

# Fuzzy ensemble clustering based on random projections for DNA microarray data analysis

Roberto Avogadri and Giorgio Valentini <sup>a</sup>

<sup>a</sup>*DSI, Dipartimento di Scienze dell' Informazione,  
Università degli Studi di Milano,  
Via Comelico 39, 20135 Milano, Italia.  
e-mail : {avogadri,valentini}@dsi.unimi.it*

---

## Abstract

### Objective:

Two major problems related the unsupervised analysis of gene expression data are represented by the accuracy and reliability of the discovered clusters, and by the biological fact that the boundaries between classes of patients or classes of functionally related genes are sometimes not clearly defined. The main goal of this work consists in the exploration of new strategies and in the development of new clustering methods to improve the accuracy and robustness of clustering results, taking into account the uncertainty underlying the assignment of examples to clusters in the context of gene expression data analysis.

### Methodology:

We propose a fuzzy ensemble clustering approach both to improve the accuracy of clustering results and to take into account the inherent fuzziness of biological and bio-medical gene expression data. We applied random projections that obey the Johnson-Lindenstrauss lemma to obtain several instances of lower dimensional gene expression data from the original high-dimensional ones, approximately preserving the information and the metric structure of the original data. Then we adopt a double fuzzy approach to obtain a consensus ensemble clustering, by first applying a fuzzy k-means algorithm to the different instances of the projected low-dimensional data and then by using a fuzzy t-norm to combine the multiple clusterings. Several variants of the fuzzy ensemble clustering algorithms are proposed, according to different techniques to combine the base clusterings and to obtain the final consensus

clustering.

Results and conclusion:

We applied our proposed fuzzy ensemble methods to the gene expression analysis of leukemia, lymphoma, adenocarcinoma and melanoma patients, and we compared the results with other state of the art ensemble methods. Results show that in some cases, taking into account the natural fuzziness of the data, we can improve the discovery of classes of patients defined at bio-molecular level. The reduction of the dimension of the data, achieved through random projections techniques, is well-suited to the characteristics of high-dimensional gene expression data, thus resulting in improved performance with respect to single fuzzy k-means and with respect to ensemble methods based on resampling techniques. Moreover, we show that the analysis of the accuracy and diversity of the base fuzzy clusterings can be useful to explain the advantages and the limitations of the proposed fuzzy ensemble approach.

*Key words:* Gene expression data clustering, ensemble clustering, fuzzy clustering, random subspace, random projections, DNA microarrays.

---

## 1 Introduction

In recent years unsupervised clustering methods have been successfully applied to DNA microarray data analysis, considering in particular two main problems: the discovery of new subclasses of diseases or functionally correlated examples [1,2] and the detection of subsets of co-expressed genes as a proxy of co-regulated genes [3]. At the same time machine learning research showed that unsupervised ensemble approaches can improve the accuracy and the reliability of clustering results [4–7].

In bioinformatics several ensemble clustering approaches have been proposed for the analysis of gene expression data [8–13]. In [8] a robust ensemble method, based on the agreement between different clustering methods in assigning genes to the same clusters, has been proposed. Dudoit and Fridlyand [9] designed an unsupervised version of the classical supervised bagged ensemble [14], showing that this approach improves existing clustering methods in the analysis of DNA microarray data. A similar approach, based on resam-

pling methods has been proposed in [10], where techniques to validate and visualize high-dimensional gene expression data are also provided. In [11] the clustering results of individual clustering algorithm applied to the analysis of gene expression data are converted into a distance matrix, a weighted graph is constructed according to the combined matrix, and a graph partitioning approach is used to cluster the graph to generate the final clusters. Consensus clustering obtained from clustering multiple times with Variational Bayes mixtures of Gaussians have been successfully applied to the unsupervised analysis of functional classes of genes in yeast [12], while a graph-based ensemble clustering algorithm has been recently proposed to discover the underlying classes of the examples in gene expression data [13].

In particular, recently proposed methods based on random projections [15] have been successfully applied to ensemble clustering of gene expression data [16] and to assess the validity of clusters discovered in bio-molecular data [17,18].

A major problem with these approaches is represented by the biological fact that classes of patients or classes of functionally related genes are sometimes not clearly defined. For instance, it is well-known that a single gene product may participate to different biological processes and as a consequence it may be at the same time expressed with different subsets of co-expressed genes.

To take into account these items we propose a fuzzy approach, in order to consider the inherent fuzziness of clusters discovered in gene expression data [19]. The main idea of this work is to combine the accuracy and the effectiveness of the ensemble clustering techniques based on random projections [15], with the expressive capacity of the fuzzy sets, to obtain clustering algorithms both reliable and able to express the uncertainty of the data.

In the next sections we briefly introduce random projections, then we present our proposed fuzzy ensemble clustering methods. In Sect. 5 the proposed method is applied to the analysis of DNA microarray data, and the results are discussed comparing the fuzzy ensemble clustering methods with other ensemble approaches, considering also the relationships between accuracy and diversity of the base clusterings, w.r.t. the overall accuracy of the ensembles. The conclusions and some remarks about future developments end the paper.

## 2 Random projections.

Let  $D$  be a  $d \times n$  real matrix, where each column represents a  $d$ -dimensional example, and let  $p$  be a vector storing the expression levels of  $d$  genes (that is, a generic column of  $D$ ). For instance  $p$  could represent the expression profile of a given patient, or the expression levels of a gene across  $d$  different experimental conditions.

Considering the usually high dimension of an expression profile of a given patient, and the relatively low cardinality of the patients, a key issue is represented by the reduction of the data dimension to contrast the well-known "curse of dimensionality" problem. [20].

Our approach proposes to reduce the dimension of the original data using random projections  $\mu : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$  from high  $d$ -dimensional spaces to lower  $d'$ -dimensional subspaces.

In this context, a key problem consists in finding a  $d'$  such that for every pair of data  $p, q \in \mathbb{R}^d$ , the distances between the projections  $\mu(p)$  and  $\mu(q)$  are approximately preserved with high probability. A natural measure of the approximation is the distortion  $dist_\mu$ :

$$dist_\mu(p, q) = \frac{\|\mu(p) - \mu(q)\|_2}{\|p - q\|_2} \quad (1)$$

If  $dist_\mu(p, q) = 1$ , the distances are preserved; if  $1 - \epsilon \leq dist_\mu(p, q) \leq 1 + \epsilon$ , we say that an  $\epsilon$ -distortion level is introduced.

