Minimizing the Variance of Cluster Mixture Models for Clustering Uncertain Objects

Francesco Gullo   Giovanni Ponti   Andrea Tagarelli

Department of Electronics, Computer and Systems Science (DEIS) - University of Calabria
Via P. Bucci, 41C — Arcavacata di Rende (CS), I87036, Italy
{fgullo,gponti,tagarelli}@deis.unical.it

***Abstract*—The increasing demand for dealing with uncertainty in data has led to the development of effective and efficient approaches in the data management and mining contexts. Clustering uncertain data objects has particularly attracted great attention in the data mining community. Most existing clustering methods however have urgently to come up with a number of issues, some of which are related to a poor efficiency mainly due to an expensive computation of the distance between uncertain objects.**

In this work, we propose a novel formulation to the problem of clustering uncertain objects, which allows for reaching accurate solutions by minimizing the variance of the mixture models that represent the clusters to be identified. We define a heuristic, *MMVar*, which exploits some analytical properties about the computation of variance for mixture models to compute local minima of the objective function at the basis of the proposed formulation. This characteristic allows MMVar to discard any distance measure between uncertain objects and, therefore, to achieve high efficiency. Experiments have shown that MMVar outperforms state-of-the-art algorithms from an efficiency viewpoint, while achieving better average performance in terms of accuracy.

I. Introduction

Uncertainty in data is typically due to a variety of phenomena such as imprecision in physical measurements, randomness implicitly present in a process of data generation/acquisition, data staling. As a result, uncertain data is naturally present in several application domains. For instance, sensor measurements may be imprecise at a certain degree due to the presence of various noisy factors (e.g., signal noise, instrumental errors, wireless transmission) [1]. Another example is given by data representing moving objects, which continuously change their location so that the exact positional information at a given time instant may be unavailable [2]. Moreover, some methods have recently been defined to handle uncertainty in gene expression data [3]. Further examples come from distributed applications, privacy preserving data mining, and forecasting or other statistical techniques used to generate data attributes [4].

The heterogeneity of application domains in which data uncertainty is significant has led to the study of various notions of uncertainty [5]–[7]. In general, uncertainty can be considered at relation, tuple or attribute level [8], and is usually specified by fuzzy models, evidence-oriented models, and probabilistic models [9], [10]. In this work, we are interested in attribute-level uncertainty modeled according to probabilistic models. Specifically, we focus on uncertain data representations based on probability distributions that are aimed at describing the likelihood that any given object appears at each position in a multidimensional space [11]–[14]. We hereinafter refer to data objects described in terms of probability distributions as *uncertain objects*.

Knowledge discovery and mining in uncertain objects lead to a number of special challenges which are mainly due to the need for developing adequately effective and efficient solutions to deal with uncertainty. Particularly, among data mining tasks, *clustering* uncertain objects has gained tremendous interest in the last few years [11]–[17]. Clustering uncertain objects is challenging due to the following major issues:

- Existing algorithms require some notion of distance between uncertain objects, whose definition is a non-trivial task. In this respect, the two main approaches to compute the distance between uncertain objects have both some weaknesses in their own. The first approach consists in computing the distance between aggregated values extracted from the probability distributions of uncertain objects (e.g., expected values), and has a complexity linear w.r.t. the number S of statistical samples used for representing distributions. The second approach involves the computation of the so-called expected distance between distributions [13], which aims to exploit the whole information available from the distributions and works in $\mathcal{O}(S^2)$. Although in principle efficient, the first approach (i.e., distance between aggregated values) may easily incur an accuracy issue, since all the information available from the distributions is collapsed into a single, representative numerical value. An opposite consideration holds instead for the expected distance based approach, which is more accurate but also inefficient.
- A further, more critical issue concerns the efficiency of existing algorithms. This is partially related to the

need for a distance measure between uncertain objects discussed above, since the use of a slow measure clearly leads to poor efficiency. However, more generally, it intrinsically depends on the specific formulations at the basis of existing algorithms, which constrain such heuristics to continuously execute critical operations, such as access to the samples of distributions (e.g., needed for integral approximations) and/or the computation of distances between objects.

