Available online at www.sciencedirect.com
ScienceDirect

Chemometrics and Intelligent Laboratory Systems 87 (2007) 208 – 217

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering

J.A.S. Almeida, L.M.S. Barbosa, A.A.C.C. Pais *, S.J. Formosinho

Departamento de Química, Universidade de Coimbra, 3004-535 Coimbra, Portugal

Received 18 November 2005; received in revised form 23 January 2007; accepted 24 January 2007
Available online 3 February 2007

Abstract

Techniques based on agglomerative hierarchical clustering constitute one of the most frequent approaches in unsupervised clustering. Some are based on the single linkage methodology, which has been shown to produce good results with sets of clusters of various sizes and shapes. However, the application of this type of algorithms in a wide variety of fields has posed a number of problems, such as the sensitivity to outliers and fluctuations in the density of data points. Additionally, these algorithms do not usually allow for automatic clustering.

In this work we propose a method to improve single linkage hierarchical cluster analysis (HCA), so as to circumvent most of these problems and attain the performance of most sophisticated new approaches. This completely automated method is based on a self-consistent outlier reduction approach, followed by the building-up of a descriptive function. This, in turn, allows to define natural clusters. Finally, the discarded objects may be optionally assigned to these clusters.

The validation of the method is carried out by employing widely used data sets available from literature and others for specific purposes created by the authors. Our method is shown to be very efficient in a large variety of situations.

Keywords: Clustering; Unsupervised pattern recognition; Hierarchical cluster analysis; Single linkage; Outlier removal

1. Introduction

Pattern recognition is a primary conceptual activity of the human being. Even without our awareness, clustering on the information that is conveyed to us is constant. This clustering activity is frequently based on a few selected properties and is not exempt from personal prejudice. Naturally, when objects are defined by a significant number of properties which are or have been made quantitative, and it is intended to obtain exempt results (natural clusters), the use of mathematical tools is mandatory. Actually, mathematical tools cannot be completely exempt. Firstly, because many algorithms rely on options made by the user. Also, because algorithms introduce some propensity for certain types of solution.

Identifying natural patterns in data is one of the most important goals of chemometrics. Specifically, clustering techniques are almost indispensable as a tool for data mining.

These techniques are often divided into unsupervised and supervised methodologies. In the former, information stems from the data and there are no pre-classified groups. Hierarchical cluster analysis usually produces a dendrogram, or other type of tree diagrams, as final output [1–3]. Each association level of the dendrogram represents a partitioning of the data set into a specific number of clusters [1]. Based on the dendrogram, it is possible to additionally define the number of clusters, but this step is often based on common sense relying on the representation of the data structure. In cases for which the data set is analyzed to produce a simple partitioning of the objects, resulting in a set of clusters, the technique is considered non-hierarchical [1]. In contrast, supervised methodologies involve classifying samples into predefined structures. Typically, there is a set of labelled objects (training set) that is used to establish decision rules so as to classify new objects [4].

In this work we propose an automated approach to find natural clusters, based on a new representation that highlights weak

* Corresponding author. Departamento de Química, Universidade de Coimbra, Rua Larga, 3004-535, Coimbra, Portugal. Tel.: +351 239854466; fax: +351 239827703.
E-mail address: pais@qui.uc.pt (A.A.C.C. Pais).

0169-7439/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2007.01.005

associations between groups. This representation is intended to complement the information provided by the dendrogram and, if the sole purpose is to separate data in an appropriate number of groups, it may even replace the dendrogram.

Situations in which data contain some amount of outliers or less characteristic data are difficult to tackle with most clustering algorithms. The data structure becomes less defined and the number of groups that are formed may be either too low or too high, depending on the actual situation and algorithm.

Identification and removal of outliers, in a preliminary step, is suggested via a self-consistent technique in which system properties are used to make an automatic specification of the necessary parameters. Additionally, the possibility of merging these initially discarded data in the formed clusters is left open. This final step is especially important when, rather than outliers, objects that are discarded correspond to less characteristic zones in the clusters.

The algorithm proposed in this work was applied to a variety of data sets, and results compared with those obtained by other algorithms. These include ROCK [5], CURE [6], DBSCAN [7], CHAMELEON [8] and FAÇADE [9]. We note that such algorithms were proposed as adequate approaches for finding clusters of different sizes, shapes, and densities in the presence of outliers.

ROCK combines, from a conceptual point of view, nearest-neighbor, relocation, and hierarchical agglomerative methods [1,5]. In this algorithm, cluster similarity is based on the number of points from different clusters that have neighbors in common [5,10]. The same authors developed another algorithm denoted as CURE that combines centroid and single linkage approaches by choosing more than one representative point from each cluster [1,6]. At each step of the algorithm, the two clusters with the closest pair of representative points (one in each cluster) are merged [6,10].

The CHAMELEON algorithm explores dynamic modeling in hierarchical methodology. In its clustering process, two clusters are merged if the respective inter-connectivity and proximity are highly related to the intra-cluster counter part [8,10].

A different approach is presented by the DBSCAN algorithm. It makes use of two external parameters, the minimum number of points in the neighborhood of a point, and the radius that defines this neighborhood. Choosing the appropriate parameters, it is then possible to identify objects located in high and low density zones. Neighboring objects in high density zones define clusters [7,10].

FAÇADE was proposed recently and is initiated by an outlier-eliminating process. Subsequently, the data is compressed, preserving original spatial patterns with a smaller number of data points. In this approach, the algorithm constructs a neighborhood graph, in which each sub-graph is regarded as a group. The clustering information of the compressed data is projected onto the original data. Finally, the groups are merged hierarchically according to the connections between each two groups [9].

2. Some concepts

There is a generic idea about what is cluster analysis and what is a cluster, which has not significantly changed in the last two decades. 'Cluster analysis is the term applied to a number of techniques that seek to divide a set of objects into several groups so that objects within the same group are more similar to each other than objects in different groups' [11]. In this context cluster is each one of the groups of similar objects.

However, this definition of cluster, which is clearly directed for well separated clusters [2], is not universal. In fact, it is not necessary in many situations that all objects belonging to a group be similar to each other [4]. Instead, it is necessary that these objects present a high connectivity among them. Connectivity is a concept that will be explored below in more detail, in the context of outlier identification. It can be regarded as the property that arises from the existence of a set of nearby objects, which allows to associate objects in a sequential mode. Thus, more dissimilar objects belonging to the same cluster can be joined by an uninterrupted chain of nearest objects.

