MMVar: Clustering Uncertain Objects via Minimization of the Variance of Cluster Mixture Models

(Discussion Paper)*

Giovanni Ponti[1] and Andrea Tagarelli[2]

[1] ENEA - UTICT-HPC – Portici Research Center, Italy
[2] DIMES, University of Calabria, Italy

Abstract. A major issue in clustering uncertain objects is related to the poor efficiency of existing algorithms, which is mainly due to expensive computation of the distance between uncertain objects. This paper discusses how we addressed this issue through an original formulation of the problem of clustering uncertain objects based on the minimization of the variance of the mixture models that represent the clusters to be discovered. The proposed partitional clustering method, named MMVar, features high efficiency since it does not need to employ any distance measure for uncertain objects. Experiments have shown that MMVar turned out to be faster than prominent state-of-the-art algorithms for clustering uncertain objects, while achieving better average accuracy in terms of both external and internal cluster validity criteria.

1 Introduction

Uncertainty in data naturally arises from a variety of real-world phenomena, such as implicit randomness in a process of data generation/acquisition, imprecision in physical measurements, and data staling [1]. In general, uncertainty can be considered at table, tuple or attribute level [14], and is formally specified in terms of fuzzy models, evidence-oriented models, or probabilistic models. This work focuses on data objects containing attribute-level uncertainty, henceforth simply referred to as *uncertain objects*. Uncertainty is modeled by probability distributions that describe the likelihood that any given object appears at each position in a multidimensional domain region [4, 5, 9–11].

In data management and mining related fields, clustering uncertain objects has attracted increasing interest in recent years. One of the earliest attempts is the partitional algorithm *UK-means* [4], which is an adaptation of the popular K-means to the context of uncertain objects. In [11], the *CK-means* algorithm is proposed as a variant of UK-means that exploits the moment of inertia of rigid bodies in order to reduce the execution time for computing object-to-centroid distances. UK-means and CK-means suffer from an accuracy issue. Indeed, cluster centroids are computed as deterministic objects using the expected values of the pdfs of the clustered objects. In [5], the *UK-medoids* algorithm is proposed to overcome the above issue. Density-based approaches to clustering uncertain

* A full version of this paper appeared in [6].

objects have been defined in [9, 10]. In [10], the $\mathcal{F}DBSCAN$ is proposed as a fuzzy version of the popular DBSCAN, mainly based on the use of fuzzy distance functions to compute the *core object* and *reachability* probabilities. A similar approach is presented in $\mathcal{F}OPTICS$ [9]. Like the well-known density-based clustering algorithm OPTICS, $\mathcal{F}$OPTICS produces an augmented ordering of the objects based on the novel notion of fuzzy object reachability-distance. The augmented ordering outputted by $\mathcal{F}$OPTICS makes that method falling into the category of hierarchical clustering algorithms as well. Another hierarchical method, i.e., *UAHC*, is proposed in [7]. UAHC exploits a cluster merging criterion based on an information-theoretic measure to compute the distance between cluster prototypes.

This paper discusses one of our latest studies on the problem of uncertain clustering that aimed to overcome the aforementioned issues [6]. Following a partitional (uncertain) clustering paradigm, the key idea underlying our study is to take into account the mixture model of the set of random variables representing the uncertain objects within a cluster in order to define a clustering objective criterion based on the minimization of the *variance* of the cluster mixture models. Such a criterion is designed to fulfill efficiency and accuracy requirements. At the same time, the proposed criterion enables the definition of a fast heuristic that does not require any distance measure between uncertain objects.

In the rest of the paper, we will present MMVar and main theoretical aspects underlying the proposed method. Finally, we will summarize our experimental findings, which revealed the better efficiency and (on average) accuracy of MMVar w.r.t. prominent state-of-the-art methods for uncertain clustering.

2 Modeling uncertainty

Uncertain objects are typically represented according to *multivariate* or *univariate* uncertainty models [5]. Although any uncertain object can be represented according to either model depending on the specific application context, the univariate model can be considered as a particular case of the multivariate one. For this reason, in the rest of the paper we will focus on the most general case of multivariate uncertainty modeling, whose formal definition is provided next.

