Membership Embedding Space Approach and Spectral Clustering

Stefano Rovetta, Francesco Masulli, and Maurizio Filippone

Department of Computer and Information Sciences, University of Genova, and CNISM, Via Dodecaneso 35, I-16146 Genova, Italy

Abstract. The data representation strategy termed "Membership Embedding" is a type of similarity-based representation that uses a set of data items in an input space as reference points (probes), and represents all data in terms of their membership to the fuzzy concepts represented by the probes. The technique has been proposed as a concise representation for improving the data clustering task. In this contribution, it is shown that this representation strategy yields a spectral clustering formulation, and this may account for the improvement in clustering performance previously reported. Then the problem of selecting an appropriate set of probes is discussed in view of this result.

1 Introduction

Exploratory analysis of large-dimensional data sets using unsupervised clustering techniques is often affected by problems due to the small cardinality and high dimensionality of the data set. These are not merely of a computational nature, but have effects on the performance of the data analysis process itself. 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 reference points termed "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 a previous paper [5] we have proposed a constructive technique based on Simulated Annealing to select sets of probes of small cardinality.

The contribution of the present work lies in proving that this representation technique is formally analogous to a spectral clustering problem. This suggests a possible explanation of the observed increase in performance, and also suggests some possible strategies for performing the selection of probes.

2 Clustering in distance-based representations

Many clustering approaches suffer from being applied in high-dimensional spaces, as common clustering algorithms often seek for areas where data is especially dense. Data sparsity observed with high-dimensional problems and a loss in significance of a large class of metrics [2, 1] are other possible issues.

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 [14] 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 be a data set of cardinality n: $X = \{x_1, x_2, \ldots, x_n\}$. We start by computing the dissimilarity matrix D:

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

As pointed out in [12], the dissimilarity measure should be a metric, since metrics preserve the *reverse of the compactness hypothesis* [12]: "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.

3 The Membership Embedding space

In [11] and [5] we proposed a specific embedding technique based on the space of memberships to fuzzy sets centered on the probes, that we have termed *Membership Embedding Space* (MES).

Let us 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.

In order to avoid the problems previously highlighted, in [5] we have proposed a different kind of embedding based on the space of memberships to fuzzy sets centered on the probes.

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 closest 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: $v_{ik} = \exp\left[-\beta d_{i,k}^2\right] / \sum_l \exp\left[-\beta d_{i,l}^2\right]$.

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 v_{ik}: $x_i \rightarrow (v_{i1}, v_{i2}, \ldots, v_{in})$.

Once a set of probes is selected, it is possible to represent each pattern in the Membership Embedding Space (MES) and to perform clustering with a suitable method, exploiting the better structure obtained in the MES to obtain results with higher confidence or better performance. We usually perform clustering using the FCM algorithm [3], but many other clustering algorithms can be employed.

4 The Membership Embedding is a normalized Laplacian

Good clustering performance has been consistently observed when representing data with the Membership Embedding technique. In this section we show that the MES representation yields a clustering problem formulation that is equivalent to a spectral clustering, i.e., the data structure to be clustered is the Laplacian of an appropriately defined data graph.

The patterns $\{x_1, \ldots, x_n\}$ mapped in the membership embedding space by the transformation $v_{ik} = h(x_i, y_k) / \sum_{l=1}^{c} h(x_i, y_l)$, for instance using the Gaussian proximity function: $v_{ik} = e^{(-\beta \|x_i - y_k\|)} / \sum_{l=1}^{c} e^{(-\beta \|x_i - y_c\|)}$, form a $n \times c$ matrix V which can be extended in the following way. Define the $n \times n$ matrix L_{rw} as:

1. $L_{\mathrm{rw}\,ij} = -v_{ik}$ for $y_k = x_j$, that is, for all patterns which have been selected as probes;
2. $L_{\mathrm{rw}\,ij} = L_{\mathrm{rw}\,ji}$, that is, the same as above applies if $y_k = x_i$;
3. $L_{\mathrm{rw}\,ii} = 1$.

