Journal of Mathematical Chemistry, Vol. 40, No. 3, October 2006 (© 2006)
DOI: 10.1007/s10910-006-9164-z

Clustering of chemical databases by means of the projection of maximum overlapping sets similarity measurements onto multidimensional spaces

Gonzalo Cerruela García, Irene Luque Ruiz and Miguel Ángel Gómez-Nieto*

Department of Computing and Numerical Analysis, University of Córdoba, Campus Universitario de Rabanales, Albert Einstein Building, E-14071 Córdoba, Spain
E-mail: mangel@uco.es

In this paper, we propose a new method for clustering of chemical databases based on the representation of measurements of structural similarity onto multidimensional spaces. The proposed method permits the tuning of the clustering process through the selection of the dimension of the projection space, the normal vectors and the sensibility of the projection process. The structural similarity of each element regarding to the database elements is projected onto the defined spaces generating clusters that represent the characteristics and diversity of the database and whose size and characteristics can be easily adjusted.

KEY WORDS: clustering, chemical databases, similarity measurements, maximum overlapping set

AMS Classification: 92E10, 91C20, 68P20, 94C15

1. Introduction

The size of all chemical databases is increasing dramatically every year; so powerful searching algorithms are necessary for these databases to be used efficiently by researchers and industry. Usually the process of information recovery consists of a screening stage where a set of molecules satisfying search criteria is recovered from the database, and a matching stage in which the molecules recovered are compared to each other *atom by atom search stage* (ABAS) [1–3].

The ABAS stage is the one with the highest computational cost required. This computational cost is inversely proportional to the size of the set of molecules recovered in the screening stage.

The behavior of the screening stage depends on: (a) the complexity of the search criteria, and (b) how the database is organized according to the variables

*Corresponding author.

used in the search criteria. Therefore, clustering techniques are absolutely necessary in order to optimize the global process [4–7].

Cluster analysis techniques are designed in order to find groups, or clusters, in data sets, where each member of a cluster is similar to the other members of the cluster based on a given set of characteristics. They have been widely applied in chemical structure handling applications, particularly for finding clusters of compounds with similar structures or physical properties. Different types of clustering algorithm can be used for chemical structure handling applications. These generally fall into one of two categories: hierarchical or non-hierarchical methods. In hierarchical methods, a hierarchical tree structure is produced, and taking slices across the hierarchy at different levels yields distinct clusterings of the data set. Non-hierarchical methods generally produce a single clustering of a data set without any hierarchical relationships.

Cluster analysis provides a number of methods to obtain an insight into data sets, and to extract relevant information from them. This analysis is the most important tool used for interpreting multivariate data containing objects and features, and sometimes also properties. Cluster analysis generically refers to different multivariate methods designed to create homogeneous sets of objects (chemical compounds) called clusters. So, each chemical compound is characterized by a set of properties, a numerical variable (in this paper – a measure of the relative structural similarity among database elements) such as concentration, temperature, or reaction time. These data are best described by a matrix, containing a row for each of the database elements, and a column for each property. Therefore, each database element corresponds to a point in the N-dimensional property space. Graphic representation of the data structure of N-dimensional picture area as two or three-dimensional drawings may be helpful in the evaluation of the results obtained.

The property space considered in the clustering method proposed is based on the calculation of the measurements of structural similarity among the elements of the chemical database and the consideration of these values in intervals of values.

Given two molecular graphs G_A and G_B representing the structure of chemical compounds A and B, respectively, it is possible to obtain a set M of all the maximum common substructures among G_A and G_B graphs [8]. Each element $M(i)$ of the M set is composed by a set of nodes and edges present in G_A and G_B graphs (atoms and bonds present in A and B molecules), for which the M set represents the maximum overlapping or the maximum matching among the molecular graphs G_A and G_B.

So, for a database of chemical compounds with db elements, a symmetrical S matrix can be obtained with information of the maximum overlapping set (MOS) between each pair of database elements (i, j), and therefore $(db^2 - db)/2$ values of structural similarity, which can be obtained in one or several of the commonly accepted similarity indexes [9,10].

In this paper, we propose a clustering method based on the MOS measurements and the projection of these values onto different spaces that represent similarity intervals representing the relative similarity of a molecule regarding to the elements of a chemical database.

The article has been organized in the following way: in section 2, the theoretical model is described on which the classification process is based, in section 3, the parameters or variables that intervene in the classification process are studied, since the tuning of these parameters determines the effectiveness of the process; two different classification methods are analyzed and the results obtained are analyzed on a database of natural products. Lastly, we present a discussion of these results and the validity of the proposed method is presented with a screening example on the database utilized.

2. Description of the clustering method

The clustering process requires a pre-processing stage in which the similarity values for each pair of the database elements is calculated.

The calculation of the structural similarity between two molecular graphs is a complex process that requires a high-computational cost [8,11]. These measures of structural similarity are based on the subgraph isomorphism calculation among the molecular graphs that are compared, and two different approaches can be considered: (a) Maximum Common Edges Subgraph (MCES) and (b) Maximum Common Subgraph (MCS).

Knowing the isomorphism among two molecular graphs (number of common nodes and edges), a distance measure or similarity index is used to obtain a measure of the structural or topological "resemblance" between two molecular graphs. Different indexes of similarity have been proposed in the literature (Tanimoto, Simpson, Kolcynski, Cosine, etc.) and their behavior and degeneracy has been studied in function of the graph, size, and density of the fingerprints, etc. [12].

The proposed method of clustering in this paper, in this preprocessing stage, uses the subgraphs isomorphism algorithm proposed by the authors [8] given its computational efficiency and the possibility of obtaining different isomorphism measures such as: MCES, MCS, AMCS (All MCS). The similarity measures considered in this paper have been obtained considering the subgraph isomorphism based on the MCS and the cosine index (Ochiai index), a commonly utilized similarity index due to its presenting little degeneration with the variation of graph size. This index is given by the following expression:

$$s = \frac{(n_c + e_c)}{\sqrt{(n_i + e_i)(n_j + e_j)}}, \tag{1}$$

where: n_i and e_i are the number of nodes and edges of the molecular graph G_i, n_j, and e_j are the number of nodes and edges of the molecular graph G_j, n_c, and e_c are the common nodes and edges to both graphs.

When this process is carried out for the db database elements, a symmetrical similarity matrix S is obtained, in which the element $S(i, j)$ stores the similarity value between the G_i and G_j molecular graphs.

When the information corresponding to the similarity values among all the elements of the database is obtained, the clustering process is carried out in four steps: (a) definition of the projection space, an N-dimensional space ($N \ll db$), (b) the projection process of the similarity values stored in S matrix in the defined N-dimensional space is carried out so that, each database element can be represented by means of a vector N size, (c) the projection vector of each element of the database is normalized in the N-dimensional space of similarity, and (d) the process of clusters generation is carried out and the database elements are assigned to the defined clusters.

2.1. *Definition of the projection space*

The similarity measures among any two database elements are given in the interval [0,1]. The proposed clustering method consists of considering different intervals of similarity inside this range and considering each of these intervals as a dimension in which the database elements can be represented. The number of defined intervals determines the dimension of the projection space and, therefore, the number of characteristic or variables in those where each database element is represented.

So, at this stage the number and range of the intervals of similarity values in which the database will be projected are selected. The similarity intervals determine the granularity of the clustering process. As the number of similarity intervals increases (and therefore are smaller in size) the granularity of the clustering process also increases, and vice versa. As the granularity increases the number of generated clusters also increases and therefore the clusters population diminishes.

