Hierarchical Clustering of Microarray Data with Probe-level Uncertainty

F. Gullo, G. Ponti, A. Tagarelli
DEIS Department
University of Calabria
Via P. Bucci, 41c
87036 Rende (CS) — Italy
{fgullo,gponti,tagarelli}@deis.unical.it

G. Tradigo, P. Veltri
Department of Experimental and Clinical Medicine
"Magna Græcia" University of Catanzaro
Viale Europa
88100 Località Germaneto (CZ) — Italy
{gtradigo,veltri}@unicz.it

Abstract

Handling microarray data is particularly challenging mainly due to the high dimensionality of such data, which demands for computer-aided methods, and to the intrinsic difficulty of devising notions of proximity between spots of array traps.

In this paper, we propose a new approach to modeling the probe-level uncertainty in microarray data that allows for a more expressive representation of the data and a more accurate processing. This approach is essentially based on a recently proposed method for uncertain data clustering. This method lies in a centroid-linkage-based agglomerative hierarchical algorithm, named U-AHC, and an information-theoretic-based distance measure between uncertain data [8]. We have conducted experiments on four large microarray datasets, in order to assess effectiveness of the proposed clustering method. Experimental results have shown high quality results in terms of compactness of the clustering solutions.

1 Background

A major goal in genomics is to discover gene relationships and their role in diseases (*functional genomics*). DNA microarray is a technology used in molecular biology and medicine which is able to trap and measure the relative quantity of a large number of genes with a single experiment. Probes that are able to trap genes (*targets*) consist of thousands of microscopic spots organized as a matrix and placed on a glass or silicon chip. The probes-target hybridization is quantified through techniques based on fluorescence. When an experiment is performed, the spots of the microarray matrix generate intensity values, which measure the *expression levels* of the genes. Recently, the Affymetrix company introduced an advanced chip version, called Human Gene 1.0 ST chip, by which the analysts can explore the whole mRNA transcript in a single experiment [18].

Microarray data analysis is challenging. In particular, there are two major issues to face. A first issue is related to the high dimensionality of microarray data, which makes it necessary to resort to computer-based methods. Another issue is that the spots of the array trap information generally do not have a straightforward meaning; rather, the spots have to be compared and analyzed by possibly using statistical techniques.

Many approaches to microarray data analysis have been proposed by the research community [6]. Most of them are essentially based on data mining techniques, in particular *clustering* methods [7, 9]. Clustering allows for understanding the huge mass of data in microarrays by grouping them in homogeneous subsets (clusters). In this way, cluster analysis aims to discover natural structures within the data and to help the analyst in identifying common structures and patterns in microarrays; for instance, finding similar expression patterns (i.e., co-expressed genes) which are related to cellular functions.

Microarray clustering approaches can be divided into three main categories [10]: (i) *gene-based clustering*, which treats genes as objects and samples as clustering features; (ii) *sample-based clustering*, where samples are the objects to be clustered and genes are the features; (iii) *co-clustering* approaches, where genes and samples are treated symmetrically (samples and genes can be both objects and features).

After several years of quantitative measurements of microarray probe-level data, new models have been proposed in order to manage the uncertainty of gene expression levels both in a single chip and across multiple chips. A novel probabilistic modeling approach is presented in [15], where the binding affinity of probe-pairs across multiple chips is modeled through a probabilistic model using Gamma distributions. In [14], a gene expression clustering algorithm has been proposed, which exploits the probabilistic modeling described above in order to improve performances w.r.t.

classic techniques.

In this paper, we propose a new approach to modeling probe-level uncertainty in microarray data. This approach is based on a study originally presented in [8], in which probabilistic models are employed to allow for a more expressive representation of the data and a more accurate processing. More precisely, uncertain data objects are modeled as probability distributions (pdfs). An information-theoretic-based distance measure is used to compare uncertain data objects, and the clustering task is performed by means of a centroid-linkage-based agglomerative hierarchical algorithm, named *U-AHC*. In U-AHC, the cluster merging step is accomplished by a centroid-linkage criterion [17] which has the following main features: (i) cluster prototypes (i.e., cluster centroids) are computed as mixture densities that summarize the pdfs of all the objects in the clusters, and (ii) the pair of closest clusters is chosen according to an information-theoretic measure that computes the distance between the cluster prototypes. The centroid-linkage-based criterion does not require a notion of distance between the objects to be clustered, unlike other traditional linkage criteria in agglomerative hierarchical clustering. This allows us to avoid defining a notion of distance between uncertain objects, which is crucial in uncertainty similarity detection; instead, the adoption of cluster prototypes as mixture densities enables a notion of information-theoretic distance measure that exploits an advantageous characteristic of the cluster prototypes: the overlaps between the cluster prototypes' domain regions are generally larger than the overlaps between the individual objects' regions.