It has been shown that using random projections that obey *Johnson-Lindenstrauss (JL) lemma* [21] we may perturb the data introducing only bounded distortions, approximately preserving the metric structure of the original data. The dimension of the projected subspace depends only on the cardinality of the original data and the desired  $\epsilon$ -distortion, and not from the dimension  $d$  of the original space (see [17] for more details).

The projections with bounded distortions can be obtained through a quite simple stochastic approach [22,23]: data are projected to lower dimensional

subspaces by random  $d' \times d$  matrices  $R = 1/\sqrt{d'}(r_{ij})$ , where  $r_{ij}$  are random variables such that:

$$E[r_{ij}] = 0, \quad \text{Var}[r_{ij}] = 1$$

Examples of random projections are the following:

- (1) *Plus-Minus-One (PMO)* or *Bernoulli* random projections: represented by  $d' \times d$  matrices  $R = 1/\sqrt{d'}(r_{ij})$ , where  $r_{ij}$  are uniformly chosen in  $\{-1, 1\}$ , such that  $\text{Prob}(r_{ij} = 1) = \text{Prob}(r_{ij} = -1) = 1/2$  (that is the  $r_{ij}$  are *Bernoulli* random variables).
- (2) *Achlioptas* random projections [22]: represented by  $d' \times d$  matrices  $P = 1/\sqrt{d'}(r_{ij})$ , where  $r_{ij}$  are chosen in  $\{-\sqrt{3}, 0, \sqrt{3}\}$ , such that  $\text{Prob}(r_{ij} = 0) = 2/3$ ,  $\text{Prob}(r_{ij} = \sqrt{3}) = \text{Prob}(r_{ij} = -\sqrt{3}) = 1/6$ .
- (3) *Normal* random projections [23,5]: this *JL lemma* compliant randomized map is represented by a  $d' \times d$  matrix  $R = 1/\sqrt{d'}(r_{ij})$ , where  $r_{ij}$  are distributed according to a gaussian with 0 mean and unit variance.
- (4) *Random Subspace (RS)* [24,25]: represented by  $d' \times d$  matrices  $R = \sqrt{d/d'}(r_{ij})$ , where  $r_{ij}$  are uniformly chosen with entries in  $\{0, 1\}$ , and with exactly one 1 per row and at most one 1 per column. Unfortunately, *RS* does not satisfy the *JL lemma*.

Using the above randomized maps (with the exception of *RS* projections), the *JL lemma* guarantees that, with high probability, the "compressed" examples of the data set represented by the matrix  $D_R = RD$  have approximately the same distance (up to an  $\epsilon$ -distortion level) of the corresponding examples in the original space, represented by the columns of the matrix  $D$ , as long as  $d' \geq c \log n/\epsilon^2$ .

### 3 Fuzzy ensemble clustering based on random projections

The ensemble algorithm applies random projection techniques to perturb the data, using a fuzzy clustering algorithm to generate multiple base clusterings. More precisely, data are perturbed through random projections to lower dimensional subspaces and multiple clusterings are performed on the projected

data; note that it is likely to obtain different clusterings, since the clustering algorithm is applied to different "views" of the data. Then the clusterings are combined, and a *consensus* ensemble clustering is computed. This approach is similar to the one proposed in [15]: the main difference consists in using a fuzzy k-means algorithm as base clustering and in applying a fuzzy approach to the combination and the consensus steps of the ensemble algorithm.

The main steps of the fuzzy ensemble clustering algorithm can be summarized as follows:

- (1) *Random projections.* Multiple instances (views) of compressed data are obtained using random projections.
- (2) *Generation of multiple fuzzy clusterings.* The fuzzy k-means algorithm is applied to the instances of data obtained from the previous step. The output of the algorithm is a membership matrix, where each element represents the membership of an example to a particular cluster.
- (3) *Aggregation.* The fuzzy clusterings are combined, using a similarity matrix [9]. The generation of each element of the matrix is obtained through fuzzy t-norms.
- (4) *Consensus clustering.* The ensemble clustering is built up by applying the fuzzy k-means algorithm to the rows of the similarity matrix obtained in the previous step.

In the *Random projections* step, for each example of the original data set with  $d$  features, only a fixed number of features  $d'$  ( $d' < d$ ) are randomly chosen to represent the same example. The value of  $d'$  is evaluated according to the JL Lemma (Sect. 2). The procedure is repeated different times to obtain different representations of the original data set. The fuzzy-k-means algorithm is then applied to each instance of the data, thus obtaining a set of fuzzy membership matrices  $\mathcal{U}$ , whose elements  $u_{ij}$  represent the membership of example  $j$  to the cluster  $i$ .

The *Aggregation* step is performed by using a square symmetric similarity matrix  $M$ , where each element represents the "level of agreement" between

each pair of examples:

$$M_{i,j} = \sum_{s=1}^k \tau(\mathcal{U}_{s,i}, \mathcal{U}_{s,j}); \quad (2)$$

where  $k$  is the number of clusters;  $i, j$  indices of the  $n$  examples,  $1 \leq i, j \leq n$ ;  $\mathcal{U}$  is a fuzzy membership matrix (where the rows refer to clusters and the columns to examples), and finally  $\tau$  is a suitable fuzzy t-norm (e.g. an algebraic product). Note that  $M_{i,j}$  can be interpreted as the "common membership" of two examples  $i$  and  $j$  to the same cluster.

The similarity matrices  $M$  obtained through  $c$  repeated applications of the fuzzy k-means clustering algorithm are aggregated simply by averaging: in this way we achieve the cumulative similarity matrix  $M^C$ :

$$M_{i,j}^C = \frac{1}{c} \sum_{t=1}^c M_{i,j}^{(t)}; \quad (3)$$

The *Consensus clustering* step is performed by applying the fuzzy-k-means clustering to the rows of  $M^C$ , thus obtaining the *consensus membership* matrix  $\mathcal{U}^C$ . Indeed note that  $i^{th}$  row of  $M^C$  represents the "common membership" to the same cluster of the  $i^{th}$  example with respect to all the other examples, averaged across multiple clusterings. In this sense the rows can be interpreted as a new "feature space" for the analyzed examples.

The *consensus clusters* can be also obtained by choosing one of two classical "crispization" techniques:

### Hard-clustering:

$$\chi_{ri}^H = \begin{cases} 1 & \Leftrightarrow \arg \max_s \mathcal{U}_{si}^C = r \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

### $\alpha$ -cut:

$$\chi_{ri}^\alpha = \begin{cases} 1 & \Leftrightarrow \mathcal{U}_{ri}^C \geq \alpha \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

where  $\chi_{ri}$  is the characteristic function for the cluster  $r$ : that is  $\chi_{ri} = 1$  if the  $i^{th}$  example belongs to the  $r^{th}$  cluster,  $\chi_{ri} = 0$  otherwise;  $1 \leq s \leq k$ ;  $1 \leq i \leq n$ ,

$0 \leq \alpha \leq 1$ , and  $\mathcal{U}^C$  is the consensus fuzzy membership matrix obtained by applying the fuzzy k-means algorithm to  $M^C$ .