In the proposed clustering method the number and size of the similarity intervals is dynamic and can be conveniently adjusted according to the results observed. So, an I array is defined corresponding to the similarity intervals, where:

- Each element $I(i)$ defines an interval of similarity $[x, y]$, where $y > x$.
- The first element $I(1)$ is an interval defined as $[0, y]$.
- The last element $I(N)$ is an interval defined as $[x, 1]$, where N represents the number of dimensions of the projection space.
- The defined intervals of similarity in the I array are disjoint intervals, so that $x_k > y_{k-1}$.

2.2. Projection of the database elements

Once the projection space is selected, for each row of the S matrix (each database element) a vector of dimensions equal to N representing the relative similarity of each element regarding the remaining database elements is obtained.

We have proposed two different methods in order to test the behavior of the clustering process regarding the consideration of the characteriztics (diversity) of the database (Method A) and the characteriztics of each database element (Method B):

Method A: Each row of the S matrix (each database element) is represented by means of a vector of N elements, as follows:

$$\left[\frac{\sum_{j}^{I(1)} S(i,j)}{M_{db}^{I(1)}}, \frac{\sum_{j}^{I(2)} S(i,j)}{M_{db}^{I(2)}}, \frac{\sum_{j}^{I(3)} S(i,j)}{M_{db}^{I(3)}}, \ldots, \frac{\sum_{j}^{I(N)} S(i,j)}{M_{db}^{I(N)}} \right], \tag{2}$$

where: $\overset{I(k)}{S}(i,j)$ represents the similarity value obtained from the MOS between the elements i and j, included in the $I(k)$ interval, and $\overset{I(k)}{M}_{db}$ represents the total number of matching of the database whose similarity value is included in the $I(k)$ interval.

Method B: In this case, the property vector for each database element is obtained as follows:

$$\left[\frac{\sum_{j}^{I(1)} S(i,j)}{M_{i}^{I(1)}}, \frac{\sum_{j}^{I(2)} S(i,j)}{M_{i}^{I(2)}}, \frac{\sum_{j}^{I(3)} S(i,j)}{M_{i}^{I(3)}}, \ldots, \frac{\sum_{j}^{I(N)} S(i,j)}{M_{i}^{I(N)}} \right], \tag{3}$$

where: $\overset{I(k)}{S}(i,j)$ represents the similarity value, obtained from the MOS between the elements i and j, included in the $I(k)$ interval, and $\overset{I(k)}{M}_{i}$ represents the number of matching of i element (with the remaining database elements) included in $I(k)$ interval.

Once the vectors representing each database element are obtained, the database can be represented by means of a B matrix of size (db, N). This B matrix can be normalized in the interval [0,1] that bears the normalization of the projection space, in the following way:

$$\forall k, \overline{B(i,j)} = \frac{B(i,j) - \min(B(k,j))}{\max(B(k,j)) - \min(B(k,j))}. \tag{4}$$

2.3. Grid definition and database clustering

In this step the clusters are defined and the database elements are assigned to them. So, we build a grid of the projection space. This grid consists of the construction of a set of N-dimensional cells or bins whose size can be equal or different in each dimension.

The grid size, together with the number of dimensions determine the maximum number of classes or clusters in which the database elements are classified, which is given by the following expression:

$$\text{Maximum number of classes} = \prod_{i=1}^{N} \frac{1}{(\text{grid})_i}. \tag{5}$$

Once the grid is built, each database element is assigned to one of the generated cells, and a series of parameters used to analyze the usefulness of the clustering process is calculated (see table 1 and 2) as follows.

- The average similarity of the database (ASDB). This value depends only on the database characteristics, and it is independent of the method and parameters used in the clustering process.
- The number of classes (CT), average of cluster population (APC), the percentage of singletons and doubletons.
- The entropy of the clustering process (CE) and the number of effective cluster (ENC) [13] are calculated from the following expression:

$$\text{CE} = -\sum_{i=1}^{q} f_i \log_2 f_i, \tag{6}$$

  where: q is the clusters number and $f_i = \frac{n_i}{db}$ is the population frequency in each cluster, calculated as the ratio between the population of each cluster (n_i) and the number of database elements (db). Knowing CE, the effective number of clusters can be calculated as follows $\text{ENC} = 2^{\text{CE}}$.
- A representative (centroid) of each cluster is selected as follows: (a) the similarity among the cluster elements is calculated (it has already been calculated previously in the preprocessing stage), (b) the average $A_k(i)$ and variance $V_k(i)$ of the similarity value obtained for each element i of the class and the average A_k and variance V_k for the class k are calculated. Then the class element whose difference $|A_k(i) - A_k|$ is the smallest, is chosen as representative (R_k) of B_k class.
- The average of similarity of the clusters (ASL). This value is calculated as ASDB value, but using the representatives of each class in the calculation instead of all the database elements.

- The Cartesian center (gravity) of the clustering (GCC). This value corresponds to a point in the projection space whose distance to the rest of the points is the smallest.
- The Cartesian center (gravity) of the clusters or classes (GCL). This value is calculated in a similar way to the GCC using in the calculation the representatives of each class instead of all the database elements.
- Some dispersion parameters are also obtained in order to evaluate the characteristics and effectiveness of the clustering process: (a) the dispersion of the representatives (RD), obtained as the sum of the Euclidean distances among all the representatives of the classes, (b) the dispersion of the clustering (CD), obtained as the sum of the Euclidean distances between all the representatives and the gravity center of the clustering and (c) the dispersion of the cluster or classes (LD), obtained as the sum of the Euclidean distances between all the representatives and the gravity center of the clusters.

The information obtained and the selected representatives are the basis of the proposed clustering method. The computational cost of this calculation is reasonably low since it is carried out in the preprocessing stage when the similarities among each database elements are calculated, that information can be used in the processing stage.

3. Analysis of the results

The tests of the proposed clustering method have been carried out using a PC Pentium II 400 MHz on a public domain database of natural compounds [14] composed of 498 elements. The average of the similarity of the database (ASDB) is 0.5700.

3.1. Influence of the clustering parameters

In the test development we have used different grids (0.05, 0.1, and 0.2) and different projection spaces (2D, 3D, and 4D) where three different values of similarity intervals have been defined for each one.

In tables 1 and 2 the values of the studied parameters in the classification process for the two proposed methods are shown, which are used to analyze the process usefulness.

As we can observed, the number of dimensions (N) of the projection space (size of the array I) considerably influences the classification process. As N increases the number of classes in those whose database elements are classified also increases. Evidently, an increase in the dimension of the projection space

generates an increase in the total number of possible clusters (expression 5), which produces a disperse projection of the database elements in the generated grid, shown by the increase in other parameters studied (CE, ENC, %S, %D, etc.) and, evidently, a decreasing in the average of the population of the clusters (APC), this is an effect that is independent of the grid size.

This behavior can be observed in figures 1–3, where the clusters number and the population of the clusters for the different values of N and intervals of similarity for the two proposed classification methods are represented.

The grid size is a decisive parameter in the behavior of the classification process. Evidently, an increase in the grid size generates a decrease in the number of cells in which the database elements are stored and, therefore, it produces a decrease in the clusters number (tables 1 and 2). Thus, an increase in the grid size generates more populated clusters (APC increases) and it produces a decrease in the singletons (%S) and doubletons (%D) percentage, diminishing the parameters that measure the clustering dispersion (RD, CD, LD), and producing a centering of the clusters in the projection space (GCL).

The value of the similarity intervals is another decisive parameter in the behavior of the classification process. The influence of this parameter depends on the database characteristics, that is, of the compounds diversity. Thus, the values of the similarity intervals affect the classification process, depending on the similarity among the database elements.

