

## Letters

## Soft ranking in clustering

Stefano Rovetta<sup>a</sup>, Francesco Masulli<sup>a,b</sup>, Maurizio Filippone<sup>c,\*</sup><sup>a</sup> Dipartimento di Informatica e Scienze dell'Informazione and CNISM - Genova Research Unit, Università di Genova, Via Dodecaneso 33, I-16146 Genova, Italy<sup>b</sup> Center for Biotechnology, Temple University, BioLife Science Bldg., 1900 N 12th Street, Philadelphia, PA 19122, USA<sup>c</sup> Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK

## ARTICLE INFO

## Article history:

Received 30 April 2008

Received in revised form

7 September 2008

Accepted 17 November 2008

Communicated by T. Heskes

Available online 3 December 2008

## Keywords:

Fuzzy rank

Clustering

Data mining

DNA microarrays

## 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.

© 2008 Elsevier B.V. All rights reserved.

## 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. In recent years, however, many data acquisition techniques able to produce large quantities of variables for each observation have been made available. 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 [12], 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 these approaches, we can mention primarily the traditional hierarchical agglomerative techniques [15] 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 [1,3] that in high dimensionality the behavior 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  $n$ ,  $X = \{x_1, x_2, \dots, x_n\}$ . 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 from experiments in the behavioral sciences, or from measurements performed through uncalibrated sensor arrays), in which case it could even be a non-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. In this work, however, we adopt a fuzzy definition of the concept of ranks. This definition has already been used to improve image reconstruction after lossy compression [11]; here it is used as a starting point for a data analysis procedure.

\* Corresponding author.

E-mail addresses: [rovetta@disi.unige.it](mailto:rovetta@disi.unige.it) (S. Rovetta), [masulli@disi.unige.it](mailto:masulli@disi.unige.it) (F. Masulli), [filippone@dcs.shef.ac.uk](mailto:filippone@dcs.shef.ac.uk) (M. Filippone).

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 the components of  $\mathbf{d}_i$  when sorted in ascending order, are then stored in the vector  $\mathbf{r}_i = [\rho_{i1}, \dots, \rho_{in}]$ . We call  $\mathbf{r}_i$  the  $D$ -ranks for the point  $x_i$ . A  $D$ -rank can be written in an algebraic fashion as

$$\rho_{ij} = \sum_{k=1}^n \theta(d_{ij} - d_{ik}) \quad (1)$$

where the function  $\theta(x)$  is an extended Heaviside step, taking the values 0 for  $x < 0$ , 1 for  $x > 0$ , and 0.5 for  $x = 0$ . This extension of the Heaviside step represents the standard way to deal with ties in rank-order statistics. It is now possible to measure the closeness of data points  $x_1, \dots, x_n$  by the concordance of their respective  $D$ -rank vectors  $\mathbf{r}_1, \dots, \mathbf{r}_n$ . 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. We will make use of a simple illustrative example to show such considerations.

#### 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 [16,17]. 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 cases (a) and (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:

$$\theta(d_{ij} - d_{ik}) \rightarrow \frac{1}{1 + e^{(d_{ik} - d_{ij})/\beta}} \quad (2)$$

where

$$\lim_{\beta \rightarrow 0^+} \frac{1}{1 + e^{(d_{ik} - d_{ij})/\beta}} = \theta(d_{ij} - d_{ik}) \quad (3)$$

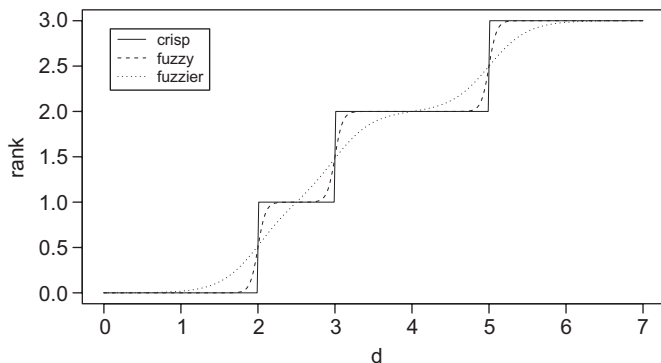


Fig. 1. A comparison between crisp and fuzzy rank functions.

