Clustering Chemical Databases Using Adaptable Projection Cells and MCS Similarity Values

Irene Luque Ruiz,* Gonzalo Cerruela García, 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

Received January 31, 2005

In this paper we propose a new method based on measurements of the structural similarity for the clustering of chemical databases. The proposed method allows the dynamic adjustment of the size and number of cells or clusters in which the database is classified. Classification is carried out using measurements of structural similarity obtained from the matching of molecular graphs. The classification process is open to the use of different similarity indexes and different measurements of matching. This process consists of the projection of the obtained measures of similarity among the elements of the database in a new space of similarity. The possibility of the dynamic readjustment of the dimension and characteristic of the projection space to adapt to the most favorable conditions of the problem under study and the simplicity and computational efficiency make the proposed method appropriate for its use with medium and large databases. The clustering method increases the performance of the screening processes in chemical databases, facilitating the recovery of chemical compounds that share all or subsets of common substructures to a given pattern. For the realization of the work a database of 498 natural compounds with wide molecular diversity extracted from SPECS and BIOSPECS B.V. free database has been used.

1. INTRODUCTION

The cluster methodology is a technique that analyzes data and, applied to a heterogeneous group of data, aims to identify homogeneous subsets, such that the similarities within the groups are significantly greater than those between the groups.[1−3]

This methodology is applied in chemistry to group molecules of similar characteristics. Using premises empirically obtained, the methodology cluster is used for the preliminary analysis of big data sets (chemical databases) as selection method (screening) and reduction (filtering), its use being of great importance in the study of properties, design of new products, etc.[4−7]

Cluster methodology is not a supervised process, that is, there are no default groupings for the cluster searches. Its objective is to obtain the outputs originating from some specific inputs, while the objective of supervised learning is to establish relationships between inputs and outputs to allow the prediction of outputs for new future inputs. Normally the cluster methodology is supplemented with a supervised process of classification whose function is to label all the items in function of the default groups, using different methods such as recursive partition, Bayesian analysis, selection of the k nearest neighbors, etc.

Cluster analysis involves three principal components: a similarity coefficient for quantifying the degree of similarity between pairs of compounds, between a compound and a cluster or between a pair of clusters; a clustering method that processes the similarity data to identify groups of structurally related compounds; and an efficient algorithm for the implementation of the method so that it can be applied to data sets of nontrivial size.

* Corresponding author phone: ma1lurui@uco.es.

A great number of classification methods have been described in the bibliography which have been applied to chemical databases.[4] Clustering techniques have been classified as (a) visual techniques, (b) hierarchical methods (agglomerative, divisive), and (c) nonhierarchical methods. Visual techniques are generally based on the observation of principal components plots. Among the other techniques, hierarchical agglomerative methods are the most popular, based on the similarity between two objects (or between two clusters).

Hierarchic agglomerative techniques start from as many clusters as there are objects. Objects are gradually joined into clusters, up to the final cluster with all the objects included. In each step, two objects, one object and one cluster, or two clusters are merged. In the first step, the two objects with the largest similarity are merged. Then in each step, the two most similar clusters are merged. The value of their similarity (or of their distance) is retained, which will be used to obtain the typical result of these techniques, the dendrogram. With N objects, the final cluster is obtained after $(N-1)$ steps. The hierarchy is a consequence of the fact that larger clusters are always obtained by the merging of smaller ones (with all their objects).

Nonhierarchical methods generally produce a single clustering of a data set without any hierarchical structure. These methods encompass a wide range of different techniques to build clusters (single pass, relocating, nearest neighbor, mixture model, etc.).[4] They have advantages in applications involving large data sets for which the construction of a dendrogram is computationally prohibitive; in turn they have the disadvantage of the choice of the number of desired output clusters.

10.1021/ci050035o CCC: $30.25
Published on Web 07/14/2005

The results of the application of these methods depend in great measure on database size, diversity of the stored molecules, and objective pursued in the later treatment of the classified information.

Although there exist clustering methods that can be applied to a great range of problems, as much in chemistry as in other areas, the cluster methodology is still an austere instrument that can be modeled for its improvement. So the proposal of new models that facilitate the cluster generation in computationally efficient ways and, mainly, whose characteristics can be adapted to the different necessities of the researchers is a research line of interest.