As tables 1 and 2 show, the similarity intervals close to ASDB value are those that most affect the behavior of the classification process. As this interval is higher, the cluster dispersion (RD) increases, whilst also increasing the value of the average similarity of the clusters (ASL).

Figures 1–4 show the behavior of the classification process for different values of similarity intervals, dimension of projection space, and grid size for the two classification methods studied.

As can be observed, method A leads to a lower number of clusters for the same values of the classification parameters than method B. These clusters are more populated (higher value of APC) and they are more grouped (lower value of the parameters in charge of measuring the dispersion), and the elements of the clusters are more diverse (lower value of ASL). This effect is because method A considers the behavior of the entire database in the calculation (see the denominator of expression 2) instead of the behavior of each element regarding to the database (see expression 3).

Studying figure 4, it can be observed that method A is more influenced by the values of the similarity intervals than method B. Method A spreads to produce a more uniform distribution of the clusters in the projection space, although this distribution is affected when partitions of intervals of similarity close to the ASDB are generated. However, method B is able to find different behaviors of the database elements, being influenced in smaller measure by this parameter (see the dispersion parameters in tables 1 and 2). This effect induces

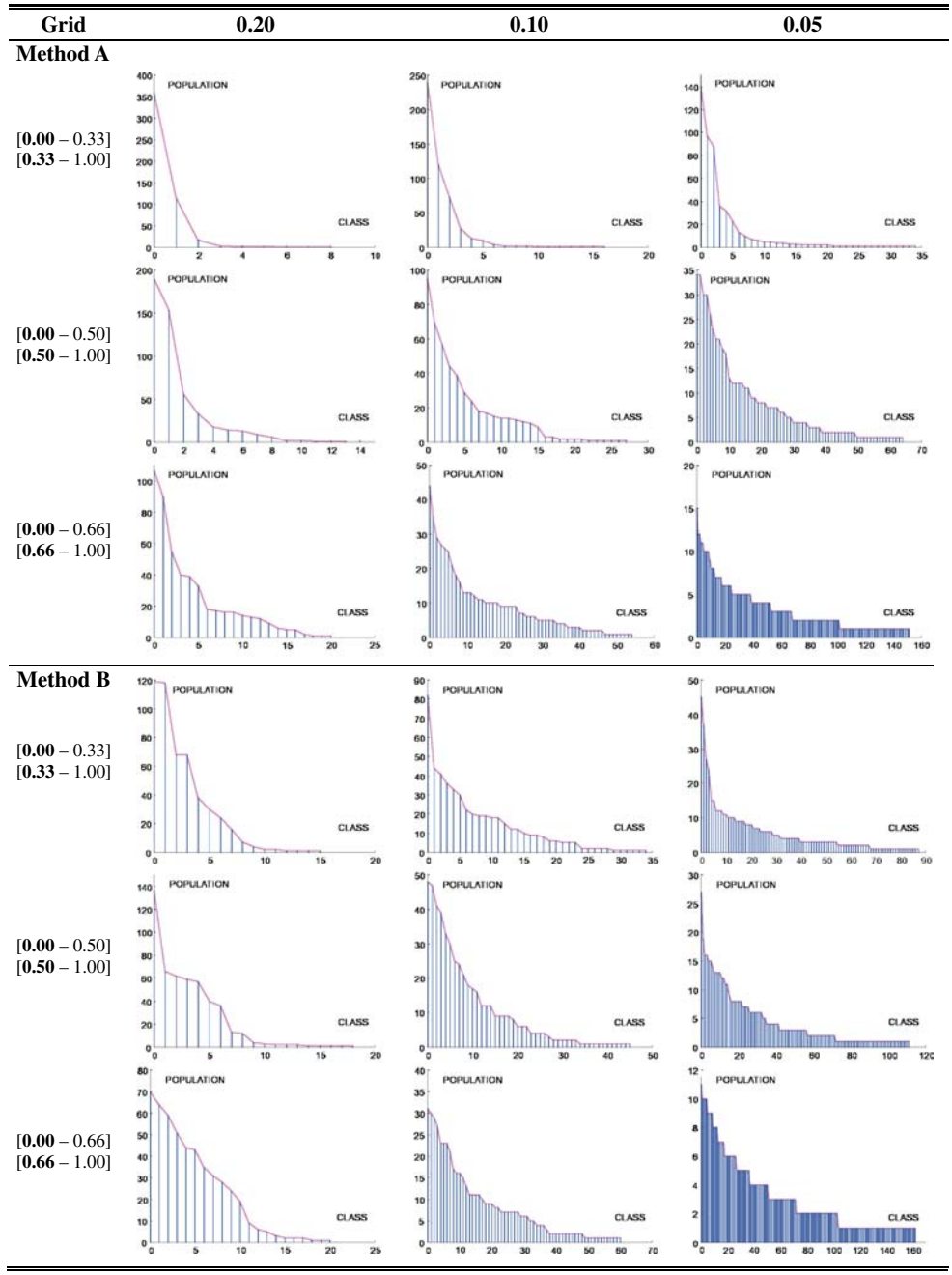

Figure 1. Behavior of the classification process using a 2D projection space for different values of cell size, intervals of similarity and for the two proposed classification methods. X-axis shows the clusters number and Y-axis shows the cluster population.