A more general definition of clustering can thus be simply 'a data analysis technique that, when applied to a set of heterogeneous items, identifies homogeneous subgroups as defined by a given model or measure of similarity' [1].

Cluster analysis is thus often based on the concept of similarity. The easiest and most intuitive way to mathematically define the similarity between two objects is based on the Euclidean distance [12] which will be used, without loss of generality, in this work. However, it is well-known that the Euclidian distance may not be totally adequate for high dimensional systems [2,13,14]. In those cases, similarity measures based on other quantities can be employed. We note that, in chemistry-related applications, the correlation coefficient is also a very common choice, minimizing scale effects (see, e.g., [15]).

A very important issue in cluster analysis (and one of the most difficult [1,16]) is to establish the number of clusters present in a data set. Since the classical semi-automatic method where a user selects and extracts manually each cluster guided by a visual inspection of a dendrogram, there has been a number of proposals for more automatic procedures [16–21].

One way to approach this problem is resorting to natural clusters i.e., clusters that are clearly (and intuitively) defined by data [1]. This type of clusters arises in situations for which inter-cluster separations are significantly higher than those found within each cluster. Also, intra-cluster separations must be close to homogeneous.

More complicated situations arise when clusters are well defined but at different levels of resolution. In this case, there are clusters with a much smaller inter-group separation, even though significantly higher than the corresponding intra-group separation. The closest groups are not natural clusters. They can be regarded as sub-clusters belonging to a larger group, of which they form some internal structure. These levels of resolution pose additional problems when trying to design automatic ways to define the number of natural clusters, and this difficulty is clearly present in HCA. Dendrograms are often an illustration of the existence of different levels of resolution, suggesting different possible clustering solutions.

A frequent approach consists in forming clusters whenever the inter-group separations are distinct enough from the intra-group distances, even if it corresponds to induce a pronounced heterogeneity in inter-group division. In this case, the tendency is to go into the fine structure of clusters, often causing an excessive division of data and a simultaneous identification of clusters and sub-clusters, ignoring the different resolution levels in the system.

This conceptual view of clustering is reflected, for instance, in some recently proposed algorithms for automatic cluster extraction [18,19,21]. These methodologies are based on reachability plots, originally produced from results obtained via the OPTICS algorithm [3,22,23], which provide more descriptive representations than dendrograms. Such representations combine information given by density analysis and selective linkage techniques, as a way to improve the HCA.

The method presented in this work intends to explore a slightly different concept. The idea is that the identification of groups should derive from an external global view of all the system comprising the relevant set of objects and inter-object space. The first set of groups must be constrained by a correct degree of homogeneity in inter-cluster separation, aiming at recovering 'natural clusters' at this resolution level. This methodology allows to, in subsequent and iterative steps, subject the identified clusters (each now constituting a new whole system) to an internal identification. A similar procedure may be followed in most hierarchical algorithms.

3. Hierarchical cluster analysis

The techniques proposed in this work are based on hierarchical cluster analysis. In what follows, we provide a brief introduction to this methodology.

HCA is a method for finding the underlying structure of objects through an iterative process that associates (agglomerative methods) or dissociates (divisive methods) object by object, and that is halted when all objects have been processed [2]. The agglomerative procedure starts with each object in a separate cluster and then combines the clusters sequentially, reducing the number of clusters at each step until all objects belong to only one cluster. The divisive methods start with all of the objects in one cluster, and then proceed to their partition into smaller clusters until there is one object per cluster [1,11]. This means that for N objects, the process involves $N-1$ clustering steps.

In HCA there are two important choices when defining a method: the type of similarity measure between objects and/or groups, and the linkage technique [11].

The first task is to determine a numerical value for the similarity between objects, constructing a similarity matrix. The most popular ways to determine the similarity between objects use the Euclidean distance and the correlation coefficient, but there are many alternatives for similarity indicators [12,24].

The next step is to group or ungroup the objects. The most common approach is an agglomerative technique, whereby single objects are gradually connected to each other in groups. The first connection corresponds necessarily to the most similar pair of objects. Once the first group is formed, it is necessary to define the similarity between the new group and the remaining objects [24]. This step requires a new choice among a variety of available techniques. Some of the most used linkage algorithms are complete-linkage (or furthest-neighbor), single linkage (or nearest-neighbor), average-linkage (between groups and within groups), centroid method and Ward′s-linkage [1,12,25].

In this work, single linkage is the underlying technique. When a new group is formed, the corresponding distance to any other group is the minimal Euclidean distance of all possible distances between each object of the former group and each object of the latter.

Once the similarity measure and the linkage method are defined, the agglomeration of objects and groups in each step of the process follows the order of larger similarity [24]. The structure obtained by hierarchical clustering is often presented in the form of a dendrogram where each linkage step in the clustering process is represented by a connection line [1,25].

The application of different methods, which may involve different similarity measures, different linkage techniques, etc., leads to dendrograms with different structures. Apparently, a good approach would be to use different methods of cluster analysis and compare the results, but due to an excessive wealth of options it is frequently more convenient to use well founded a priori choices.

It is widely accepted that the average-linkage, centroid and Ward′s methods are sensitive to the shape and size of clusters. Thus, they can easily fail when clusters have complicated forms departing from the hyperspherical shape [1]. Complete-linkage is not strongly affected by outliers, but can break large clusters, and has trouble with convex shapes [2]. The single linkage methodology, on the other hand, displays total insensibility to shape and size of clusters [1]. However, there are also shortcomings associated with single linkage, which is the sensitivity to the presence of outliers and the difficulty in dealing with severe differences in the density of clusters.

It is apparent that each method has its own limitations and scope of application. We show in this work how to considerably improve agglomerative HCA based on single linkage. This improvement uses two new analysis tools. One of these tools is aimed at the self-consistent identification of outliers in data and its consequent remotion. The other consists in a representation of the data structure, complementary to the dendrogram, through a descriptive function that stresses low connectivities among objects, defining potential zones of cluster division, and pinpoints the inhomogeneity present. Most of all, the number of clusters arises naturally in our method.

4. Computational procedure

This section deals with the algorithm for the modified HCA proposed in this work. The algorithm comprises three tasks: (i) outlier-removal, (ii) identification of groups (including building the descriptive function, and establishing the clusters), and (iii) classification of the objects discarded in the first step.