In this paper, we propose a novel formulation to the problem of clustering uncertain objects, which aims to overcome both the above issues, while maintaining high accuracy. The key idea of the proposed formulation is to take into account the mixture model of a set (cluster) of uncertain objects and to define a criterion based on the minimization of the *variance* of the mixture models that represent the clusters to be discovered. This criterion is devised to fulfill both accuracy and efficiency requirements: accuracy is motivated as minimizing the variance of cluster mixture models leads to discover highly homogeneous clusters; at the same time, some analytical properties we derive about the computation of mixture models and their variances lead to the definition of a fast heuristic that does not require any distance measure between uncertain objects. Within this view, the main contributions of this work can be summarized as follows:

1) We propose a novel formulation to the problem of clustering uncertain objects, essentially based on the minimization of the variance of the mixture models that represent the clusters of uncertain objects to be discovered.
2) We derive results based on analytical properties about the objective function at the basis of the proposed formulation, which enable the design of fast heuristics for clustering uncertain objects.
3) We define *MMVar*, a heuristic algorithm to compute good approximations of the proposed formulation. Major features of MMVar include: (i) high efficiency, (ii) no need for using distance measures between uncertain objects, (iii) capability of discovering local minima of the proposed objective function.
4) We have conducted an experimental evaluation on several datasets to assess our MMVar from both efficiency and accuracy viewpoints, as well as to compare it with prominent state-of-the-art algorithms. In this respect, MMVar revealed to be always faster and on average more accurate, particularly achieving the best maximum accuracy in more than half cases we have considered.

II. Related Work

We briefly review the main state-of-the-art algorithms for clustering uncertain objects, paying special attention to their computational complexities (Table I). In this respect, we adopt the following notation: n is the size of the input set of uncertain objects, m is the dimensionality of the uncertain objects (i.e., number of features), k is the desired number of clusters, I is the number of iterations for convergence required by partitional clustering algorithms, and S denotes the number of statistically independent samples employed for representing probability distributions.

Table I
Computational complexities of prominent state-of-the-art algorithms for clustering uncertain objects

algorithm	*total*	*online*	*offline*
UK-means	$\mathcal{O}(I\ S\ k\ n\ m)$	$\mathcal{O}(I\ S\ k\ n\ m)$	—
CK-means	$\mathcal{O}(n\ m\ (I\ k + S))$	$\mathcal{O}(I\ k\ n\ m)$	$\mathcal{O}(S\ n\ m)$
UK-medoids	$\mathcal{O}(n^2(I + S^2\ m))$	$\mathcal{O}(I\ n^2)$	$\mathcal{O}(S^2\ n^2\ m)$
$\mathcal{F}$DBSCAN	$\mathcal{O}(S^2\ n^2\ m)$	$\mathcal{O}(S^2\ n^2\ m)$	—
$\mathcal{F}$OPTICS	$\mathcal{O}(S\ n^2\ m)$	$\mathcal{O}(S\ n^2\ m)$	—
U-AHC	$\mathcal{O}(n^2(S\ m + \log n))$	$\mathcal{O}(n^2(S\ m + \log n))$	—

One of the earliest attempts to solve the problem of clustering uncertain objects is the partitional algorithm *UK-means* [12], which is essentially an adaptation of the popular K-means to the context of uncertain objects. UK-means has a computational complexity of $\mathcal{O}(I\,S\,k\,n\,m)$.

In order to improve the efficiency of UK-means, [16] and [18] propose some pruning techniques to avoid the calculation of redundant object-to-centroid distances, whereas the *CK-means* algorithm [17] exploits the moment of inertia of rigid bodies in order to reduce the execution time for computing the aforementioned distances. CK-means comprises two main steps: in the first one (*offline* phase), the distances between each object and its mass center are computed in $\mathcal{O}(S\ n\ m)$, whereas the second step carries out a classic partitional relocation scheme; in this step, the distances computed in the first step are exploited to obtain a K-means-like strategy working in $\mathcal{O}(I\ k\ n\ m)$. A further partitional algorithm is *UK-medoids* [13], which employs proper distances between uncertain objects that are pre-computed offline in $\mathcal{O}(S^2\ n^2\ m)$; these distances are then employed in a classic K-medoids scheme working in $\mathcal{O}(I\ n^2)$.

Density-based approaches to clustering uncertain objects have also been developed [11], [15]. In [11], the *$\mathcal{F}$DBSCAN* is proposed as a fuzzy version of the popular DBSCAN, whose computational complexity, in the worst case, is $\mathcal{O}(S^2\ n^2\ m)$. A similar approach is presented in *$\mathcal{F}$OPTICS* [15], which resorts to the well-known density-based clustering algorithm OPTICS. By exploiting proper data structures (i.e., core object arrays and reachability lists), $\mathcal{F}$OPTICS can be executed in $\mathcal{O}(S\ n^2\ m)$.