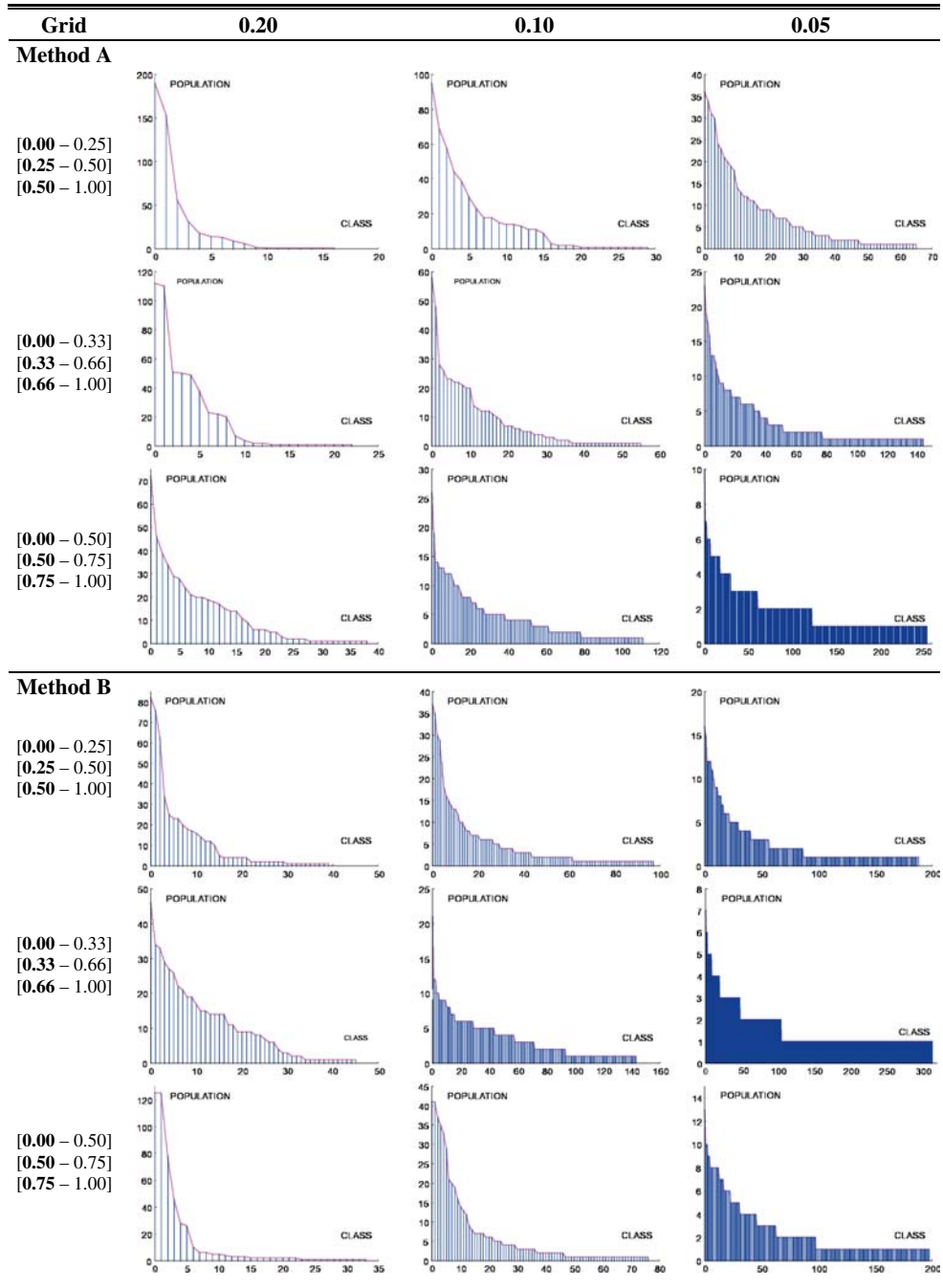

Figure 2. Behavior of the classification process using a 3D projection space for different values of cell size, intervals of similarity for the two proposed classification methods. X-axis shows the clusters number and Y-axis shows the cluster population.

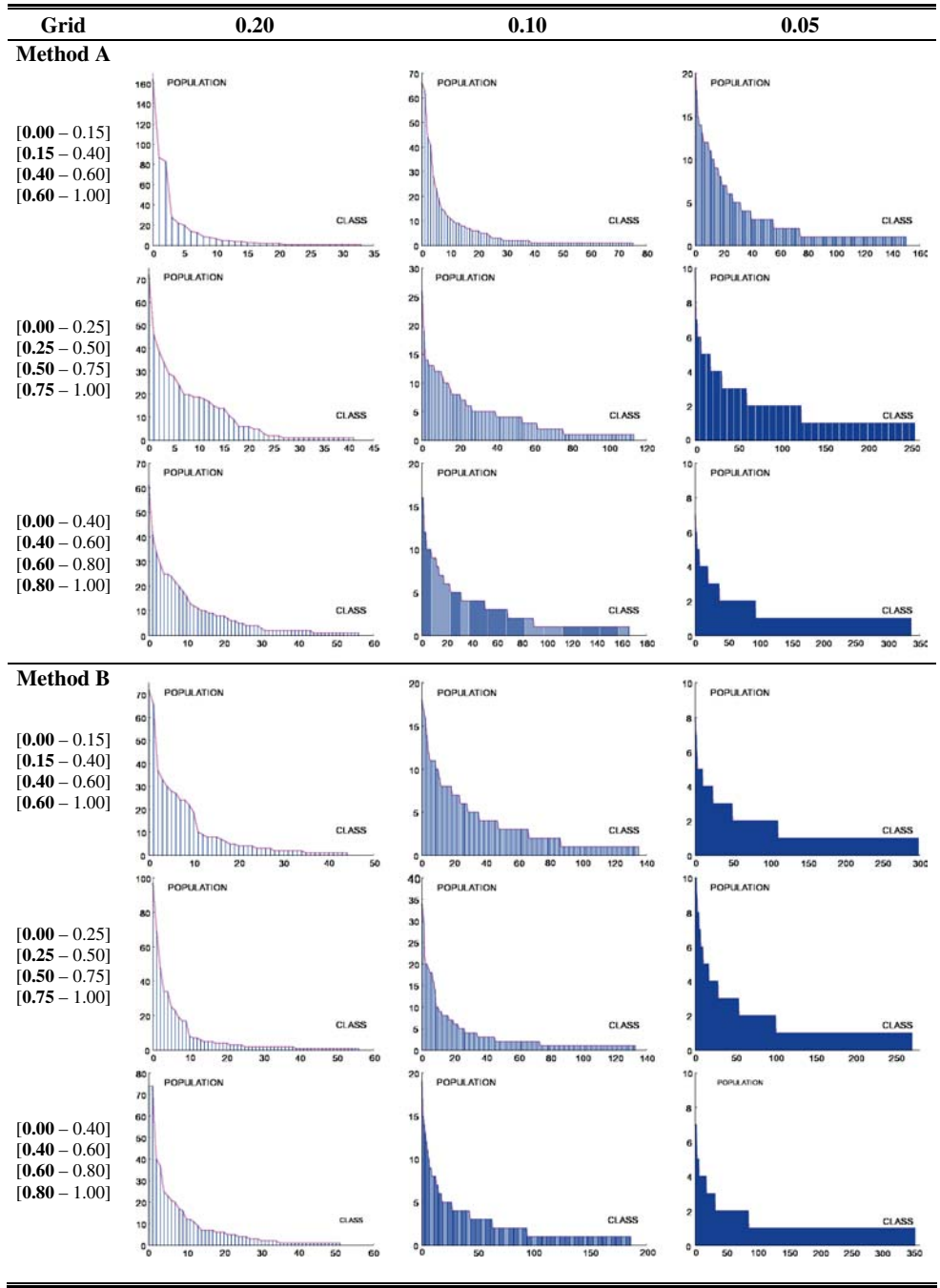

Figure 3. Behavior of the classification process using a 4D projection space for different values of cell size, intervals of similarity for the two proposed classification methods. X-axis shows the clusters number and Y-axis shows the cluster population.

Table 1
Study of the classification parameters in the clustering process using method A.

