Fuzzy ensemble clustering for DNA microarray data analysis

Roberto Avogadri and Giorgio Valentini

DSI, Dipartimento di Scienze dell' Informazione,
Università degli Studi di Milano,
Via Comelico 39, 20135 Milano, Italia.
{avogadri,valentini}@dsi.unimi.it

Abstract. 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 classes of examples or classes of functionally related genes are sometimes not clearly defined. To face these items, we propose a fuzzy ensemble clustering approach to both improve the accuracy of clustering results and to take into account the inherent fuzziness of biological and bio-medical gene expression data. Preliminary results with DNA microarray data of lymphoma and adenocarcinoma patients show the effectiveness of the proposed approach.

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 and the detection of subsets of co-expressed genes as a proxy of co-regulated genes [1]. Different unsupervised ensemble approaches have been proposed to improve the accuracy and the reliability of clustering results [2, 3, 4]. In bioinformatics applications, recently proposed methods based on random projections [5] have been also successfully applied to gene expression data analysis [6].

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 [7]. The main idea of this work is to combine the accuracy and the effectiveness of the ensemble clustering techniques based on random projections [5], 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 section we briefly introduce random projections, then we present our proposed fuzzy ensemble clustering method, and we show some preliminary results with two DNA microarray data sets.

2 Random projections.

Our proposed method perturb the original data using random projections $\mu : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ from high d-dimensional spaces to lower d'-dimensional subspaces.

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 $dist_\mu$:

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

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 ϵ-*distortion* level is introduced.

It has been shown that using random projections that obey *Johnson-Lindenstrauss (JL) lemma* [8] we may perturb the data introducing only bounded distortions, approximately preserving the metric structure of the original data (see [9] for more details). Examples of random projections related with the JL Lemma can be found in [9, 5].

3 Fuzzy ensemble clustering based on random projections

The general structure of the algorithm is similar to the one proposed in [5]: 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. The main difference of our proposed method 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 [2]. 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.

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:

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

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 are clusters and the columns examples), and finally τ 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 application of the fuzzy k-means clustering algorithm are aggregated simply by averaging: in this way we achieve the cumulative similarity matrix M^C:

$$M^C_{i,j} = \frac{1}{c} \sum_{t=1}^{c} M^{(t)}_{i,j}; \tag{3}$$

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 obtained by choosing one of two classical "crispization" techniques:

Hard-clustering:

$$\chi^H_{ri} = \begin{cases} 1 \;\; \Leftrightarrow \text{arg } max_s \; \mathcal{U}^C_{si} = r \\ 0 \text{ otherwise.} \end{cases} \tag{4}$$

α-cut:

$$\chi^\alpha_{ri} = \begin{cases} 1 \;\; \Leftrightarrow \mathcal{U}^C_{ri} \geq \alpha \\ 0 \text{ otherwise.} \end{cases} \tag{5}$$

where χ_{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 $X = \{x_1, x_2, \ldots, x_n\}$, 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 the 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 $i, j \in \{1, \ldots, n\}$ do $M_{ij} = 0$
    (2) Repeat for $t = 1$ to $c$
        (3) $R_t =$ 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 $i, j \in \{1, \ldots, n\}$
            $M_{ij}^{(t)} = \sum_{s=1}^{k} \tau(\mathcal{U}_{si}^{(t)}, \mathcal{U}_{sj}^{(t)})$
    end repeat
    (7)$M^C = \frac{\sum_{t=1}^{c} M^{(t)}}{c}$
    (8) $< A_1, A_2, \ldots, A_k >= \mathcal{C}_f(M^C, k, m)$
end algorithm.
Output:
- the final clustering $C =< A_1, A_2, \ldots, A_k >$
- 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 μ [5], 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).

4 Experimental results

4.1 Experimental environment

We considered two 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 [10]. The second one, the *Primary-Metastasis* (PM) data set, contains expression values in Affymetrix's scaled average difference units for 64 primary adenocarcinomas and 12 metastatic adenocarcinomas (lung, breast, prostate, colon, ovary, and uterus) from unmatched patients prior to any treatment [11]. In both cases we followed the same preprocessing and normalization steps described in [10] and [11].

For each ensemble we randomly repeated the randomized projections 20 times, and each time we built fuzzy ensembles composed by 20 base clusterings (choosing projections with bounded 1 ± 0.2 distortion, according to the

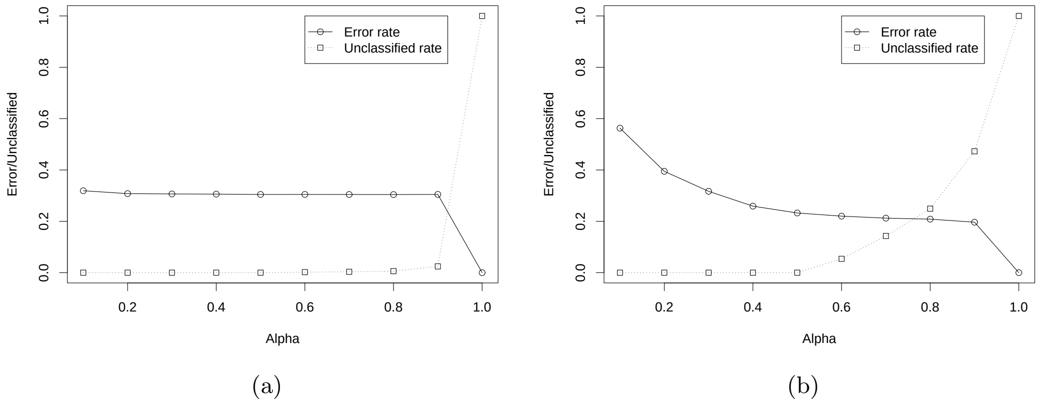

Fig. 1. *Fuzzy-Alpha* ensemble clustering error and unclassified rate with respect to α. (a) *Primary-Metastasis*; (b) *DLBCL-FL* data sets.

