Soft Rank Clustering

Stefano Rovetta[1,2], Francesco Masulli[3,2], and Maurizio Filippone[1,2]

[1] Dipartimento di Informatica e Scienze dell'Informazione, Università di Genova,
Via Dodecaneso, 35, I -16146 Genova, Italy
[2] Istituto Nazionale per la Fisica della Materia, Unità di Genova,
Via Dodecaneso, 33, I -16146 Genova, Italy
[3] Dipartimento di Informatica, Università di Pisa,
Largo B. Pontecorvo, 3, I-56127 Pisa, Italy

Abstract. Clustering methods provide an useful tool to tackle the problem of exploring large-dimensional data. However many common approaches suffer from being applied in high-dimensional spaces. Building on a dissimilarity-based representation of data, we propose a dimensionality reduction technique which preserves the clustering structure of the data. The technique is designed for cases in which data dimensionality is large compared to the number of available observations. In these cases, we represent data in the space of soft D-ranks, by applying the concept of fuzzy ranking. A clustering procedure is then applied. Experimental results show that the method is able to retain the necessary information, while considerably reducing dimensionality.

1 Introduction

The exploration of large-dimensional data has always been an ubiquitous problem in science and information technologies. Clustering methods provide an useful tool to tackle this problem. Several clustering algorithms have been modified in the direction of incorporating fuzzy concepts (starting with the Fuzzy c-Means algorithm [1, 2]).

Many common approaches suffer from being applied in high-dimensional spaces. For instance, one of the most common methods, k-means 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-neighbour criterion which is the basis of the method may even become useless, in the sense that the distances of a given query point from its nearest and farthest neighbours tend to converge [3].

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 even easy to define the concept of volumetric density, let alone estimating it.

A further problem is again related to distances in high space dimensionality. Defining clusters on the basis of distance requires that distances can be estimated. However there are results [4] stating that, 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

B. Apolloni et al. (Eds.): WIRN/NAIS 2005, LNCS 3931, pp. 207–213, 2006.

to become equal. Therefore the evaluation of distances, and the concept of "nearest neighbor" itself, become less and less meaningful with growing dimension.

A notable complexity reduction in the presence of large-dimensional data sets is provided by representations 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 [5] may be used to represent data sets in a more compact way. Pękalska and Duin[6] 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.

We adopt the same 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, in which case it could not even be a symmetric matrix (for instance when obtained from subjective measurements by a panel of experts, or in the behavioral sciences) and no 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. Note that k-means type algorithms only work in metric spaces, and usually their extensions to non-metric cases are somewhat arbitrary. By representing data with this dissimilarity-based technique, we could apply this family of clustering algorithms even to non-metric data, e.g. categorical or mixed.

2 Clustering with Fuzzy Ranks

To tackle the dimensionality problem, a typical countermeasure 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 [7] of the concept of ranks.

Let ρ_{ij} be the rank of data point j with respect to data point i according to the set of dissimilarities $\{d(x_i, \cdot)\}$ when sorted in decreasing order with respect to values. This value is termed D-rank. It 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.

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 $\rho(x_1), \ldots, \rho(x_n)$. We can therefore represent a data point x_i by the vector of its D-ranks:

$$x_i \longrightarrow \rho(x_i) = [\rho_{i1}, \ldots, \rho_{in}] \tag{2}$$

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 procedure illustrated in the following section.

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 [8] and, more recently, by Wang and Kerre [9, 10].

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:

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

where:

$$\lim_{\beta \to 0} \frac{1}{1 + e^{(d_{ij} - d_{ik})/\beta}} = \theta\left(d_{ij} - d_{ij}\right) \tag{4}$$

So the computation of fuzzy rank can be expressed as

$$\rho_{ij} = \sum_{k=1}^{n} \frac{1}{1 + e^{(d_{ij} - d_{ik})/\beta}} \tag{5}$$

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 (5) for the rank function $\rho_{\cdot\cdot}$ are compared in a simple example, illustrated in Figure 1, where the following set of values is used: $\{d, 2, 3, 5\}$. The diagram is a plot of $\rho_{d\cdot}$ (in the two expressions, crisp and fuzzy) for d in the range $[0, 7]$. Two plots are shown for the fuzzy expression, one for $\beta = 0.05$ and another for $\beta = 0.25$ (smoother).