We finally mention *U-AHC* [14], the first hierarchical algorithm for clustering uncertain objects, whose complexity is $\mathcal{O}(n^2(S\ m + \log n))$.

III. Modeling uncertainty

Uncertain data objects are usually represented using the *multivariate uncertainty model* [14]. A *multivariate uncertain object* o is defined as a pair $(\mathcal{R}, f)$, where $\mathcal{R} \subseteq \Re^m$ is the m-dimensional region in which o is defined and $f : \Re^m \to \Re_0^+$ is the probability density function of o at each point $\vec{x} \in \Re^m$, such that:

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

Note that the above definition also includes the case where uncertain objects are represented by probability mass functions, as well as the case where the pdfs are approximated by a set of statistical samples. For the sake of brevity, we hereinafter refer only to the continuous uncertainty model, as the corresponding discrete version can be trivially obtained by replacing integrals with sums.

Given any (multivariate) uncertain object $o = (\mathcal{R}, f)$, the corresponding expected value ($\vec{\mu}$), second order moment ($\vec{\mu}_2$), and variance ($\vec{\sigma}^{\,2}$) are defined as follows:

$$\vec{\mu} \ = \ (\mu_1, \ldots, \mu_m) \ = \int_{\vec{x} \in \mathcal{R}} \vec{x}\ f(\vec{x})\ \mathrm{d}\vec{x} \tag{1}$$

$$\vec{\mu}_2 \ = \ ((\mu_2)_1, \ldots, (\mu_2)_m) \ = \int_{\vec{x} \in \mathcal{R}} \vec{x}^{\,2} f(\vec{x})\ \mathrm{d}\vec{x} \tag{2}$$

$$\vec{\sigma}^{\,2} = ((\sigma^2)_1, \ldots, (\sigma^2)_m) = \int_{\vec{x} \in \mathcal{R}} (\vec{x} - \mu)^2\ f(\vec{x})\ \mathrm{d}\vec{x} \ = \ \vec{\mu}_2 - \vec{\mu}^2 \tag{3}$$

where $\vec{\mu}^2$ denotes the vector $(\mu_1^2, \ldots, \mu_m^2)$. Although the notion of variance for multivariate distributions is meaningful only along a particular dimension, we consider the sum of variances along each dimension to resemble a notion of "global" variance in terms of a single numerical value σ^2. Formally, given any vector $\vec{\sigma}^{\,2}$ of variances, we have:

$$\sigma^2 \ = \|\vec{\sigma}^{\,2}\|_1 = \sum_{j=1}^{m} |(\sigma^2)_j| = \sum_{j=1}^{m} (\sigma^2)_j = \sum_{j=1}^{m} (\mu_2)_j - \mu_j^2 \tag{4}$$

If f is either a probability mass function or a pdf approximated by a set $\mathcal{S}$ of statistical samples, (1) and (2) can be rewritten as follows:

$$\vec{\mu} = \left(\sum_{\vec{y} \in \mathcal{S}} f(\vec{y})\right)^{-1} \sum_{\vec{y} \in \mathcal{S}} \vec{y}\ f(\vec{y}) \qquad \vec{\mu}_2 = \left(\sum_{\vec{y} \in \mathcal{S}} f(\vec{y})\right)^{-1} \sum_{\vec{y} \in \mathcal{S}} \vec{y}^{\,2} f(\vec{y}) \tag{5}$$

IV. Clustering Uncertain Objects via Cluster Variance Minimization

Our proposed formulation to the problem of clustering uncertain objects is based on the notion of *uncertain prototype* for a given set of uncertain objects. An uncertain prototype is essentially described by the mixture model of the random variables representing the objects in a set. Formally, given a set C of uncertain objects, we define the *uncertain prototype* $\mathcal{P}_C$ of C as the pair $(\mathcal{R}_C, f_C)$, where

$$\mathcal{R}_C = \bigcup_{o=(\mathcal{R},f) \in C} \mathcal{R} \qquad \text{and} \qquad f_C(\vec{x}) = \frac{1}{|C|} \sum_{o=(\mathcal{R},f) \in C} f(\vec{x}) \tag{6}$$

The rationale behind the definition of uncertain prototype as mixture model lies in the intuition that the compactness of a set of uncertain objects is higher as the variance of the mixture model (i.e., uncertain prototype) representing that set is lower. In fact, it can be proved that the variance of an uncertain prototype computed according to (4) is equivalent to the expected distance between any uncertain object and the centroid of the cluster represented by that prototype.