4.1. Removal of outliers

Identification of outliers is a relatively new concept in cluster analysis. The presence of outliers may have different consequences in different algorithms, and we will focus on the methodology used in this work. Specifically, in the single linkage approach, outliers promote both an excess of divisions prompted by the existence of isolated objects or small groups of objects, and under-divisions due to 'bridges' of outliers connecting what would otherwise be 'natural clusters' [2].

Outliers can be viewed as objects or small groups of objects located in low density zones, contrasting with the denser intra-cluster structure. In a similar view, but a slightly different perspective, outliers can be regarded as objects with low connectivity in opposition to higher connectivity in the intra-cluster region.

To identify low connectivity zones we use a D-dimensional target of a specific hyperradius. The center of the target is fixed in each object and the number of nearest-neighbors enclosed in this target establishes the connectivity of that object, c_i. If this procedure is repeated for every object, potential zones of outliers are pinpointed. This is not a new concept to identify outliers and it was previously proposed with density algorithms [14]. However, in these methods, it is necessary to provide external parameters to define the size of the target and the lower limit for density, below which an object is regarded as an outlier [9,14]. To avoid external parameters, we use some characteristics of the system and an iterative algorithm for removal of outliers that converges to a data set with more homogeneous connectivity. The internal parameters are established on the basis of the average nearest-neighbor distance of all objects (first parameter) and the average connectivity for all objects (second parameter). The latter depends on the previous parameter. In a convergence process, these parameters are automatically adjusted each time an elimination process is carried out, until the characteristics of the system stabilize i.e., there are no significant connectivity variations from object to object.

In practice, the convergence process is split in two. In the first one, the radius of the target is taken as $4\overline{d}_j$, where $\overline{d}_j$ is the average nearest-neighbor distance previous to iteration j. In the second, a smaller multiplier is used, and the radius is $2\overline{d}_j$.

In both processes, points with a connectivity lower than 1/3 of the average value for connectivity, $\overline{c}_j$, are discarded in each iteration. The value of $\overline{d}_j$ is recalculated and the process repeated, until the number of objects discarded is zero. The rationale for two convergence processes is simple. In the first one the objective is to remove both scattered objects or small groups and thin bridges of noise linking 'natural clusters'. However, the use of a $4\overline{d}_j$ radius has a consequence, related to the fact that a larger target is less sensitive to outliers present close to the boundaries of the clusters. The second iterative process overcomes this difficulty.

The identification of outliers may be summarised as follows:

```
Set iteration counter $j = 1$
Repeat until number of discarded objects = 0
    Calculate $\overline{d}_j$
    Set $R = 4\overline{d}_j$
    Calculate $\overline{c}_j\ (R)$
    Discard objects $i$ if $c_i < 1/3\overline{c}_j\ (R)$
    Increase $j$
```

followed by a new process in which $R = 2\overline{d}_j$.

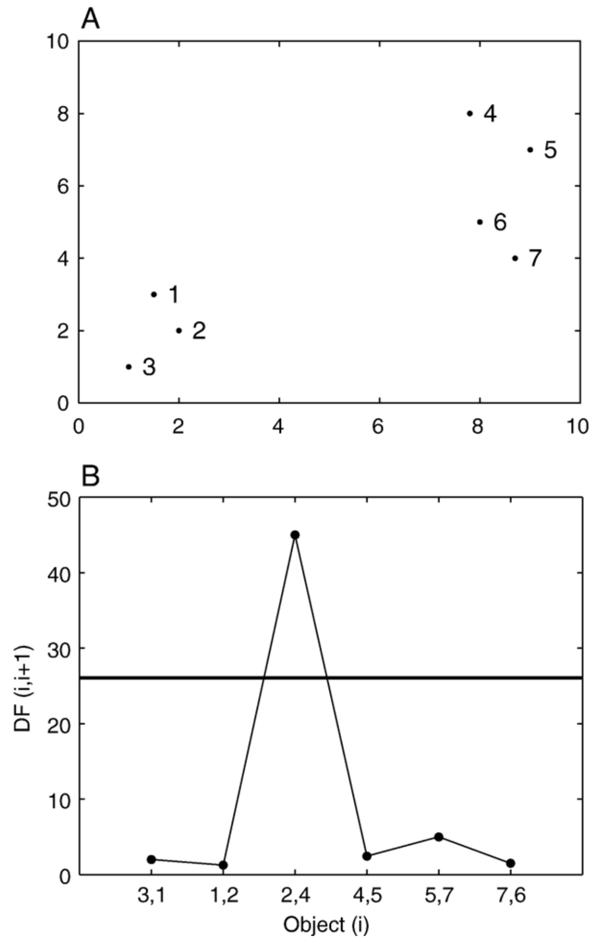

Fig. 1. Original data set (A) and descriptive function (B) for the example outlined in Tables 1 and 2.

The use of the average nearest-neighbor distance as system metric is one of the strongest points in our method. Firstly, because the average value reflects the presence of non-characteristic values, which then allows to converge the internal parameters so as to obtain more homogeneous intra-cluster zones. Moreover, it is very important to stress that the specific values of the multipliers used to define the dimensions of the target are not critical. As the procedure relies on an overall comparison of the connectivity values for all objects (to eliminate those with lower connectivity values), the tendency is to obtain similar results for a large range of multipliers. Naturally, it is not possible to use targets so small that the connectivity is zero for most of the objects in the system. This confines the choice to values higher than one, corresponding to radii above the average nearest-neighbor distance. Also, large targets tend to establish overlapping regions around contiguous objects and not produce an adequate characterisation of the system. As a conclusion, the choice of low values, although larger than unity, is appropriate for most cases. The choice of these values is even less critical given the fact that the classification procedure proposed for the discarded objects (see below) corrects most of the incorrect assumptions from this preliminary treatment.

It was referred that objects with a connectivity lower than 1/3 of the average value are discarded. This value was obtained

Table 1
Schematic description of the association process in the classification algorithm proposed in this work

Linkage step	Association vector block	Nearest-neighbor distance
1	1, 2	d_1
2	7, 6	d_2
3	3, 1, 2	d_3
4	4, 5	d_4
5	4, 5, 7, 6	d_5
6	3, 1, 2, 4, 5, 7, 6	d_6

from an extensive battery of tests that have shown the procedure to attain a degree of intra-cluster homogeneity that clearly facilitates clustering.

