Clustering Genomic Data in the Membership Embedding Space

Francesco Masulli
INFM and Dept of Computer Science
University of Pisa
Largo B. Pontecorvo, 3 I-56125 Pisa, Italy
masulli@di.unipi.it

Stefano Rovetta and Maurizio Filippone
INFM and Department of Computer and
Information Sciences
University of Genova
Via Dodecaneso 35 I-16146 Genova, Italy
{rovetta,filippone}@disi.unige.it

Abstract

In this paper we proposed a method to face clustering problems due to small cardinality and high dimensionality that are typical of many genomic data. We use an embedding space where each data point is represented by a vector containing memberships to fuzzy sets centered on a set of probes selected from the data base. The proposed approach leads to significant improvements with respect the application of clustering algorithms in the original space and in the distance embedding space

I. Introduction

Often genomic data, such as those of gene expression obtained from DNA microarrays, are characterized by small cardinality and high dimensionality. Clustering methods provide an useful tool to explore those data, but the crude application of classical clustering algorithms leads to poor results. In fact, many clustering approaches suffer from being applied in high-dimensional spaces, as clustering algorithms often seek for areas where data is especially dense. However it is often the case that the cardinality of the data sets available is not only small with respect to the size of the data space, which would lead to insufficient sampling of the space: sometimes it is even less than the number of variables. This means that the data span only a subspace within the data space. In these conditions, it is not easy to define the concept of volumetric density.

A further problem is again related to distances in high space dimensionality. when space dimensionality is high or even moderate (as low as 10-15), the distance of a point to its farthest neighbor and to its nearest neighbor tend to become equal [3], [1]. Therefore the evaluation of distances, and the concept of “nearest neighbor” itself, become less and less meaningful with growing dimension. Defining clusters on the basis of distance requires that distances can be estimated. For instance, one of the most common methods, c-means (CM) clustering, is based on iteratively computing distances and cluster averages. Increasing the data space dimensionality may introduce a large number of suboptimal solutions (local minima), and the nearest-neighbor criterion which is the basis of the method may even become useless. This problem is not avoided even when CM is modified in the direction of incorporating fuzzy concepts, e.g. as for the FCM (Fuzzy c-Means) algorithm [5], [2].

II. Representations in embedding spaces

A notable complexity reduction in the presence of large-dimensional data sets can be provided by representations in an embedding space based on mutual distances between points. If the cardinality of the data set is small compared to the input space dimensionality, then the matrix of mutual distances or other pairwise pattern evaluation methods such as kernels [10] may be used to represent data sets in a more compact way. Pękalska and Duin [9] have developed a set of methods based on representing each pattern according to a set of similarity measurements with respect to other patterns in the data set.

Following this approach, the data matrix is replaced by a pairwise dissimilarity matrix D. Let X be a data set of cardinality n.

$$X = \{x_1, x_2, \ldots, x_n\} \tag{1}$$

We start by computing the dissimilarity matrix D:

$$d_{ik} = d(x_i, x_k) \ \ \forall i, k \tag{2}$$

according to an assigned dissimilarity measure $d(x, y)$ between points x and y (e.g., using Euclidean distance).

Applications of projection into dissimilarity embedding spaces to clustering are reported in [6], [8].

As pointed out in [9], the dissimilarity measure should be a metric, since metrics preserve the *reverse of the compactness hypothesis* [9]: ”objects that are similar in their representation are also similar in reality and belong, thereby, to the same class”. Often non-metric distances are used as well.

Let we consider now the Euclidean distance as the dissimilarity measure. In case of a data set with points in general position and dimensionality of the original data set N there is the upper bound of $N+1$ probes (or support data) that we can use in order to build the dissimilarity matrix. This upper bound is often un-realistic (as in the case of genomic data), but for data having some structure we only require that the dimension of the embedding space is large enough to preserve the reverse of the compactness hypothesis.

Note that, if the embedding dimension n is lower than $N + 1$, some points could have an ambiguous representation and, moreover, clustering will be affected by the high metrical contribution of farthest points.

III. Membership Embedding Space

In order to avoid the problems highlighted in the previous sections, we propose a different kind of embedding based on the space of memberships to fuzzy sets centered on the probes.

Following this approach in the embedding space a point will be represented by a vector containing only few non-null components (depending on the width of the membership function), in correspondence of the closer probes in the original feature space.

In our experiments, the memberships of fuzzy sets centered on the probes were modeled using the following normalized function:

$$\nu_{ik} = \frac{\exp\left[-\beta d_{i,k}^2\right]}{\sum_l \exp\left[-\beta d_{i,l}^2\right]} \tag{3}$$

Probes were extracted at random (using an uniform pdf) from the data set without replacement.