The clustering method proposed in this paper is based on the initial calculation of the measurements of structural similarity among the elements of a chemical database.[8] Given two molecular graphs G_A and G_B representing the structure of the chemical compounds *A* and *B*, respectively, it is possible to obtain, by means of an algorithm proposed by the authors, the maximum common subgraph (MCS) and the maximum edge subgraph (MCES)—called in the chemical information community the maximum overlapping set (MOS).

A common induced subgraph between two graphs G_A and G_B is a subgraph G_C with vertices $V(G_C) \subseteq V(G_A)$ and $V(G_C) \subseteq V(G_B)$ such that the subgraphs induced in both G_A and G_B are isomorphic. The subgraph is maximal if it is not a subgraph of a larger subgraph possessing this property. The subgraph is referred to as a maximum common subgraph (MCS) if there is no other subgraph of greater cardinality meeting the aforementioned criteria. Related to the MCS is the maximum common edge subgraph (MCES), which is a set of edges where $E(G_C) \subseteq E(G_A)$ and $E(G_C) \subseteq E(G_B)$ such that the edges induce isomorphic subgraphs in G_A and G_B of maximum cardinality.[9]

The classification model proposed in this paper uses similarity values based on the maximum common subgraphs (MCS), the results obtained being compared with the generally utilized similarity values based on the maximum overlapping set (MOS). The classification process is based on the consideration of the MCS measurements between all the elements of the database and the generation of clusters by means of the projection of these values (normalized) over adaptive cells defined in an *N*-dimensional space. This multidimensional space is defined as a set of disjoint intervals of similarity in which the range of similarity ([0,1]) is divided.

The method described in the manuscript is a nonhierarchical method. The clusters are generated by means of the projection of the similarity measurements among the database elements (molecular graphs) on previously defined cells through the selection of the number and characteristic of the intervals of similarity. However the proposed method benefits from the advantages of the hierarchical methods, since the redefinition of the cell size allows for the grouping/disaggregating of the clusters. This characteristic allows the proposed classification method to adapt to the different necessities and researcher objectives and the database characteristics.

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. Last, we present a discussion of these results, and the validity of the proposed method is presented with a screening example on the database utilized, comparing the results obtained using the MCS and MOS.

2. CLUSTERING METHOD

Using the algorithm described by the authors,[8] it is possible to obtain for each pair of molecular graphs G_A and G_B, the MCS (and the MOS) between both molecular graphs. So, given a chemical database with *db* elements, for each pair of elements *(i, j)* the similarity value between the elements *i* and *j* can be obtained using the MCS (maximum common substructure) which can be calculated by any one of the similarity indexes proposed in the literature[9−11] (equally, a similarity value can be obtained using the MOS).

2.1. Preprocessing Stage. The proposed clustering process requires a preprocessing stage in which the similarity values for each pair of the database elements are calculated. We have used as similarity index, the cosine index, calculated as

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

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, and n_c and e_c are the number of nodes and edges in the MCS.

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

2.2. Processing Stage. As in most of the database clustering methods, the preprocessing stage is the most computationally expensive. Having obtained the information corresponding to the similarity values among all the elements of the database, the clustering process is carried out in four steps.

(1) Definition of the projection space. The projection is a process that allows us to reduce the number of variables (dimensions) in those where the objects (molecules) are represented. If we observe the matrix *S*, each element of the database (each row of the matrix *S*) is defined by *db* variables representing the similarity (in the interval [0,1]) of each molecule with the rest of the molecules of the database. The proposed projection model consists of reducing this *db* dimensional space to an *N*-dimensional space ($N \ll db$), where *N* represents the number of disjoint intervals of similarity in those where the range [0,1] is divided. Therefore, we select the number and range of the intervals of similarity values in which the database information will be projected. It is an *N*-dimensional space (called space of similarity) where each element of the database will be represented in function of its similarity (MCS) with the rest of the database elements.

(2) The projection process of the similarity values stored in *S* matrix in the defined *N*-dimensional space of similarity

is carried out. So, the database elements can be represented by means of a *P* matrix of *dbN* size.

(3) The projection vector of each element of the database (each row of the *P* matrix) is normalized in the *N*-dimensional space of similarity.