The pseudo-code of the algorithm is reported below:

**Fuzzy ensemble clustering algorithm :**

Input:

- a data set  $X = \{x_1, x_2, \dots, x_n\}$ , stored in a  $d \times n$   $D$  matrix.
- an integer  $k$  (number of clusters)
- an integer  $c$  (number of clusterings)
- the fuzzy k-means clustering algorithm  $\mathcal{C}_f$
- a procedure that realizes the randomized map  $\mu$
- an integer  $d'$  (dimension of the projected subspace)
- a function  $\tau$  that defines the t-norm

begin algorithm

- (1) For each  $i, j \in \{1, \dots, n\}$  do  $M_{ij} = 0$
- (2) Repeat for  $t = 1$  to  $c$ 
  - (3)  $R_t = \text{Generate\_projection\_matrix}(d', \mu)$
  - (4)  $D_t = R_t \cdot D$
  - (5)  $\mathcal{U}^{(t)} = \mathcal{C}_f(D_t, k, m)$
  - (6) For each  $i, j \in \{1, \dots, n\}$ 

$$M_{ij}^{(t)} = \sum_{s=1}^k \tau(\mathcal{U}_{si}^{(t)}, \mathcal{U}_{sj}^{(t)})$$
- end repeat
- (7)  $M^C = \frac{\sum_{t=1}^c M^{(t)}}{c}$
- (8)  $\langle A_1, A_2, \dots, A_k \rangle = \mathcal{C}_f(M^C, k, m)$

end algorithm.

Output:

- the final clustering  $C = \langle A_1, A_2, \dots, A_k \rangle$
- the cumulative similarity matrix  $M^C$ .

Note that the dimension  $d'$  of the projected subspace is an input parameter of the algorithm, but it may be computed according to the *JL* lemma (Sect. 2), to approximately preserve the distances between the examples. Inside the mean loop (steps 2-6) the procedure `Generate_projection_matrix` produces a  $d' \times d$

$R_t$  matrix according to a given random map  $\mu$  [15], that it is used to randomly project the original data matrix  $D$  into a  $d' \times n$   $D_t$  projected data matrix (step 4). In step (5) the fuzzy k-means algorithm  $\mathcal{C}_f$  with a given fuzziness  $m$  is applied to  $D_t$  and a  $k$ -clustering represented by its  $\mathcal{U}^{(t)}$  membership matrix is achieved. Hence the corresponding similarity matrix  $M^{(t)}$  is computed, using a given  $t$ -norm (step 6). In (7) the "cumulative" similarity matrix  $M^C$  is obtained by averaging across the similarity matrices computed in the main loop. Finally, the *consensus* clustering is obtained by applying the fuzzy k-means algorithm to the rows of the similarity matrix  $M^C$  (step 8).

We can observe that in both the steps (5) and (8) the clusterings obtained are fuzzy clusterings (represented by membership matrices). In some applications we need to obtain a partition of the data, or more in general we need to obtain crisp results. Using our algorithm both the results can be obtained applying respectively the hard-clustering and the alpha-cut techniques to the output of the consensus clustering on the step (8) (eq. 4 and 5). In the rest of the paper we call the first algorithm *fuzzy-max*, and the second one *fuzzy-alpha*.

Note that we may obtain crisp clusters just from the fuzzy base clusterings, by applying the above defuzzification techniques just at a "base clustering level". For instance, we can process the the membership matrix  $\mathcal{U}^{(t)}$  of the fuzzy ensemble algorithm to achieve crisp clusters by adding one of the following lines after step (5) of the ensemble algorithm:

$$(5\_hard) \chi^{(t)} = Crisp_{hard}(\mathcal{U}^{(t)})$$

$$(5\_alpha) \chi^{(t)} = Crisp_{\alpha}(\mathcal{U}^{(t)})$$

By using step (5\_hard) we obtain a "hard-clustering" crispization (eq. 4), and by using step (5\_alpha) we obtain an  $\alpha$ -cut crispization (eq. 5). In both cases the result is a "characteristic matrix"  $\chi$  whose binary elements  $\chi_{si} \in \{0, 1\}$  denote whether element  $i$  belongs to cluster  $s$ . The next step (6) of the fuzzy ensemble algorithm needs to be modified by replacing the fuzzy t-norm operator with the product of the characteristic vectors of the examples  $i$  and  $j$ :

$$(6\_crisp) M_{ij}^{(t)} = \sum_{s=1}^k \chi_{si}^{(t)} * \chi_{sj}^{(t)}$$

One may observe that by using the alpha-cut as "defuzzification" method

each example of the data set can belong to more than one cluster, and hence a different normalization method (step 7) has to be used. In this case we need to replace in (7) the number of clusterings  $c$  with  $v = k * c$ :

We refer to the algorithm obtained by applying the hard-clustering in both the (5) and (8) steps with *max-max*, while we name *max-alpha* the algorithm that applies the hard-clustering at step (5) and the  $\alpha$ -cut at step (8).

## 4 Experimental environment

We test our proposed fuzzy ensemble algorithms with four DNA microarray data sets, comparing the results with other related cluster ensemble methods.

### 4.1 Gene expression data

We considered four DNA microarray data sets available on the web. The first one (*DLBCL-FL* data set) is composed by tumor specimens from 58 Diffuse Large B-Cell Lymphoma (DLBCL) and 19 Follicular Lymphoma (FL) patients [26] in a gene expression space with 6285 genes. The second one, the *Primary-Metastasis* (PM) data set, contains expression values in Affymetrix's scaled average difference units (14925 genes) for 64 primary adenocarcinomas and 12 metastatic adenocarcinomas (lung, breast, prostate, colon, ovary, and uterus) from unmatched patients prior to any treatment [27]. In both cases we followed the same preprocessing and normalization steps described in [26] and [27]. The third one, *Leukemia*, contains gene expression levels of 7129 genes in Affymetrix's scaled average difference units relative to 47 patients with Acute Lymphoblastic Leukemia (ALL) and 25 cases of Acute Myeloid Leukemia (AML) [28]. The fourth one, *melanoma*, is composed by 31 melanoma samples and 7 control samples with 6971 genes [29]. Also for these data sets we applied the same pre-processing procedures described respectively in [28] and [29].