Based on the above considerations, we propose to formulate the problem of clustering uncertain objects by minimizing the variance of the uncertain prototypes of the clusters to be identified. Formally, given a set $\mathcal{D}$ of uncertain objects, the objective is to find a partition $\mathcal{C}$ of $\mathcal{D}$ that minimizes the following objective function:

$$J(\mathcal{C}) = \sum_{C \in \mathcal{C}} \sigma^2(\mathcal{P}_C) \tag{7}$$

where $\sigma^2(\mathcal{P}_C)$ is the variance of the prototype $\mathcal{P}_C$ of cluster C, which is computed according to (4).

A. The MMVar Algorithm

The objective function in (7) refers to a classic NP-hard clustering problem. In order to define a fast heuristic, the proposed formulation to the problem aims to exploit analytical properties about the computation of the variance for mixture models. We now discuss such properties in detail.

Proposition 1: Let $\mathcal{D}$ be a set of m-dimensional uncertain objects, where each $o \in \mathcal{D}$ has expected value and second order moment denoted by $\vec{\mu}(o)$ and $\vec{\mu}_2(o)$, respectively. Also, let $\mathcal{C}$ be a partition of $\mathcal{D}$, $\mathcal{P}_C$ be the prototype of any cluster $C \in \mathcal{C}$, and $\vec{\mu}(\mathcal{P}_C)$, $\vec{\mu}_2(\mathcal{P}_C)$ and $\sigma^2(\mathcal{P}_C) = \|\vec{\mu}_2(\mathcal{P}_C) - \vec{\mu}(\mathcal{P}_C)^2\|_1$ the expected value, the second order moment and the variance of $\mathcal{P}_C$, respectively. Let us consider a new partition $\mathcal{C}'$ of $\mathcal{D}$ obtained from $\mathcal{C}$ by moving an object $\tilde{o}$ from cluster $C \in \mathcal{C}$ to cluster $\widehat{C} \in \mathcal{C}$. If we denote $C' = C \setminus \{\tilde{o}\}$ and $\widehat{C}' = \widehat{C} \cup \{\tilde{o}\}$, it holds that the value $J_{\mathcal{C}}(C, \tilde{o}, \widehat{C}) := J(\mathcal{C}')$ of the objective function J for the new partition $\mathcal{C}'$ is computed as:

$$J_{\mathcal{C}}(C, \tilde{o}, \widehat{C}) = J(\mathcal{C}) - (\sigma^2(\mathcal{P}_C) + \sigma^2(\mathcal{P}_{\widehat{C}})) + (\sigma^2(\mathcal{P}_{C'}) + \sigma^2(\mathcal{P}_{\widehat{C}'})) \tag{8}$$

where

$$\sigma^2(\mathcal{P}_{C'}) = \|\vec{\mu}_2(\mathcal{P}_{C'}) - \vec{\mu}(\mathcal{P}_{C'})^2\|_1$$

$$\sigma^2(\mathcal{P}_{\widehat{C}'}) = \|\vec{\mu}_2(\mathcal{P}_{\widehat{C}'}) - \vec{\mu}(\mathcal{P}_{\widehat{C}'})^2\|_1$$

and

$$\vec{\mu}(\mathcal{P}_{C'}) = \frac{|C| \times \vec{\mu}(\mathcal{P}_C) - \vec{\mu}(\tilde{o})}{|C| - 1} \qquad \vec{\mu}_2(\mathcal{P}_{C'}) = \frac{|C| \times \vec{\mu}_2(\mathcal{P}_C) - \vec{\mu}_2(\tilde{o})}{|C| - 1}$$

Algorithm 1 MMVar

```
Input: A set $\mathcal{D}$ of uncertain objects; the number $k$ of output clusters
Output: A partition $\mathcal{C}$ of $\mathcal{D}$
 1: compute $\vec{\mu}(o)$, $\vec{\mu}_2(o)$, $\forall o \in \mathcal{D}$                          {(1)–(2), (5)}
 2: $\mathcal{C} \leftarrow randomPartition(\mathcal{D}, k)$
 3: compute $\vec{\mu}(\mathcal{P}_C)$, $\vec{\mu}_2(\mathcal{P}_C)$, $\forall C \in \mathcal{C}$                     {Prop. 1}
 4: $v \leftarrow J(\mathcal{C})$                                                   {(7)}
 5: repeat
 6:     for all $o \in \mathcal{D}$ do
 7:         let $C \in \mathcal{C}$ be the cluster s.t. $o \in C$
 8:         $C^* \leftarrow \arg\min_{\widehat{C}} J_{\mathcal{C}}(C, o, \widehat{C})$                          {(8)}
 9:         if $C^* \neq C$ then
10:             $v = J_{\mathcal{C}}(C, o, \widehat{C})$                                   {(8)}
11:             recompute $\mathcal{C}$ by moving $o$ from $C$ to $C^*$
12:             recompute $\vec{\mu}(\mathcal{P}_C), \vec{\mu}_2(\mathcal{P}_C), \vec{\mu}(\mathcal{P}_{C^*}), \vec{\mu}_2(\mathcal{P}_{C^*})$   {Prop. 1}
13:         end if
14:     end for
15: until no object in $\mathcal{D}$ is relocated
```