Grid	Projection	CT	APC	%S	%D	CE	ENC	GCC			GCL			ASL	RD	CD	LD
0.05	0.00-0.33-1.00	35	14.29	40.00	14.29	3.30	9.86	0.0319	0.8208		0.1991	0.5721		0.4729	229.96	11.91	9.42
0.05	0.00-0.50-1.00	65	7.69	23.08	16.92	5.27	38.68	0.3002	0.7218		0.4456	0.5659		0.5148	912.79	22.72	20.81
0.05	0.00-0.66-1.00	152	3.29	32.89	22.37	6.82	112.68	0.4964	0.4696		0.5160	0.3983		0.5450	4894.21	48.02	46.92
0.05	0.00-0.25-0.50-1.00	66	7.58	27.27	13.64	5.27	38.49	0.0128	0.2982	0.7218	0.0573	0.4398	0.5579	0.5085	1055.09	24.75	23.44
0.05	0.00-0.33-0.66-1.00	145	3.45	46.21	17.93	6.50	90.69	0.0319	0.6618	0.4696	0.0751	0.6366	0.3860	0.5293	4306.40	44.13	42.52
0.05	0.00-0.50-0.75-1.00	255	1.96	51.76	24.31	7.71	208.87	0.3002	0.6325	0.3507	0.3445	0.5899	0.3496	0.5573	16588.98	94.23	94.38
0.05	0.00-0.15-0.40-0.60-1.00	151	3.31	50.33	12.58	6.55	93.53	0.0060	0.1024	0.5488	0.5613	0.0198	0.1856	0.5487	0.4549	0.5293	4975.83	47.33	46.48
0.05	0.00-0.25-0.50-0.75-1.00	255	1.96	51.76	24.71	7.71	208.72	0.0128	0.2982	0.6325	0.3507	0.0199	0.3430	0.5882	0.3479	0.5563	16925.98	95.51	95.75
0.05	0.00-0.40-0.60-0.80-1.00	337	1.48	72.11	16.62	8.17	287.69	0.1024	0.5488	0.5210	0.2932	0.1193	0.5469	0.5057	0.2684	0.5606	28107.68	119.13	118.79
0.10	0.00-0.33-1.00	17	29.41	41.18	17.65	2.17	4.49	0.0319	0.8208		0.2684	0.5012		0.4626	69.44	7.67	6.02
0.10	0.00-0.50-1.00	28	17.86	21.43	14.29	3.85	14.37	0.3002	0.7218		0.4976	0.5056		0.4908	191.29	11.77	10.39
0.10	0.00-0.66-1.00	55	9.09	14.55	10.91	5.17	36.07	0.4964	0.4696		0.4861	0.3893		0.5290	670.29	18.15	17.65
0.10	0.00-0.25-0.50-1.00	30	16.67	33.33	10.00	3.84	14.34	0.0128	0.2982	0.7218	0.1154	0.4820	0.4898	0.4863	264.46	14.27	13.33
0.10	0.00-0.33-0.66-1.00	56	8.93	33.93	7.14	4.89	29.75	0.0319	0.6618	0.4696	0.1279	0.6000	0.3473	0.5068	745.16	20.15	18.96
0.10	0.00-0.50-0.75-1.00	112	4.46	29.46	15.18	6.27	76.98	0.3002	0.6325	0.3507	0.3725	0.5572	0.3653	0.5495	3381.15	44.42	43.75
0.10	0.00-0.15-0.40-0.60-1.00	76	6.58	48.68	13.16	4.86	29.01	0.0060	0.1024	0.5488	0.5613	0.0394	0.2599	0.5233	0.3907	0.5070	1481.35	30.64	27.59
0.10	0.00-0.25-0.50-0.75-1.00	114	4.39	33.33	12.28	6.27	77.31	0.0128	0.2982	0.6325	0.3507	0.0366	0.3800	0.5437	0.3531	0.5412	3714.64	47.60	46.85
0.10	0.00-0.40-0.60-0.80-1.00	167	2.99	46.11	12.57	6.87	116.75	0.1024	0.5488	0.5210	0.2932	0.1513	0.5340	0.4773	0.2889	0.5503	7577.50	65.06	65.07
0.20	0.00-0.33-1.00	9	55.56	33.33	22.22	1.16	2.23	0.0319	0.8208		0.3594	0.4545		0.4333	24.51	5.05	4.30
0.20	0.00-0.50-1.00	14	35.71	21.43	14.29	2.42	5.35	0.3002	0.7218		0.4731	0.5277		0.4932	47.48	5.74	5.22
0.20	0.00-0.66-1.00	21	23.81	14.29	4.76	3.58	11.92	0.4964	0.4696		0.4372	0.4326		0.5243	108.70	7.81	7.74
0.20	0.00-0.25-0.50-1.00	17	29.41	41.18	5.88	2.45	5.46	0.0128	0.2982	0.7218	0.1791	0.4793	0.4706	0.4778	90.37	8.79	8.13
0.20	0.00-0.33-0.66-1.00	23	21.74	43.48	8.70	3.22	9.30	0.0319	0.6618	0.4696	0.2279	0.5337	0.3312	0.4892	158.21	10.81	10.07
0.20	0.00-0.50-0.75-1.00	39	12.82	28.21	10.26	4.43	21.56	0.3002	0.6325	0.3507	0.4331	0.4967	0.3647	0.5321	436.22	17.15	16.35
0.20	0.00-0.15-0.40-0.60-1.00	34	14.71	38.24	11.76	3.27	9.67	0.0060	0.1024	0.5488	0.5613	0.0622	0.3329	0.4723	0.3631	0.4853	352.47	17.17	14.81
0.20	0.00-0.25-0.50-0.75-1.00	42	8.77	31.58	24.56	4.46	21.97	0.0128	0.2982	0.6325	0.3507	0.0840	0.4356	0.4706	0.3465	0.5150	560.39	20.42	19.35
0.20	0.00-0.40-0.60-0.80-1.00	57	9.62	32.69	9.62	4.92	30.29	0.1024	0.5488	0.5210	0.2932	0.1936	0.5068	0.4570	0.3076	0.5404	991.28	25.31	25.13

Table 2
Study of the classification parameters in the clustering process using method B.