So the computation of fuzzy rank can be expressed as

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

The parameter  $\beta$  is a fuzziness parameter: for large  $\beta$  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 Fig. 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  $\rho_{i1}$  (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 [7] that is an agglomerative hierarchical clustering algorithm.

As an illustrative example, we show the resulting crisp and fuzzy rank matrices in a simple case, where data points are  $\{5, 4, 4.5, 1\}$ . The distance matrix and the crisp rank matrix are

$$D = \begin{pmatrix} 0 & 1 & 0.5 & 4 \\ 1 & 0 & 0.5 & 3 \\ 0.5 & 0.5 & 0 & 3.5 \\ 4 & 3 & 3.5 & 0 \end{pmatrix}, \quad \rho = \begin{pmatrix} 0.5 & 2.5 & 1.5 & 3.5 \\ 2.5 & 0.5 & 1.5 & 3.5 \\ 2 & 2 & 0.5 & 3.5 \\ 3.5 & 1.5 & 2.5 & 0.5 \end{pmatrix}$$

The fuzzy rank matrices with  $\beta = 0.5$  and  $\beta = 1$  are, respectively,

$$\rho_{\beta=0.5} = \begin{pmatrix} 0.89 & 2.11 & 1.50 & 3.50 \\ 2.13 & 0.89 & 1.51 & 3.47 \\ 1.73 & 1.73 & 1.04 & 3.49 \\ 3.11 & 1.89 & 2.50 & 0.50 \end{pmatrix},$$

$$\rho_{\beta=1} = \begin{pmatrix} 1.16 & 1.90 & 1.53 & 3.41 \\ 1.97 & 1.19 & 1.58 & 3.26 \\ 1.67 & 1.67 & 1.28 & 3.38 \\ 2.84 & 2.10 & 2.47 & 0.59 \end{pmatrix}$$

#### 5. Experiments

The proposed soft rank clustering algorithm was tested on two synthetic problems and then applied to a publicly available bioinformatics data set: the Colon data by Alon et al. [2]. Data sets are described in Table 1. 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  $\beta$ , and results were compared for several linkage methods. Specifically, the linkage methods used are [6,18]: single (or nearest neighbor linkage); average (UPGMA); complete (or farthest neighbor linkage);

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

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  $\beta$  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 criteria. In all experiments presented in the paper we have used the Euclidean distance.

The first synthetic problem consists in classifying two-dimensional points obtained from two Gaussian distributions. Gaussian perturbation with standard deviation  $\sigma$  is applied to each point coordinates to obtain a sequence of increasingly random data sets. The method is applied using several values of  $\beta$  and on the distance matrices. We obtain a number of experimental results, one for each standard deviation  $\sigma$  of the perturbation, representing the error percentage as a function of  $\beta$ . These results are shown in Fig. 2 for selected values of noise level, namely  $\sigma = 0.5$  and 1. The values are not shown directly, but as ratio of error reduction with respect to the results on distances, used as a baseline. Results on distances pairwise are obtained by clustering the points after representing them using the distances with respect to the other points. The baseline error lies on the line Err% ratio = 1. The box-and-whiskers plot has been obtained by considering the results over 100 different generation of the perturbation. From the plot, it can be seen the error reduction over the baseline.

The second synthetic data set is composed by two 10-dimensional Gaussian distribution centered, respectively, in  $(-1, \dots, -1)$  and  $(1, \dots, 1)$ . Both the distributions have a diagonal

covariance matrix with entries equal to 4. Ten uniformly 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-whiskers plot obtained for different numbers of selected features over 100 trials. Fig. 3 shows the results for two values of  $\beta$ , namely  $\beta = 0.1$  and 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 are 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 on the crisp ranks, we decided to set the number of cluster to 3. We scaled the data setting the maximum of the norm of the patterns to one. 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 (dotted line). Eventually, the error achieved by clustering data in the input space is the same as in the clustering of the distances.