$$\vec{\mu}(\mathcal{P}_{\widehat{C}'})=\frac{|C|\times\vec{\mu}(\mathcal{P}_C)+\vec{\mu}(\tilde{o})}{|C|+1} \quad \vec{\mu}_2(\mathcal{P}_{\widehat{C}'})=\frac{|C|\times\vec{\mu}_2(\mathcal{P}_C)+\vec{\mu}_2(\tilde{o})}{|C|+1}$$

Proposition 1 states that, given any partition $\mathcal{C}$ of $\mathcal{D}$ and any other partition $\mathcal{C}'$ obtained from $\mathcal{C}$ by moving an object from a cluster to another one, the value of the objective function J for $\mathcal{C}'$ can be computed in $\mathcal{O}(m)$ from $J(\mathcal{C})$ according to (8). This result puts the basis for our proposed heuristic algorithm, called *MMVar*, whose major feature lies in the capability of efficiently finding local minima of function J, without requiring any distance measure between uncertain objects.

The outline of MMVar is reported in Alg. 1. To exploit the result of Proposition 1, MMVar has only to store the expected values and second order moments of both the objects within $\mathcal{D}$ (i.e., vectors $\vec{\mu}(o)$ and $\vec{\mu}_2(o)$) and the prototypes that are identified at each iteration (i.e., vectors $\vec{\mu}(\mathcal{P}_C)$ and $\vec{\mu}_2(\mathcal{P}_C)$). MMVar consists of two main steps. In the first step, an initialization phase (Lines 1-4) is carried out to compute: (i) expected values and second order moments of each object within $\mathcal{D}$, according to either the exact (cf. (1)-(2)) or approximated (cf. (5)) formulas (Line 1), (ii) an initial random partition $\mathcal{C}$ of $\mathcal{D}$ (Line 2), (iii) expected values and second order moments of the prototypes of the clusters in $\mathcal{C}$ according to Proposition 1 (Line 3), and (iv) the value v of function J for $\mathcal{C}$ according to (7) (Line 4). In the second step, the main cycle of the algorithm (Lines 5-15) is repeated until convergence (i.e., no object is relocated during any specific iteration). At any iteration of the main cycle, for each object $o \in \mathcal{D}$, the cluster C^* is discovered (Lines 7-8) so that the maximum decrease in the objective function J is obtained if o is moved to it. C^* is discovered by applying (8) to the current partition $\mathcal{C}$; note that (8) is computed by taking into account the value v of the objective function J for the current partition $\mathcal{C}$. If C^* is not the same as the cluster C which o currently belongs to (i.e., there exists at least one cluster but C such that function J decreases if o is moved to it), then o is moved to C^* and both the new value v of function J and the expected values and second order moments of the prototypes of clusters C and C^* are recomputed (Lines 9-13).

Table II
DATASETS USED IN THE EXPERIMENTS

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

It can be proved that Alg. 1 converges to a local minimum of function J defined in (7) in a finite number of steps and its computational complexity is $\mathcal{O}(n\ m\ (I\ k + S))$, where I is the number of iterations needed for convergence. This complexity reduces to $\mathcal{O}(I\ k\ n\ m)$ when the moments of the various uncertain objects are computed according to some closed-form expression in $\mathcal{O}(m)$, which typically happens in a wide number of real cases. Thus, looking at Table I, it is easy to note that the complexity of the proposed MMVar algorithm is at worst equal to and very often (far) lower than that of other existing algorithms for clustering uncertain objects. This result also supports a major claim of this work, which concerns the efficiency in solving the problem of clustering uncertain objects.

V. Experimental Evaluation

The proposed MMVar algorithm was evaluated in performing effective and efficient clustering of uncertain objects. Moreover, MMVar was compared with existing partitional algorithms, i.e., UK-means (UKM), CK-means (CKM), and UK-medoids (UKmed), density-based algorithms, i.e., $\mathcal{F}$DBSCAN ($\mathcal{F}$DB) and $\mathcal{F}$OPTICS ($\mathcal{F}$OPT), and the hierarchical algorithm U-AHC.