We have tested the U-AHC algorithm on four large microarray datasets, and evaluated performance in achieving effective clustering. Experimental results have shown that U-AHC achieves high results in terms of compactness of the clustering solution.

The rest of the paper is organized as follows. Section 2 introduces the data uncertainty modeling. Section 3 describes the clustering strategy, (i) the definition of cluster prototypes as new uncertain objects that summarize the features of all the objects in each cluster, (ii) an information-theoretic-based distance measure, which is particularly suitable for uncertain objects, and (iii) the scheme of a hierarchical clustering algorithm for uncertain objects (U-AHC). Section 4 describes experimental analysis and shows the clustering results obtained by U-AHC on microarray datasets. Finally, Section 5 concludes the paper.

2 Modeling uncertainty

Uncertain data objects are traditionally represented by using either a *multivariate* probabilistic model or a *univariate* probabilistic model.

In a multivariate uncertainty model, an m-dimensional uncertain object is defined in terms of an m-dimensional region and a multivariate probability density function, which stores the probability according to which the exact representation of the object coincides with any point in the region. In a univariate uncertainty model, an m-dimensional uncertain object has, for each attribute, an interval and a univariate probability density function that assigns a probability value to any point within the interval. Formally, this is expressed by the following definitions.

Definition 1 (multivariate uncertain object) *A* multivariate uncertain object *o is a pair (R, f), where $R = [l_1, u_1] \times \cdots \times [l_m, u_m]$ 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 $\vec{x} \in R$, such that:*

$$\int\limits_{\vec{x} \in R} f(\vec{x}) \mathrm{d}\vec{x} = 1 \qquad \textit{and} \qquad \int\limits_{\vec{x} \in \Re^m \setminus R} f(\vec{x}) \mathrm{d}\vec{x} = 0$$

Definition 2 (univariate uncertain object) *A* univariate uncertain object *o is a tuple $(a^{(1)}, \ldots, a^{(m)})$. Each attribute $a^{(h)}$ is a pair $(I^{(h)}, f^{(h)})$, for each $h \in [1..m]$, where $I^{(h)} = [l^{(h)}, u^{(h)}]$ is the interval of definition of $a^{(h)}$, and $f^{(h)} : \Re \to \Re_0^+$ is the probability density function that assigns a probability value to each $x \in I^{(h)}$, such that:*

$$\int\limits_{x \in I^{(h)}} f^{(h)}(x) \mathrm{d}x = 1 \qquad \textit{and} \qquad \int\limits_{x \in \Re \setminus I^{(h)}} f^{(h)}(x) \mathrm{d}x = 0$$

3 Clustering uncertain objects

3.1 Uncertain prototype

An *uncertain prototype*, or simply *prototype*, is seen as a new uncertain object computed from a set of uncertain objects, which properly summarizes the features of all the objects in the set. More precisely, an uncertain prototype is represented by mixture densities from the pdfs associated to each object in the set to be summarized.

Definition 3 (multivariate uncertain prototype) *Let $\mathcal{C} = \{o_1, ..., o_n\}$ be a set of multivariate uncertain objects, where $o_i = (R_i, f_i)$, $R_i = [l_{i_1}, u_{i_1}] \times \ldots \times [l_{i_m}, u_{i_m}]$, for each $i \in [1..n]$. The* multivariate uncertain prototype *of $\mathcal{C}$ is a multivariate uncertain object $\mathcal{P}_\mathcal{C} = (R_\mathcal{C}, f_\mathcal{C})$, where*

$$R_\mathcal{C} = \Big[\min_{i \in [1..n]} l_{i_1}, \max_{i \in [1..n]} u_{i_1}\Big] \times \cdots \times \Big[\min_{i \in [1..n]} l_{i_m}, \max_{i \in [1..n]} u_{i_m}\Big],$$

$$f_\mathcal{C}(\vec{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\vec{x})$$