## 4.2 Methods

We tested the performance of the fuzzy ensemble algorithms *fuzzy-max*, *fuzzy-alpha*, *max-max* and *max-alpha* using the previously described data sets. We compared the results with *Randclust*, the corresponding "crisp" version of our proposed ensemble methods [15] and with *Bagclust1*, based on an unsupervised version of bagging [9], and with the "single" fuzzy k-means clustering algorithm. The *Randclust* ensemble method is similar to the algorithm presented in this paper, but it uses the hierarchical clustering algorithm to produce the base clusterings, and a crisp approach to combine the resulting clusters. *Bagclust1* generates multiple instances of perturbed data through bootstrap techniques, and then it applies to each instance the base clustering algorithm (we used in our experiments k-means); the final clustering is obtained by majority voting.

Each ensemble is composed by 50 base clusterings and each ensemble method has been repeated 30 times. Regarding to ensemble methods based on random projections, we chose projections with bounded  $1 \pm 0.2$  distortion, according to the *JL* lemma, while for *Bagclust1* we randomly drew with replacement a number of examples equal to the number of the available data.

## 4.3 Assessment of the error

Since clustering does not univocally associate a label to the examples, but only provides a set of clusters, we evaluated the error by choosing for each clustering the permutation of the classes that best matches the "a priori" known "true" classes. More precisely, considering the following clustering function:

$$f(x) : \mathcal{R}^d \rightarrow \mathcal{Y}, \text{ with } \mathcal{Y} \subseteq \{1, \dots, k\} \quad (6)$$

where  $x$  is the sample to classify,  $d$  its dimension,  $k$  the number of the classes; the error function we applied is the following:

$$\mathcal{L}_{0/1}(Y, t) = \begin{cases} 0 & \text{if } (|Y| = 1 \wedge t \in Y) \vee Y = \{\lambda\} \\ 1 & \text{otherwise.} \end{cases} \quad (7)$$

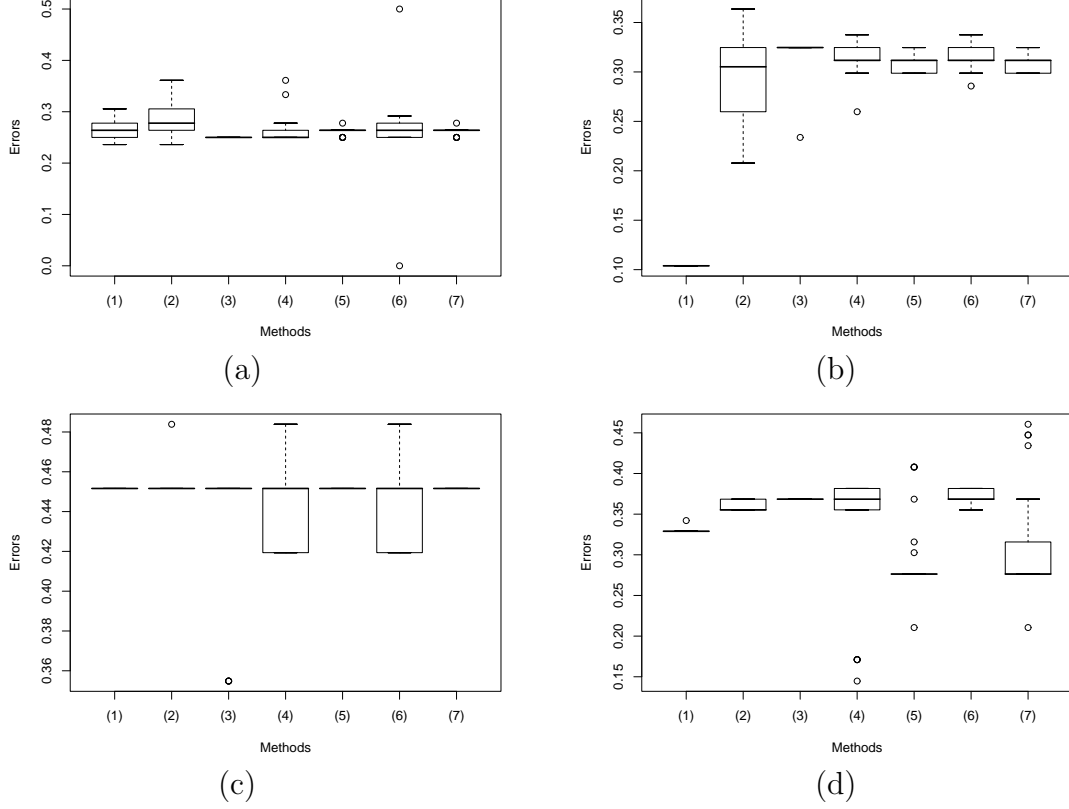


Fig. 1. Boxplot of the results (fuzziness 1.1). (a) *Leukemia* data set, (b) *DLBCL-FL* data set, (c) *Melanoma* data set and (d) *Primary-Metastasis* data set. (1)..(7) in abscissa refer to the results obtained respectively with (1) Randclust, (2) Bag-clust1, (3) Single fuzzy k-means, (4) max-max, (5) fuzzy-max, (6) max-alpha and (7) fuzzy-alpha ensemble algorithms.

with  $t$  the “real” label of the sample  $x$ ,  $Y \in \mathcal{Y}$  the predicted set of class labels, and  $\{\lambda\}$  the empty set. Note that when  $Y = \{\lambda\}$  the loss is 0, but the example is registered as unassigned. Other loss functions or measures of the performance of clustering algorithms may be applied, but we chose this modification of the 0/1 loss function to take into account the multi-label output of fuzzy k-means algorithms.