4.2. Main clustering step

The first step of the algorithm corresponds, as usual when resorting to single linkage, to the calculation of the nearest-neighbor Euclidean distance matrix. The minimal Euclidian distance (major similarity) is used to associate the first two objects. This association methodology will pursue associating objects with objects, objects with groups of objects, or groups with groups of objects and is halted when all objects are associated in just one vector. For each association step we preserve the similarity value and the index of the objects that constitute the new cluster. This is the data required to construct the proposed descriptive function.

Let us clarify the order of the objects present in the final association vector. When two objects are first associated, they are always placed in consecutive positions. When an additional object or group is further associated into an existing structure, being either a single object or a formed grouped, it will appear in the vector immediately before or after the original group. This implies that, during the hierarchical association procedure, a formed block suffers no changes in subsequent association steps.

In a second phase, looking at each pair of consecutive objects, i, $i+1$, of this association vector we calculate the value of the descriptive function, $\mathrm{DF}_{i,i+1}$. This corresponds to the squared minimal distance measure of all linkage steps in which both objects participate,

$$\mathrm{DF}_{i,i+1} = d^2_{i,i+1}. \tag{1}$$

The mathematical function given by $\mathrm{DF}_{i,i+1}$ for each pair of sequential objects in the association vector will produce localised higher peaks, corresponding to a high probability of inter-cluster separation, and low value regions indicating a high probability of intra-cluster association. The use of a squared distance is intended to emphasise inter-cluster separations.

The presented descriptive function is obviously not the only possible, but is one of the simplest that can be used to produce results of quality comparable to the most sophisticated approaches.

To separate data into clusters, it is necessary that we recognise peaks corresponding to the mentioned separations. For this, we simply identify values of $\mathrm{DF}_{i,i+1}$ based on a modified outlier scheme for the corresponding distributions, as extracted from the overall system. Thus, an inter-cluster separation $\mathrm{DF}_{\mathrm{sep}}$ is found in the descriptive function for

$$\mathrm{DF}_{\mathrm{sep}} > 6 \times (Q_3 - Q_1) \tag{2}$$

where Q_1 and Q_3 are the upper limits of the first and third quartile of the distribution of values in the descriptive function. The above equation further emphasises the usual definition of severe outlier. We note that the use of a single cut-line for the whole system corresponds to the concept of an overall assessment of cluster structure, as discussed in Section 2.

To illustrate the procedure we use a simple example based in the small data set of Fig. 1(A), for which the association steps are depicted in Table 1. This table describes the growth of the association vector, and its last element of the second column displays the final form of this vector. Analysis of the vector, and the building-up of the respective descriptive function [shown in Fig. 1(B) together with the separation line from Eq. (2)] is schematically presented in Table 2.

The order of the objects in the final association vector does not apparently retain any relevant information on the system structure. We use the first pair of consecutive objects (3 and 1) in the example to explain how the structural information is still present. When this pair of objects is analysed, the information that is extracted does not correspond to the Euclidean distance of the pair, but rather to the distance in which these two *objects* were first associated in the same group, d_3. This distance is, in fact, that between objects 2 and 3 following a nearest-neighbor scheme. This means that using the association vector and the information contained in Table 1, obtained during the associative process, it is possible to recover relevant structural information on the system and build-up the descriptive function. It should be stressed that the proposed method places objects from the same group contiguously in the descriptive function, separated by the higher peaks, without any reorganisation of the final association vector. This placement corresponds, simply, to one of the possible orderings for which a dendrogram would be built without line crossings. It also further clarifies that each value of $\mathrm{DF}_{j,j+1}$ can be directly extracted from the dendrogram organised as stated above.

4.3. Classification of the discarded objects

An optional final step after identifying 'natural clusters' is the classification of the initially discarded objects. In the pretreatment for cleaning outliers, some scattered data points are

Table 2
Analysis of the final association vector and distances to build the descriptive function

Pair of consecutive objects	Distance for linkage steps where both objects participate	Value of descriptive function
3, 1	d_3, d_6	d_3^2
1, 2	d_1, d_3, d_6	d_1^2
2, 4	d_6	d_6^2
4, 5	d_4, d_6	d_4^2
5, 7	d_5, d_6	d_5^2
7, 6	d_2, d_5, d_6	d_2^2

identified as outliers and therefore eliminated so as to improve the main clustering process. However, these data may not be necessarily outliers. These can, for instance, stem from a 'dilution' phenomenon in which boundary objects have a larger scatter, when compared to more characteristic objects closer to the center of the cluster.

In such cases, a scheme based on supervised pattern recognition is usually adequate. A simple K-nearest-neighbor approach (see [24]) has shown to produce good results in the examples studied in the scope of this work.

5. Results and discussion

The algorithm described in the previous section has been applied to a variety of systems. In Fig. 2 we present the results obtained with a two dimensional set of objects created by the authors. This relatively simple set comprises clusters with different number of objects. It includes non-challenging circular shapes, but also concentric circular crowns, which represent an additional difficulty for several clustering algorithms. Scattered outlier points were superimposed on the underlying data.

The system was previously subject to the 'outlier-removal' process. As can be seen in panels (A) and (B) the denser zones have remained essentially untouched, while the scattered and less densely distributed objects have been discarded. After this filtering, the system gives rise to the descriptive function shown in panel (C). The threshold defined in Eq. (2) is also depicted. The seven clusters defined by the identified peaks are represented in panel (D). It can be seen that the algorithm is not affected by the size of the clusters. Concentric, but clearly separated motifs are recognised as different clusters, unlike what would be expected from a centroid method.

Fig. 3 presents a data set [26] that has been used to test a variety of clustering algorithms. It comprises dense areas of varying shapes and sizes. There are clearly scattered objects, but also stripes of outliers (disposed horizontally and vertically) connecting what visually impacts as clusters. Once again, the removal process acts essentially in the inter-cluster region, cleaning both scattered objects and the thin 'bridges' that connect the clusters [see panels (A) and (B)]. Panel (C) represents the nearest-neighbor distributions for the original system and after cleaning the data, respectively. It is seen that, although the overall appearance of the distribution remains unaltered, the longer distance tail significantly differs in these two systems.