Definition 4 (univariate uncertain prototype) *Let* $\mathcal{C} = \{o_1, ..., o_n\}$ *be a set of univariate uncertain objects, where* $o_i = ((I_i^{(1)}, f_i^{(1)}), \ldots, (I_i^{(m)}, f_i^{(m)}))$, $I_i^{(h)} = [l_i^{(h)}, u_i^{(h)}]$, *for each* $h \in [1..m]$, $i \in [1..n]$. *The* univariate uncertain prototype *of* $\mathcal{C}$ *is a univariate uncertain object* $\mathcal{P}_{\mathcal{C}} = ((I_{\mathcal{C}}^{(1)}, f_{\mathcal{C}}^{(1)}), \ldots, (I_{\mathcal{C}}^{(m)}, f_{\mathcal{C}}^{(m)}))$ *such that, for each* $h \in [1..m]$*:*

$$I_{\mathcal{C}}^{(h)} = \Big[\min_{i\in[1..n]} l_i^{(h)}, \max_{i\in[1..n]} u_i^{(h)}\Big],$$

$$f_{\mathcal{C}}^{(h)}(x) = \frac{1}{n}\sum_{i=1}^{n} f_i^{(h)}(x)$$

3.2 Distance between uncertain prototypes

To define a distance measure between uncertain prototypes, we employ a function that exploits the full information stored in the pdfs. Two of the most frequently used distance measures between probability densities are the Kullback-Leibler divergence [13, 12] and the Chernoff distance [5]. These measures fall into the Ali-Silvey class of *information-theoretic* distance measures [2] and have been widely used in several application contexts, such as signal processing, pattern recognition, and speech recognition [1, 3]. However, Kullback-Leibler divergence and Chernoff distance suffer from some drawbacks—e.g., Kullback-Leibler divergence is not symmetric, whereas Chernoff distance is typically hard to compute; also, both the measures do not satisfy the triangle inequality.

Within this view, the adopted definition of distance between prototypes exploits a measure based on the *Bhattacharyya coefficient* [4, 11], which is defined as follows:

$$\rho(p(\vec{x}), q(\vec{x})) = \int_{\vec{x}\in\Re^m} \sqrt{p(\vec{x})\, q(\vec{x})}\, \mathrm{d}\vec{x} \qquad (1)$$

The original definition of ρ in [4] considers the pdfs p and q as two multinomial populations, each one consisting of k classes with associated probabilities; also, ρ has a geometric interpretation: it can be seen as the cosine between the two vectors for p and q, whose components are the square root of the probabilities of the k classes that compose p and q. This interpretation also holds in the extended definition reported in Equation (1), which aims to define the Bhattacharyya coefficient for continuous pdfs.

Based on the Bhattacharyya coefficient, several distance measures can be defined [11]. In this work, we use the following measure

$$\mathrm{B}(p(\vec{x}), q(\vec{x})) = \sqrt{1 - \rho(p(\vec{x}), q(\vec{x}))} \qquad (2)$$

which has a number of advantages w.r.t. other Bhattacharyya distances, such as the commonly used $-\log \rho$ definition. In particular, the Bhattacharyya distance in Equation (2) obeys the triangle inequality, ranges within the interval [0,1], and unlike the Chernoff distance (which is a more general case), it is easier to compute and satisfies the additive property even if the random variables are not identically distributed.

We now provide the definition of distance measure in both cases of multivariate and univariate uncertain prototype [8].

Definition 5 (multivariate uncertain prototype distance) *Given a set* $\mathcal{D}$ *of multivariate uncertain objects, let* $\mathcal{P}_{\mathcal{C}_i} = (R_{\mathcal{C}_i}, f_{\mathcal{C}_i})$ *and* $\mathcal{P}_{\mathcal{C}_j} = (R_{\mathcal{C}_j}, f_{\mathcal{C}_j})$ *be the multivariate uncertain prototypes of the sets* $\mathcal{C}_i, \mathcal{C}_j \subseteq \mathcal{D}$*, respectively. The* multivariate uncertain prototype distance *between* $\mathcal{P}_{\mathcal{C}_i}$ *and* $\mathcal{P}_{\mathcal{C}_j}$ *is defined as*