Grid	Projection	CT	APC	%S	%D	CE	ENC	GCC				GCL				ASL	RD	CD	LD
0.05	0.00-0.33-1.00	88	5.68	22.73	14.77	5.72	52.99	0.5924	0.7010			0.7397	0.6274			0.5380	1748.83	34.51	29.36
0.05	0.00-0.50-1.00	112	4.46	35.71	13.39	6.16	71.55	0.7979	0.5519			0.7601	0.4855			0.5294	2134.80	28.10	27.46
0.05	0.00-0.66-1.00	163	3.07	36.20	19.63	6.94	123.13	0.7398	0.4589			0.6907	0.4422			0.5437	5193.24	46.36	46.05
0.05	0.00-0.25-0.50-1.00	189	2.65	53.97	15.87	6.97	125.80	0.2353	0.7295	0.5519		0.3540	0.6823	0.5340		0.5444	11109.44	91.38	92.86
0.05	0.00-0.33-0.66-1.00	314	1.59	66.56	18.15	8.05	265.92	0.5924	0.6647	0.4589		0.6453	0.6484	0.4263		0.5608	30029.65	149.35	144.54
0.05	0.00-0.50-0.75-1.00	198	2.53	50.51	17.68	7.17	143.79	0.7979	0.5307	0.8413		0.7852	0.5098	0.8156		0.5496	6612.43	45.82	46.25
0.05	0.00-0.15-0.40-0.60-1.00	299	1.67	62.88	20.40	7.97	250.67	0.0194	0.8466	0.5980	0.4113	0.0325	0.8266	0.5895	0.3874	0.5635	22798.76	105.37	106.47
0.05	0.00-0.25-0.50-0.75-1.00	271	1.85	62.73	16.97	7.72	211.19	0.2353	0.7295	0.5307	0.8413	0.2831	0.7081	0.5269	0.8241	0.5538	21924.50	122.17	124.94
0.05	0.00-0.40-0.60-0.80-1.00	353	1.42	75.64	15.30	8.25	303.46	0.8464	0.5980	0.4566	0.8277	0.8339	0.5926	0.4448	0.8041	0.5630	33132.20	128.82	130.85
0.10	0.00-0.33-1.00	35	14.29	17.14	14.29	4.34	20.28	0.5924	0.7010			0.6751	0.5999			0.5267	320.70	14.71	13.81
0.10	0.00-0.50-1.00	46	10.87	26.09	13.04	4.65	25.10	0.7979	0.5519			0.7259	0.4861			0.5092	425.56	13.81	13.42
0.10	0.00-0.66-1.00	61	8.20	19.67	18.03	5.29	39.22	0.7398	0.4589			0.6542	0.4456			0.5331	832.14	20.27	19.96
0.10	0.00-0.25-0.50-1.00	98	5.10	36.73	18.37	5.70	52.04	0.2353	0.7295	0.5519		0.4564	0.6546	0.5247		0.5279	3133.06	52.86	50.10
0.10	0.00-0.33-0.66-1.00	144	3.47	34.72	15.28	6.73	106.33	0.5924	0.6647	0.4589		0.6171	0.6408	0.4260		0.5553	6683.08	70.72	69.82
0.10	0.00-0.50-0.75-1.00	77	6.49	38.96	12.99	5.18	36.27	0.7979	0.5307	0.8413		0.7598	0.4817	0.7649		0.5282	1347.48	24.32	24.49
0.10	0.00-0.15-0.40-0.60-1.00	136	3.68	36.03	14.71	6.56	94.28	0.0194	0.8466	0.5980	0.4113	0.0644	0.7796	0.5811	0.3999	0.5601	5586.90	56.55	57.89
0.10	0.00-0.25-0.50-0.75-1.00	134	3.73	44.78	20.90	6.18	72.67	0.2353	0.7295	0.5307	0.8413	0.3430	0.6789	0.5099	0.8002	0.5354	6051.37	68.41	69.18
0.10	0.00-0.40-0.60-0.80-1.00	187	2.67	49.73	16.58	7.00	127.89	0.8464	0.5980	0.4566	0.8277	0.7945	0.5928	0.4276	0.7720	0.5552	11055.83	81.65	83.82
0.20	0.00-0.33-1.00	16	31.25	25.00	12.50	2.94	7.67	0.5924	0.7010			0.5811	0.5698			0.5039	74.50	7.33	7.26
0.20	0.00-0.50-1.00	19	26.32	26.32	15.79	3.11	8.62	0.7979	0.5519			0.6484	0.5185			0.4923	79.79	6.54	6.17
0.20	0.00-0.66-1.00	21	23.81	14.29	14.29	3.67	12.73	0.7398	0.4589			0.6019	0.4528			0.5100	105.06	7.74	7.45
0.20	0.00-0.25-0.50-1.00	40	12.50	25.00	20.00	4.14	17.63	0.2353	0.7295	0.5519		0.5029	0.6140	0.5541		0.5095	548.52	23.12	21.19
0.20	0.00-0.33-0.66-1.00	46	10.87	26.09	6.52	4.87	29.19	0.5924	0.6647	0.4589		0.5524	0.6266	0.4209		0.5404	732.06	24.01	24.14
0.20	0.00-0.50-0.75-1.00	34	14.71	32.35	23.53	3.23	9.38	0.7979	0.5307	0.8413		0.7339	0.5278	0.6649		0.5153	334.59	14.33	14.42
0.20	0.00-0.15-0.40-0.60-1.00	45	11.11	22.22	15.56	4.43	21.53	0.0194	0.8466	0.5980	0.4113	0.1082	0.7265	0.5378	0.4324	0.5504	741.27	23.52	23.81
0.20	0.00-0.25-0.50-0.75-1.00	57	8.77	31.58	24.56	4.35	20.38	0.2353	0.7295	0.5307	0.8413	0.3449	0.6553	0.5276	0.7460	0.5200	1222.12	32.45	32.32
0.20	0.00-0.40-0.60-0.80-1.00	52	9.62	32.69	9.62	4.53	23.11	0.8464	0.5980	0.4566	0.8277	0.7371	0.5516	0.4058	0.6833	0.5342	1030.45	28.59	28.91

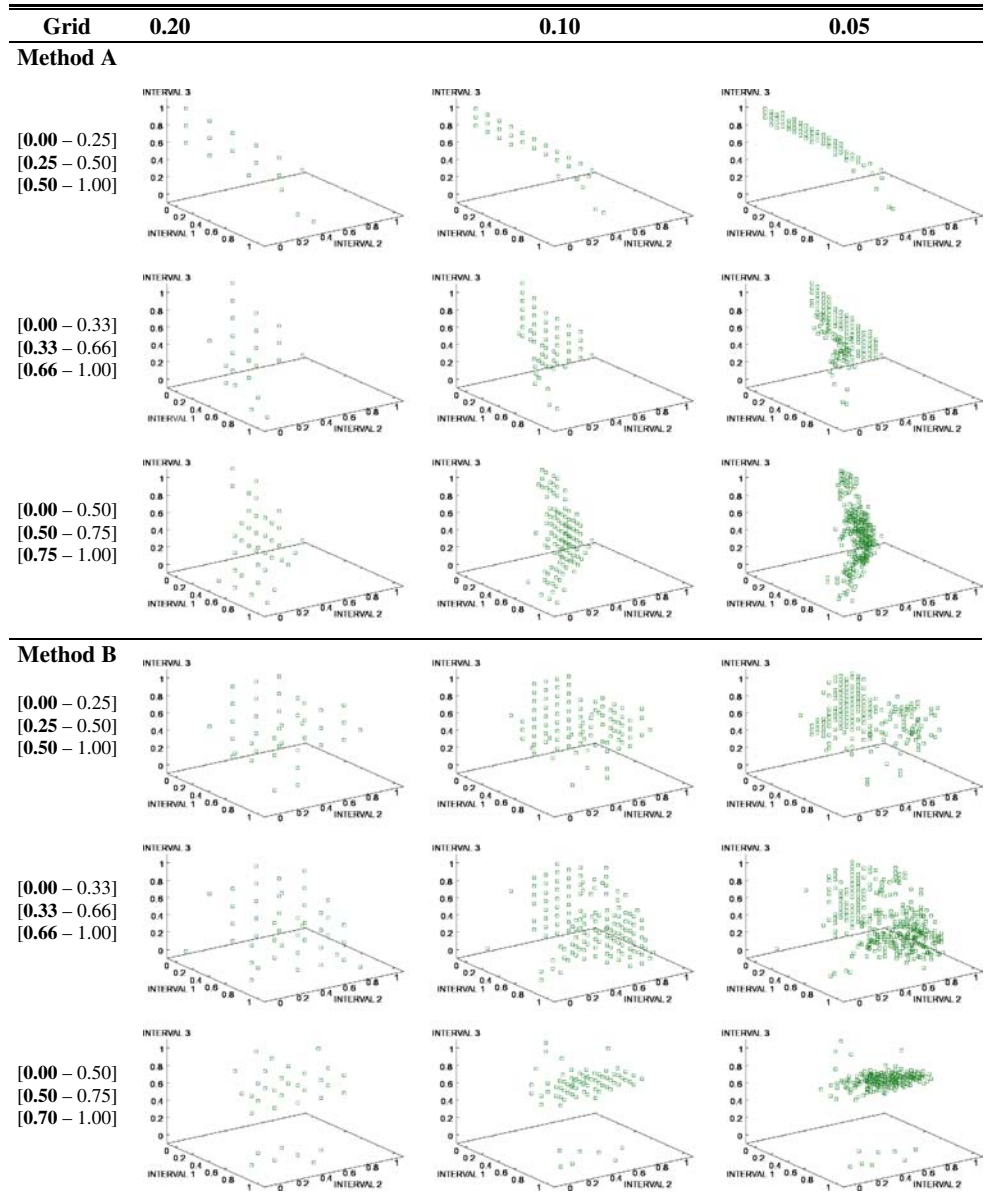

Figure 4. Cluster Distribution in a 3D projection space for the two proposed projection methods and for different values of cell size and intervals of similarity.

method B to produce clusters whose elements are more similar to each other (higher value of ASL) and the clusters are more distributed throughout the projection space, independently of the dimension of this space. Thus, the distance among the clusters generated by method B is markably higher than the generated

by method A for a similar number of clusters and equal values of the projection parameters, which allows us to consider this method useful for applications in which the diversity of groups of molecules in databases is necessary.