(4) The process of classes (clusters) generation is carried out, and the database elements are assigned to the defined clusters.

2.2.1. Definition of the Projection Space. The similarity intervals will determine the granularity of the cluster process. As the interval number of similarity increases (and therefore the intervals decrease in size) the granularity of the clustering process will be higher and vice versa.

As the granularity increases the generated number of classes also increases, and therefore the population of the classes diminishes. Evidently, the performance of the screening processes or search of elements of the database is directly dependent on the number of classes defined in the clustering process. When the number of classes increases the performance of this process diminishes. But, when the number of clusters diminishes the population of the classes increases, and, evidently, a high population of the classes are affected negatively in the screening and recovery processes of the information of the database.

This fact forces us to take a solution of commitment that guarantees a good level of grouping of the database elements in classes in the clustering processes. This solution of commitment is directly dependent on the characteristics (diversity) of the database, information that is difficult to know a priori. Our proposal seeks to avoid this inconvenience by proposing a classification method adaptive to the characteristics and future changes in the database.

The similarity values $S(i,j)$ are limited between 1 (only when *i* and *j* elements are equal) and 0 when any common substructure exists between *i* and *j* elements. Similarity values close to zero show the existence of common substructures of small size and vice versa. These substructures are very common among very different chemical compounds (i.e. short acyclic chains), and they have little importance from the chemist's point of view in the processes of searches in databases. Also, small variations in low similarity values have little relevance, since they represent small variations in the size of the common substructure.[8,9,11] However similarities close to the unit denote common substructures of considerable size and small variations in these values are chemically significant.

So, in the proposed clustering method the number and size of the similarity intervals is dynamic, and they can be adjusted conveniently in function of the observed performance in the recovery processes.

For that, we define a *V* vector corresponding to the similarity intervals, where (1) each element $V(i)$ defines a similarity interval $[x, y[$, where $y > x$, (2) the first element $V(1)$ is an interval defined as $[0, y[$, (3) the last element $V(N)$ is an interval defined as $[x, 1]$, and (4) the defined similarity intervals in the *V* vector are disjoint intervals, so that $x_k > y_{k-1}$.

2.2.2. Projection of the Similarity Values. Different approaches can be considered in the projection process, which will determine the clustering process characteristics. In this paper two projection approaches have been studied.

Method 1. For each row of the *S* matrix (each database element) a vector in the *N*-dimensional space of similarity is obtained by means of the following expression

$$p_i = \left[\left(\frac{\sum_j S(i,j)}{M_{db}}\right)_{V(1)}, \left(\frac{\sum_j S(i,j)}{M_{db}}\right)_{V(2)}, \left(\frac{\sum_j S(i,j)}{M_{db}}\right)_{V(3)}, \ldots, \left(\frac{\sum_j S(i,j)}{M_{db}}\right)_{V(N)}\right] \quad (2)$$

where $S(i,j)$ represents the similarity value obtained from the MCS among the elements *i* and *j*, included in the interval $V(k)$, and M_{db} represents the total number of matching of the database whose similarity value is included in the $V(k)$ interval.

Method 2. In this case, the vector for each database element is obtained as follows

$$p_i = \left[\left(\frac{\sum_j S(i,j)}{M_i}\right)_{V(1)}, \left(\frac{\sum_j S(i,j)}{M_i}\right)_{V(2)}, \left(\frac{\sum_j S(i,j)}{M_i}\right)_{V(3)}, \ldots, \left(\frac{\sum_j S(i,j)}{M_i}\right)_{V(N)}\right] \quad (3)$$

where $S(i,j)$ represents the similarity value obtained from the MCS between the elements *i* and *j*, included in the interval $V(k)$, and M_i represents the number of matching of *i* element (with the remaining database elements) included in the $V(k)$ interval.

With these two methods we wish to study two different approaches in the clustering process of the database elements. With method 1 we consider the information corresponding to the following: how different/similar each database element is with regard to how different/similar the rest of database elements are to each other. So, each vector element $p_i(k)$ contributes a value of the similarity of the molecule *i* in the interval of similarity $V(k)$ with regard to how many values in this interval exist among the rest of database elements.