Definition 1 (multivariate uncertain object). *A* multivariate uncertain object o *is a pair* $(\mathcal{R}, f)$*, where* $\mathcal{R} \subseteq \Re^m$ *is the* m*-dimensional region in which* o *is defined and* $f : \Re^m \rightarrow \Re_0^+$ *is the probability density function of* o *at each point* $\boldsymbol{x} \in \Re^m$*, such that* $f(\boldsymbol{x}) = 0,\ \forall \boldsymbol{x} \in \Re^m \setminus \mathcal{R}$*, and* $f(\boldsymbol{x}) > 0,\ \forall \boldsymbol{x} \in \mathcal{R}$*.*

The above definition refers to the most general case of uncertain objects modeled as continuous random variables and described by pdfs. Nevertheless, this definition also includes the case where uncertain objects are represented by discrete 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 to the continuous uncertainty model, as the corresponding discrete version can be trivially obtained by replacing integrals with sums. The expected value ($\boldsymbol{\mu}$), second order moment ($\boldsymbol{\mu}_2$), and variance ($\boldsymbol{\sigma}^2$) vectors of any given uncertain object $o = (\mathcal{R}, f)$ are defined as follows:

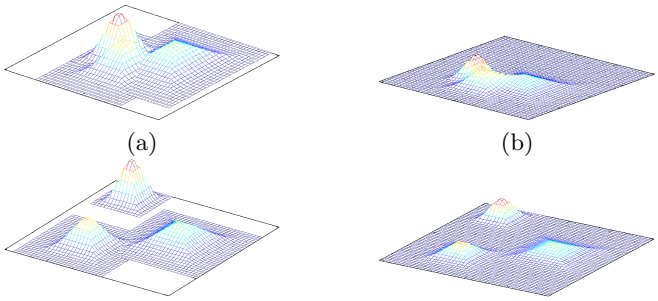

Fig. 1. Uncertain prototype variance and cluster compactness: (a) compact set of objects and (b) its low-variance uncertain prototype; (c) less compact set of objects and (d) its higher-variance uncertain prototype.

$$\boldsymbol{\mu}(o) = \int_{\boldsymbol{x}\in\mathcal{R}} \boldsymbol{x}\ f(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} \qquad \boldsymbol{\mu}_2(o) = \int_{\boldsymbol{x}\in\mathcal{R}} \boldsymbol{x}^2 f(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x}$$
$$\boldsymbol{\sigma}^2(o) = \int_{\boldsymbol{x}\in\mathcal{R}} (\boldsymbol{x}-\boldsymbol{\mu}(o))^2 f(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} \;=\; \boldsymbol{\mu}_2(o) - \boldsymbol{\mu}^2(o)$$

The j-th component ($j \in [1..m]$) of the $\boldsymbol{\mu}$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\sigma}^2$ vectors is defined as:

$$\mu_j(o) = \int_{\boldsymbol{x}\in\mathcal{R}} x_j\ f(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} \qquad (\mu_2)_j(o) = \int_{\boldsymbol{x}\in\mathcal{R}} x_j^2 f(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x}$$
$$(\sigma^2)_j(o) = \int_{\boldsymbol{x}\in\mathcal{R}} (x_j-\mu_j(o))^2 f(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} = (\mu_2)_j(o) - \mu_j^2(o)$$

Moreover, given any vector $\boldsymbol{\sigma}^2$ of variances, the "global" variance is defined as the sum of variances along each dimension:

$$\sigma^2(o) = \|\boldsymbol{\sigma}^2(o)\|_1 = \sum_{j=1}^{m} (\sigma^2)_j = \int_{\boldsymbol{x}\in\mathcal{R}} \|\boldsymbol{x}-\boldsymbol{\mu}(o)\|^2 f(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x}$$

3 The MMVar Algorithm

A key notion in the proposed formulation to the problem of clustering uncertain objects is that of *uncertain prototype* (or simply *prototype*) of a set of uncertain objects. Any uncertain prototype is defined as the *mixture model* of the random variables representing the objects in a given set.