A. Methodology

Experiments were performed on both benchmark and real-world datasets. Due to the space limits of this paper, here we present experiments conducted on benchmark datasets only [19] (Table II). These datasets were originally established as collections of data with deterministic values, therefore we generated uncertainty in these collections synthetically; for this purpose, we followed the method as described in [14].

To assess the quality of clustering solutions, we exploited the availability of reference classifications for each dataset. This allowed us to evaluate how well a clustering fits a predefined scheme of known classes. For this purpose, we resorted to the well-known *F-measure* (F), which is defined as the harmonic mean of two standard notions in Information Retrieval, namely precision and recall. If we denote with $\widetilde{\mathcal{C}} = \{\widetilde{C}_1, \ldots, \widetilde{C}_h\}$ a reference classification and

Table III
ACCURACY RESULTS (F-MEASURE)

		F-measure ($F \in [0,1]$)						
data	*pdf*	UKM	CKM	UKmed	$\mathcal{F}$DB	$\mathcal{F}$OPT	UAHC	**MMVar**
Iris	U	0.84	0.93	0.925	0.8	0.907	0.934	0.975
	N	0.853	0.876	0.873	0.8	0.907	0.854	0.893
	B	0.634	0.503	0.838	0.5	0.906	0.544	0.871
Wine	U	0.5	0.726	0.854	0.5	0.853	0.933	0.83
	N	0.5	0.708	0.599	0.499	0.714	0.76	0.723
	B	0.5	0.581	0.604	0.5	0.714	0.693	0.496
Glass	U	0.65	0.663	0.668	0.286	0.596	0.724	0.609
	N	0.549	0.586	0.504	0.534	0.438	0.75	0.609
	B	0.389	0.318	0.629	0.286	0.438	0.47	0.801
Ecoli	U	0.653	0.786	0.677	0.318	0.477	0.542	0.778
	N	0.593	0.734	0.507	0.523	0.477	0.509	0.748
	B	0.556	0.413	0.682	0.333	0.477	0.537	0.699
Yeast	U	0.503	0.562	0.598	0.198	0.535	0.465	0.623
	N	0.476	0.462	0.282	0.396	0.316	0.484	0.715
	B	0.413	0.31	0.456	0.2	0.316	0.402	0.722
Image	U	0.811	0.807	0.774	0.25	0.42	0.579	0.604
	N	0.623	0.633	0.44	0.25	0.42	0.628	0.529
	B	0.533	0.353	0.644	0.25	0.42	0.537	0.701
Abal.	U	0.323	0.295	0.582	0.111	0.494	0.287	0.742
	N	0.282	0.217	0.257	0.343	0.209	0.307	0.386
	B	0.373	0.209	0.318	0.111	0.608	0.306	0.789
Letter	U	0.528	0.633	0.757	0.182	0.319	0.546	0.685
	N	0.446	0.44	0.483	0.182	0.319	0.559	0.649
	B	0.411	0.214	0.646	0.182	0.319	0.572	0.646
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	—

with $\mathcal{C} = \{C_1, \ldots, C_k\}$ a clustering solution, F-measure is defined as:

$$F(\mathcal{C}, \widetilde{\mathcal{C}}) = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{h} |\widetilde{C}_i| \max_{j \in [1..k]} F_{ij}$$

where $F_{ij} = (2\ P_{ij}\ R_{ij})/(P_{ij} + R_{ij})$ such that $P_{ij} = |C_j \cap \widetilde{C}_i|/|C_j|$ and $R_{ij} = |C_j \cap \widetilde{C}_i|/|\widetilde{C}_i|$, for each $j \in [1..k]$ and $i \in [1..h]$. F-measure ranges within $[0, 1]$, where higher values correspond to better quality results.

B. Results

1) Accuracy: Table III shows accuracy results on benchmark datasets for Uniform (U), Normal (N), and Binomial (B) distributions. The last rows of this table also report, for each method, (i) the score for each type of pdf averaged over all datasets (for short, *average score*), (ii) the score averaged over all datasets and pdfs (for short, *overall average score*), and (iii) the *overall average gain* of MMVar computed as the difference between the overall average score of MMVar and the overall average scores of the other algorithms.