$$\Delta(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j}) = \gamma\, \Delta'(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j}) + (1-\gamma)\, \Delta''(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j}) \quad (3)$$

where

$$\Delta'(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j}) = \mathrm{B}(f_{\mathcal{C}_i}, f_{\mathcal{C}_j}),$$

$$\Delta''(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j}) = \frac{1}{E_{max}(\mathcal{D})} d(E[f_{\mathcal{C}_i}], E[f_{\mathcal{C}_j}])$$

$$\gamma = \frac{\mathcal{V}(R_{\mathcal{C}_i} \cap R_{\mathcal{C}_j})}{\min\{\mathcal{V}(R_{\mathcal{C}_i}), \mathcal{V}(R_{\mathcal{C}_j})\}}$$

In Definition 5, d is a function that measures the distance between m-dimensional points (e.g., Euclidean norm), $E[f]$ denotes the expected value of the pdf f, $\mathcal{V}(R)$ is the hypervolume of the m-dimensional region R, and E_{max} is a normalization term, which is defined as:

$$E_{max}(\mathcal{D}) = \max_{o_u, o_v \in \mathcal{D}} d(E[f_u], E[f_v])$$

It should be noted that Δ ranges within $[0, 1]$, since Δ' and Δ'' range within $[0, 1]$ in turn.

Let us now explain the reasons for introducing the two terms Δ' and Δ'' in Equation (3). The Bhattacharyya distance (Equation (2)) compares two pdfs by considering their portions defined over a common event space (i.e., common domain region). Thus, if the event spaces of the two pdfs do not have any intersection, the Bhattacharyya distance does not work, i.e., it is always equal to one. Although these cases are quite infrequent because of the way uncertain prototypes are defined, we introduce the term Δ'' in Equation (3) to discriminate among those cases by considering the distance between the expected values of the prototype pdfs. We weight the terms Δ' and Δ'' by involving the coefficient γ (ranging within $[0, 1]$), which aims to quantify the importance of Δ' and Δ'' in the definition of Δ. In particular, γ is proportional to the width of the domain region shared between the prototypes to be compared. This definition of γ represents a reasonable choice, since the larger the portion of the pdfs involved into the Bhattacharyya distance calculation, the smaller the need for comparing the pdfs by also considering the corresponding expected values, and vice versa.

Definition 6 (univariate uncertain prototype distance) *Given a set $\mathcal{D}$ of univariate uncertain objects, let $\mathcal{P}_{\mathcal{C}_i} = ((I^{(1)}_{\mathcal{C}_i}, f^{(1)}_{\mathcal{C}_i}), \ldots, (I^{(m)}_{\mathcal{C}_i}, f^{(m)}_{\mathcal{C}_i}))$ and $\mathcal{P}_{\mathcal{C}_j} = ((I^{(1)}_{\mathcal{C}_j}, f^{(1)}_{\mathcal{C}_j}), \ldots, (I^{(m)}_{\mathcal{C}_j}, f^{(m)}_{\mathcal{C}_j}))$ be the univariate uncertain prototypes of the sets $\mathcal{C}_i, \mathcal{C}_j \subseteq \mathcal{D}$, respectively. The* univariate uncertain prototype distance *between $\mathcal{P}_{\mathcal{C}_i}$ and $\mathcal{P}_{\mathcal{C}_j}$ is defined as*

$$\Delta(\mathcal{P}_{\mathcal{C}_i}, \mathcal{P}_{\mathcal{C}_j}) = f_{dist}(\delta^{(1)}, \ldots, \delta^{(m)}) \qquad (4)$$

where

$$\delta^{(h)} = \gamma^{(h)}\ \mathrm{B}(f^{(h)}_{\mathcal{C}_i}, f^{(h)}_{\mathcal{C}_j}) +$$

$$+(1-\gamma^{(h)})\left(\frac{1}{E^{(h)}_{max}(\mathcal{D})}\left|E\left[f^{(h)}_{\mathcal{C}_i}\right] - E\left[f^{(h)}_{\mathcal{C}_j}\right]\right|\right)$$

and

$$\gamma^{(h)} = \frac{\mathcal{V}(I^{(h)}_{\mathcal{C}_i} \cap I^{(h)}_{\mathcal{C}_j})}{\min\{\mathcal{V}(I^{(h)}_{\mathcal{C}_i}), \mathcal{V}(I^{(h)}_{\mathcal{C}_j})\}},$$

$$E^{(h)}_{max}(\mathcal{D}) = \max_{o_u, o_v \in \mathcal{D}} |E[f^{(h)}_u] - E[f^{(h)}_v]|$$

for each $h \in [1..m]$, and $f_{dist} : \Re^m \rightarrow \Re$ is a function that computes a scalar value from the components of an m-dimensional vector.