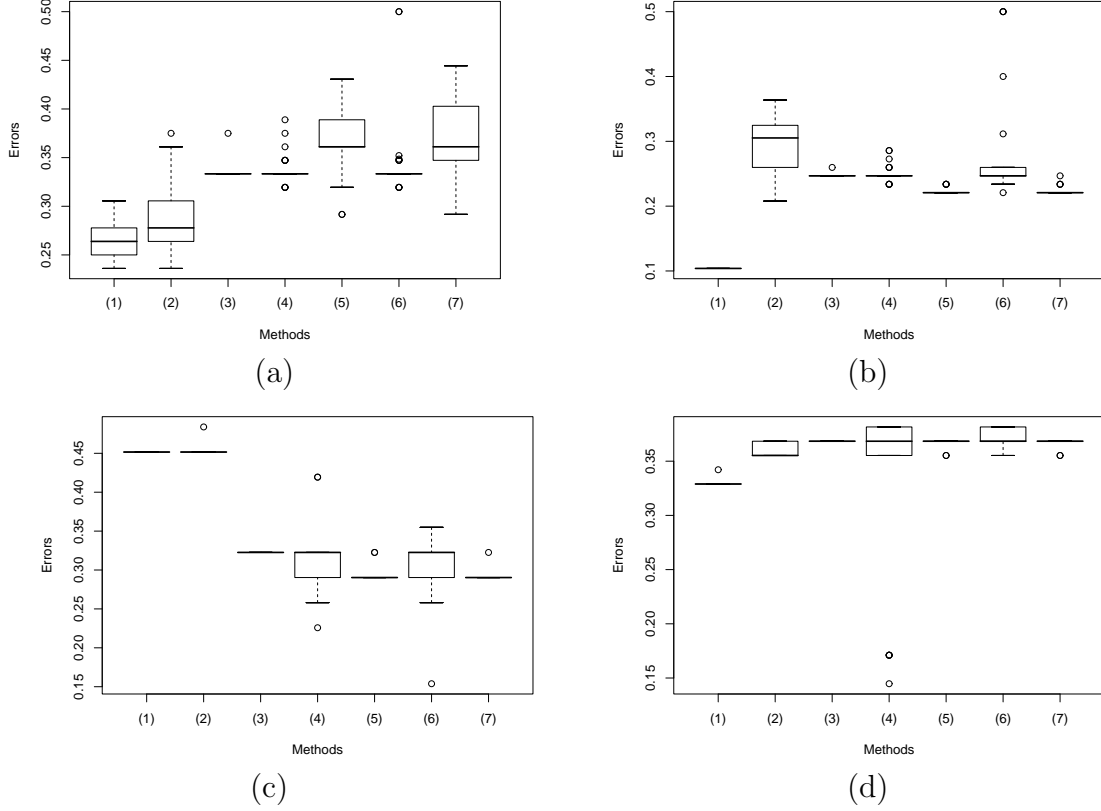


Fig. 2. Boxplot of the results (fuzziness 2.0). (a) *Leukemia* data set, (b) *DLBCL-FL* data set, (c) *Melanoma* data set and (d) *Primary-Metastasis* data set. (1)..(7) in abscissa refer to the results obtained respectively with (1) *Randclust*, (2) *Bagclust1*, (3) *Single fuzzy k-means*, (4) *max-max*, (5) *fuzzy-max*, (6) *max-alpha* and (7) *fuzzy-alpha* ensemble algorithms.

## 5 Results and discussion

### 5.1 Assessment of the accuracy of the clusterings

The boxplots in Fig. 1 and 2 represent the distribution of the error across multiple repetitions of the fuzzy ensemble algorithms compared with the “single” fuzzy k-means, the *Randclust* and the *Bagclust1* ensemble methods for all the data sets used in the experiments. In particular in Fig. 1 are shown the results obtained with fuzziness equal to 1.1, while in Fig. 2 the results have been obtained with a fuzziness level equal to 2.0. With *max-alpha* and *fuzzy-alpha* ensemble algorithms the results shown in Fig. 1 and 2 correspond to a choice of *alpha* value equal to 0.5.

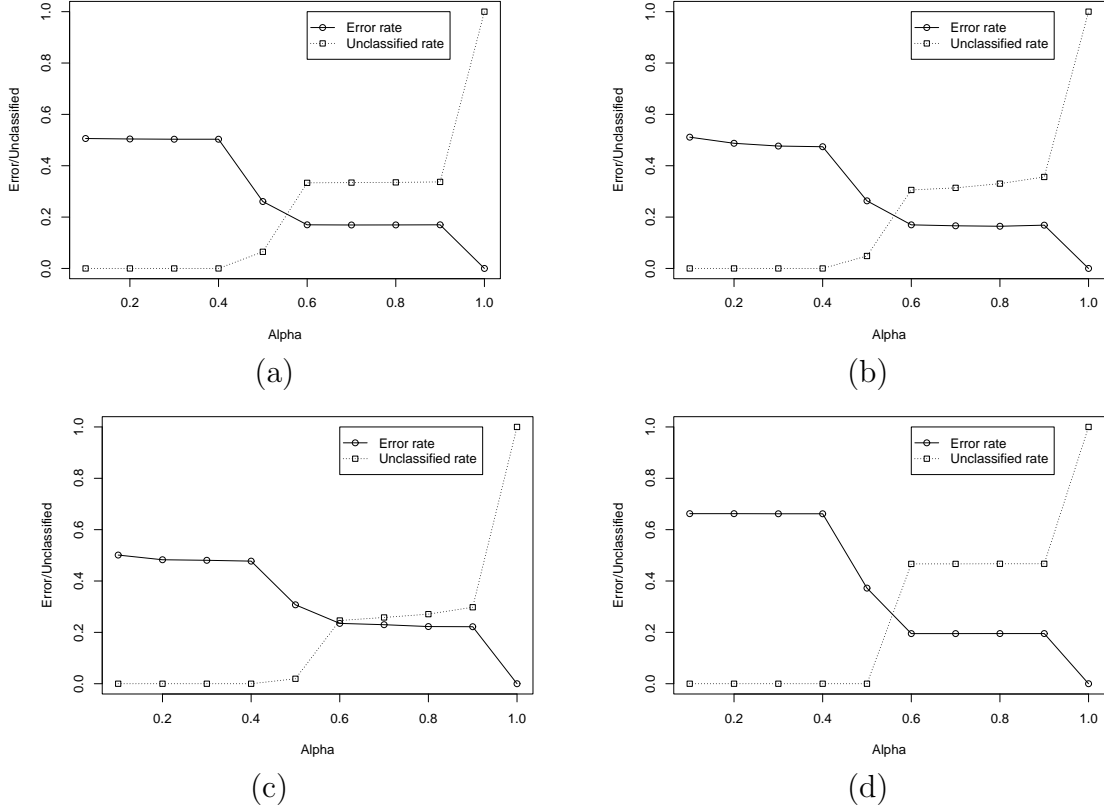


Fig. 3. Results of the fuzzy ensemble max-alpha. The solid line refers to the error rate and the dotted line corresponds to the unclassified rate. (a) *Leukemia* data set, (b) *DLBCL-FL* data set, (c) *Melanoma* data set and (d) *Primary-Metastasis* data set.

We can observe that, while with the *Primary-Metastasis* (Fig. 1, 2 (d)) and *Leukemia* (Fig. 1, 2 (a)) data sets larger levels of fuzziness reduce the performances of the fuzzy ensemble algorithms, in the *DLBCL-FL* (Fig. 1, 2 (b)) and *Melanoma* (Fig. 1, 2 (c)) data sets the increment of the fuzziness level leads to an improvement of the overall performance.