The results on the Colon data set are summarized 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 in many cases the use of ranks (crisp and fuzzy) performs better with respect to the clustering of patterns in input space and with respect to the clustering in the

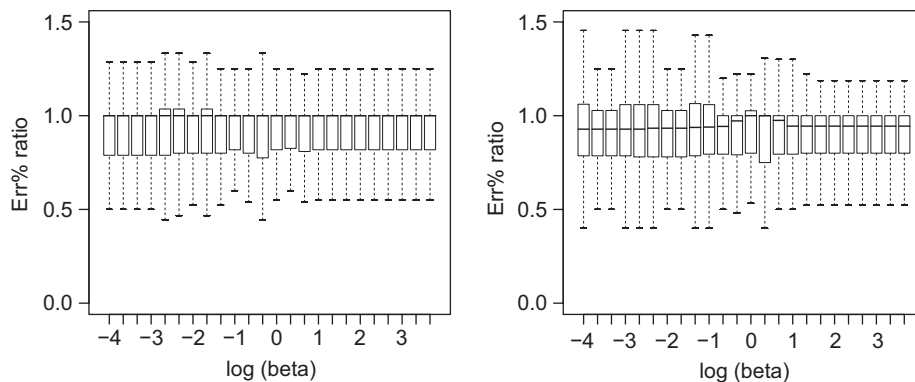


Fig. 2. Results on the Synthetic data 1—Left: standard deviation of the Gaussian perturbation  $\sigma = 0.5$ ; right: standard deviation of the Gaussian perturbation  $\sigma = 1$ .

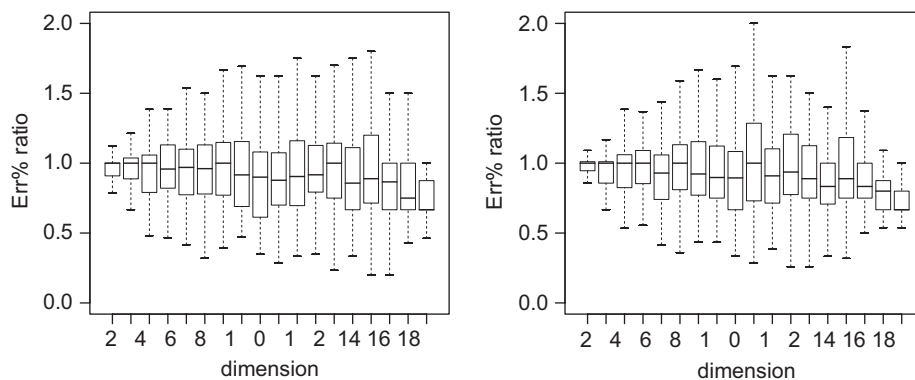
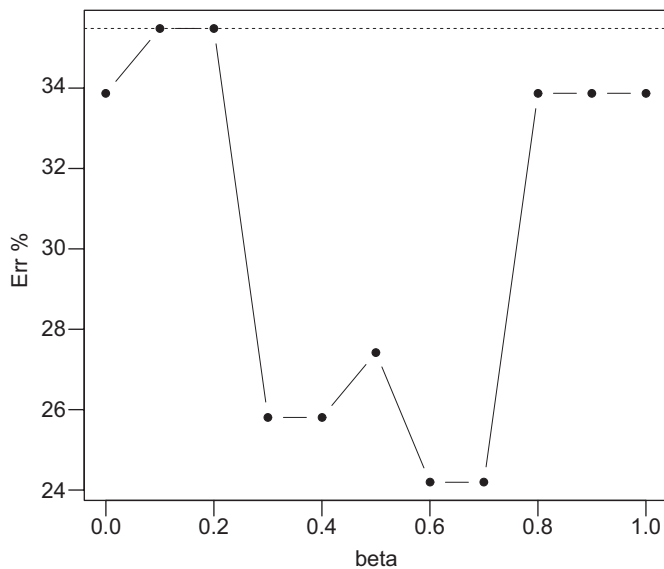


Fig. 3. Results on the Synthetic data 2—Left:  $\beta = 0.1$ ; right:  $\beta = 1$ .



**Fig. 4.** Results on the Colon data set. Error level as a function of the fuzziness parameter  $\beta$ . The dotted line represents the error obtained by clustering in the data space directly and the error obtained by clustering the distance matrix.

distance space (both the errors overlap on the dotted line). Here the use of the fuzzy ranks improves the classification error achieved by the crisp ranks (corresponding to  $\beta = 0$ ).