3.3 The U-AHC algorithm

In this section we present the AHC-based algorithm for clustering uncertain objects, named *U-AHC*. The outline of U-AHC is given in Algorithm 1.

The input for U-AHC algorithm is a dataset $\mathcal{D}$ of n uncertain objects, whereas the output is a hierarchy of clusters $\mathbf{D}$. The algorithm follows the classic AHC scheme. Initially, every object in $\mathcal{D}$ forms a cluster (line 1). In the main cycle of the algorithm (lines 5-8), the two closest clusters are merged to form a new partition $\mathbf{C}$ (lines 5-6). $\mathbf{C}$ is then added to the set $\mathbf{D}$ as a new level (clustering) of the hierarchy (line 6). The cycle is iteratively repeated until the whole hierarchy has been built, i.e., the number of clusters in the current clustering $\mathbf{C}$ is equal to one (line 9).

The merge score used to decide for the pair of clusters to be merged at each step of the U-AHC algorithm (line 5) employs the notions of distance between uncertain prototypes (Definition 6). In particular, for any pair of clusters $\mathcal{C}_i, \mathcal{C}_j$ belonging to the current clustering $\mathbf{C}$, we compute the prototype of the cluster given by the union of the objects in $\mathcal{C}_i$ and $\mathcal{C}_j$, and evaluate the uncertain distances between this prototype and the prototypes of $\mathcal{C}_i$ and $\mathcal{C}_j$. We use the mean of these distances as a merge score, since intuitively the smaller these distances, the smaller the error of merging $\mathcal{C}_i$ and $\mathcal{C}_j$ to form a new cluster.

Algorithm 1 U-AHC

```
Input: a set of uncertain objects $\mathcal{D} = \{o_1, \ldots, o_n\}$
Output: a set of partitions $\mathbf{D}$
1: $\mathbf{C} \leftarrow \{\{\mathcal{C}_1\}, \ldots, \{\mathcal{C}_n\}\}$ such that $\mathcal{C}_i = \{o_i\}, \forall i \in [1..n]$
2: $\mathcal{P}_{\mathcal{C}_i} \leftarrow o_i, \forall i \in [1..n]$, as initial cluster prototypes
3: $\mathbf{D} \leftarrow \{\mathbf{C}\}$
4: repeat
5:     let $\mathcal{C}_i, \mathcal{C}_j$ be the pair of clusters in $\mathbf{C}$ such that $\frac{1}{2}(\Delta(\mathcal{P}_{\mathcal{C}_i \cup \mathcal{C}_j}, \mathcal{P}_{\mathcal{C}_i}) + \Delta(\mathcal{P}_{\mathcal{C}_i \cup \mathcal{C}_j}, \mathcal{P}_{\mathcal{C}_j}))$ is minimum
6:     $\mathbf{C} \leftarrow \{\mathcal{C} \in \mathbf{C} : \mathcal{C} \neq \mathcal{C}_i, \mathcal{C} \neq \mathcal{C}_j\} \cup \{\mathcal{C}_i \cup \mathcal{C}_j\}$
7:     $\mathbf{D} \leftarrow \mathbf{D} \cup \{\mathbf{C}\}$
8:     update prototypes $\mathcal{P}_{\mathcal{C}},\ \mathcal{C} \in \mathbf{C}$
9: until $|\mathbf{C}| = 1$
```

We compute the integrals involved into the distances calculation by taking into account lists of samples (s) derived from the pdfs. For this purpose, we employed the classic *Monte Carlo* sampling method. [1]

4 Experimental evaluation

The U-AHC algorithm was evaluated in performing effective clustering of microarray data with a probe-level uncertainty. In this section we first discuss the evaluation methodology used in this work, which includes a description of the datasets, the uncertainty modeling, and the measures to assess the quality of the clustering solutions. Then, we present preliminary experimental results.