The fuzzy ensemble approach improves the performance of the single fuzzy k-means applied to the entire set of gene expression levels: for all the data sets the fuzzy ensemble methods obtain equal or better results with respect to the single fuzzy k-means algorithm (Fig. 1 and 2).

The performances obtained with *BagClust1* (marked with number (2) in the boxplot figures) are in general lower w.r.t to the fuzzy ensemble methods (except with the *Leukemia* data set, when a fuzziness equal to 2.0 is applied). In

the context of high dimensional gene expression data, when we have a large number of features (gene expression levels) and a low number of examples (patients) an approach that reduces the dimension instead of the cardinality of the data is likely to produce better results. Indeed, exploiting the inherent redundancy of information and the approximate preservation of the metric space that characterizes random projections obeying the JL lemma, our approach largely outperforms resampling based methods (Fig. 1 and 2). This hypothesis is confirmed also by the fact that also the other ensemble method based on random projections (*Randclust*, marked with number (1) in the box-plots of Fig. 1 and 2) achieved always better or equal results with respect to *BagClust1*.

Considering the data sets *Melanoma* and *Primary-Metastasis* (Fig. 1, 2 (c) and (d)) fuzzy ensemble clustering methods achieve better results with respect to *Randclust*. Nevertheless, considering the *Leukemia* gene expression data, there is no significant difference between the proposed methods and *Randclust* (Fig. 1 (a)), while with the *DLBCL* data *Randclust* largely outperforms the fuzzy ensemble clustering methods (Fig. 1, 2 (b)). These results largely depend on the different characteristics of the base clusterings used in the two methods. Indeed, with the *DLBCL* data set, the hierarchical clustering algorithm (Ward’s method [30]) largely outperforms the fuzzy k-means algorithm: using a single hierarchical clustering algorithm applied to the high dimensional gene expression space (6285 genes) we obtain an error slightly below 0.11, while with the single fuzzy k-means algorithm we largely double the error to about 0.25. In this case the choice of the base clustering algorithm is determinant to achieve ”good” results.

*Max-alpha* and *fuzzy-alpha* ensemble results depend on the choice of the  $\alpha$  value. Fig. 3 reports the error and unclassified rate as a function of  $\alpha$  for the *max-alpha* algorithm. We can observe that error and unclassified rate follow, as expected, opposite trends: we may obtain good accuracy results at the expenses of high unclassified rates.

The tables 1 and 2 summarize the results obtained respectively with the *Leukemia*, *DLBCL-FL*, *Melanoma*, and *Primary-Metastasis* data sets. From

Table 1

*Primary-Metastasis* and *DLBCL-FL* gene expression data: compared results between fuzzy ensemble clustering methods and other ensembles and "single" clustering algorithms.

Primary Metastasis				DLBCL-FL			
Algorithms	Mean error	Median error	Std. dev.	Algorithms	Mean error	Median error	Std. dev.
Fuzzy-Max	0.2925	0.2763	0.0451	Fuzzy-Max	0.2247	0.2208	0.0141
Fuzzy-Alpha	0.3083	0.2763	0.0613	Fuzzy-Alpha	0.2273	0.2208	0.0131
Max-Max	0.3364	0.3684	0.07816	Max-Max	0.2498	0.2468	0.0228
Max-Alpha	0.3724	0.3684	0.0092	Max-Alpha	0.2633	0.2468	0.0585
Rand-Clust	0.3294	0.3289	0.0024	Rand-Clust	0.1039	0.1039	0
Bagclust1	0.3605	0.3552	0.0066	Bagclust1	0.2957	0.3052	0.0442
Fuzzy "single"	0.3684	0.3684	0	Fuzzy "single"	0.2472	0.2468	0.0024

a general standpoint the proposed fuzzy ensemble methods achieve results competitive with other clustering ensemble methods, by exploiting the redundancy and the high dimension of gene expression data. The results depend on the proper choice of the fuzziness, and on the characteristics of the data sets. Moreover note that the choice of the data is not favourable to our proposed methods, because in all cases we assessed the performance of the clustering algorithms by assuming crisp partitions of the data, according to a bio-medical diagnosis of patients that does not admit fuzzy memberships to different sub-

Table 2

*Melanoma* and *Leukemia* gene expression data: compared results between fuzzy ensemble clustering methods and other ensembles and "single" clustering algorithms.

Melanoma				Leukemia			
Algorithms	Mean error	Median error	Std. dev.	Algorithms	Mean error	Median error	Std. dev.
Fuzzy-Max	0.2925	0.2903	0.0082	Fuzzy-Max	0.262	0.2639	0.006
Fuzzy-Alpha	0.2914	0.2903	0.0059	Fuzzy-Alpha	0.262	0.2639	0.006
Max-Max	0.3097	0.3226	0.0394	Max-Max	0.2639	0.25	0.025
Max-Alpha	0.3073	0.3226	0.0347	Max-Alpha	0.2606	0.2639	0.0668
Rand-Clust	0.4516	0.4516	0	Rand-Clust	0.2657	0.2639	0.0162
Bagclust1	0.4527	0.4516	0.0059	Bagclust1	0.2852	0.2778	0.0289
Fuzzy "single"	0.3226	0.3226	0	Fuzzy "single"	0.25	0.25	0

classes of diseases.

## 5.2 Relationships between accuracy and diversity of the base learners

To understand the behaviour of ensemble methods and the reasons why an ensemble approach is more effective than others, we can study the trade-off between accuracy and diversity of the base clusterings. Indeed in the literature several authors tried to analyze the results of ensemble methods taking into account accuracy and diversity of the base learners [31,32,5].

In particular, we analyze the relationships between accuracy and diversity of the base clusterings, considering both fuzzy ensembles and *Randclust*. In particular we analyzed the *max-\** ensembles, where "star" stands for any fuzzy or crisp consensus clustering (Sect. 3): indeed in this context we are interested only in the characteristics of the base clusterings.

To evaluate the relationships between accuracy and diversity of the base learners we adopt two measures based on the Normalized Mutual Information (*NMI*), proposed in [31] and [5].

We briefly introduce below the definitions of Mutual Information (*MI*) and *NMI* and the measures of accuracy and diversity based in *NMI* used in our experiments. Let be  $X$  and  $Y$  two random variables; the *MI* is defined as:

$$MI(X, Y) = H(X) - H(X/Y) = H(X) - H(X, Y) + H(Y) \quad (8)$$

where  $H(X)$ ,  $H(Y)$  represents respectively the entropy of  $X$  and  $Y$ ,  $H(X/Y)$  is the entropy of  $X$  conditioned to  $Y$  and  $H(X, Y)$  is the joint entropy of  $X$  and  $Y$ . The *NMI* between  $X$  and  $Y$  can be obtained from *MI* and  $H$  in the following way:

$$NMI(X, Y) = \frac{MI(X, Y)}{\sqrt{H(X)H(Y)}} \quad (9)$$

It is easy to see that  $0 \leq NMI(X, Y) \leq 1$ .