Looking at the overall average scores and overall average gains, it can be noted that MMVar performed better than any other competing method. In particular, the maximum gain obtained by MMVar was equal to 0.345 as compared to $\mathcal{F}$DB; this algorithm revealed to be the least accurate method, probably due to the difficulty in setting the parameters ϵ and μ. Among the other competitors, UKmed and UAHC achieved the best results, with gaps from MMVar equal to 0.093 and 0.121, respectively. Moreover, UKM and CKM were comparable to each other, and in general were less accurate than the best competing methods (i.e., UKmed and UAHC).

Considering the average scores on the various distributions, accuracy of MMVar remained on average higher than those of all competitors. The maximum average gain over all competing algorithms corresponded to Binomial pdf (0.254), whereas the minimum average gain to Normal pdf (0.134).

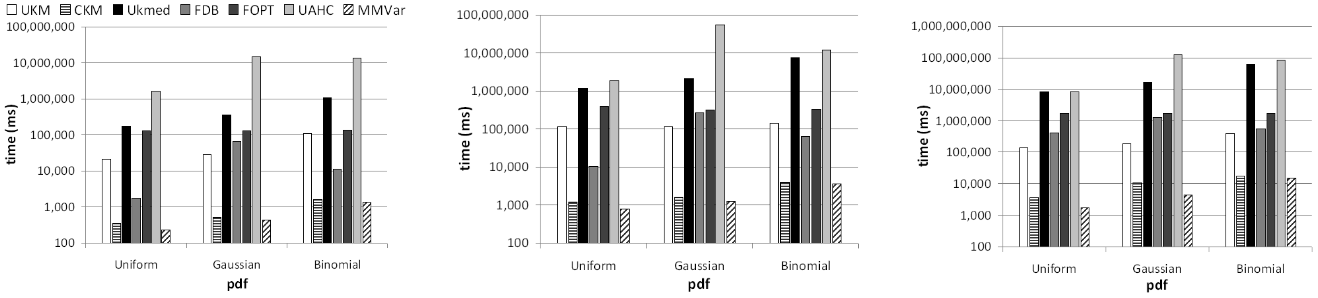
The results obtained on the single dataset-by-pdf configurations further confirmed the high accuracy of clustering solutions obtained by MMVar in relation to the other algorithms. In fact, MMVar achieved the best absolute results on 13 out of 24 dataset-by-pdf configurations; on nine more configurations (i.e., all the remaining ones except Wine-Binomial and Image-Uniform), MMVar remained comparable to the relative best method, with gaps below 0.15. Finally, we point out that MMVar was in general much more accurate than the method having the lowest computational complexity among the competitors, which is CKM (cf. Table I); more precisely, MMVar outperformed CKM on 19 out of 24 dataset-by-pdf configurations, with maximum gain of 0.483 obtained on Abalone-Binomial.

2) Efficiency: We evaluated time performance of our MMVar and the other algorithms on the selected datasets.[1] We only present results for the three largest datasets, namely Image, Abalone, and Letter; aside from the space limits of this paper, this choice is motivated by the fact that the performance trends observed on the remaining datasets were roughly similar to those of the datasets we report here.

Figure 1 shows total (i.e., offline plus online) execution times; for this analysis, we considered the computationally most expensive version of our MMVar, which corresponds to setting the moments of the distributions as approximated according to a set of statistical samples (cf. Section IV).

MMVar always performed faster than all the competing algorithms. Particularly, with the only exception of CKM, all the other algorithms obtained execution times at least one order of magnitude higher than MMVar. UAHC and UKmed were mostly the slowest method (3-5 and 3-4 orders of magnitude slower than MMVar, respectively), which should be ascribed to the intrinsic complexity of hierarchical approaches (in the case of UAHC) and to the slow offline computation of expected distances between every pair of uncertain objects (in the case of UKmed). Apart CKM, the best average performance among the competitors was obtained by $\mathcal{F}$DB, which was 1-2 orders slower than MMVar. $\mathcal{F}$DB performed as good as or better than the other density-based algorithm $\mathcal{F}$OPT (which was, in turn, 2-3 orders slower than MMVar), although the computational complexity of $\mathcal{F}$DB (in the worst case) is greater than that of

[1] Experiments were conducted on a quad-core platform Intel Pentium IV 3GHz with 4GB memory and running Microsoft WinXP Pro.



(a) Image (b) Abalone (c) Letter

Figure 1. Execution times (milliseconds)

$\mathcal{F}$OPT (cf. Table I); this was probably due to the procedure employed by $\mathcal{F}$DB for pruning unnecessary distance calculations which behaved pretty well on the selected datasets. The partitional algorithm UKM was always 2 orders slower than our MMVar, comparable to or slightly faster than $\mathcal{F}$OPT, and comparable to, one order slower than, or even slightly faster than $\mathcal{F}$DB.