Note that, using this membership function, the parameter β regulates the spread of the membership function and can be related to the average distance between the data points. Its value must be selected in order to improve the overall result (model selection).

In the Membership Embedding Space each data point x_i is represented as a row of ν_{ik}:

$$x_i = (\nu_{i1}, \nu_{i2}, \ldots, \nu_{in}) \tag{4}$$

IV. Fuzzy C-Means Algorithm

Let we consider the following fuzzy C-Means Functional:

$$J_m(\mathbf{U}, Y) = \sum_{i=1}^{n} \sum_{k=1}^{c} (u_{ik})^m E_k(x_i) \tag{5}$$

where

- $X = \{x_1, x_2, \ldots, x_n\}$ is a data set containing n unlabeled sample points;
- $Y = \{y_1, y_2, \ldots, y_c\}$ is the set of the centers of clusters;
- $E_k(x_i)$ is a dissimilarity measure (distance or cost) between data point x_i and the center y_k of a specific cluster k (e.g., $E_k(x_i) = \|x_i - y_k\|^2$);
- $\mathbf{U} = [u_{ik}]$ is the $c \times n$ fuzzy c-partition matrix, containing the membership values of all samples to all clusters;
- $m \in (1, \infty)$ is the fuzziness control parameter.

The clustering problem can be formulated as the minimization of J_m with respect to Y, under the normalization

$$\sum_{k=1}^{c} u_{ik} = 1. \tag{6}$$

The Fuzzy C-Means (FCM) algorithm proposed by Bezdek [2] starts with random initialization of the fuzzy c-partition matrix $\mathbf{U}$ (or of the centroids y_k) and then iterates until convergence the following Eq.s 7 and 8:

$$y_k = \frac{\sum_{i=1}^{n} (u_{ik})^m x_i}{\sum_{i=1}^{n} (u_{ik})^m} \qquad \text{for all } k, \tag{7}$$

$$u_{ik} = \begin{cases} \left(\sum_{l=1}^{c} \frac{E_k(x_i)}{E_l(x_i)}\right)^{\frac{2}{1-m}} & \text{if } E_k(x_i) > 0\ \forall k, i; \\ 1 \quad \text{if} \quad E_k(x_i) = 0 & \text{and} \quad u_{il} = 0\ \ \forall\ l \neq k \end{cases} \tag{8}$$

It is worth to underline that if one chooses $m = 1$ the fuzzy C-Means Functional J_m (Eq. (5)) reduces to the expectation of the C-Means global error (that we shall denote as $< E >$):

$$< E >= \sum_{i=1}^{n} \sum_{k=1}^{c} u_{ik} E_k(x_i), \tag{9}$$

and the FCM becomes the classic crisp C-Means algorithm [4].

V. Experimental setup

The method was tested on the publicly available Leukemia data by Golub et al. [7]. The Leukemia problem consists in characterizing two forms of acute leukemia, Acute Lymphoblastic Leukemia (ALL) and Acute Mieloid Leukemia (AML). The original work proposed both a supervised classification task ("class prediction") and an unsupervised characterization task ("class discovery"). Here we obviously focus on the latter, but we exploit the diagnostic information on the type of leukemia to assess the goodness of the clustering obtained.

The training data set contains 38 samples for which the expression level of 7129 genes has been measured with the DNA microarray technique (the interesting human genes are 6817, and the other are controls required by the technique). These expression levels have been scaled by a factor of 100. Of these samples, 27 are cases of ALL and 11 are cases of AML. Moreover, it is known that the ALL class is in reality composed of two different diseases, since they are originated from different cell lineages (either T-lineage or B-lineage). In the data set, ALL cases are the first 27 objects and AML cases are the last 11. Therefore, in the presented results, the object identifier can also indicate the class (ALL if id $\leq$ 27, AML if larger).

We have performed an extended experimentation comparing the following approaches:

1) FCM on the original data set (RD);
2) FCM in the Distance Embedding Space (DES) with different probe/data ratios;
3) FCM in the Membership Embedding Space (MES) with different probe/data ratios.

Each experiment corresponds to 1000 independent trials, each of them using a different random initialization of the membership in the FCM algorithm.

In all trials, the number of clusters was set to 3, and the fuzziness parameter m of FCM was set to 2.