According to [31] accuracy and diversity can be evaluated using the *NMI*. Given a clustering ensemble  $\mathcal{C} = \{C_1, \dots, C_n\}$ , composed by  $n$  base clusterings

$C_i$ , the similarity  $s(C_i, C_j)$  between two base clustering  $C_i$  and  $C_j$  can be computed in the following way:

$$s(C_i, C_j) = NMI(C_i, C_j) \quad (10)$$

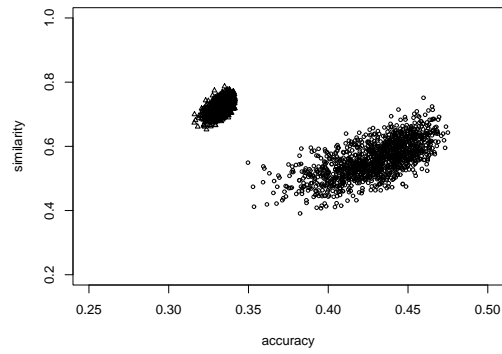
Note that if  $s(C_i, C_j) \simeq 1$  the two clusterings are very similar (and hence their diversity is very low), on the contrary if  $s(C_i, C_j) \simeq 0$  they are strongly diverse.

The average accuracy  $a(C_i, C_j)$  of the same pair  $C_i$  and  $C_j$  of base clusterings can be computed considering the average  $NMI$  of each base clustering with respect to the "true" clustering  $C$ :

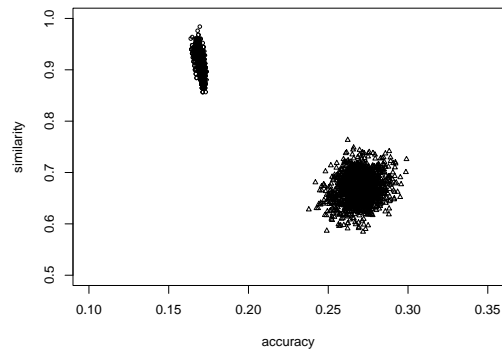
$$a(C_i, C_j) = \frac{NMI(C_i, C) + NMI(C_j, C)}{2} \quad (11)$$

In our experiments we computed the similarity and accuracy for each pair of base clusterings, and we represented their distribution through scatter plots of  $\frac{n*(n-1)}{2}$  points.

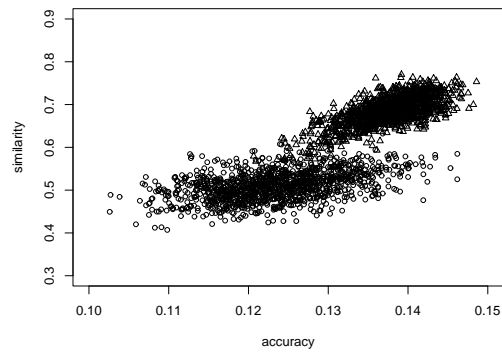
Fig. 4 (a) represents the relationships between accuracy and diversity w.r.t. the *DLBCL-FL* data set; the better results of the *Randclust* ensembles are due to the higher accuracy and higher diversity of its hierarchical base clusterings. Indeed the accuracy (estimated according to eq. 11) is between 0.35 and 0.50, while the accuracy of the base fuzzy k-means clusterings is always below 0.35; moreover also the diversity is quite higher for the *Randclust* ensemble. These results explain the better performance of *Randclust* on this data set (Table 1). The opposite situation can be observed in Fig. 4 (b) (*Melanoma* data set): here the base clusterings of the fuzzy ensemble approach are both more accurate (their average  $NMI$  ranges approximately between 0.25 and 0.30, while in *Randclust* the accuracy is below 0.20), and more diverse (their  $NMI$  in the y axis are between 0.55 and 0.80, while in *Randclust* are above 0.80: recall that a low value of  $NMI$  between base clusterings reveals a high diversity and viceversa). In this case the fuzzy ensembles largely outperform *Randclust* (Table 2). With the *Leukemia* data set the results between *Randclust* and our proposed fuzzy ensemble methods are comparable (Table 2). Indeed the slightly better accuracy of the base clusterings of *max-\** fuzzy ensembles are



(a)



(b)



(c)

Fig. 4. Relationships between accuracy and similarity of the base clusterings. Abscissa represents accuracy and ordinate similarity. Triangles refer to base clusterings of *max-\** ensembles, circles to base clusterings of *Randclust* ensembles. (a) *DLB-CL-FL* (b) *Melanoma* (c) *Leukemia* data sets.

counter-balanced by the higher diversity of the base hierarchical clusterings of *Randclust*, thus resulting in a comparable accuracy of the ensembles (Fig. 4 (c)).

## 6 Conclusions

In this paper we proposed several variants of fuzzy unsupervised ensemble methods to analyze gene expression data. The proposed methods on one hand exploit the accuracy and the effectiveness of the ensemble clustering techniques based on random projections, and on the other hand the expressive capacity of the fuzzy sets, to obtain clustering algorithms both reliable and able to express the uncertainty of the data. In our experiments we applied our proposed method to the analysis of gene expression data, but in principle this approach can be applied to the analysis of any bio-molecular data characterized by high dimensionality (e.g. mass-spectrometry data).

The experimental results show that our proposed fuzzy ensemble approach is competitive with other ensemble methods and it may be successfully applied to the analysis of gene expression data, even when we consider data sets with a single "crisp" label for each example.

As a future application of these methods, we are planning experiments to discover functional classes of genes to exploit the structure of unlabeled data when the boundaries of the clusters are uncertain and to analyze data characterized by multi-labels and with partial membership to different clusters. Indeed genes may belong to different biological processes or different pathways and as a consequence they may belong to different sets of co-expressed genes.

Several open problems need to be considered for future research work. For instance, we may consider the choice of the t-norm to be used in the fuzzy aggregation of multiple clusters. In our experiments we applied the algebraic product, but we need to experiment with other t-norms [33]. Moreover we may experiment with other random projections that obey the *JL* lemma, such as normal or Achlioptas random projections [34,17]. Another possible development of this work consists in studying if we can embed recently proposed stability-based methods based on random projections [35,18] into ensemble clustering methods to guide the construction of the consensus clustering.

