Soft Ranking in Clustering

Stefano Rovetta , Francesco Masulli , Maurizio Filippone

CNISM and Dipartimento di Informatica e Scienze dell'Informazione, Università di Genova, Via Dodecaneso 33, I-16146 Genova, Italy

Abstract

Due to the diffusion of large-dimensional data sets (e.g., in DNA microarray or document organization and retrieval applications), there is a growing interest in clustering methods based on a proximity matrix. These have the advantage of being based on a data structure whose size only depends on cardinality, not dimensionality. In this paper we propose a clustering technique based on fuzzy ranks. The use of ranks helps to overcome several issues of large-dimensional data sets, whereas the fuzzy formulation is useful in encoding the information contained in the smallest entries of the proximity matrix. Comparative experiments are presented, using several standard hierarchical clustering techniques as a reference.

Key words: Fuzzy rank; clustering; data mining; DNA Microarrays

1 Introduction

Data clustering is traditionally studied in the hypothesis, sometimes implied, that data themselves are available in large quantities, or, at least, that cardinality is larger than dimensionality. However in recent years many data acquisition techniques have been made available that produce large quantities of observed variables for each observation. Therefore, the cardinality-dimensionality relation has sometimes been inverted. Two prominent examples of high-throughput techniques are DNA microarrays and document collections for information retrieval. In this situation, as it had already happened for classification with featureless approaches [10] and kernel methods [11], there is a growing interest in clustering methods based on a proximity matrix, which share the obvious advantage of being based on a data structure whose size only depends on cardinality, not dimensionality. Among

Email addresses: rovetta@disi.unige.it (Stefano Rovetta), masulli@disi.unige.it (Francesco Masulli), filippone@disi.unige.it (Maurizio Filippone).

these approaches we can mention primarily the traditional hierarchical agglomerative techniques [14] and the more recent spectral methods [9].

Proximity- (or affinity-)based methods rely on an appropriate definition of metrics or distance, and it is known [3,1] that in high dimensionality the behaviour of large classes of metrics is not as intuitive as it may appear. For this reason, we propose to base our data representation on distance ranks rather than distance values.

2 The proximity matrix approach

We adopt a proximity matrix representation, whereby 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 $d_{ik} = d(x_i, x_k) \ \forall i,k$ according to the dissimilarity measure $d(x,y)$ between points x and y (e.g, the Euclidean distance). The dissimilarity matrix may as well be given as input (for instance when obtained from subjective measurements by a panel of experts, or in the behavioral sciences, or with uncalibrated sensor arrays), in which case it could not even be a symmetric matrix and no explicit function $d(x,y)$ may exist. The matrix D may now be used as the representation of all points of the set X in a space with dimension n. Dissimilarity-based clustering algorithms can be applied even to non-metric data, e.g. categorical or mixed.

3 Clustering with ranks

Problems may arise in high dimensionality also for other effects, especially (but not exclusively) with Minkowski norms as dissimilarity [1]. A typical countermeasure also found in traditional statistics is moving from the analysis of values (in our case, distances) to the analysis of their *ranks*. Rank is the position of a given value in the ordered list of all values. However in this work we adopt a fuzzy definition of the concept of ranks. This definition has already been used to improve image reconstruction after lossy compression [8]; here it is used as a starting point for a data analysis procedure.

Let $\mathbf{d}_i$ be the i-th row of D (storing the dissimilarities of points in X from $x_i \in X$). The corresponding ranks, or positions in the list of components of $\mathbf{d}_i$ when sorted in ascending order, are then stored in the vector $\mathbf{r}_i = [\rho_{i1}, \ldots, \rho_{in}]$. We call these the *D-ranks* for point x_i. A D-rank can be written in an algebraic fashion as:

$$\rho_{ij} = \sum_{k=1}^{n} \theta\left(d_{ij} - d_{ik}\right), \tag{1}$$

where the function $\theta(x)$ is an extended Heaviside step, taking on the values 0 for $x < 0$, 1 for $x > 0$, and 0.5 for $x = 0$, so $\rho_{ij} \in [0, \ldots, n-1]\ \forall i \in \{1, \ldots, n\}$. This extension of the Heaviside step represents the standard way to deal with ties in rank-order statistics. (Note that we define ranks to start at 0, so that D-rank 1 refers to the nearest neighbour.) It is now possible to measure the closeness of data points $x_1, \ldots, x_n$ by the concordance of their respective D-rank vectors $\mathbf{r}_1, \ldots, \mathbf{r}_n$. Therefore for this purpose a data point x_i is represented by the vector of its D-ranks.

