GENE EXPRESSION DATA ANALYSIS IN THE MEMBERSHIP EMBEDDING SPACE: A CONSTRUCTIVE APPROACH*

M. FILIPPONE, F. MASULLI AND S. ROVETTA

Department of Computer and Information Sciences
University of Genova and CNISM
Via Dodecaneso 35, I-16146 Genova, Italy
{*filippone,masulli,rovetta*}*@disi.unige.it*

Exploratory analysis of genomic data sets using unsupervised clustering techniques is often affected by problems due to the small cardinality and high dimensionality of the data set. A way to alleviate those problems lies in performing clustering in an embedding space where each data point is represented by a vector of its memberships to fuzzy sets characterized by a set of probes selected from the data set. This approach has been demonstrated to lead to significant improvements with respect the application of clustering algorithms in the original space and in the distance embedding space. In this paper we propose a constructive technique based on Simulated Annealing able to select sets of probes of small cardinality and supporting high quality clustering solutions.

1. Introduction

Clustering methods provide an useful tool to explore genomic data sets, but often the crude application of classical clustering algorithms leads to poor results. Actually, many clustering approaches suffer from being applied in high-dimensional spaces, as clustering algorithms often seek for areas where data is especially dense. However, sometimes the cardinality of the data sets available 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.

Moreover, 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,

*Work funded by the MIUR grant code 2004062740

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].

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[13] may be used to represent data sets in a more compact way. Pękalska and Duin[12] 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. In this framework the data set is embedded in a lower dimensional space called *embedding space*, in which, in the presence of large-dimensional data sets, a notable complexity reduction is achieved.

Following this approach, the data matrix is replaced by a pairwise dissimilarity matrix D. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a data set of cardinality n. We start by computing the dissimilarity matrix D:

$$d_{ik} = d(x_i, x_k) \;\; \forall i, k \tag{1}$$

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[7,10].

As pointed out in[12], the dissimilarity measure should be a metric, since metrics preserve the *reverse of the compactness hypothesis*: "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. In the following, we will adopt the Euclidean distance as the dissimilarity measure.

In case of a data set with dimensionality N there is the upper bound of $N + 1$ *probes* (or *support data*)[12] that we can use in order to build the dissimilarity matrix. In the case of genomic data this upper bound is often un-realistic, since the cardinality is much lower than the dimensionality. However, for data having some structure, it is not necessary to reach this

upper bound for good representation. We only require that the dimension of the embedding space is large enough to preserve the reverse of the compactness hypothesis. On the other hand, if the embedding dimension n is lower than $N+1$, some points could have an ambiguous representation and, moreover, clustering could be affected by the high metrical contribution of farthest points.

In order to avoid those problems, in[6] we proposed a different kind of embedding based on the space of memberships to fuzzy sets centered on the probes, that we will call *Membership Embedding Space* (MES) .

Following this approach, a point in the embedding space 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:

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

where $i = 1, \ldots, n$ and $k = 1, \ldots, s$. Note that the parameter β regulates the spread of the membership function and it is related to the average distance between the data points. For large values of β the memberships tends more rapidly to zero than for little β. In the MES each data point x_i is represented as a row of ν_{ik}.

In this paper, we propose a constructive method to obtain the set of probes leading to optimal clustering in the MES using *Simulated Annealing*.

2. Simulated Annealing for Probe Selection

The proposed method for probe selection makes use of the Simulated Annealing (SA) technique[9] that is a global search method technique derived by Statistical Mechanics. SA is based on the work by Metropolis et al.[11] aimed to simulate the behavior and small fluctuations of a system of atoms starting from an initial configuration, by the generation of a sequence of iterations. In the Metropolis algorithm each iteration is composed by a random perturbation of the actual configuration and the computation of the corresponding energy variation (ΔE). If $\Delta E < 0$ the transition is unconditionally accepted, otherwise the transition is accepted with probability given by the Boltzmann distribution:

$$P(\Delta E) = exp\left(\frac{-\Delta E}{KT}\right) \tag{3}$$

where K is the Boltzmann constant and T the temperature.

In SA this approach is generalized to the solution of general optimization problems[9] by using an *ad hoc* selected cost function (*generalized energy*), instead of the physical energy. SA works as a probabilistic hill-climbing procedure searching for the global optimum of the cost function. The temperature T takes the role of a control parameter of the search area (while K is usually set to 1), and is gradually lowered until no further improvements of the cost function are noticed. SA can work in very high-dimensional searches, given enough computational resources.