Definition 2. *The* uncertain prototype $\mathcal{P}_C$ *of any given set* C *of uncertain objects is a pair* $(\mathcal{R}_C, f_C)$*, where:*

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

Proposition 1. *The uncertain prototype* $\mathcal{P}_C = (\mathcal{R}_C, f_C)$ *of any set* C *of uncertain objects is a multivariate uncertain object satisfying Def. 1.*

Defining uncertain prototypes as mixture models has a number of key advantages. First, it allows for maintaining information about the uncertainty of the objects to be summarized, which makes the probabilistic representation particularly accurate. This contrasts to traditional definitions of prototypes (or centroids) of uncertain objects, which collapse all the information about uncertainty into a single, representative numerical value, like that exploited by UK-means and CK-means algorithms. Second, computing the mixture model of a set of random variables is fast as it can be performed in linear time w.r.t. the size of that set. Third, focusing on statistical properties of the uncertain prototypes defined as mixture models enables the definition of an effective criterion to properly formulate clustering of uncertain objects. In this respect, an interesting remark about the variance of mixture models can be drawn based on the following intuition: the lower the variance of the mixture model (i.e., uncertain prototype) of a set of uncertain objects, the higher the compactness of that set, i.e., the higher the probability that the set is a high-quality cluster, as graphically depicted in Fig. 1. Moreover, it should be noted that the variance of the mixture model of a set of uncertain objects is strictly related to the uncertainty of those objects; thus, it allows for taking into account uncertainty carefully, increasing the probability of achieving higher accuracy in a clustering process.

Based on the above considerations, we formulated 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 of $\mathcal{D}$ that minimizes the following objective function:

$$J(\mathcal{C}) = \sum_{C \in \mathcal{C}} \sigma^2(\mathcal{P}_C) \tag{1}$$

where $\sigma^2(\mathcal{P}_C)$ is the variance of the prototype $\mathcal{P}_C$ of cluster C.

We show next some analytical properties about the computation of the variance of mixture models, which are at the basis of the MMVar algorithm.

Lemma 1. *Let C be a set of uncertain objects, where each $o \in C$ is a pair $(\mathcal{R}, f)$, and $\mathcal{P}_C$ be the uncertain prototype of C. The expected value $\boldsymbol{\mu}(\mathcal{P}_C)$ and the second order moment $\boldsymbol{\mu}_2(\mathcal{P}_C)$ of prototype $\mathcal{P}_C$ are as follows:*

$$\boldsymbol{\mu}(\mathcal{P}_C) = \frac{1}{|C|} \sum_{o \in C} \boldsymbol{\mu}(o) \qquad \boldsymbol{\mu}_2(\mathcal{P}_C) = \frac{1}{|C|} \sum_{o \in C} \boldsymbol{\mu}_2(o)$$

Lemma 2. *Let $\mathcal{P}_C$ be the uncertain prototype of any set C of uncertain objects, C' (resp. C'') be the set defined by deleting (resp. adding) object o' (resp. o'') from (resp. to) C, i.e., $C' = C \backslash \{o'\}$, $C'' = C \cup \{o''\}$. The expected values $\boldsymbol{\mu}(\mathcal{P}_{C'})$, $\boldsymbol{\mu}(\mathcal{P}_{C''})$ and second order moments $\boldsymbol{\mu}_2(\mathcal{P}_{C'})$, $\boldsymbol{\mu}_2(\mathcal{P}_{C''})$ of prototypes $\mathcal{P}_{C'}$, $\mathcal{P}_{C''}$ of sets C', C'' are as follows:*

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

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

Proposition 2. *Let $\mathcal{D}$ be a set of m-dimensional uncertain objects, $\mathcal{C}$ be a partition of $\mathcal{D}$, $\mathcal{P}_C$ be the prototype of any cluster $C \in \mathcal{C}$, and $\boldsymbol{\mu}(\mathcal{P}_C)$, $\boldsymbol{\mu}_2(\mathcal{P}_C)$ and*

Algorithm 1 MMVar

Input: A set $\mathcal{D}$ of uncertain objects; $k = \sup(\mathbb{N})$, or $k \leq |\mathcal{D}|$ as a user-specified parameter (i.e., number of output clusters).
Output: A partition $\mathcal{C}$ of $\mathcal{D}$.

```
compute μ(o), μ2(o), ∀o ∈ D
select o* ∈ D, randomly or by a predefined criterion
C ← {{o*}}
repeat
  for all o ∈ D do
    if o is not assigned to any cluster yet then
      if |C| < k then
        C ← {o}, C ← C ∪ {C}
        compute μ(P_C), μ2(P_C)                                  {Lemma 1}
      else
        select C ∈ C, randomly or by a predefined criterion
        C ← C ∪ {o}
        recompute μ(P_C), μ2(P_C)                                {Lemma 2}
      end if
      V ← J(C)                                                   {(1)}
    end if
    C* ← arg min_Ĉ J_{C,Ĉ}(C),  with o ∈ C                       {(2)}
    if C* ≠ C ∧ (|C| > 1 ∨ k = sup(N))  then
      V ← J_{C,C*}(C)                                            {(2)}
      recompute C: move o from C to C*, and remove C from C in case C = ∅
      recompute μ(P_C), μ2(P_C), μ(P_C*), μ2(P_C*)               {Lemma 2}
    end if
  end for
until no o ∈ D is relocated
```