On the other hand, method 2 uses as clustering approaches the measure of the disparity/similarity of each database element with regard to the rest of database elements. So, each vector element $p_i(k)$ contributes a value of the similarity of the molecule *i* in the interval of similarity $V(k)$ with regard to the rest of the molecules of the database.

2.2.3. Normalization of the Similarity Vectors. Having obtained the vectors representing each database element in the *N*-dimensional space of similarity (p_i), the database elements can be represented by means of a *P* matrix of equal size to *dbN*. This *P* matrix must be normalized in the interval [0,1] which bears the normalization of the projection space.

The normalization is carried out as follows:

$$\forall k, \overline{P(i,j)} = \frac{P(i,j) - \min(P(k,j))}{\max(P(k,j)) - \min(P(k,j))} \quad (4)$$

That is, the maximum and minimum value is obtained for each column of the P matrix, and the elements $P(i,j)$ are normalized by means of the expression (4). In the process of obtaining the $\bar{P}$ matrix, a set of information and statistical values is obtained in order to analyze the effectiveness of the selected classification method and, in another case, its refinement, whose study will be described in the following section of this manuscript.

2.2.4. Generation of the Classes and Clustering of the Database. Once the $\bar{P}$ matrix is generated, the rows maintain information of the elements of the database represented in a reduced and normalized space. That is, each element of the database is identified by a normalized vector $\overline{p_i}$ of size N (number of interval of similarity defined) and whose values are real numbers of six decimal precision. Next we process in the following way.

Grid of the Representation Space. A grid of the projection space is built, which consists of the construction of a set of N-dimensional cells whose size can be equal or different in each dimension.

Initially the size of the cells is 10^{-7} units in each dimension (since the considered precision is 6 decimals), which supposes that two elements i and j of the database have the same affix in the space if and only if $\overline{p_i} = \overline{p_j}$.

With such a "fine" grid, the number of clusters that will be generated will be very high, since there exists a great probability of $\overline{p_i} \neq \overline{p_j}$ for most of the database elements.

That is why in the classification process we can define a cell size (bin) equal or different for each dimension of the representation space. This cell size is defined by means of a vector C, of size equal to N (number of dimensions) and whose elements $C(i)$ store the value (smaller than 1) of cell size in the dimension i.

The grid size, together with the number of dimensions, determines the maximum number of classes or clusters in which the database elements can be classified, which can be expressed by the following expression for a cell size equal in all dimensions:

$$\text{maximum number of clusters} = \left(\frac{1}{\text{cell size}}\right)^N \qquad (5)$$

Generation of the Clusters. Once the grid is generated, each element of the database is assigned to one of the generated cells. Empty cells are not considered, and those cells with population are defined as clusters. In the stage of classification of the database elements in classes we proceed to obtain a series of values used later on in two senses: (a) the analysis of the effectiveness of the method of realized clustering and (b) to serve as measures for the process of refinement of the clustering being carried out. These values are the following, some of which are shown in Tables 1 and 2.

(1) The mean value of the similarity of the database (ASDB), only considering the value of the MCS in the calculation among all the database elements. This value only depends on the characteristics of the databases and is independent of the method and parameters used in the clustering process.

(2) The number of clusters (CT) and the populations of each one of the generated classes.

Table 1. Study of the Classification Parameters in the Clustering Process Using Method 1