4.1 Evaluation methodology

Datasets. Experiments were performed on four large microarray datasets, each of which describes the expressions of thousands of genes in biological tissues, as shown in Table 1.

Three datasets, namely Leukaemia, Neuroblastoma and Myelodysplastic are cancer tissue data of humans,[2] while Mouse is about mouse tissues.[3] Leukaemia de-

[1] We used the SSJ library, available at http://www.iro.umontreal.ca/∼simardr/ssj/

[2] Cancer Program dataset page of the Broad Institute of MIT and Harvard, available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi

[3] Microarray Data resource page of the European Bioinformatics Institute (EMBL-EBI), available at http://www.ebi.ac.uk/microarray-as/ae/browse.html

Table 1. Microarray datasets used in the experiments

dataset	*# of genes*	*# of attributes*
Leukaemia	22,690	21
Neuroblastoma	22,282	14
Myelodysplastic	22,277	25
Mouse	45,101	10

scribes the transformation process of leukaemia stem cells initiated by MLL-AF9 fusion gene. Neuroblastoma contains expression-based screening results for neuroblastoma differentiation. In Myelodysplastic, somatic chromosomal deletions in cancer are measured by means of an RNA-mediated interference (RNAi)-based approach to discovery of the $5q^-$ disease gene, which is a subtype of the myelodysplastic syndrome characterized by a defect in erythroid differentiation. Mouse contains a transcription profiling of mouse cochlea Reissner's membrane (RM). This is grown as explants and treated with dexamethasone and then subject to RNA extraction to investigate gene expressions.

Uncertainty models. The probe-level uncertainty for microarray datasets was extracted by exploiting the multi-mgMOS method [16].[4] For each dimension, the multi-mgMOS method yields a set of information that includes mean, standard deviation and principal percentiles (i.e., 5%, 25%, 50%, 75%, 95%).

In this work, the information outputted by multi-mgMOS was exploited to model uncertainty according to the univariate model (Section 2). In particular, the univariate pdfs of each uncertain object (i.e., each row in the microarray matrix) was built by employing two different methods:

- *Normal* method, where Normal pdfs were easily derived from a combination of mean values with standard deviations;
- *Percentiles-based* method, where suitable statistical models were involved to fit pdfs to percentiles [19].

Clustering validity criteria. We performed a *gene-based clustering* in such a way that each group describes a particular macroscopic phenotype, such as cancer expressions or biological states [10]. Since there is no available reference classification for such data, we resorted to internal validity criteria based on the *cophenetic correlation coefficient* [20], which ranges between [0,1] and evaluates a dendrogram according to how it preserves the pairwise distances between the original data points. Intuitively, the higher the cophenetic correlation value for a dendrogram, the higher is the compactness and the better is the quality. This measure is particularly suitable for biological data as it is widely used in biostatistic fields, e.g., to assess the cluster-based models of DNA sequences or to evaluate taxonomic models.

Table 2. Accuracy results for univariate models

dataset	*pdf form*	*cophenetic value*
Leukaemia	Normal	0.76
	Percentiles-based	0.82
Neuroblastoma	Normal	0.67
	Percentiles-based	0.75
Myelodysplastic	Normal	0.80
	Percentiles-based	0.89
Mouse	Normal	0.84
	Percentiles-based	0.92

Formally, let $\mathcal{D} = \{o_1, \ldots, o_n\}$ be a dataset of n objects and let $\mathbf{D}$ be the dendrogram solution produced by a hierarchical clustering algorithm (i.e., a hierarchy of clusters). The *cophenetic correlation coefficient* ($c(\mathcal{D}, \mathbf{D})$) is defined as

$$c(\mathcal{D}, \mathbf{D}) = \frac{\sum_{i<j}(d_E(o_i, o_j) - \overline{d}_E)(t(o_i, o_j) - \overline{t})}{\sqrt{\left[\sum_{i<j}(d_E(o_i, o_j) - \overline{d}_E)^2\right]\left[\sum_{i<j}(t(o_i, o_j) - \overline{t})^2\right]}}$$

for all $i, j \in [1..n]$. In the formula above, $d_E(o_i, o_j)$ denotes the Euclidean distance between the objects o_i and o_j, while $t(o_i, o_j)$ is the dendrogrammatic distance of such objects, which indicates the level of the dendrogram at which the objects o_i and o_j are first joined together. The values $\overline{d}_E$ and $\overline{t}$ represent the average of the $d_E(o_i, o_j)$ and the average of the $t(o_i, o_j)$, respectively.