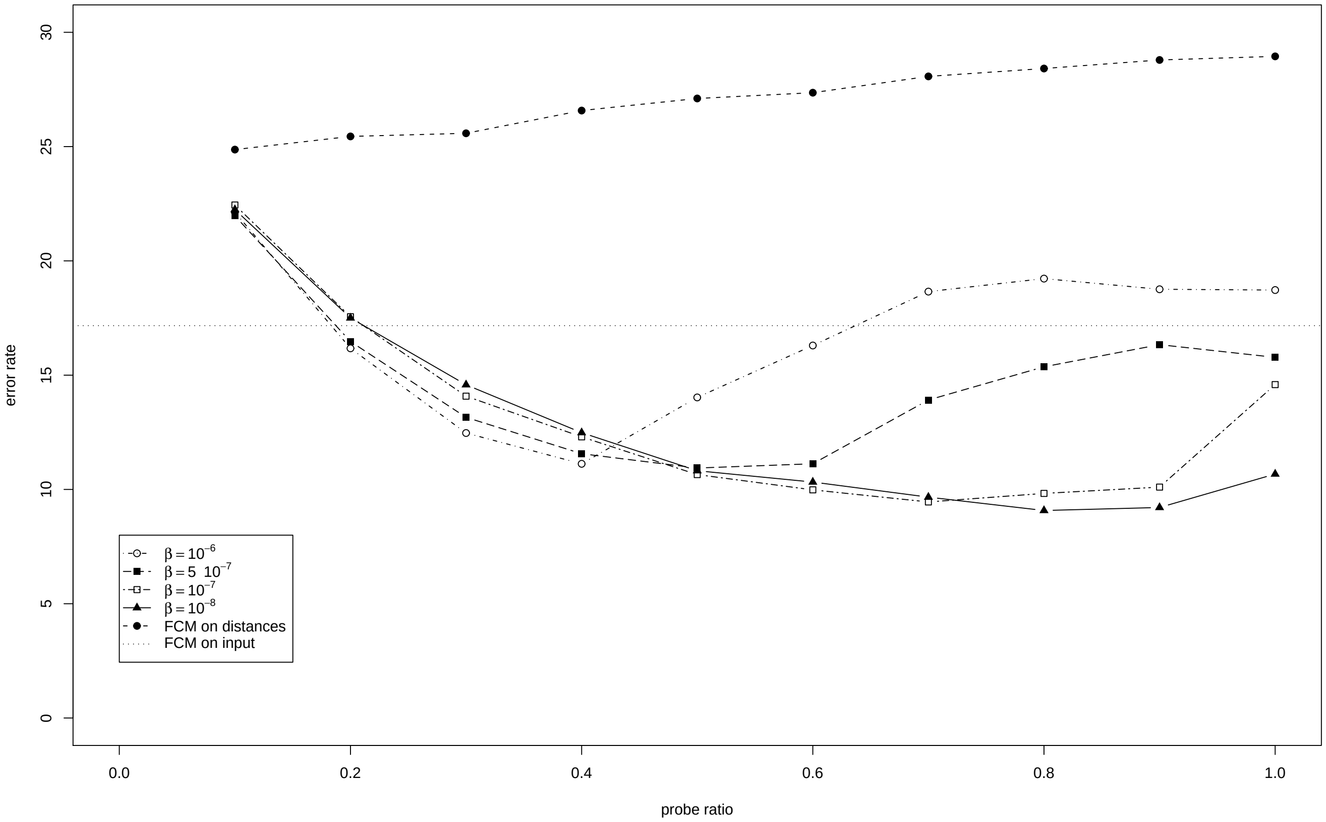

Fig. 1. Error rate for the tested methods.

VI. Results

In Fig. 1 we show the error rate versus the probe/data ratio averaged over 1000 independent trials for each experimental point.

The first approach (standard FCM on original data) obtains a mean error rate of 17.2%.

The projection into the distance embedding space (second approach) leads to worse results than the previous one: the error rate is more than 25.0% for all probe/data ratios in the range $[.1, 1.0]$.

The last approach, projecting the data set into the membership embedding space, leads to better results.

Moreover, increasing the parameter β from 10^{-8} to 10^{-6} we obtain for increasing probe/data ratios (from $.8$ to $.4$) the shift of the optimal error ratio.

The average distance between the data points is 10^6.

A reasonable choice is then to take $\beta = 10^{-8}$ that is about one hundred times the inverse of the average distance between the data points.

The membership vectors (the rows of the ν_{ik} matrix) have a number of non null components related to the spread of the membership function. The minimum of the error rate is achieved for situations for which we have a good compromise between the number of probes and the width of the membership function.

A comparison of the best mean error rate for the tested methods is reported in Tab. I.

Finally we made model selection on the value of the fuzziness m calculating the mean error over 1000 trials.