Then L_{rw} is equivalent to the normalized Laplacian as defined in the Shi and Malik [15] spectral clustering algorithm, and its eigenvectors, the solutions of the eigenproblem $L_{\mathrm{rw}} z = \lambda z$, can be used as indicators of clusters in the data. To prove this, L_{rw} can be decomposed as follows:

$$L_{\mathrm{rw}} = D^{-1} L = D^{-1}(D - W) = I - D^{-1} W \tag{1}$$

where D and W are defined as follows. For W:

$$w_{ii} = 0 \tag{2}$$

$$w_{ij} = w_{ji} = h(x_i, x_j) \quad \text{if} \quad \exists y_k : x_j = y_k \ , \tag{3}$$

that is, data points are vertices in a graph and there is an undirected edge connecting each probe to all other points, with weight given by the proximity function adopted;

$$w_{ij} = w_{ji} = 0 \quad \text{if} \quad \forall k : x_j \neq y_k \tag{4}$$

that is, there is no edge between data points which are not probes. The matrix W thus defined is the adjacency matrix of the graph described.

D is a diagonal matrix whose diagonal terms are defined by $d_{ii} = \sum_{j=1, j\neq i}^{n} w_{ij}$. D is therefore the degree matrix of the graph.

It is easy to see that these definitions are equivalent to the membership embedding, that is, for point x_i and probe $y_k = x_j$, $L_{\mathrm{rw}\,ij} = L_{\mathrm{rw}\,ji} = v_{ik}$

Therefore, with respect to the reduced connectivity graph defined above, the membership embedding carries the same information that is used to perform clustering with the Shi and Malik spectral algorithm.

This equivalence is interesting because this particular definition of a normalized Laplacian is related to random walk probabilities and to circuit-theoretical properties of networks, for instance Kirchhoff's current law (KCL). These in turn highlight information that is not purely local. One proof of this is provided in [10] where it is shown that the *resistance distance*, a measure of connectivity not just for pairs, but for sets of nodes, can be directly computed from L_{rw}.

Given the equivalence just proven, an explanation of the good clustering performance observed with the MES technique can be stated observing that the first eigenvalues of a Laplacian incorporate the most significant information for clustering. If the Laplacian itself is subject to clustering, as proposed, this is equivalent to highlighting the block structure of the Laplacian. The reasons for the effectiveness of this procedure have been studied in [4].

5 The probe selection problem

Once proven that a *suitable* set of probes can yield an efficient MES transformation, the problem that immediately follows is how to select the patterns that should act as probes. In principle, probes need not be parts of the training set. However, to obtain a computational advantage, this is a sensible choice, because otherwise the probe selection problem itself turns out to be a clustering problem in the original data space, which is what we want to avoid from the start.

Probes should also be reasonably related to clusters, so that the pattern of memberships can discriminate well between points belonging to different clusters. This does not mean that they should belong or be close to clusters themselves, as suggested by results presented in [13], neither that their number should approach the number of clusters.

Several selection techniques have been proposed, including also random selection. We have previously proposed a method based on the Simulated Annealing (SA) technique [9], a well-known global search method technique derived from Statistical Mechanics. SA models the behavior and small fluctuations of a system of atoms starting from an initial configuration, by the generation of a sequence of iterations.

This model is generalized to the solution of quite arbitrary optimization problems [9] by using an *ad hoc* selected cost function (*generalized energy*), instead of the physical energy. In our case, the generalized energy E is computed as a linear combination between an assigned clustering quality measure ε and the number of selected probes s: $E = \varepsilon + \lambda s$. The 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 disagree-

ing with the majority label in that cluster, summed over all clusters and expressed as a percentage.

The introduction of the number of selected probes s in the computation 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.

This method has the advantage to be able to suggest the whole structure of the probe set, i.e., both the number of probes and the data points to be used as probes. Performance is usually good. A drawback is the computational cost of the SA technique itself, so that the whole procedure is justified only in view of the performance improvement in the final clustering solution.

An alternate technique that we are starting to investigate consists in actually performing clustering in the original data space, but with a very efficient and suboptimal algorithm. Experiments are ongoing with a variation on K-Means clustering where centroids are selected from the training set. Runs are limited to a low number of iterations, which nevertheless provides sensible solutions.

A clustering algorithm which lends itself well to this task is the recently proposed "Affinity Propagation Clustering" algorithm [7], which inherently selects significant data points as cluster prototypes, again suggesting both number and position of probes. This technique is not as computationally light as the simpler K-Means, and there is an on-going project about finding solutions efficiently by providing approximations to the original algorithm.

6 Experiments and results

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 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).

The experimentation compares 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.