(1) Initialize parameters (see list in Tab. 1);
(2) Initialize the binary mask **g** at random;
(3) Perform clustering and evaluate the generalized system energy E;
(4) **do**
(5) Initialize $f = 0$ (number of iterations), h=0 (number of success);
  (a) **do**
  (b) Increment number of iterations f;
  (c) Perturb mask **g**;
  (d) Perform clustering and evaluate the generalized system energy E;
  (e) Generate a random number rnd in the interval [0,1];
  (f) **if** $rnd < P(\Delta E)$ **then**
    i. Accept the new **g** mask;
    ii. Increment number of success h;
  (g) **endif**
  (h) **loop until** $h \leq h_{min}$ **and** $f \leq f_{max}$;
(6) update $T = \alpha T$;
(7) **loop until** $h > 0$;
(8) **end**.

Figure 1. Simulated Annealing Probe Selection (SA-PS) Algorithm.

In Fig. 1 the proposed Simulated Annealing Probe Selection (SA-PS) algorithm is shown. In our approach the state of the system is represented by a binary mask $\mathbf{g} = (g_1, g_2, \ldots, g_n)$, where each bit g_i (with $i = 1, \ldots, n$) corresponds to the selection ($g_i = 1$) / deselection ($g_i = 0$) of a probe. The initialization of the vector mask **g** (Step 2) is done by generating s_0 integer

numbers with uniform distribution in the interval $[1, n]$ and setting the corresponding bits to 1 of $\mathbf{g}$ and the remaining ones to 0. At each step only s probes are selected from the original set of n patterns. A perturbation or move is done in the following way: (1) Chose randomly $w \in [w_{\min}, w_{\max}]$ and $v \in [v_{\min}, v_{\max}]$; (2) w bits of $\mathbf{g}$ set to 1 are switched to 0; (3) v bits of $\mathbf{g}$ set to 0 are switched to 1.

The values $w_{\min}, w_{\max}, v_{\min}, v_{\max}$ can be used to reduce or to increase the variability of each perturbation.

Once a set of probes is selected, it is possible to represent each pattern in the Membership Embedding Space (MES) and to perform clustering. In the experiments reported in the remainder of this paper, we performed clustering using the FCM algorithm[2], but many other clustering algorithms can be employed.

The generalized energy E is computed as a linear combination between an assigned clustering quality measure ε and the number of selected probes s:

$$E = \varepsilon + \lambda s \tag{4}$$

The clustering quality measure ε can be a function of either the cost function associated to the clustering algorithm, a clustering validation index, or, in the case of labeled data sets, the *Representation Error* (RE). RE is the count of data points in each cluster disagreeing with the majority label in that cluster, summed over all clusters and expressed as a percentage.

Note that the introduction of the number of selected probes s in the computation of E penalizes situations in which the number of selected probes is high. This choice of E leads to the minimization of the cardinality of the set of probes able to achieve a good clustering quality measure. The balance between these two terms is controlled by λ (*regularization coefficient*).

3. Experimental setup

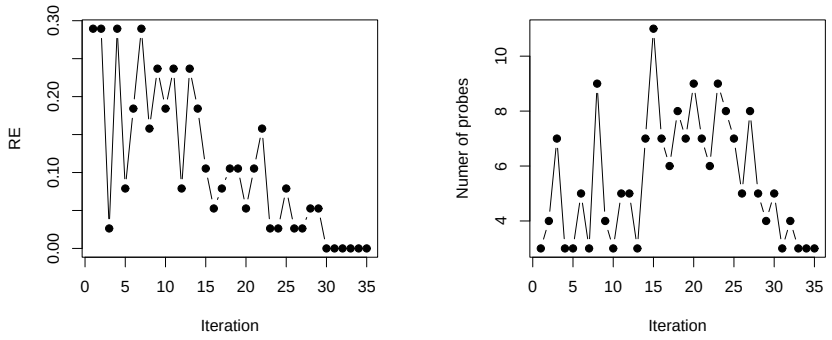
The method was tested on the publicly available Leukemia data by Golub et al.[8]. The Leukemia problem consists in characterizing two forms of acute leukemia, Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML). The original work proposed both a supervised classification task ("class 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.

Table 1. Choice of parameters.

Meaning	*Symbol*	*Value*
Number of random perturbations of **g** used to estimate the initial value of T	p	10000
Number of probes to be initially selected	s_0	3
Cooling parameter	α	0.9
Membership width parameter	β	10^{-6}
Maximum number of iteration at each T	f_{max}	2000
Minimum number of success for each T	h_{min}	200
Regularization coefficient	λ	10^{-2}
Minimum number of bits to be switched	$w_{\min}$, $v_{\min}$	1,1
Maximum number of bits to be switched	$w_{\max}$, $v_{\max}$	s, 5
Number of clusters	c	3
FCM fuzziness parameter	m	2
FCM trials	r	10

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).