This definition has several advantages. It embeds the problem into a space of dimension n, which, by assumption, is smaller than the cardinality of the original data. Metric and non-metric cases are treated alike, since the new measure is numeric in both cases. Using this representation of data, any metric clustering algorithm can be applied. In the experiments, we will refer to the specific procedure illustrated in the following section.

Obviously, ranks also discard information, and this is more evident when the distribution of points is uneven. In this case, points in a dense region and points in a sparse region may be represented with similar rank patterns. In the following section, we introduce a fuzzy definition of rank that is able to preserve more information in these cases.

4 Fuzzy ranks

In a fuzzy set-theoretic perspective, it is more natural to define the relation "larger" among two numbers as a degree to which one number is larger than another. The problem of ranking fuzzy quantities has been reviewed for instance by Bortolan and Degani [4] and, more recently, by Wang and Kerre [15,16]. For instance, suppose that we are to compare (a) $d_1 = 3$ with $d_2 = 4$, and (b) $d_1 = 3$ with $d_2 = 3.01$. Clearly in both case (a) and case (b) we can rightfully say that $d_2 > d_1$, but it is also clear that in (a) this is "more true" than in (b). Therefore, we can make the following substitution:

$$\theta\left(d_{ij} - d_{ik}\right) \longrightarrow \frac{1}{1 + e^{(d_{ij} - d_{ik})/\beta}} \tag{2}$$

where:

$$\lim_{\beta \to 0^+} \frac{1}{1 + e^{(d_{ij} - d_{ik})/\beta}} = \theta\left(d_{ij} - d_{ik}\right) \tag{3}$$

So the computation of fuzzy rank can be expressed as

$$\rho_{ij} = \sum_{k=1}^{n} \frac{1}{1 + e^{(d_{ij} - d_{ik})/\beta}} \tag{4}$$

The parameter β is a fuzziness parameter: for large β the ranking function is definitely fuzzy, while for $\beta = 0$ we obtain the original, crisp ranking function. The two expressions (1) and (4) for the rank function are compared in a simple example, illustrated in Figure 1, where we assume $\mathbf{d}_i = [d, 2, 3, 5]$ and the first value d_{i1} sweeps from 0 to 7. We plot the corresponding ρ_{i1} (in the two expressions, crisp and fuzzy). Two fuzzy plots are shown, one for $\beta = 0.05$ and another for $\beta = 0.25$ (smoother).

This new definition of rank allows us to integrate into a rank-based clustering approach the notion that two ranks may be clearly defined (this happens when comparing very different values), and in this case the soft rank behaves similarly to the standard, crisp definition of ranks; or they may be less clearly defined (when the values to be compared are not very different), and in this case the soft rank takes into account the degree of closeness between the values. The clustering algorithm we applied in the new representation is the *agnes* procedure [6] that is an agglomerative hierarchical clustering algorithm.

5 Experiments

The proposed soft rank clustering algorithm was tested on two synthetic problems and then applied to a publicly available bioinformatics dataset: the Colon data by Alon et al. [2]. Data sets are described in Table 5. The synthetic problems have been used to check the properties and consistency of the approach.

The tests were performed according to the proposed method for a number of different fuzziness levels β. and diagrams were compared for several linkage methods. Specifically, the linkage methods used are: single (or nearest neighbor linkage); average (UPGMA); complete (or farthest neighbor linkage); weighted (WPGMA); ward (Ward's method with analysis of cluster variance). We focus on the results obtained by the Ward method only, since it is known to yield small and compact clusters not affected by the "chaining" effect, and this was fully confirmed by the experimental analysis.

For the purpose of this study, values of β have been selected by hand-tuning on the basis of experimental results (we used the available test data). More sound techniques may be based on the quality of clustering obtained, as measured for instance by the coefficient of agglomeration or other citeria.

Table 1
Data sets used for experiments

	Dimensionality	**Cardinality**	**Class balance**
Synthetic data 1	2	100	1:1
Synthetic data 2	20	100	1:1
Colon cancer	2000	62	2:1

The first synthetic problem consists in classifyng two-dimensional points obtained from two Gaussian distributions. Noise is added to each point coordinates to obtain a sequence of increasingly random data sets. The method was applied for several values of β and for distances (not distance ranks). We obtain a number of experimental results, one for each noise level, representing the error percentage as a function of β. These results are shown in Fig. 2 for selected, values of noise level, namely $\sigma = 0.5$ and $\sigma = 1$. The values are not shown directly, but as ratio of error reduction with respect to the results on distances, used as a baseline. The box-and-wiskers plot has been obtained on the basis of 100 different runs, with varying random generation. From the plot, it can be seen the error reduction over the baseline.