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 1. Comparison of the best mean error rate for the tested methods: FCM on row data (RD), FCM on Distance Embedding Space (DES), FCM on Membership Embedding Space (MES)

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. 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.

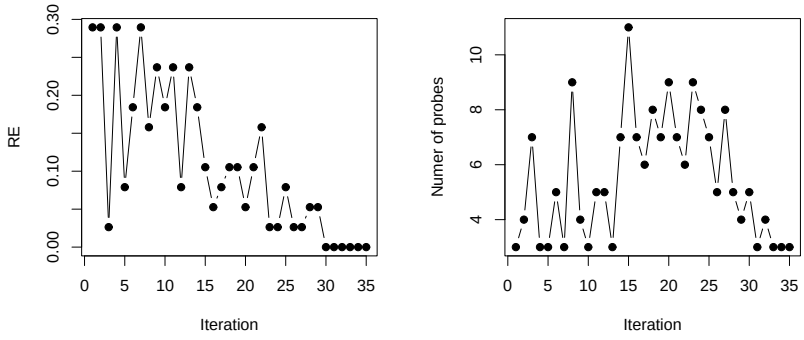
The membership vectors (the rows of the v_{ik} matrix) have a number of non null components related to the spread of the membership function. The minimum of the error rate is achieved for situations for which we have a good compromise between the number of probes and the width of the membership function.

A comparison of the best mean error rate for the tested methods is reported in Tab. 1.

The following results are related to experiments about probe selection. 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. 1, 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.

The preliminary experiments with the “rough clustering” approach show that, as expected, the whole procedure takes a much shorter time, of the order of tenths of second (the implementation is in C) which is much better than the time required for SA.

The experimental results are still to be analyzed. However, an interesting fact has been observed: the performance of the clustering algorithm used is not directly related to the number of probes. This can open the way to iterative procedures whereby the number of probes (and with it the dimensionality of the embedding space) is progressively reduced, while maintaining a given accuracy level. An example of this behaviour is presented in Table 2.



(a) (b)

Fig. 1. *RE* (a) and number of probes selected (b) during a run of the SA-PS algorithm.

Number of probes	RE
50	73.68%
25	81.58%
3	73.68%

Table 2. Example of error rate with varying number of probes.

7 Conclusions

A way to alleviate dimensionality problems in clustering 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, and a constructive technique based on Simulated Annealing has been proposed to select sets of probes for clustering in the embedding space of fuzzy memberships.

In the present contribution, the MES approach has been proved to yield a spectral clustering problem, and the general problem of proble selection has been recast to allow future solutions with a computational, in addition to performance, advantage.

References

1. Charu C. Aggarwal and Philip S. Yu. Redefining clustering for high-dimensional applications. *IEEE Transactions on Knowledge and Data Engineering*, 14(2):210–225, March/April 2002.
2. K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbor meaningful? In *7th International Conference on Database Theory Proceedings (ICDT'99)*, pages 217–235. Springer-Verlag, 1999.
3. James C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Plenum, New York, 1981.

4. Mattehw Brand and Kun Huang. A unifying theorem for spectral embedding and clustering. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.
5. Maurizio Filippone, Francesco Masulli, and Stefano Rovetta. Gene expression data analysis in the membership embedding space: A constructive approach. In *Proceedings of the 7th International FLINS Conference on Applied Artificial Intelligence, Genova, Italy*, August 2006.
6. A. L. N. Fred and J. M. N. Leitao. A new cluster isolation criterion based on dissimilarity increments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):944–958, August 2003.
7. Brendan J J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, January 2007.
8. T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, October 1999.
9. C. D. Kirkpatrick, S. and Gelatt and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.
10. D. J. Klein and M. Randi. Resistance distance. *Journal of Mathematical Chemistry*, 12(1):81–95, December 1993.
11. Francesco Masulli, Stefano Rovetta, and Maurizio Filippone. Clustering genomic data in the membership embedding space. In *Proceedings of the International Joint Conference on Neural Networks 2005*, August 2005.
12. Elżbieta Pękalska, Pavel Paclík, and Robert P. W. Duin. A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, 2:175–211, 2001.
13. Stefano Rovetta and Francesco Masulli. Shared farthest neighbor approach to clustering of high dimensionality, low cardinality data. *Pattern Recognition*, 39(12):2415–2425, December 2006.
14. John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
15. Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
16. Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*. (to appear).