In[6] we presented an extended experimentation using the FCM algorithm[2] and 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. For each experiment we made 1000 independent trials, each of them using a different random initialization of the membership in the FCM algorithm. In all trials probes were extracted at random (using an uniform pdf) from the data set without replacement, the number of clusters was set to 3, and the fuzziness parameter m of FCM was set to 2. The last approach (3), projecting the data set into the membership embedding space, lead to better results. Moreover, increasing the parameter β from 10^{-8} to 10^{-6} we obtained for increasing probe/data ratios (from .8 to .4) a shift of the optimal error ratio.



(a) (b)

Figure 2. *RE* (a) and number of probes selected (b) during a run of the SA-PS algorithm.

Starting from those previous results, we ran the SA-PS algorithm in the MES with the assumptions shown in Tab. 1. The value of the parameter β used in the experiments ($\beta = 10^{-6}$) was about the reciprocal of the mean distance between patterns. As a clustering quality measure we used the *Representation Error* (RE) evaluated as the best value obtained on $r = 10$ independent trials of FCM.

Each independent run of the SA-PS algorithm finds a different small subset of probes leading to a clustering Representation Error equal 0. In Fig. 2, the Representation Error and the number of selected bits of $\mathbf{g}$ are plotted versus the iteration number during a run of the SA-PS algorithm, where each iteration corresponds to a different value of temperature T. In this case, at iterations 31, 33, 34 and 35 we obtained 4 different sets of 3 probes giving clustering RE equal 0.

4. Conclusions

Exploratory analysis of genomic data sets using unsupervised clustering techniques, are often affected by problems due to the small cardinality and high dimensionality of the data set. A way to alleviate those problems lies in performing clustering in an embedding space where each data point is represented by a vector of its memberships to fuzzy sets centered on a set of probes selected from the data set. In previous work, this approach has been demonstrated to lead to significant improvements with respect the

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

In this paper we have presented a constructive technique based on Simulated Annealing able to select sets of probes for clustering in the embedding space of fuzzy memberships. The application of the proposed probe selection algorithm combined with FCM to the Leukemia data by Golub et al[8] leads to high quality clustering solutions.

References

1. C.C. Aggarwal and P.S. Yu, Redefining clustering for high-dimensional applications. IEEE Transactions on Knowledge and Data Engineering **14** 210–225 (2002).
2. J.C. Bezdek, Pattern recognition with fuzzy objective function algorithms. Plenum, New York (1981).
3. K. Beyer, J. Goldstein, R. Ramakrishnan and U. Shaft, When is nearest neighbor meaningful? In: 7th International Conference on Database Theory Proceedings (ICDT'99), Springer-Verlag 217–235 (1999).
4. R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York (1973).
5. J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics **3** 32–57 (1974).
6. M. Filippone, F. Masulli and S. Rovetta, Clustering Genomic Data in the Membership Embedding Space. In: CI-BIO Workshop on Computational Intelligence Approaches for the Analysis of Bioinformatics Data, Montreal-Canada, IEEE, Piscataway, NJ, USA (2005), *http://ci-bio.disi.unige.it/CI-BIO-booklet/CI-BIO.html*.
7. A. Fred, J. Leitão, A new cluster isolation criterion based on dissimilarity increments. IEEE Trans. on PAMI, **25(8)** 944–958 (2003).
8. T. Golub, et al., Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science **286** 531–537 (1999).
9. S. Kirkpatrick, C.D. Gelatt and M.P. Vecchi, Optimization by simulated annealing. Science, **220** 661–680 (1983).
10. F. Masulli and S. Rovetta, A New Approach to Hierarchical Clustering for the Analysis of Genomic Data. In: Proc. I.J.C. on Neural Networks, Montreal-Canada, IEEE, Piscataway, NJ, USA (2005).
11. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller, Equation of state calculations for fast computing machines. Journal of Chemical Physics, **21** 1087–1092 (1953).
12. E. Pękalska, P. Paclík and R.P.W. Duin, A generalized kernel approach to dissimilarity-based classification. Journal of Machine Learning Research **2** 175–211 (2001).
13. J. Shawe-Taylor and N. Cristianini, Kernel Methods for Pattern Analysis. Cambridge University Press (2004).