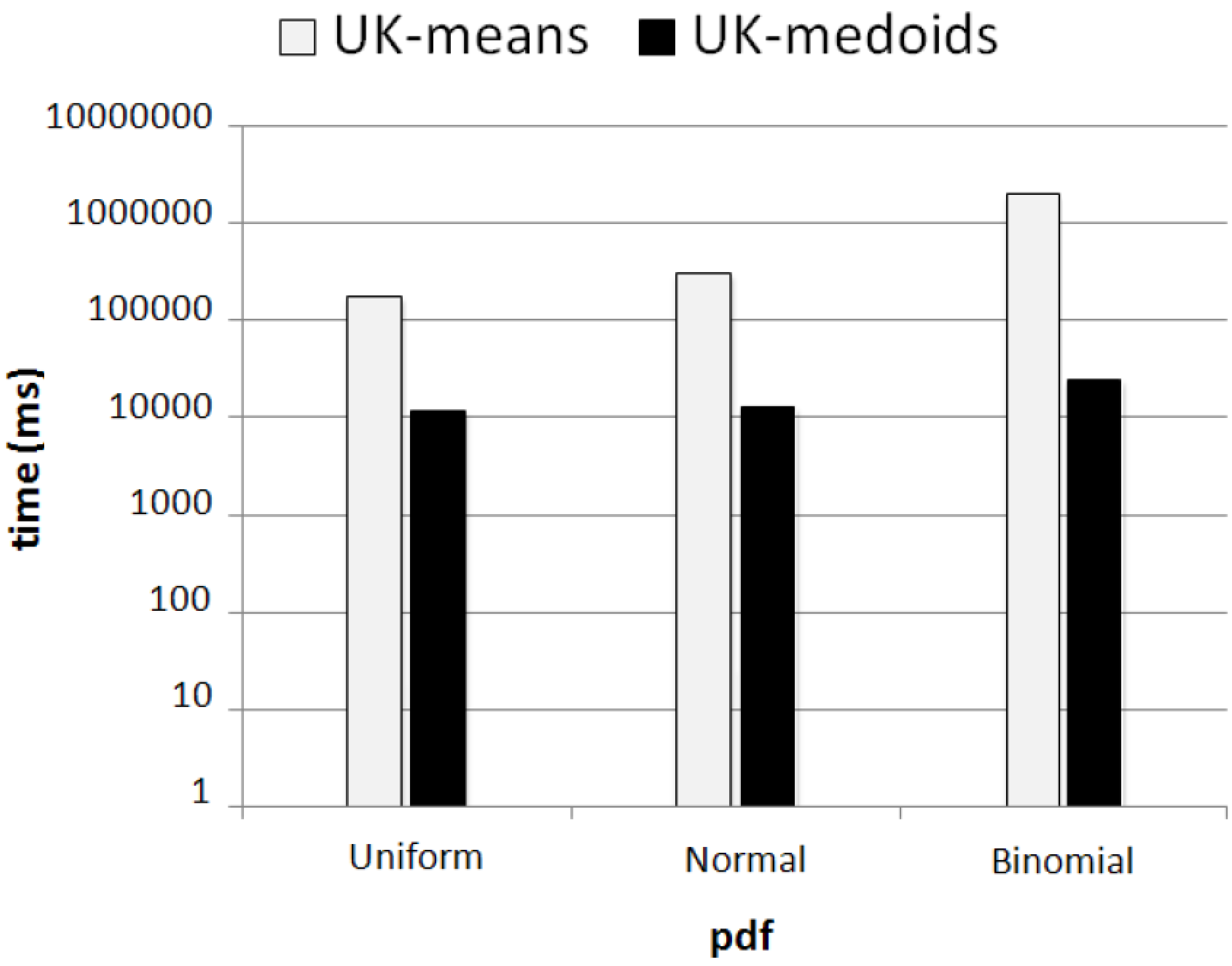
Wine

Experiments: *efficiency results*



Glass

Experiments: *efficiency results*



Ecoli