This new definition of rank allows us to integrate into a clustering algorithm 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. We want to exploit this added capability, and in the next section we present one possible proposal.

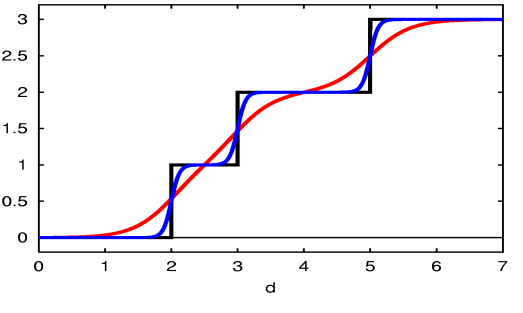

Fig. 1. Comparing crisp and fuzzy rank functions

3 Clustering Algorithms Using D-Ranks

To exploit the concept of soft-ranks, we apply a standard agglomerative hierarchical clustering algorithm to the soft D-ranks. We adopt the *agnes* procedure [11], which is available in the R language and environment [12], because it provides the agglomerative coefficient, which is defined as the average height of the mergers in a dendrogram.

We use the agglomerative coefficient to assess the value of β for which the clustering is best defined. However, with growing β the clusters in the soft D-rank space tend to collapse into one, since soft D-ranks tend all to their middle value $n/2$. Since we need to compare these values obtained on different scales, we use a different index which takes into account this problem. We define the weighted agglomerative coefficient a_w as follows:

$$a_w = a\left(\frac{max\{\rho_{ij}\} - min\{\rho_{ij}\}}{n}\right) \tag{6}$$

that equals zero for trivial clusters (one cluster for the whole data set).

4 Experiments

The method was tested on the publicly available Leukemia data by Golub et al. [13]. 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). 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 test was performed according to the proposed method for a number of different fuzziness levels β. The weighted agglomeration coefficient a_w was used to assess the "best" fuzziness level, 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).

The method which can be expected to find the most interesting clusters is Ward, since it is known to yield small and compact clusters not affected by the "chaining" effect, and this is fully confirmed by the experimental analysis.

Figure 2 shows the weighted agglomeration coefficient for the various linkage methods, with varying fuzziness level. The diagram shows that a peak is visible for a given

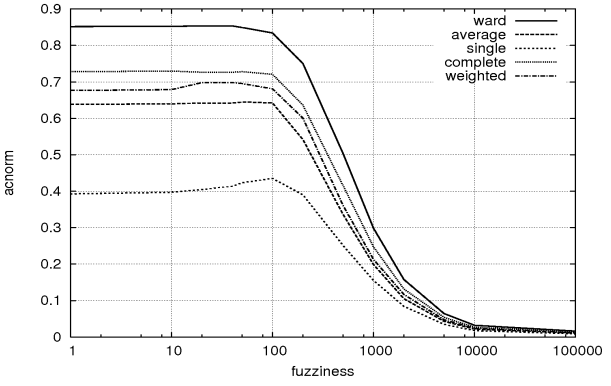

Fig. 2. Weighted agglomeration coefficient as a function of the fuzziness level β, for various linkage methods

value of β, especially for the less discriminative methods. For the method which finds the clearest clusters, the peak is hardly noticeable. Over a certain fuzziness threshold, the rank vectors tend to collapse in a single big cluster, and the agglomerative coefficient a rises; however the *weighted* agglomerative coefficient a_w plotted in the diagram correctly shows that this clustering is less and less significant.

Figure 3 shows the dendrograms obtained for some values of β with Ward's linkage method. The weighted agglomerative coefficient has a slight peak of 0.86 around $\beta = 50$, and inspection of the corresponding dendrogram reveals that indeed the clusters found are very adherent to the available domain knowledge. For instance, the first split in the dendrogram is between a cluster containing all ALL cases plus one AML case (observation 35), on the left, and a cluster containing only AML cases, on the right.

The subtypes of ALL are also fairly well evidenced, since all T-cell cases are in a major sub-cluster of the ALL cluster (on the right), with only two B-cell cases in the T-cell cluster.