$\sigma^2(\mathcal{P}_C) = \|\boldsymbol{\mu}_2(\mathcal{P}_C) - \boldsymbol{\mu}^2(\mathcal{P}_C)\|_1$ *the expected value, second order moment and 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* o *from cluster* $C \in \mathcal{C}$ *to cluster* $\widehat{C} \in \mathcal{C}$*; it holds that the value* $J_{C,\widehat{C}}(\mathcal{C})$ *of the objective function* J *for the new partition* $\mathcal{C}'$ *is:*

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

where

$$C' = C \setminus \{o\},\ \widehat{C}' = \widehat{C} \cup \{o\},\ \sigma^2(\mathcal{P}_{C'}) = \|\boldsymbol{\mu}_2(\mathcal{P}_{C'}) - \boldsymbol{\mu}^2(\mathcal{P}_{C'})\|_1,\ \sigma^2(\mathcal{P}_{\widehat{C}'}) = \|\boldsymbol{\mu}_2(\mathcal{P}_{\widehat{C}'}) - \boldsymbol{\mu}^2(\mathcal{P}_{\widehat{C}'})\|_1$$

and

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

$$\boldsymbol{\mu}(\mathcal{P}_{\widehat{C}'}) = \frac{|\widehat{C}| \times \boldsymbol{\mu}(\mathcal{P}_{\widehat{C}}) + \boldsymbol{\mu}(o)}{|\widehat{C}| + 1} \qquad \boldsymbol{\mu}_2(\mathcal{P}_{\widehat{C}'}) = \frac{|\widehat{C}| \times \boldsymbol{\mu}_2(\mathcal{P}_{\widehat{C}}) + \boldsymbol{\mu}_2(o)}{|\widehat{C}| + 1}$$

Proposition 2 hence states that, given any partition $\mathcal{C}$ of $\mathcal{D}$ and any other partition $\mathcal{C}'$ obtained from $\mathcal{C}$ by moving a single 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 (2). 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. MMVar can either automatically discover the number of clusters or, alternatively, can be constrained to form a user-specified number of clusters. In the former case, k should be set to $\sup(\mathbb{N})$. To exploit the result of Proposition 2, MMVar only needs to maintain expected

values and second order moments of the objects within $\mathcal{D}$ (i.e., vectors $\boldsymbol{\mu}(o)$ and $\boldsymbol{\mu}_2(o)$) and of the prototypes that are identified at each iteration (i.e., vectors $\boldsymbol{\mu}(\mathcal{P}_C)$ and $\boldsymbol{\mu}_2(\mathcal{P}_C)$). Expected values and second order moments are computed according to either the exact (or approximated) formulas (Line 1). The main cycle of the algorithm (Lines 4–24) is in charge of iteratively process all objects within $\mathcal{D}$, until no decrement in the objective function has been observed. We can distinguish two phases: the initialization phase and the refinement of the clustering produced at the previous iteration. The initialization phase consists in an incremental scheme where the initial cluster of any single object is chosen in a greedy fashion according to the assignments already performed for the objects that have been processed earlier. The refinement step is carried out, where objects are relocated until cluster stability (Line 24) is reached. Note that once all objects have been clustered, the only steps of the main cycle actually performed are the ones in Lines 17–22, while Lines 6–16 are entirely skipped.

The proposed MMVar has been proved to converge to a local optimum of the objective function therein involved, and to work linearly w.r.t. both the size of the input dataset and the dimensionality of the input uncertain objects [6].