3.2. Comparison with other classification methods

However, as the literature describes [2,3,15], it is difficult to carry out a comparative study of different classification methods due to each method being based on the use of different variables, measures of distance or similarity, etc., which leads to the generation of a different number of clusters and grouping characteristic, we have carried out a comparative study of the clustering method proposed regarding to two classic and traditionally utilized methods: (a) the hierarchical method and (b) the K-means method.

As well as this, we have carried out a principal components analysis (PCA) with the database under study. This technique is based on the projection of database elements onto a multidimensional and orthogonal space (principal components). When the classification parameter is the measure of similarity among the database elements, PCA uses similarity matrix S.

Usually, PCA is used as a preparatory technique (unsupervised) of the clustering process. This technique helps the researcher to find groups of elements, which can then be used to carry out the database classification using a supervised technique [3,15].

As can be observe in figure 5, by means of the use of PCA technique it is difficult to find groups of compounds, or these groups would be composed of a high number of very diverse elements. The PCA has demonstrated that 32 principal components are necessary to explain the 95% of the variance, explaining the first three principal components (represented in figure 5) only 74.72%.

For an objective comparison with the hierarchical and K-means methods we have used in both cases the similarity matrix S as input data and we have imposed that the same clusters number is generated as those generated for our classification model for some given classification parameters (size of the projection space, values of intervals of similarity, cell size), analyzing for each method the similarity of the clusters and the dispersion of the clustering.

We have observed K-means and hierarchical methods generate clusters with a similar average similarity to the proposed methods. However, our proposed methods generate more dispersed clusters than K-means and hierarchical methods, as figure 6 shows.

In the graphs of figure 6, the sum of the normalized distance among the clusters has been represented for the K-means, hierarchical methods and the classification methods A and B proposed in this paper. As can be observed, for different classification parameters and, therefore, different number and characteristic of clusters, our classification models (blue-solid lines) generate, in all cases,

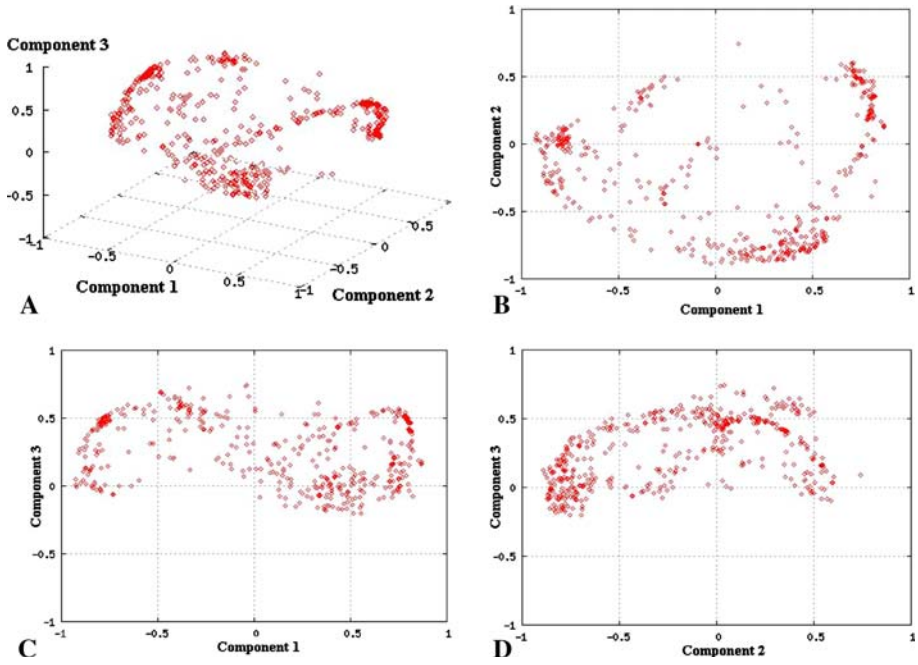

Figure 5. Results of the principal component analysis with the database in study. *A*: 3D representation of the first three principal components. *B*: PC1 versus PC2, *C*: PC1 versus PC3, *D*: PC2 versus PC3.

more dispersed clusters than the *K*-means (red-dot lines) and hierarchical (line green-dashed lines).

This characteristic shows that clusters are more dispersed in the classification space, they are more varied, and therefore the projection of *S* matrix onto spaces defined by intervals of similarity (model used in our proposal) improves the classification of chemical databases regarding to the use of the similarity matrix (used by other clustering models as *K* means or hierarchical).

4. Discussion and remarks

With a low-computational cost the classification method proposed in this article presents an appropriate behavior for its use in chemical databases. This computational cost is of the same order and even smaller than other classification methods studied, without considering the preprocessing phase in which the similarity matrix *S* is generated and the stage of extraction of the studied statistics (equal for all the methods).

For instance, in the case of a cell size equal to 0.05, intervals of similarity equal to (0.00 – 0.33.1.00) and the classification method A are needed only 1.2 s while hierarchical methods need 1.9 s and *K*-means method needs 25.9 s (using