## References

- [1] L. Dyrskjøt, T. Thykjaer, M. Kruhøffer, J. Jensen, N. Marcussen, S. Hamilton-Dutoit, H. Wolf, T. Ørntoft, Identifying distinct classes of bladder carcinoma using microarrays, *Nature Genetics* 33 (jan.) (2003) 90–96.
- [2] M. Onken, L. Worley, J. Ehlers, J. Harbour, Gene expression profiling in uveal melanoma reveals two molecular classes and predicts metastatic death, *Cancer Research* 64 (2004) 7205–7209.
- [3] J. Dopazo, Functional interpretation of microarray experiments, *OMICS* 3 (10).
- [4] A. Strehl, J. Ghosh, Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions, *Journal of Machine Learning Research* 3 (2002) 583–618.
- [5] X. Fern, C. Brodley, Random projections for high dimensional data clustering: A cluster ensemble approach, in: T. Fawcett, N. Mishra (Eds.), *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003)*, AAAI Press, Washington D.C., USA, 2003.
- [6] A. Topchy, A. Jain, W. Puch, Clustering Ensembles: Models of Consensus and Weak Partitions, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (12) (2005) 1866–1881.
- [7] L. Kuncheva, D. Vetrov, Evaluation of stability of k-means cluster ensembles with respect to random initialization, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28 (11) (2006) 1798–1808.
- [8] S. Swift, A. Tucker, V. Vinciotti, N. Martin, C. Orengo, X. Liu, P. Kellam, Consensus clustering and functional interpretation of gene-expression data, *Genome Biology* 5:R94.
- [9] S. Dudoit, J. Fridlyand, Bagging to improve the accuracy of a clustering procedure, *Bioinformatics* 19 (9) (2003) 1090–1099.
- [10] S. Monti, P. Tamayo, J. Mesirov, T. Golub, Consensus Clustering: A Resampling-based Method for Class Discovery and Visualization of Gene Expression Microarray Data, *Machine Learning* 52 (2003) 91–118.

- [11] X. Hu, I. Yoo, Cluster ensemble and its applications in gene expression analysis, in: Proc. 2nd Asia-Pacific Bioinformatics Conference, Dunedin, New-Zealand, 2004, pp. 297–302.
- [12] T. Grotkjaer, O. Winther, B. Regenberg, J. Nielsen, L. Hansen, Robust multi-scale clustering of large DNA microarray data sets with the consensus algorithm, *Bioinformatics* 22 (1) (2006) 58–67.
- [13] Z. Yu, H. Wong, H. Wang, Graph based consensus clustering for class discovery from gene expression data, *Bioinformatics* Advance Access published online on September 14, 2007.
- [14] L. Breiman, Bagging predictors, *Machine Learning* 24 (2) (1996) 123–140.
- [15] A. Bertoni, G. Valentini, Ensembles based on random projections to improve the accuracy of clustering algorithms, in: *Neural Nets, WIRN 2005*, Vol. 3931 of *Lecture Notes in Computer Science*, Springer, 2006, pp. 31–37.
- [16] A. Bertoni, G. Valentini, Randomized embedding cluster ensembles for gene expression data analysis, in: *SETIT 2007 - IEEE International Conf. on Sciences of Electronic, Technologies of Information and Telecommunications*, Hammamet, Tunisia, 2007.
- [17] A. Bertoni, G. Valentini, Randomized maps for assessing the reliability of patients clusters in DNA microarray data analyses, *Artificial Intelligence in Medicine* 37 (2) (2006) 85–109.
- [18] A. Bertoni, G. Valentini, Model order selection for bio-molecular data clustering, *BMC Bioinformatics* 8 (Suppl.3).
- [19] P. Gasch, M. Eisen, Exploring the conditional regulation of yeast gene expression through fuzzy k-means clustering, *Genome Biology* 3 (11).
- [20] R. Bellman, *Adaptive Control Processes: a Guided Tour*, Princeton University Press, New Jersey, 1961.
- [21] W. Johnson, J. Lindenstrauss, Extensions of Lipschitz mapping into Hilbert space, in: *Conference in modern analysis and probability*, Vol. 26 of *Contemporary Mathematics*, Amer. Math. Soc., 1984, pp. 189–206.
- [22] D. Achlioptas, Database-friendly random projections., in: P. Buneman (Ed.), *Proc. ACM Symp. on the Principles of Database Systems*, Contemporary Mathematics, ACM Press, New York, NY, USA, 2001, pp. 274–281.

- [23] E. Bingham, H. Mannila, Random projection in dimensionality reduction: Applications to image and text data, in: Proc. of KDD 01, ACM, San Francisco, CA, USA, 2001.
- [24] M. Smolkin, D. Gosh, Cluster stability scores for microarray data in cancer studies, *BMC Bioinformatics* 36 (4).
- [25] T. Ho, The random subspace method for constructing decision forests, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (8) (1998) 832–844.
- [26] M. Shipp, K. Ross, P. Tamayo, A. Weng, J. Kutok, R. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. Pinkus, T. Ray, M. Koval, K. Last, A. Norton, T. Lister, J. Mesirov, D. Neuberg, E. Lander, J. Aster, T. Golub, Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning., *Nature Medicine* 8 (1) (2002) 68–74.
- [27] S. Ramaswamy, K. Ross, E. Lander, T. Golub, A molecular signature of metastasis in primary solid tumors, *Nature Genetics* 33 (2003) 49–54.
- [28] T. Golub, et al., Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science* 286 (1999) 531–537.
- [29] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, V. Sondak, Molecular classification of malignant melanoma by gene expression profiling, *Nature* 406 (2000) 536–540.
- [30] J. Ward, Hierarchical grouping to optimize an objective function, *J. Am. Stat. Assoc.* 58 (1963) 236–244.
- [31] T. Dietterich, An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization, *Machine Learning* 40 (2) (2000) 139–158.
- [32] S. Hadjitodorov, L. Kuncheva, L. Todorova, Moderate Diversity for Better Cluster Ensembles, *Information Fusion* 7 (3) (2006) 264–275.
- [33] E. P. Klement, R. Mesiar, E. Pap, *Triangular Norms*, Kluwer Academic, 2000.

- [34] D. Achlioptas, Database-friendly random projections: Johnson-lindenstrauss with binary coins, *Journal of Comp. & Sys. Sci.* 66 (4) (2003) 671–687.
- [35] G. Valentini, Clusterv: a tool for assessing the reliability of clusters discovered in DNA microarray data, *Bioinformatics* 22 (3) (2006) 369–370.