4 Experimental Evaluation

MMVar was compared to partitional algorithms, i.e., UK-means, CK-means, and UK-medoids, density-based algorithms, i.e., $\mathcal{F}$DBSCAN and $\mathcal{F}$OPTICS, and the hierarchical algorithm UAHC. Experiments were carried out on a CentOS 5.5 platform, with Linux 2.6.18 kernel, 64GB memory, 4 Intel(R) Xeon(R) CPU E7330, 2.40GHz quadcore, which hosted by the CRESCO HPC system [2], integrated into the ENEA-GRID infrastructure and located in the ENEA Portici Research Center.[3]

We selected eight benchmark datasets from the UCI repository,[4] and also we used microarray data collections from [3]. To assess the quality of clustering solutions, we employed the well-known *F-measure* as external validity criterion, as well as the *Silhouette index* [13] and the intra/inter-cluster similarity as internal validity criteria. According to an approach already employed by previous works [4], we synthetically generated uncertainty in benchmark datasets, as they originally contained deterministic values, by considering *Uniform*, *Normal*, and *Binomial* distributions; conversely, this was not necessary for real microarray datasets since they inherently have *probe-level* uncertainty, which can easily be modeled in the form of Normal pdfs according to the *multi-mgMOS* method [12]. We focused on two evaluation cases: (i) the clustering task was performed by considering only the observed (i.e., non-uncertain) representations of the various data objects, (ii) the clustering task was performed by involving an uncertainty model. The ultimate goal was to assess whether results obtained in Case 2 were better than those obtained in Case 1. In Case 1, we generated a *perturbed dataset* $\mathcal{D}'$ from $\mathcal{D}$ by adding to each point $\boldsymbol{w} \in \mathcal{D}$ random noise sampled from its assigned pdf $f_{\boldsymbol{w}}$ according to the classic Monte Carlo and Markov Chain Monte

[3] http://www.cresco.enea.it
[4] http://archive.ics.uci.edu/ml/

Carlo methods. As a result, $\mathcal{D}'$ still contained deterministic data, which can be interpreted as observed representations of the input uncertain objects. In our evaluation, each of the selected clustering methods was carried out on $\mathcal{D}'$ so that it produced an output clustering solution denoted by $\mathcal{C}'$. An F-measure score $F(\mathcal{C}', \widetilde{\mathcal{C}})$ was hence obtained by comparing the output clustering $\mathcal{C}'$ to the reference classification of $\mathcal{D}$ (denoted by $\widetilde{\mathcal{C}}$). In Case 2, when uncertainty was taken into account, we further created an *uncertain dataset* $\mathcal{D}''$ from $\mathcal{D}$ which is the one designed to contain uncertain objects. In particular, for each $\boldsymbol{w} \in \mathcal{D}$, we derived an uncertain object $o = (\mathcal{R}, f)$ in such a way that $f = f_{\boldsymbol{w}}$, while $\mathcal{R}$ was defined as the region containing most of the area (e.g., 95%) of $f_{\boldsymbol{w}}$. Again, we ran each of the selected clustering methods on $\mathcal{D}''$ as well, in order to obtain a clustering solution $\mathcal{C}''$ and a score $F(\mathcal{C}'', \widetilde{\mathcal{C}})$. Finally, we compared the scores obtained in Case 1 and Case 2 by computing $\Theta(\mathcal{C}', \mathcal{C}'', \widetilde{\mathcal{C}}) = F(\mathcal{C}'', \widetilde{\mathcal{C}}) - F(\mathcal{C}', \widetilde{\mathcal{C}})$; the higher Θ, the better the quality of $\mathcal{C}''$ w.r.t. $\mathcal{C}'$, and, therefore, the better the performance of the clustering method when the uncertainty is taken into account w.r.t. the case where no uncertainty is employed.

Summary of Clustering Results. For the sake of brevity, we provide here a qualitative summary of the results we obtained in our experiments; the interested reader is referred to [6] for a detailed description. MMVar was always faster than all competing algorithms which, with the exception of CKMeans, were found to be up to 5 orders of magnitude slower than MMVar. CKmeans was the fastest competing method which was of the same order of magnitude of MMVar, although not so fast as MMVar.

