An unsupervised fuzzy ensemble algorithmic scheme for gene expression data analysis

Roberto Avogadri[1], Giorgio Valentini*[1]

[1]DSI, Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano,Via Comelico 39, 20135 Milano, Italy

Email: Roberto Avogadri - avogadri@dsi.unimi.it; Giorgio Valentini*- valentini@dsi.unimi.it;

*Corresponding author

Abstract

Background: In recent years unsupervised ensemble clustering methods have been successfully applied to DNA microarray data analysis to improve the accuracy and the reliability of clustering results. Nevertheless, a major problem is represented by the fact that classes of functionally correlated examples (e.g. subclasses of diseases characterized at bio-molecular level) are not in general clearly separable, and in many cases the same gene may belong to different functional classes (e.g. may participate to different biological processes).

Results: We propose an ensemble clustering algorithm scheme, based on a fuzzy approach, that directly permit to deal with overlapping classes or with genes or samples that may belong to more clusters at the same time. From our algorithmic scheme several fuzzy ensemble clustering algorithms may be derived, according to the way the multiple clusterings are combined and the consensus clustering is generated. We test some of the proposed ensemble algorithms with two DNA microarray data sets available on the web, comparing the results with other single and ensemble clustering methods.

Conclusions: Our proposed fuzzy ensemble approach may be applied to discover classes of co-expressed genes or subclasses of functionally related examples, and in principle it may be applied for the unsupervised analysis of different types of complex bio-molecular data. Fuzzy ensemble algorithms can assign each gene/sample to multiple classes and can estimate and improve the accuracy and the reliability of the discovered clusterings, as shown by our experimental results.

Background

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 [1–3]. Different unsupervised ensemble methods have been proposed to improve the accuracy and the reliability of clustering results in bioinformatics applications [4–7].

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 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 [8]. The main idea of this work is to combine the accuracy and the effectiveness of the ensemble clustering techniques with the expressive capacity of the fuzzy sets, to obtain clustering algorithms both reliable and able to express the uncertainty of the data.

Methods

We propose a fuzzy ensemble algorithmic scheme, from which different ensemble clustering algorithms may be derived. At first we perturb the data according to a given perturbation procedure: resampling or noise injection techniques could be in principle applied, but we

choose random projections to lower dimensional subspaces [9, 10] in order to exploit the high dimensionality of DNA microarray data. Then, multiple fuzzy k-means clusterings [11] 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. The obtained multiple clusterings are successively aggregated using a fuzzy approach. According to the way the multiple clustering are combined, the ”consensus” ensemble clustering may result in crisp clusters that may overlap or in a fuzzy partition by which each example may belong to each cluster with a certain fuzzy membership.
The structure of the *fuzzy ensemble algorithmic scheme* can be summarized as follows:

1. *Random projections*: generation of multiple instances (views) of the data through random projections

2. *Generation of multiple fuzzy clusterings*: the fuzzy k-means algorithm is applied to the projected 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. *“Crispization” of the base clusterings.* This step is executed if a ”crisp” aggregation is performed. The fuzzy clusterings obtained in the previous step can be “defuzzified” through *hard-clustering* techniques, by which each example is assigned to the cluster with the largest membership, or through α*-cut* techniques, where an example is assigned to a cluster only if its membership is larger than a prefixed value α.

4. *Aggregation.* If a fuzzy aggregation is performed, the base clusterings are combined, using a square similarity matrix [4] whose elements are generated through fuzzy t-norms applied to the membership functions of each pair of examples. If a crisp aggregation is performed, the similarity matrix is built using the product of the characteristic function between each pair of examples.

5. *Clustering in the ”embedded” similarity space.* The similarity matrix induces a new representation of the data based on the pairwise similarity between pairs of examples: the fuzzy k-means clustering algorithm is applied to the rows (or equivalently to the columns) of the similarity matrix.

6. *Consensus clustering.* The consensus clustering could be represented by the overall consensus membership matrix, resulting in a fuzzy representation of the consensus clustering. Alternatively, we may apply the same crispization techniques used at step 3 to transform the fuzzy consensus clustering to a crisp one.

We may observe that considering the possibility of applying crisp or fuzzy approaches at steps 3, 4 and 6, we can obtain 9 different algorithms, exploiting different combinations of aggregation and consensus clustering techniques. For instance, combining a fuzzy aggregation with a consensus clustering obtained trough α-cut we obtain from the algorithmic scheme a *fuzzy-alpha* ensemble clustering algorithm, while using a hard-clustering crispization technique for aggregation and a fuzzy consensus we obtain a *max-fuzzy* ensemble clustering. The two names separated by a hyphen (e.g. *fuzzy-alpha*) refer respectively to the type of aggregation and consensus steps. As an example of the fuzzy ensemble algorithms that may be derived from the general algorithmic scheme, we provide here the high-level pseudo-code of the algorithm based on fuzzy aggregation and hard-clustering consensus (*fuzzy-max* clustering ensemble algorithm).

Fuzzy-max ensemble clustering:

```
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 $m$ (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 $m$
```

```
      (3) $R_t =$ Generate_projection_matrix $(d', \mu)$
      (4) $D_t = R_t \cdot D$
      (5) $\mathcal{U}^{(t)} = \mathcal{C}_f(D_t, k, z)$
      (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)$M^C = \frac{\sum_{t=1}^{c} M^{(t)}}{m}$
   (8) $< A_1, A_2, \ldots, A_k >= \mathcal{C}_f(M^C, k, z)$
end algorithm.
Output:
- the final clustering $C =< A_1, A_2, \ldots, A_k >$
- the cumulative similarity matrix $M^C$.
```

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 [9], 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 z 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* (e.g. the algebraic product) (step 6). Note that $\mathcal{U}$ is a fuzzy membership matrix (where the rows are clusters and the columns examples). 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 and by assigning each example to the cluster with the maximum membership (step 8).