cell size	projection	CT	CE	ENC	%S	%D	APC	GCC				GCL				ASL	RD	CD	LD
0.05	0.00−0.33−1.00	88	5.89	59.12	20.45	12.50	5.68	0.3748	0.6243			0.4384	0.5388			0.3721	1582.14	27.41	26.32
0.05	0.00−0.50−1.00	151	6.76	108.62	35.76	19.21	3.31	0.4484	0.4781			0.4573	0.4303			0.3805	4762.90	46.63	46.17
0.05	0.00−0.66−1.00	169	6.98	126.10	32.54	23.08	2.96	0.4807	0.3713			0.4640	0.3682			0.3988	5944.72	51.81	51.77
0.05	0.00−0.25−0.50−1.00	281	7.81	223.84	59.43	24.91	1.78	0.3334	0.4199	0.4781		0.3417	0.4136	0.4669		0.3970	20944.63	108.47	108.72
0.05	0.00−0.33−0.66−1.00	300	7.96	249.28	62.00	24.00	1.67	0.3748	0.5547	0.3853		0.3844	0.5513	0.3507		0.3985	23992.15	119.74	119.00
0.05	0.00−0.50−0.75−1.00	328	8.13	280.79	68.60	19.51	1.52	0.4484	0.4073	0.2842		0.4306	0.4413	0.2624		0.4147	27956.19	125.73	124.31
0.05	0.00−0.15−0.40−0.60−1.00	336	8.13	280.92	71.73	18.45	1.49	0.0688	0.3081	0.4673	0.3713	0.0752	0.3128	0.4607	0.3584	0.3967	29290.49	127.65	127.88
0.05	0.00−0.25−0.50−0.75−1.00	376	8.38	333.74	78.46	15.16	1.33	0.3334	0.4199	0.4073	0.2842	0.3298	0.4114	0.4331	0.2502	0.4089	43687.60	171.73	170.19
0.05	0.00−0.40−0.60−0.80−1.00	376	8.37	331.15	79.79	12.50	1.33	0.3275	0.4673	0.2575	0.2293	0.3198	0.4705	0.2840	0.1996	0.4120	37638.13	148.05	146.55
0.10	0.00−0.33−1.00	38	4.44	21.69	21.05	7.89	13.16	0.3748	0.6243			0.4894	0.4823			0.3467	318.83	13.51	12.23
0.10	0.00−0.50−1.00	61	5.30	39.28	21.31	13.11	8.20	0.4484	0.4781			0.4513	0.4196			0.3670	830.62	20.23	19.96
0.10	0.00−0.66−1.00	64	5.56	47.15	12.50	9.38	7.81	0.4807	0.3713			0.4437	0.3945			0.3963	907.91	20.86	20.72
0.10	0.00−0.25−0.50−1.00	120	6.40	84.28	27.50	20.83	4.17	0.3334	0.4199	0.4781		0.3805	0.4020	0.4295		0.3743	4057.80	49.69	49.48
0.10	0.00−0.33−0.66−1.00	145	6.66	91.13	35.17	20.00	3.45	0.3748	0.5547	0.3853		0.4099	0.5177	0.3746		0.3906	5835.14	60.07	59.71
0.10	0.00−0.50−0.75−1.00	161	6.83	114.05	41.61	16.77	3.11	0.4484	0.4073	0.2842		0.3955	0.4526	0.2883		0.4108	6974.20	63.98	63.19
0.10	0.00−0.15−0.40−0.60−1.00	179	6.92	121.05	46.93	18.99	2.79	0.0688	0.3081	0.4673	0.3713	0.1056	0.3404	0.4207	0.3191	0.3678	8749.42	71.99	70.98
0.10	0.00−0.25−0.50−0.75−1.00	223	7.40	168.97	48.43	24.22	2.24	0.3334	0.4199	0.4073	0.2842	0.3543	0.3822	0.4373	0.2619	0.4047	15596.89	103.01	102.35
0.10	0.00−0.40−0.60−0.80−1.00	215	7.28	155.34	52.56	20.93	2.33	0.3275	0.4673	0.2575	0.2293	0.3140	0.4650	0.2987	0.1966	0.4112	12487.06	85.50	84.34
0.20	0.00−0.33−1.00	16	2.95	7.74	18.75	0.00	31.25	0.3748	0.6243			0.4971	0.4887			0.3510	63.27	6.32	5.78
0.20	0.00−0.50−1.00	20	3.68	12.86	15.00	0.00	25.00	0.4484	0.4781			0.4582	0.4503			0.3816	94.23	7.10	7.04
0.20	0.00−0.66−1.00	21	3.94	15.34	14.29	0.00	23.81	0.4807	0.3713			0.4420	0.4204			0.3930	101.07	7.18	7.06
0.20	0.00−0.25−0.50−1.00	39	4.66	25.20	15.38	10.26	12.82	0.3334	0.4199	0.4781		0.3842	0.4048	0.4156		0.3654	450.15	17.19	17.06
0.20	0.00−0.33−0.66−1.00	49	4.78	27.56	18.37	14.29	10.20	0.3748	0.5547	0.3853		0.4253	0.4908	0.4132		0.3832	689.63	21.08	20.77
0.20	0.00−0.50−0.75−1.00	52	4.97	31.24	25.00	9.62	9.62	0.4484	0.4073	0.2842		0.3850	0.4623	0.3594		0.4350	739.80	21.49	20.85
0.20	0.00−0.15−0.40−0.60−1.00	59	4.92	30.35	33.90	10.17	8.47	0.0688	0.3081	0.4673	0.3713	0.1723	0.3256	0.3759	0.3297	0.3669	1018.94	26.10	24.89
0.20	0.00−0.25−0.50−0.75−1.00	83	5.77	54.57	24.10	16.87	6.02	0.3334	0.4199	0.4073	0.2842	0.3632	0.3566	0.4404	0.3233	0.4060	2180.61	39.05	38.42
0.20	0.00−0.40−0.60−0.80−1.00	65	5.25	38.12	27.69	13.85	7.69	0.3275	0.4673	0.2575	0.2293	0.3086	0.4304	0.2992	0.2485	0.3930	1255.51	28.26	27.93