Nevertheless, CKmeans was also much less accurate than MMVar. Indeed, on benchmark datasets, the overall average scores and gains in terms of Θ showed that the proposed MMVar was more accurate than any other competing method. Among the competitors, UKmedoid achieved the best results, while UAHC overall performance was closer to that of UKMeans and CKMeans than to the performance of density-based algorithms. UKMeans and CKMeans were comparable to each other and performed better than UAHC and the density-based methods. Moreover, MMVar was in general much more accurate than the method having the lowest computational complexity among the competitors, which is CKMeans. In terms of intra/inter-cluster similarity, MMVar was the best performing method in terms of average scores on the microarray datasets, and also achieved the best overall average performance, particularly outperforming CKMeans. In general, UAHC (resp. $\mathcal{F}$DBSCAN) was found as the best (resp. worst) method, whereas the performance of $\mathcal{F}$OPTICS and UKmedoids significantly decreased w.r.t. the benchmark datasets. As far as the Silhouette cluster validity criterion, the average accuracy gains achieved by MMVar w.r.t. UKMeans, CKMeans, UKmedoids, and UAHC were even larger than those observed in terms of the intra/inter-cluster similarity scores.

5 Conclusion

We addressed the problem of clustering uncertain objects by focusing on the minimization of the variance of the mixture models that represent the clusters

to be discovered. The MMVar approach has the merit of taking into account the variance of the uncertain objects to model a cluster prototype, which brings the important benefit of avoiding the computation of uncertain object distances, and hence represents a notable advantage of higher accuracy with respect to existing distance-dependent centroid-based clustering methods for uncertain objects. We believe however that a purely variance-based notion of uncertain cluster prototype can still be inadequate to correctly summarize all cluster structures, and certainly improved by developing uncertain cluster prototypes in which central tendency and variability of the clustered objects are both taken into account in the definition of the cluster mixture models, while still avoiding to compute costly uncertain object-to-prototype distances [8].

References

1. C. C. Aggarwal. *Managing and Mining Uncertain Data*. Springer, 2009.
2. G. Bracco, S. Podda, S. Migliori, P. D'Angelo, A. Quintiliani, D. Giammattei, M. De Rosa, S. Pierattini, G. Furini, R. Guadagni, F. Simoni, A. Perozziello, A. De Gaetano, S. Pecoraro, A. Santoro, C. Scio', A. Rocchi, A. Funel, S. Raia, G. Aprea, U. Ferrara, D. Novi, and G. Guarnieri. CRESCO HPC System Integrated into ENEA-GRID Environment. In *Proc. of the Final Workshop of the Grid Projects of the Italian National Operational Program 2000-2006 – Call 1575*, pages 151–155, 2009.
3. Broad Institute of MIT and Harvard. Cancer program dataset page, http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.
4. M. Chau, R. Cheng, B. Kao, and J. Ng. Uncertain data mining: An example in clustering location data. In *Proc. PAKDD Conf.*, pages 199–204, 2006.
5. F. Gullo, G. Ponti, and A. Tagarelli. Clustering uncertain data via k-medoids. In *Proc. SUM Conf.*, pages 229–242, 2008.
6. F. Gullo, G. Ponti, and A. Tagarelli. Minimizing the variance of cluster mixture models for clustering uncertain objects. *Statistical Analysis and Data Mining*, 6(2):116–135, 2013.
7. F. Gullo, G. Ponti, A. Tagarelli, and S. Greco. A hierarchical algorithm for clustering uncertain data via an information-theoretic approach. In *Proc. IEEE ICDM Conf.*, pages 821–826, 2008.
8. F. Gullo, and A. Tagarelli. Uncertain Centroid based Partitional Clustering of Uncertain Data. *PVLDB* 5(7):610–621, 2012.
9. H. P. Kriegel and M. Pfeifle. Hierarchical density-based clustering of uncertain data. In *Proc. IEEE ICDM Conf.*, pages 689–692, 2005.
10. H. P. Kriegel and M. Pfeifle. Density-based clustering of uncertain data. In *Proc. ACM KDD Conf.*, pages 672–677, 2005.
11. S. D. Lee, B. Kao, and R. Cheng. Reducing uk-means to k-means. In *Proc. IEEE ICDM Workshops*, pages 483–488, 2007.
12. X. Liu, M. Milo, N. D. Lawrence, and M. Rattray. A tractable probabilistic model for affymetrix probe-level analysis across multiple chips. *Bioinformatics*, 21(18):3637–3644, 2005.
13. P. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Computational and Applied Mathematics*, 20(1):53–65, 1987.
14. Y. Tao, X. Xiao, and R. Cheng. Range search on multidimensional uncertain data. *ACM Transactions on Database Systems*, 32(3):15–62, 2007.