Table 1: Primary-metastasis gene expression data: compared results between fuzzy ensemble clustering methods (Fuzzy-Max, Fuzzy-Alpha, Max-Max and Max-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
Max-Max	0.3684	0.0854
Max-Alpha	0.3684	0.0910
Rand-Clust	0.3289	0.0088
Fuzzy ”single”	0.3684	–
Hierarchical ”single”	0.3553	–

Results

To show the effectiveness of the proposed approach, we analyzed two DNA microarray data sets available on the web: the *DLBCL-FL* data set, composed by tumor specimens from 58 Diffuse Large B-Cell Lymphoma (DLBCL) and 19 Follicular Lymphoma (FL) patients [12]; the *Primary-Metastasis* (PM) data set, that 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 [13]. For each ensemble method we randomly repeated the randomized projections 20 times, and each time we built fuzzy ensembles composed by 20 base clusterings. Since clustering does not univocally associate a label to the examples we evaluated the error by choosing for each clustering the permutation of the classes that best matches the ”a priori” known ”true” classes.

We compared our proposed *fuzzy-max* and *fuzzy-alpha* ensemble clusterings, both characterized by a fuzzy aggregation with *max-alpha* and *max-max* methods, both characterized by a crisp aggregation of multiple clusterings. As baseline methods we considered ”crisp” ensemble methods based on random projections (*Rand-clust*) [9], and ”single” clustering algorithms (hierarchical clustering and fuzzy k-means).

The Tables 1 and 2 show the compared numerical results of the experiments on the *PM* and *DLBCL-FL* data sets respectively. Fuzzy ensemble methods obtain significantly better

Table 2: DLBCL-FL gene expression data: compared results between fuzzy ensemble clustering methods (Fuzzy-Max, Fuzzy-Alpha, Max-Max and Max-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
Max-Max	0.2987	0.0157
Max-Alpha	0.2987	0.0127
Rand-Clust	0.1039	0.0023
Fuzzy ”single”	0.2987	–
Hierarchical ”single”	0.1039	–

results with respect to the other methods (considering the median error). 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. It is also worth noting that the fuzzy ensemble approach significantly outperforms ”single” fuzzy k-means runs. Moreover results with the *Max-max* and *Max-alpha* ensemble clustering algorithms, where a crispization step is performed in the aggregation phase (see the algorithmic scheme), show that the fuzzy aggregation is essential to improve the accuracy of clustering results (Table 1 and 2).

Conclusions

We proposed a fuzzy ensemble clustering algorithmic scheme from which several ensemble clustering methods may be derived. By this approach we can identify clusters of genes/examples characterized by uncertain boundaries, or we can assign gene/examples to multiple clusters, according to the characteristics of gene expression data. Results with DNA microarray data show the effectiveness of the proposed approach, but fuzzy ensemble clustering methods may be also applied to the unsupervised analysis of other types of complex bio-molecular data.

References

1. Dyrskjøt L, Thykjaer T, Kruhøffer M, Jensen J, Marcussen N, Hamilton-Dutoit S, Wolf H, Ørntoft T: **Identifying distinct classes of bladder carcinoma using microarrays**. *Nature Genetics* 2003, **33**(jan.):90–96.

2. Onken M, Worley L, Ehlers J, Harbour J: **Gene Expression Profiling in Uveal Melanoma Reveals Two Molecular Classes and Predicts Metastatic Death**. *Cancer Research* 2004, **64**:7205–7209.

3. Dopazo J: **Functional Interpretation of Microarray Experiments**. *OMICS* 2006, **3**(10).

4. Dudoit S, Fridlyand J: **Bagging to improve the accuracy of a clustering procedure**. *Bioinformatics* 2003, **19**(9):1090–1099.

5. Monti S, Tamayo P, Mesirov J, Golub T: **Consensus Clustering: A Resampling-based Method for Class Discovery and Visualization of Gene Expression Microarray Data**. *Machine Learning* 2003, **52**:91–118.

6. Grotkjaer T, Winther O, Regenberg B, Nielsen J, Hansen L: **Robust multi-scale clustering of large DNA microarray data sets with the consensus algorithm**. *Bioinformatics* 2006, **22**:58–67.

7. 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.

8. Gasch P, Eisen M: **Exploring the conditional regulation of yeast gene expression through fuzzy k-means clustering**. *Genome Biology* 2002, **3**(11).

9. 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.

10. Bertoni A, Valentini G: **Model order selection for bio-molecular data clustering**. *BMC Bioinformatics* 2007, **8**(Suppl.3).

11. Bezdek J: *Pattern Recognition with Fuzzy Objective Function Algorithms.* New York: Plenum 1981.

12. Shipp M, Ross K, Tamayo P, Weng A, Kutok J, Aguiar R, Gaasenbeek M, Angelo M, Reich M, Pinkus G, Ray T, Koval M, Last K, Norton A, Lister T, Mesirov J, Neuberg D, Lander E, Aster J, Golub T: **Diffuse large B-cell Lymphoma outcome prediction by gene-expression profiling and supervised machine learning.** *Nature Medicine* 2002, **8**:68–74.

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