If we look at the corresponding descriptive function in Fig. 4, we see that there is a dominant peak, coexisting with a set of much smaller ones. These apparently small peaks, in turn, contrast very markedly with the intra-cluster background (note the insert in this panel). The separation, as assessed from the visual interpretation [panel (B)] seems to underdivide the set of three clusters comprising the external contour

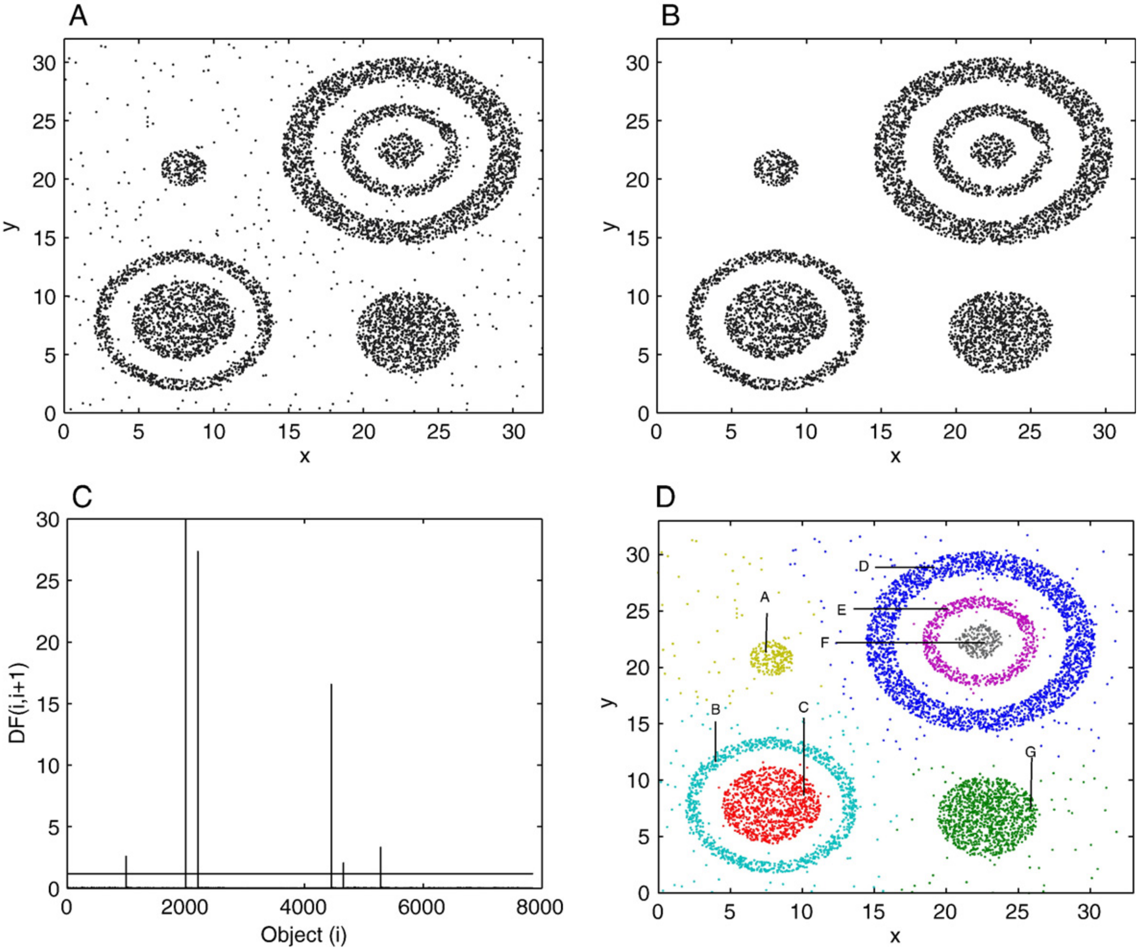

Fig. 2. Original data set (A), and result of the cleaning procedure (B), the corresponding descriptive function (C), and the overall classification result (D). In this and what follows arbitrary labels are assigned to identified clusters. This set is characterised by dense areas of varying shapes (circles and circular crowns) and sizes. It also comprises clusters surrounded by other clusters and a significant amount of outliers. Results pertain to the algorithm depicted in this work.

of an oval plus the two inserted circles. One of these circles is considered as associated to the external oval. In a sense, this fact results simply from the fact that the density of the bridging points is high enough to promote the connection between the two clusters. Also, these two clusters are very close. The direct use of the CURE algorithm [6] on the set with outliers, met with serious difficulties, irrespective of the choice of parameters [8,27]. A similar situation was encountered [8,26] with the ROCK algorithm [5]. In both cases the algorithms tend to produce an excessive number of clusters, although for ROCK other choices of parameters have led to some degree of underdivision [27]. The CHAMELEON [8] and FAÇADE [9] algorithms have divided the system adequately [8,27]. CHAMELEON has been used directly in the original system, originating some spurious groups. The FAÇADE algorithm has a similar behavior in the presence of outliers, but yielded the correct division when these are removed. Finally, the DBSCAN algorithm [7] may produce very good results, but requires an appropriate choice of parameters [8,27]. In fact, the final results are similar to those obtained using the algorithm proposed in this work, but minor changes in the values of parameters may significantly affect the analysis [27]. Application of other algorithms to the data set depicted in Fig. 3(A) can be found in Ref. [27].