VI. Conclusion

We presented a novel formulation to the problem of clustering uncertain objects, which essentially consists in minimizing the variance of the mixture models that represent the clusters to be discovered. The rationale of the proposed criterion is twofold: on the one hand, it allows for effectively recognizing clusters of uncertain objects, as the variance of the mixture model of any set of uncertain objects is inversely proportional to the compactness of that set; on the other hand, computing the variance of mixture models can be carried out in a very efficient way by exploiting some analytical properties. This led us to the development of a fast algorithm, called MMVar, to compute local optima of the objective function at the basis of the proposed formulation, which does not require any distance measure between uncertain objects. Based on experiments conducted on benchmark datasets, MMVar revealed to be faster than prominent state-of-the-art algorithms for clustering uncertain objects, while achieving better average accuracy.

References

[1] V. Cantoni, L. Lombardi, and P. Lombardi, "Challenges for Data Mining in Distributed Sensor Networks," in *Proc. ICPR Conf.*, 2006, pp. 1000–1007.

[2] Y. Li, J. Han, and J. Yang, "Clustering Moving Objects," in *Proc. KDD Conf.*, 2004, pp. 617–622.

[3] M. Milo, A. Fazeli, M. Niranjan, and N. D. Lawrence, "A probabilistic model for the extraction of expression levels from oligonucleotide arrays," *Biochemical Society Transactions*, vol. 31, pp. 1510–1512, 2003.

[4] C. C. Aggarwal and P. S. Yu, "A Survey of Uncertain Data Algorithms and Applications," *TKDE*, vol. 21, no. 5, pp. 609–623, 2009.

[5] T. Imielinski and W. Lipski Jr., "Incomplete Information in Relational Databases," *Journal of the ACM*, vol. 31, no. 4, pp. 761–791, 1984.

[6] S. Abiteboul, P. Kanellakis, and G. Grahne, "On the Representation and Querying of Sets of Possible Worlds," in *Proc. SIGMOD Conf.*, 1987, pp. 34–48.

[7] F. Sadri, "Modeling Uncertainty in Databases," in *Proc. ICDE Conf.*, 1991, pp. 122–131.

[8] Y. Tao, X. Xiao, and R. Cheng, "Range Search on Multidimensional Uncertain Data," *TODS*, vol. 32, no. 3, pp. 15–62, 2007.

[9] S. K. Lee, "An Extended Relational Database Model for Uncertain and Imprecise Information," in *Proc. VLDB Conf.*, 1992, pp. 211–220.

[10] A. D. Sarma, O. Benjelloun, A. Halevy, and J. Widom, "Working Models for Uncertain Data," in *Proc. ICDE Conf.*, 2006, pp. 7–18.

[11] H. P. Kriegel and M. Pfeifle, "Density-Based Clustering of Uncertain Data," in *Proc. KDD Conf.*, 2005, pp. 672–677.

[12] M. Chau, R. Cheng, B. Kao, and J. Ng, "Uncertain Data Mining: An Example in Clustering Location Data," in *Proc. PAKDD Conf.*, 2006, pp. 199–204.

[13] F. Gullo, G. Ponti, and A. Tagarelli, "Clustering Uncertain Data via K-medoids," in *Proc. SUM Conf.*, 2008, pp. 229–242.

[14] F. Gullo, G. Ponti, A. Tagarelli, and S. Greco, "A Hierarchical Algorithm for Clustering Uncertain Data via an Information-Theoretic Approach," in *Proc. ICDM Conf.*, 2008, pp. 821–826.

[15] H. P. Kriegel and M. Pfeifle, "Hierarchical Density-Based Clustering of Uncertain Data," in *Proc. ICDM Conf.*, 2005, pp. 689–692.

[16] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, and K. Y. Yip, "Efficient Clustering of Uncertain Data," in *Proc. ICDM Conf.*, 2006, pp. 436–445.

[17] S. D. Lee, B. Kao, and R. Cheng, "Reducing UK-means to K-means," in *Proc. ICDM Workshops*, 2007, pp. 483–488.

[18] B. Kao, S. D. Lee, D. W. Cheung, W.-S. Ho, and K. F. Chan, "Clustering Uncertain Data using Voronoi Diagrams," in *Proc. ICDM Conf.*, 2008, pp. 333–342.

[19] A. Asuncion and D. Newman, "UCI Machine Learning Repository," http://archive.ics.uci.edu/ml/.