JL lemma). We compared results with corresponding "crisp" ensemble methods based on random projections proposed in [5] and with "single" clustering algorithms (hierarchical clustering and fuzzy-k-means).

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:

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

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

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

with t the "real" label of the sample x, $Y \in \mathcal{Y}$ and $\{\lambda\}$ is the empty set. 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.

4.2 Results

To test the performance of the ensemble fuzzy algorithms proposed in this paper, we compare the results of two versions of the proposed fuzzy ensemble clustering method with other types of clustering algorithms. The two version are the *fuzzy-max* ensemble clustering, where the "defuzzifaction" of the consensus clustering is obtained through hard clustering (eq. 4), and the *fuzzy-alpha* ensemble

Table 1. *Primary-metastasis* gene expression data: compared results between fuzzy ensemble clustering methods (Fuzzy-Max and Fuzzy-Alpha) and other ensemble and "single" clustering algorithms.

Algorithms	**Median error**	**Std. Dev.**
Fuzzy-Max	0.2763	0.0477
Fuzzy-Alpha	0.2763	0.0560
Rand-Clust	0.3289	0.0088
Fuzzy "single"	0.3684	–
Hierarchical "single"	0.3553	–

clustering, where the final consensus clustering is "crispized" through the α-cut operation (eq. 5). The other clustering algorithms considered for comparison are *Rand-clust*, a crisp ensemble algorithm based on random projections proposed in [5], and other "single" clustering algorithms (hierarchical agglomerative and fuzzy k-means).

The tables 1 and 2 show the compared numerical results of the experiments on the PM data set and the DLBCL-FL data set respectively. Fuzzy ensemble methods obtain better results with respect to the other methods (considering the median error, see Tab. 1 and 2). Anyway note that the larger standard deviation (with respect to the *Rand-clust* ensemble algorithm) denotes a higher instability of the fuzzy approach, and with the *DLBCL-FL* data set *Fuzzy-Alpha* achieves significantly worse results than *Fuzzy-Max* and *Rand-clust* ensemble methods.

The graphics 1 (a) and 1 (b) represent the performance of the "fuzzy-alpha" ensemble algorithm (error rate and unclassified rate for every level of alpha-cut analyzed). The figure shows that we may obtain acceptable results with the *Fuzzy-Alpha* method too if we accept a certain rate of unclassified examples (Fig. 1(b)).

5 Conclusions

The experimental results show that our proposed fuzzy ensemble approach may be successfully applied to the analysis of gene expression data, even when we

Table 2. *DLBCL-FL* gene expression data: compared results between fuzzy ensemble clustering methods (Fuzzy-Max and Fuzzy-Alpha) and other ensemble and "single" clustering algorithms.

Algorithms	**Median error**	**Std. Dev.**
Fuzzy-Max	0.0779	0.1163
Fuzzy-Alpha	0.2727	0.1142
Rand-Clust	0.1039	0.0023
Fuzzy "single"	0.2987	–
Hierarchical "single"	0.1039	–

consider data sets with a single certain label for each example. Nevertheless we know that genes may belong to different biological processes or different pathways and as a consequence they may belong to different sets of co-expressed genes. We are planning new experiments with multi-label genes or examples to show more clearly the effectiveness of the proposed approach and to analyze the structure of unlabeled data when the boundaries of the clusters are uncertain.

References

[1] Dopazo, J.: Functional interpretation of microarray experiments. OMICS **3** (2006)

[2] Dudoit, S., Fridlyand, J.: Bagging to improve the accuracy of a clustering procedure. Bioinformatics **19** (2003) 1090–1099

[3] Fern, X., Brodley, C.: Random projections for high dimensional data clustering: A cluster ensemble approach. In Fawcett, T., Mishra, N., eds.: Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), Washington D.C., USA, AAAI Press (2003)

[4] Topchy, A., Jain, A., Puch, W.: Clustering Ensembles: Models of Consensus and Weak Partitions. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 1866–1881

[5] Bertoni, A., Valentini, G.: Ensembles based on random projections to improve the accuracy of clustering algorithms. In: Neural Nets, WIRN 2005. Volume 3931 of Lecture Notes in Computer Science., Springer (2006) 31–37

[6] Bertoni, A., Valentini, G.: 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)

[7] Gasch, P., Eisen, M.: Exploring the conditional regulation of yeast gene expression through fuzzy k-means clustering. Genome Biology **3** (2002)

[8] Johnson, W., Lindenstrauss, J.: Extensions of Lipshitz mapping into Hilbert space. In: Conference in modern analysis and probability. Volume 26 of Contemporary Mathematics., Amer. Math. Soc. (1984) 189–206

[9] Bertoni, A., Valentini, G.: Randomized maps for assessing the reliability of patients clusters in DNA microarray data analyses. Artificial Intelligence in Medicine **37** (2006) 85–109

[10] Shipp, M. et al.: Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine **8** (2002) 68–74

[11] Ramaswamy, S., Ross, K., Lander, E., Golub, T.: A molecular signature of metastasis in primary solid tumors. Nature Genetics **33** (2003) 49–54