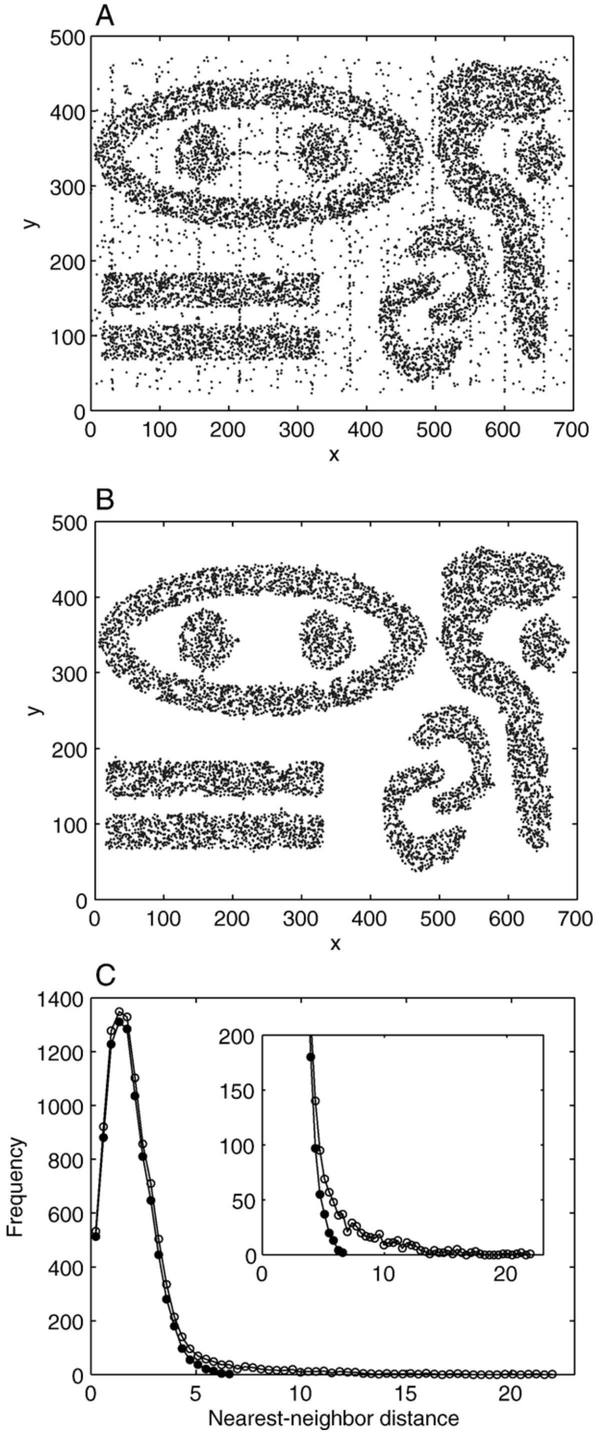

Fig. 3. Procedure for the cleaning of outliers in a set commonly used in the literature: clusters differ in size and shape and scattered points coexist with horizontal and vertical thin connecting structures. Panel (A) represents the original data structure, (B) the data set without outliers, and (C) the corresponding nearest-neighbor distributions. Key: ○, including outliers; ●, after the respective removal.

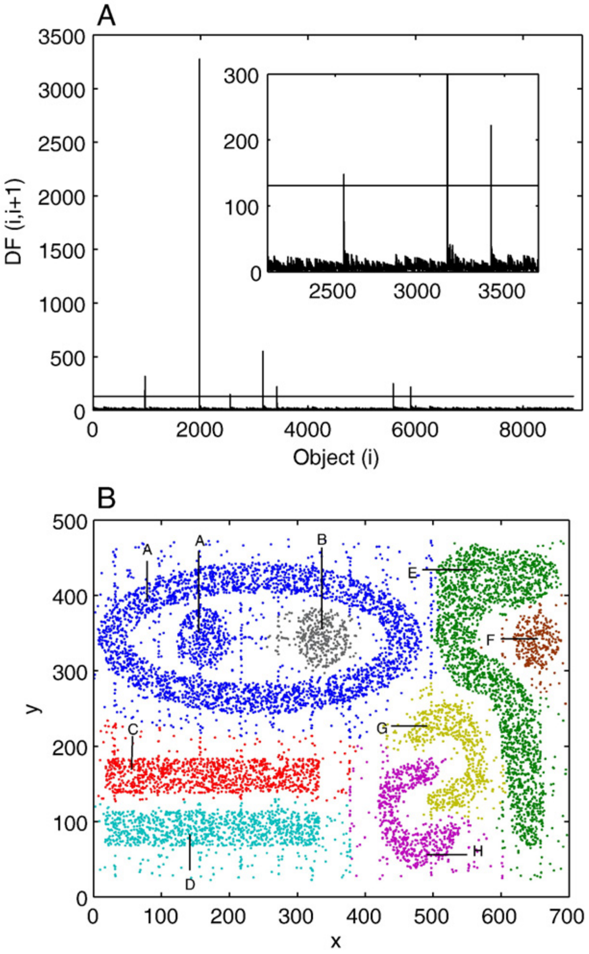

Fig. 4. Descriptive function (A) and resulting clusters (B) for the set of Fig. 3 using the procedure from this work.

In Fig. 5 a system with denser connecting bridges is presented [26]. The removal of scattered points is adequate, and the six large groups are clearly defined in the descriptive function of panel (A). Also, panel (B) shows the main clustering results and the partitioning of objects discarded in the previous step. They are ascribed to larger clusters in a sensible way. The CURE algorithm [6] is very dependent on the choice of external parameters [8]. Generally, the tendency is to produce an excessive number of clusters. This behavior results from the interpenetrating clusters. At the same time, it is not strongly affected by the

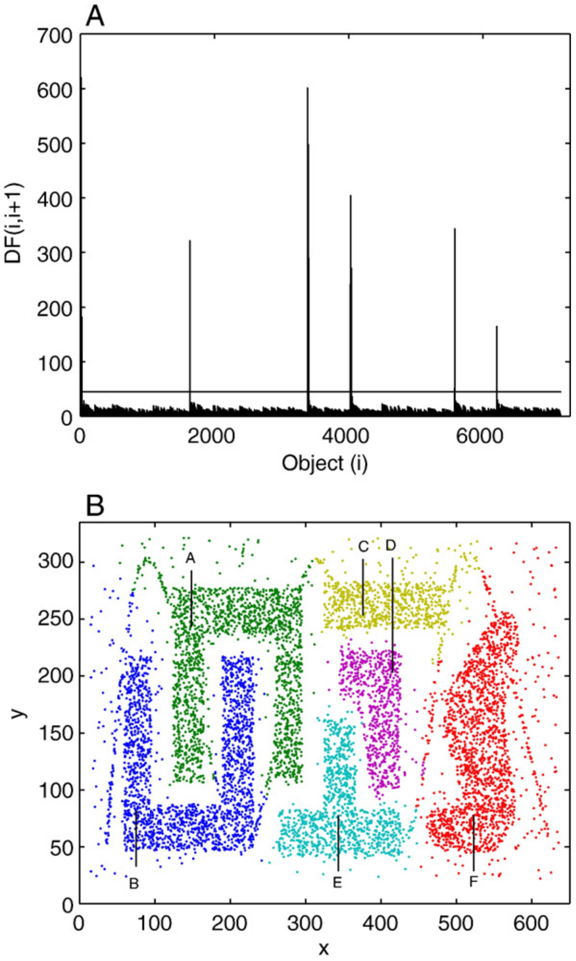

Fig. 5. Same as Fig. 2, for another set commonly used in the literature, in which clusters depart from spherical shape and interpenetrate. Strong connecting bridges are visible in conjunction with a large amount of scattered points.

presence of outliers. ROCK [5], in the other hand, is not adapted to divide this data set [8,26]. Once again, the more sophisticated CHAMELEON [8] follows what should be expected from the visual judgement. Some of the connecting bridges and scattered points are considered as individual clusters, but even these choices are appropriate [8]. A similar behavior is found for FAÇADE [9].