The second synthetic data set is composed of two 10-dimensional Gaussian distribution centered respectively in $(-1,\ldots,-1)$ and $(1,\ldots,1)$. Both the distributions have standard deviation 2. Ten unformly distributed features have been added to the data set. We tested the proposed method on different subsamples of the features. In Fig. 3 we report the box-and-wiskers plot obtained for different numbers of selected features over 100 trials. Fig. 3 shows the results for two values of β, namely $\beta = 0.1$ and $\beta = 1$. Also in this situation, it is possible to see that the use of fuzzy ranks outperforms the clustering of the distances on different subsamples of the data set.

The Colon cancer data is again a collection of expression levels for 2000 genes measured with the DNA microarray technique. They have been selected as those having the highest minimum intensity across all samples. The ratio between classes is approximately 2 tumor to 1 normal. Here we focus on what has been called the "class discovery" problem (unsupervised categorization, i.e., clustering), but we exploit the diagnostic information to assess the goodness of the clustering obtained, thereby evaluating clustering in a transductive setting. This diagnostic information is the distinction between normal and cancer tissue labels. From the analysis of the dendrogram, we decided to set the number of cluster to 3. In Fig. 4 we show the reduction in the error achieved by the use of fuzzy ranks in comparison with the clustering of the distances (thick dotted line) and the clustering using crisp ranks (light dotted line).

The results on the Colon data set are resumed in fig. 4 where the clustering using fuzzy ranks achieves better classification errors with respect to the methods used for comparison. In particular the figure shows that the use of ranks (crisp and fuzzy)

performs better with respect to the clustering of patterns in input space and in many cases with respect to the clustering of distances. Here the use of the fuzzy ranks improves the classification error achieved by the crisp ranks.

6 Conclusions

We have presented a technique to perform clustering of high-dimensional data sets by mapping these data in a lower dimensional space, the space of fuzzy D-rank vectors. Several clustering techniques can be applied, and we used the standard *agnes* procedure to obtain an indication of the best value for the fuzziness parameter β. The analysis confirms the quality of the proposed procedure by comparison to the knowledge available in the literature, and its superiority to the other methods experimented. The overall method is closely related to non-metric multidimensional scaling (MDS) techniques based only on dissimilarity ranks as opposed to metric distances[7,12,13]. The fuzzy rank mapping itself in the proposed method plays the role of a multidimensional embedding. One difference is that MDS provides a mapping by pointing out possibly interesting relationships, and the subsequent data analysis is left to the researcher, whereas clustering aims directly at outlining a structure in the form of suggested categories.

Acknowledgements

Work funded by the the Italian Ministry of Education, University and Research (code 2004062740), and the Biopattern EU Network of Excellence.

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] U. Alon, N. Barkai, D. A. Notterman, K. Gish, D. Mack S. Ybarra, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Science USA*, 96:6745–6750, June 1999.

[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)*, pages 217–235. Springer-Verlag, 1999.

[4] G. Bortolan and R. Degani. A review of some methods for ranking fuzzy sets. *Fuzzy Sets and Systems*, 15:1–19, 1985.

[5] 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.

[6] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data*. John Wiley & Sons, New York, USA, 1990.

[7] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, (29):1–27, 1964.

[8] Francesco Masulli and Stefano Rovetta. Fuzzy variations in the training of vector quantizers. In *Proceedings of the 2003 International Workshop on Fuzzy Logics, Napoli, Italy*, October 2003.

[9] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.

[10] 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.

[11] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[12] R. N. Shepard. The analysis of proximities: multidimensional scaling with unknown distance function Part I. *Psychometrika*, (27):125–140, 1962.

[13] R. N. Shepard. The analysis of proximities: multidimensional scaling with unknown distance function Part II. *Psychometrika*, (27):219–246, 1962.

[14] R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438, 1958.

[15] W. Wang and E.E. Kerre. Reasonable properties for the ordering of fuzzy quantities (I). *Fuzzy Sets and Systems*, 118:375–385, 2001.

[16] W. Wang and E.E. Kerre. Reasonable properties for the ordering of fuzzy quantities (II). *Fuzzy Sets and Systems*, 118:386–405, 2001.