4.2 Preliminary results

Table 2 summarizes the quality results in terms of cophenetic correlation for each dataset and for each pdf. It can be noted that the U-AHC algorithm obtained good accuracy results on all the datasets, from 67% to 84% with Normal pdfs. Also, the uncertainty generation based on percentiles generally led to higher quality results than the previous case (about 8% on average); this improvement can be easily explained by the fact that percentiles provide a more refined representation of the uncertainty than the summarized in-

[4] We used the Bioconductor package PUMA (*Propagating Uncertainty in Microarray Analysis*), available at http://www.bioinf.manchester.ac.uk/resources/puma/

formation of mean value and standard deviation used for Normal pdf modeling.

5 Conclusion

We addressed the problem of clustering microarray data by adopting a probabilistic approach which is conceived to model the (probe-level) uncertainty in the data. The proposed approach lies in a centroid-linkage-based agglomerative hierarchical algorithm, named U-AHC. The U-AHC algorithm is equipped with a notion of uncertain cluster prototype represented as a mixture of the probability distributions associated to the objects belonging to any given cluster. Also, the cluster merging criterion in U-AHC exploits a new information-theoretic-based distance between uncertain prototypes.

The U-AHC algorithm has experimentally shown to achieve significant accuracy in identifying clustering solutions that are well-suited to capture the underlying gene-based patterns of microarray data.

References

[1] B. P. Adhikari and D. D. Joshi. Distance discrimination resume exhaustif. *Publ. Inst. Stat.*, 5:57–74, 1956.

[2] S. M. Ali and S. D. Silvey. A General Class of Coefficients of Divergence of One Distribution from Another. *J. Roy. Stat. Soc.*, 28(1):131–142, 1966.

[3] M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18(4):349–369, 1989.

[4] A. Bhattacharyya. On a Measure of Divergence Between Two Statistical Populations Defined by their Probability Distributions. *Bull. Calcutta Math. Soc.*, 35:99–110, 1943.

[5] H. Chernoff. A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations. *Ann. Math. Stat.*, 23(4):493–507, 1952.

[6] S. Draghici. *Data Analysis Tools for DNA Microarrays*. Chapman & Hall/CRC Mathematical Biology and Medicine Series, 2003.

[7] G. Gan, C. Ma, and J. Wu. *Data Clustering: Theory, Algorithms, and Applications*. ASA-SIAM Series on Statistics and Applied Probability, 2007.

[8] F. Gullo, G. Ponti, A. Tagarelli, and S. Greco. A Hierarchical Algorithm for Clustering Uncertain Data via an Information-Theoretic Approach. In *Proc. IEEE Int. Conf. on Data Mining (ICDM)*, pages 821–826, 2008.

[9] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[10] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data: a survey. *TKDE*, 16(11):1370–1386, 2004.

[11] T. Kailath. The Divergence and Bhattacharyya Distance Measures in Signal Selection. *IEEE Trans. on Comm. Tech.*, 15(1):52–60, 1967.

[12] S. Kullback. *Information theory and statistics*. Wiley, 1959.

[13] S. Kullback and R. A. Leibler. On Information and Sufficiency. *Ann. Math. Stat.*, 22(1):79–86, 1951.

[14] X. Liu, K. K. Lin, B. Andersen, and M. Rattray. Including probe-level uncertainty in model-based gene expression clustering. *Bioinformatics*, 8:98, 2007.

[15] 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.

[16] 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.

[17] F. Murtagh. A survey of recent advances in hierarchical clustering algorithm. *The Computer Journal*, 26(4):354–359, 1983.

[18] S. Pradervand, A. Paillusson, J. Thomas, J. Weber, P. Wirapati, O. Hagenbuchle, and K. Harshman. Affymetrix Whole-Transcript Human Gene 1.0 ST array is highly concordant with standard 3' expression arrays. *Biotechniques*, 44(6):759–762, 2008.

[19] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[20] R. R. Sokal and F. J. Rohlf. The comparison of dendrograms by objective methods. *Taxon*, 11:33–40, 1962.