In Chemistry, we are often faced with small data sets (see, e.g. [28]), characterised by less well defined structures, which may induce a certain degree of overlap that directly results from a gradual variation of underlying factors. In Fig. 6 we approach this type of problem with simulated data comprising three slightly overlapping clusters. Each one possesses a well defined core of objects, but the characteristics become less and less marked as we travel towards the periphery. The algorithm presented here has formed the clusters in a very adequate manner, in spite of the fact that there was no a priori suggestion on the number of clusters. In such data sets, the algorithm for removal of outliers has a different purpose. It clears the areas with more scattered, i.e., less characteristic objects, emphasising the central core and thus the inter-group separations. Note, as well, the relevance of the classification procedure that allocates the discarded objects, clearly not outliers, to the groups previously formed.

The final example illustrates, thus, one of the remaining difficulties of major automatic cluster extraction algorithms, which is the recognition of patterns with overlapping structures that gradually evolve from a well defined cluster to another, through a set of objects with intermediate characteristics. This type of problem further justifies the use of a pre-treatment step. Also, it clearly improves the efficiency of algorithms based on the evaluation of local peaks used in some methodologies [19] and that usually face some problems with slowly declining peaks.

Despite any similarity that seems to exist between the descriptive function presented in this work and the reachability plots [18,19,22,23], there is a very different information represented in each one of them. Firstly, we are not representing a characteristic value of each object. We are instead representing a value that establishes the ”propensity” for two objects to belong to the same cluster. Our representation does not also include any information concerning density, using solely that obtainable from a single linkage association technique.

The descriptive function can also be directly used upon data sets not subjected to the procedure for the removal of outliers. In Fig. 7 we present the descriptive function corresponding to the data set of Fig. 3(A), without previously removing outliers. It contrasts with its counterpart in Fig. 4(A), being less well defined. Weaker associations are, however, still clearly marked.

All data sets presented in this work are two dimensional, so that the solution may be assessed through visual inspection. It should be

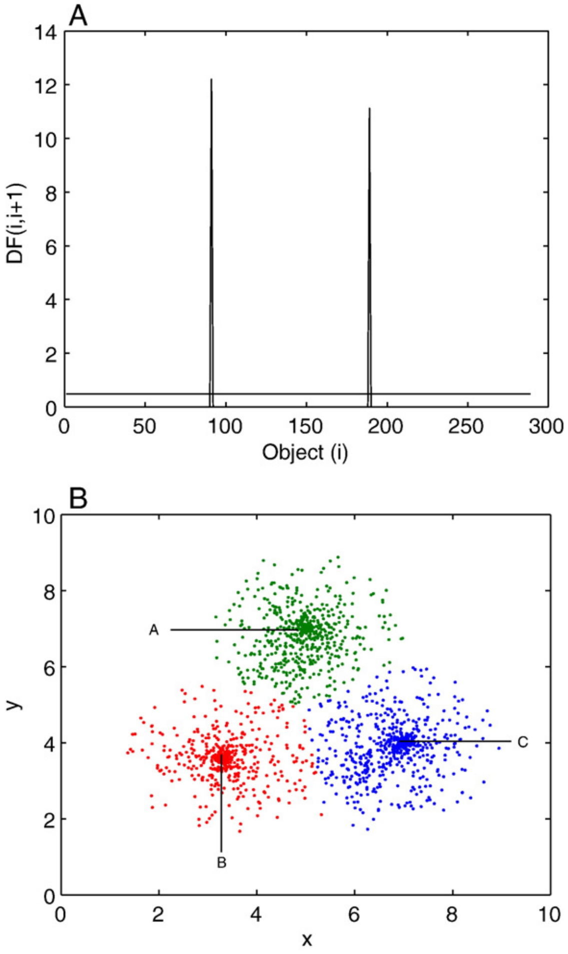

Fig. 6. Data consisting of three overlapping clusters, each one significantly less dense in the periphery. Panel (A) depicts the descriptive function, and panel (B) the final classification using the algorithm proposed in this work.

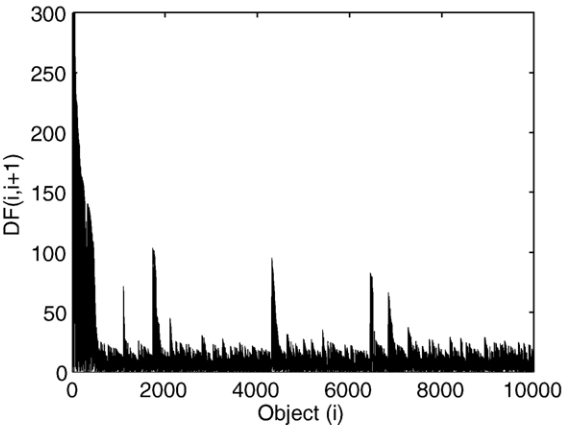

Fig. 7. Aspect of the descriptive function corresponding to the data set of Fig. 3(A), without removing the outliers.

remarked that we have conducted successful tests on a number of higher dimensional sets (see, e.g., Ref. [29]).

6. Conclusion

In this work we have proposed to use information usually obtained in common HCA procedures in order to identify 'natural clusters', without resorting to external parameters. In fact, the procedure described here is completely automated and attains a degree of reliability comparable to the most sophisticated approaches. It complements the information usually obtained from HCA, often summarised in a dendrogram, with a descriptive function that leads to the number of 'natural clusters' and the corresponding separation. This feature is obviously absent in classical HCA. The methodology presented also includes two additional steps. A previous one, in which outliers are removed, which may be applied in a variety of other filtering applications, and an optional classification after the main clustering analysis, that allows to enlarge the clusters with the whole set of objects.