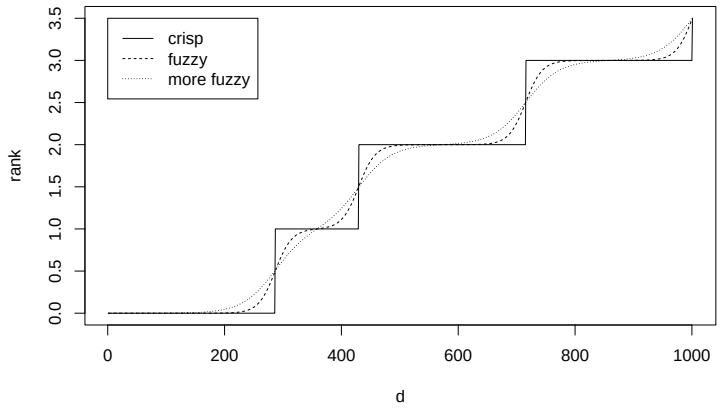

Fig. 1. Comparing crisp and fuzzy rank functions.

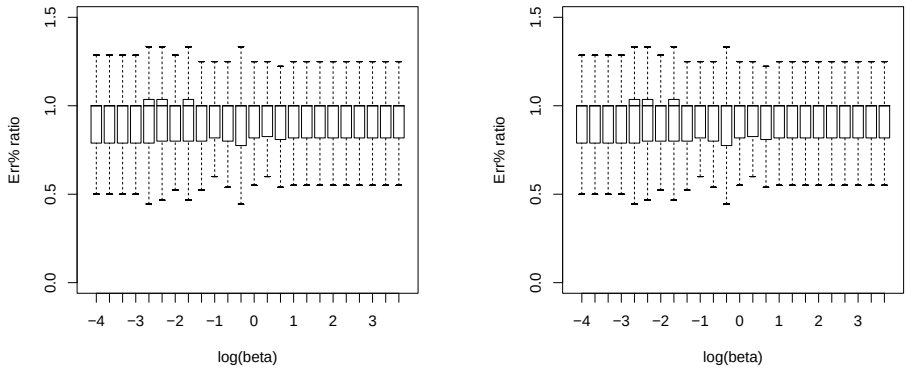

Fig. 2. Synthetic 1

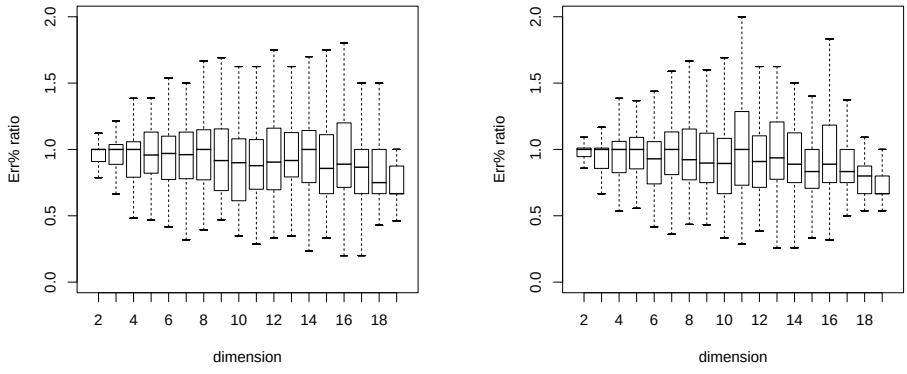

Fig. 3. Synthetic 2

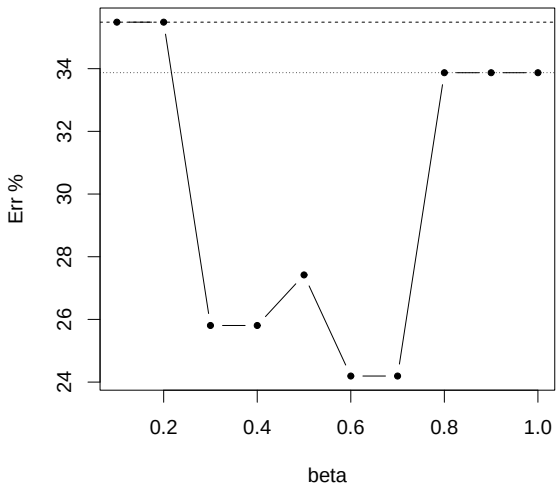

Fig. 4. Results on the Colon data set. Vertical bars: error level as a function of the fuzziness parameter β. Thin dotted line: the error obtained by clustering in the data space directly. Thick dotted line: the error obtained by clustering in the space of distances.