## 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  $\beta$ . 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 [8,13,14]. 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.

## Acknowledgments

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

## References

- [1] C.C. Aggarwal, P.S. Yu, Redefining clustering for high-dimensional applications, *IEEE Transactions on Knowledge and Data Engineering* 14 (2) (2002) 210–225.

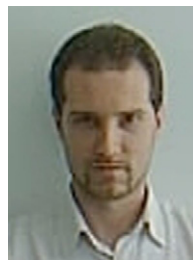
- [2] U. Alon, N. Barkai, D.A. Notterman, K. Gish, D. Mack, S. Ybarra, A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proceedings of the National Academy of Science USA* 96 (1999) 6745–6750.
- [3] K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, When is nearest neighbor meaningful?, in: 7th International Conference on Database Theory Proceedings (ICDT'99), Springer, Berlin, 1999, pp. 217–235.
- [4] G. Bortolan, R. Degani, A review of some methods for ranking fuzzy sets, *Fuzzy Sets and Systems* 15 (1985) 1–19.
- [6] A.K. Jain, R.C. Dubes, *Algorithms for Clustering Data*, Englewood Cliffs, NJ, USA, 1988.
- [7] L. Kaufman, P.J. Rousseeuw, *Finding Groups in Data*, Wiley, New York, USA, 1990.
- [8] J.B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika* 29 (1964) 1–27.
- [9] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: analysis and an algorithm, in: T.G. Dietterich, S. Becker, Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems*, vol. 14, MIT Press, Cambridge, MA, 2002.
- [10] E. Pekalska, P. Paclík, R.P.W. Duin, A generalized kernel approach to dissimilarity-based classification, *Journal of Machine Learning Research* 2 (2001) 175–211.
- [11] S. Rovetta, F. Masulli, Vector quantization and fuzzy ranks for image reconstruction, *Image and Vision Computing* 25 (2006) 204–213.
- [12] J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge, 2004.
- [13] R.N. Shepard, The analysis of proximities: multidimensional scaling with unknown distance function Part I, *Psychometrika* 27 (1962) 125–140.
- [14] R.N. Shepard, The analysis of proximities: multidimensional scaling with unknown distance function Part II, *Psychometrika* 27 (1962) 219–246.
- [15] R.R. Sokal, C.D. Michener, A statistical method for evaluating systematic relationships, *University of Kansas Science Bulletin* 38 (1958) 1409–1438.
- [16] W. Wang, E.E. Kerre, Reasonable properties for the ordering of fuzzy quantities (I), *Fuzzy Sets and Systems* 118 (2001) 375–385.
- [17] W. Wang, E.E. Kerre, Reasonable properties for the ordering of fuzzy quantities (II), *Fuzzy Sets and Systems* 118 (2001) 386–405.
- [18] J.H. Ward, Hierarchical Grouping to optimize an objective function, *Journal of American Statistical Association* 58 (1963) 236–244.



**Maurizio Filippone** received a Master's degree in Physics in 2004 and a PhD in Computer Science in 2008, from the University of Genova. In 2007, he has been Research Scholar at the Information and Software Engineering Department, at George Mason University. In 2008, he joined the Machine Learning group at the University of Sheffield as a Research Associate. His research interests are focused on Machine Learning techniques and their applications to pattern recognition, bioinformatics, and time series analysis and forecasting.



**Francesco Masulli** received the Laurea degree in Physics in 1976 from the University of Genova. He is currently an Associate Professor in Computer Science at the University of Genova. From 1983 to 2001 he has been an Assistant Professor at the University of Genova and from 2001 to 2005 an Associate Professor at University of Pisa. He authored or co-authored more than 120 scientific papers in Machine Learning, Neural Networks, Fuzzy Systems, Pattern Recognition and Bioinformatics.



**Stefano Rovetta** received a Laurea degree in Electronic Engineering in 1992 and a PhD degree in Models, Methods and Tools for Electronic and Electromagnetic Systems in 1996, both from the University of Genova. He is currently an Assistant Professor in Computer Science at the University of Genova. His current research interests include Machine Learning and Clustering techniques for high-dimensional biomedical data analysis and document analysis.