The algorithm presented has been applied to a variety of data sets. The behavior was almost flawless in most of them, with an adequate recognition of the 'natural clusters'.

The descriptive function has shown to be very adequate in the visualisation of large data sets, for which dendrograms become extremely cumbersome.

It was also specifically employed in sets comprising 'diluting' clusters, i.e., clusters that possess scattered objects in their periphery. Recognition has been very satisfactory, in spite of the significant degree of overlap between adjacent clusters.

Acknowledgements

The authors would like to thank Fundação para a Ciência e a Tecnologia (Portugal) and FEDER for the financial support through project POCTI/QUI/45331/2002. JASA also acknowledges FCT for a jr. research grant in the scope of the above project.

References

[1] G.M. Downs, J.M. Barnard, Clustering methods and their uses in computational chemistry, in: K.B. Lipkowitz, D.B. Boyd (Eds.), Reviews in Computational Chemistry, vol. 18, Wiley, United Kingdom, 2002, pp. 1–40.

[2] M. Steinbach, L. Ertoz, V. Kumar, Challenges of clustering in high dimensional data, University of Minnesota Supercomputing Institute Research Report, vol. 213, 2003, pp. 1–33.

[3] M. Daszykowski, B. Walczak, D.L. Massart, Density-based clustering for exploration of analytical data, Anal. Bioanal. Chem. 380 (2004) 370–372.

[4] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264–323.

[5] S. Guha, R. Rsatogi, K. Shim, Rock: a robust clustering algorithm for categorical attributes, Inf. Syst. 25 (2) (2000) 345–366.

[6] S. Guha, R. Rsatogi, K. Shim, Cure: an efficient clustering algorithm for large databases, Inf. Syst. 26 (1) (2001) 35–58.

[7] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density based algorithm for discovering clusters in large spatial databases with noise, Proceedings of ACM Knowledge Discovery and Data Mining Conference, 1996, pp. 226–231.

[8] G. Karypis, E.-H. Han, V. Kumar, Chameleon: a hierarchical clustering algorithm using dynamic modeling, IEEE Comput. 32 (8) (1999) 68–75.

[9] Y. Qian, G. Zhang, K. Zhang, FAÇADE: a fast and effective approach to the discovery of dense clusters in noisy spatial data, Proceedings of the ACM SIGMOD International Conference on Management of Data, ACM, Paris, 2004, pp. 921–922.

[10] J. Han, M. Kamber, Data Mining: Concepts and Techniques, 2nd edition, Kaufmann, 2006 Ch. 7.

[11] N. Bratchell, Cluster analysis, Chemometr. Intell. Lab. Syst. (1989) 105–125.

[12] R. Kellner, J. Mermet, M. Otto, M. Valcárcel, H.M. Widmer (Eds.), Analytical Chemistry: a Modern Approach to Analytical Science, 2nd edition, Wiley, 2004, pp. 176–189, Ch. 8.

[13] L. Ertöz, M. Steinbach, V. Kumar, A new shared nearest neighbor clustering algorithm and its applications, Proceedings of the Workshop on Clustering High Dimensional Data and its Applications at 2nd SIAM International Conference on Data Mining, Arlington, 2002.

[14] L. Ertöz, M. Steinbach, V. Kumar, Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data, Proceedings of the 3rd SIAM International Conference on Data Mining, 2003, pp. 47–58.

[15] F.O. Costa, J.J.S. Sousa, A.A.C.C. Pais, S.J. Formosinho, Comparison of dissolution profiles of ibuprofen pellets, J. Control. Release 89 (2003) 199–212.

[16] C.A. Sugar, G.M. James, Finding the number of clusters in a data set: an information theoretic approach, J. Am. Stat. Assoc. 98 (2003) 750–763.

[17] S. Salvador, P. Chan, Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms, Proceedings of the 16th IEE International Conference on Tools with AI, 2004, pp. 576–584.

[18] J. Sander, X. Qin, Z. Lu, N. Niu, A. Kovarsky, Automatic extraction of clusters from hierarchical clustering representations, Proceedings of the 6th Pacific–Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2003), Springer, Seoul, 2003, pp. 75–87.

[19] H.-P. Kriegel, S. Brecheisen, E. Januzaj, P. Kröger, M. Pteifle, Visual mining of cluster hierarchies, Proceedings of the 3rd International ICDM Workshop on Visual Data Mining, Melbourne, 2003, pp. 151–165.

[20] S. Brecheisen, H.-P. Kriegel, P. Kröger, M. Pfeifle, Visually mining through cluster hierarchies, Proceedings of the 4th SIAM International Conference on Data Mining, Orlando, 2004, pp. 400–411.

[21] M. Daszykowski, B. Walczak, D.L. Massart, Looking for natural patterns in data part 1. density-based approach, Chemom. Intell. Lab. Syst. 56 (2001) 83–92.

[22] M. Ankerst, M.M. Breunig, H.-P. Kriegel, J. Sander, Optics: ordering points to identify the clustering structure, Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, vol. 28, Philadelphia, 1999, pp. 49–60.

[23] M. Daszykowski, B. Walczak, D.L. Massart, Looking for natural patterns in analytical data. 2. tracing local density with optics, J. Chem. Inf. Comput. Sci. 42 (2002) 500–507.

[24] R.G. Brereton, Chemometrics, Data Analysis for the Laboratory and Chemical Plant, 1st edition, Wiley, London, 2004.

[25] A. Smoliński, B. Walczak, J.W. Einax, Hierarchical clustering extended with visual complements of environmental data set, Chemom. Intell. Lab. Syst. 64 (2002) 45–54.

[26] G. Karypis, E.-H. H., V. Kumar, Chameleon: other experimental results, in: http://www.cs.umn.edu/~han/chameleon.html, 1999.

[27] O.R. Zaïane, A. Foss, C.-H. Lee, W. Wang, On data clustering analysis: scalability, constraints and validation, Proceedings of the 6th Pacific–Asia Conference on Knowledge Discovery and Data Mining, Springer, 2002, pp. 28–39.

[28] D. Coomans, D.L. Massart, Potential methods in pattern recognition: Part 2. clupot — an unsupervised pattern recognition technique, Anal. Chim. Acta 133 (1981) 225–239.

[29] J.M.G. Sarraguca, R.S. Dias, A.A.C.C. Pais, Coil–globule coexistence and compaction of DNA chains, J. Biol. Phys. 32 (2006) 421–434.