In fig. 2 we can see that the best value for the m parameter of FCM algorithm is $m = 1.8$ for which we obtain an error rate equal to 8.8%. We can see that for $m > 2$ we obtain a rapid increasing in the error rate. So increasing the fuzziness when the membership function are wider (for low values of β), leads to worse situation in terms of local minima for FCM. On the other hand for low values of m, FCM tends to behave like C-Means algorithm and in this case performs worse than FCM.

VII. Conclusions

Clustering methods provide an useful tool to explore genomic data, but often they get poor results due to small cardinality and high dimensionality of data sets.

In this paper we proposed a method to face those clustering problems using an embedding space where each data point is represented by a vector containing memberships to fuzzy sets centered on a set of probes selected from the data base.

We tested out approach on the data by Golub et al [7]. The proposed approach leads to significant improvements with

Method	β	Mean error rate	probe/data ratio
RD	-	17.2	/
DES	-	24.9	0.1
MES	10^{-6}	11.1	0.4
MES	$5 \cdot 10^{-7}$	10.9	0.5
MES	10^{-7}	9.5	0.7
MES	10^{-8}	9.1	0.8

TABLE I

COMPARISON OF THE BEST MEAN ERROR RATE FOR THE TESTED METHODS: FCM ON ROW DATA (RD), FCM ON THE DISTANCE EMBEDDING SPACE (DES), FCM ON THE MEMBERSHIP EMBEDDING SPACE (MES)

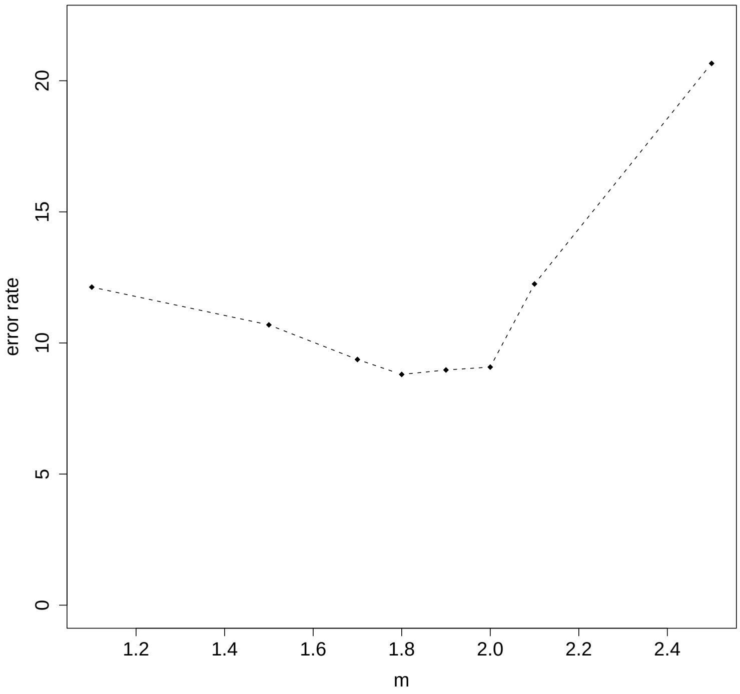

Fig. 2. The behavior of the best error rate (achieved with $\beta = 10^{-8}$, probe/data ratio $= 0.8$) vs the fuzziness parameter m.

respect the application of clustering algorithms in the original space and in the distance embedding space.

ACKNOWLEDGMENT

Work funded by the Italian Ministry of Education, University and Research (2004 "Research Projects of Major National Interest", code 2004062740), and the Biopattern EU Network of Excellence.

REFERENCES

[1] Aggarwal, C.C., Yu, P.S.: Redefining clustering for high-dimensional applications. IEEE Transactions on Knowledge and Data Engineering **14** (2002) 210–225

[2] Bezdek, J.C.: Pattern recognition with fuzzy objective function algorithms. Plenum, New York (1981)

[3] Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is nearest neighbor meaningful? In: 7th International Conference on Database Theory Proceedings (ICDT'99), Springer-Verlag (1999) 217–235

[4] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[5] Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics **3** (1974) 32–57

[6] Fred, A., Leitão, J.: A new cluster isolation criterion based on dissimilarity increments. IEEE Trans. on Pattern Analysis and Machine Intelligence, **25(8)** (2003) 944–958.

[7] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., Lander, E.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science **286** (1999) 531–537.

[8] Masulli, F., Rovetta, S.: A New Approach to Hierarchical Clustering for the Analysis of Genomic Data. In: Proceedings of the International Joint Conference on Neural Networks, Montreal-Canada, IEEE, Piscataway, NJ, USA (2005), *in press*.

[9] Pękalska, E., Paclík, P., Duin, R.P.W.: A generalized kernel approach to dissimilarity-based classification. Journal of Machine Learning Research **2** (2001) 175–211

[10] Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press (2004)