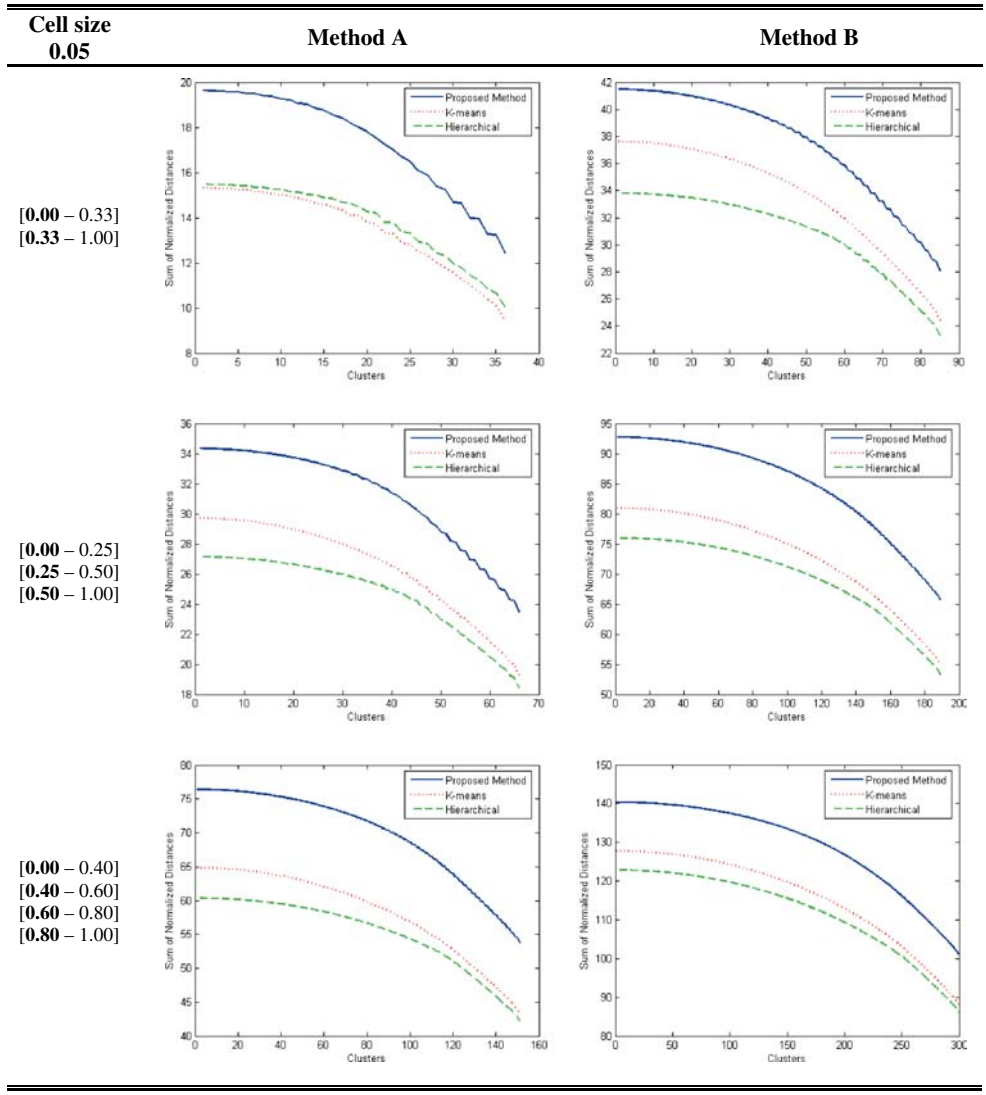

Figure 6. Behavior of the proposed methods regarding K-means and hierarchical methods for different clustering conditions.

Matlab [16] software and a PC Pentium II 400 MHz). Also, the computational cost of the proposed methods is slightly influenced by the number of clusters generated, while the hierarchical method is more influenced and K-means methods is the most influenced.

Moreover, the proposed method allows the generation of appropriate clusters depending on the database characteristics and of the objectives pursued in the classification process. Even for databases with few diverse elements in those

where other statistical techniques are applied with difficult; the proposed method can generate clusters with an acceptable average similarity.

So, the proposed clustering method permits building homogeneous groups of molecules, allowing the development of useful screening process, as can be observed in figure 7 where a screening example using the clustering methods *A* and *B* on the database used in the description of this paper is shown, considering different projection spaces and using a grid of 0.05. In this example, we have selected a chance molecule (in yellow frame in figure 7) as search pattern, and for each of the different values of similarity intervals (in a 3D projection space) we have recovered the cluster elements in those where the pattern molecule has been classified.

The versatility of the proposed classification method allows us to readjust the number of dimensions of the projection space, grid size, and the similarity interval values, carrying out adjustments in the clustering until the desired state is obtained, with a low computational cost.

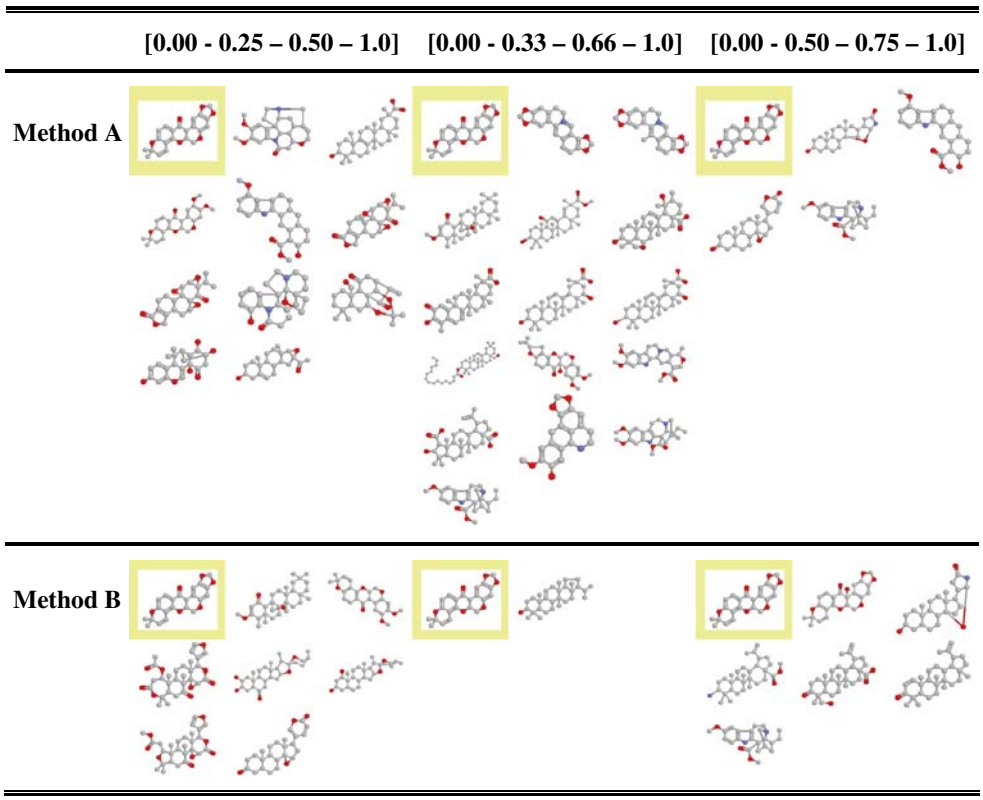

Figure 7. Example of screening for different intervals of similarity for the two clustering methods proposed.

Furthermore, method B allows finding groups of database elements with different behavior to the half behavior of the database. This fact allows us to use this classification method searching diversity in chemical databases.

Acknowledgments

We would like to thank the Comisión Interministerial de Ciencia y Tecnología (CICyT) and FEDER for their financial support (Projects: TIN2004-04114-C02-01 and TIN2006-02071)

References

[1] P. Willett, *Similarity and Clustering in Chemical Information Systems* (Research Studies Press: Letchworth, 1987).
[2] G.M. Downs and J.M. Barnard, Clustering and their uses in *Computational Chemistry, In Reviews in Computational Chemistry*, Vol. 18, eds. K.B. Lipkowitz and D.B. Boyd (Wiley-VCH, New York, 2003) pp. 1–39.
[3] K. Jajuga, A. Sokoowski, and A. Hermann Bock, *Classification, Clustering and Data Analysis* (Springer-Verlag, Berlin, 2002).
[4] T.R. Cundari and M. Russo, J. Chem. Inf. Comput. Sci. 41 (2001) 281–287.
[5] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms* (Wiley-IEEE Computer Society Pr, New York, 2002).
[6] L. Kaufman and P.J. Rousseeew, *Finds Group in Data: An Introduction to Clustering Analysis* (J Wiley New York, 1990).
[7] B.S. Everitt, *Cluster Analysis*, 3rd edn. (Edward Arnold, London, 1993).
[8] G. Cerruela García, I. Luque Ruiz, and M.A. Gómez-Nieto, J. Chem. Inf. Comput. Sci. 44 (2004) 30–41.
[9] P. Willett, J.M. Barnard, and G. Downs, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996.
[10] D.H. Rouvray and A.T. Balaban, In: *Chemical Applications of Graph Theory. Applications of Graph Theory*, eds. R.J. Wilson and L.W. Beineke (Academic Press, New York, 1979) pp. 177–221.
[11] J.W. Raymond, E.J. Gardiner, and P. Willett, J. Chem. Inf. Comput. Sci. 42 (2002) 305–316.
[12] J.D. Holliday, N. Salim, M. Whittle, and P. Willett, J. Chem. Inf. Comput. Sci. 43 (2003) 819–828.
[13] S.L. Taraviras, O. Ivanciuc, and D. Carbol-Bass, J. Chem. Inf. Comput. Sci. 40 (2000) 1128–1146.
[14] SPECS and BioSPECS B.V. http://www.specs.net
[15] B.S. Everett, S. Landau, and M. Leese, *Cluster Analysis*, (Arnold Publishers, 2001).
[16] The MathWorks, Inc. http://www.mathworks.com