For AML cases also prognostic informations are available, however the original work found no strong correlation between the molecular profile and therapy outcome (success or failure). This can also be found in the dendrogram we present here, since for cases 28-33 the outcome is failure and for cases 34-38 it is success, but no clustering is perceivable among these cases.

The analisys of clustering results thus reveals that the same conclusion which were obtained by several research step in ref. [13] are also found in a single application of our proposed method.

For comparison, we performed a cluster analysis according to three different methods: 1) agglomerative clustering on the original, 7129-dimensional data; 2) agglomerative clustering on the distance matrix; and 3) agglomerative clustering on the conventional rank matrix. In all cases, the obtained dendrograms do not show this level of match with respect to the knowledge available in the literature, thus confirming the superiority of the proposed method.

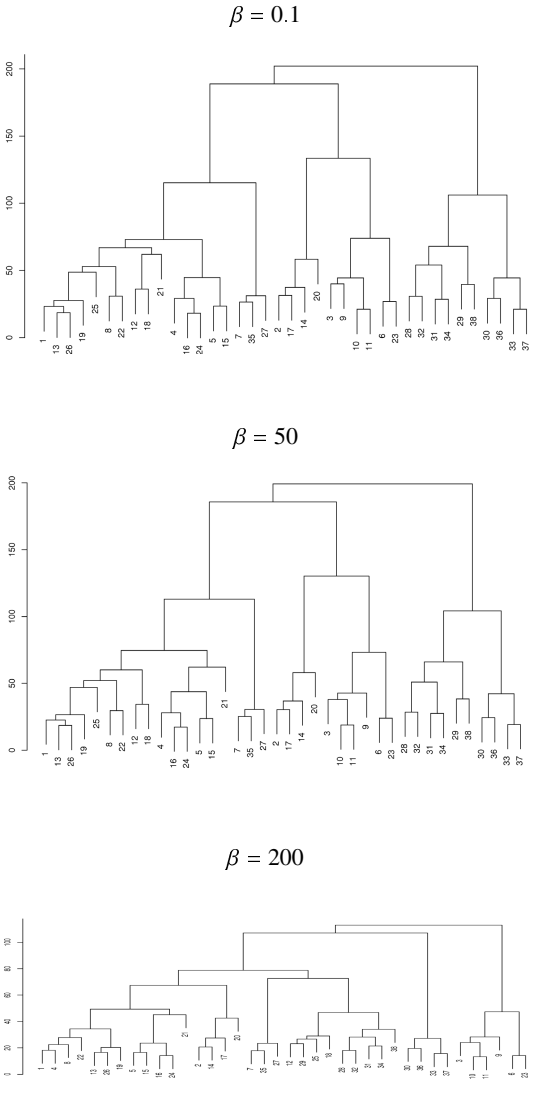

Fig. 3. Dendrograms obtained with different levels of fuzzines β with Ward's linkage method

5 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.

Acknowledgment

Work funded by the Italian National Institute for the Physics of Matter (INFM), 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. 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
2. Bezdek, J.C.: Pattern recognition with fuzzy objective function algorithms. Plenum, New York (1981)
3. Aggarwal, C.C., Yu, P.S.: Redefining clustering for high-dimensional applications. IEEE Transactions on Knowledge and Data Engineering **14** (2002) 210–225
4. 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
5. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press (2004)
6. 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
7. Masulli, F., Rovetta, S.: Fuzzy variations in the training of vector quantizers. In: Proceedings of the 2003 International Workshop on Fuzzy Logics, Napoli, Italy. (2003)
8. Bortolan, G., Degani, R.: A review of some methods for ranking fuzzy sets. Fuzzy Sets and Systems **15** (1985) 1–19
9. Wang, W., Kerre, E.: Reasonable properties for the ordering of fuzzy quantities (I). Fuzzy Sets and Systems **118** (2001) 375–385
10. Wang, W., Kerre, E.: Reasonable properties for the ordering of fuzzy quantities (II). Fuzzy Sets and Systems **118** (2001) 386–405
11. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data. John Wiley & Sons, New York, USA (1990)
12. Ihaka, R., Gentleman, R.: R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics **5** (1996) 299–314
13. 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