Table 2. Study of the Classification Parameters in the Clustering Process Using Method 2

cell size	projection	CT	CE	ENC	%S	%D	APC		GCC			GCL			ASL	RD	CD	LD
0.05	0.00−0.33−1.00	156	6.73	106.28	39.10	21.15	3.21	0.5803	0.6046			0.5533	0.5565		0.3662	4801.98	45.52	45.11
0.05	0.00−0.50−1.00	86	5.77	54.75	29.07	9.30	5.81	0.5734	0.7985			0.5136	0.7614		0.3515	1275.96	21.48	21.15
0.05	0.00−0.66−1.00	87	5.64	49.96	37.93	9.20	5.75	0.6260	0.7311			0.5449	0.6789		0.3438	1460.32	24.02	23.99
0.05	0.00−0.25−0.50−1.00	286	7.81	224.73	61.54	23.43	1.75	0.5418	0.6029	0.7985		0.5489	0.5890	0.7854	0.3912	15929.11	80.26	80.31
0.05	0.00−0.33−0.66−1.00	278	7.78	219.21	61.51	22.30	1.80	0.5803	0.6258	0.7769		0.5813	0.6057	0.7468	0.3880	16456.92	83.13	84.19
0.05	0.00−0.50−0.75−1.00	162	6.68	102.23	46.30	18.52	3.09	0.5734	0.8368	0.7817		0.5534	0.8294	0.6965	0.3671	6403.01	55.18	58.34
0.05	0.00−0.15−0.40−0.60−1.00	355	8.24	301.37	76.62	16.06	1.41	0.6436	0.4713	0.6163	0.7311	0.6450	0.4620	0.6145	0.7170	0.3910	39604.60	171.06	170.74
0.05	0.00−0.25−0.50−0.75−1.00	343	8.14	282.14	74.05	17.78	1.46	0.5418	0.6029	0.8368	0.7817	0.5559	0.5886	0.8359	0.7553	0.3955	28126.79	115.26	116.85
0.05	0.00−0.40−0.60−0.80−1.00	348	8.21	296.12	74.43	15.80	1.44	0.5745	0.6163	0.8280	0.7797	0.5757	0.6072	0.8181	0.7406	0.3907	31937.20	128.75	132.46
0.10	0.00−0.33−1.00	64	5.24	37.76	21.88	15.63	7.81	0.5803	0.6046			0.5200	0.5309			0.3436	885.34	20.89	20.34
0.10	0.00−0.50−1.00	37	4.20	18.37	21.62	24.32	13.51	0.5734	0.7985			0.5107	0.7446			0.3333	268.45	10.42	10.38
0.10	0.00−0.66−1.00	40	4.28	19.37	37.50	7.50	12.50	0.6260	0.7311			0.4986	0.6309			0.3201	364.37	13.48	13.17
0.10	0.00−0.25−0.50−1.00	129	6.35	81.42	37.21	17.05	3.88	0.5418	0.6029	0.7985		0.5170	0.5512	0.7783		0.3642	3691.73	41.59	41.27
0.10	0.00−0.33−0.66−1.00	126	6.40	84.62	37.30	18.25	3.97	0.5803	0.6258	0.7769		0.5724	0.5531	0.7085		0.3615	4061.84	45.67	46.23
0.10	0.00−0.50−0.75−1.00	66	4.63	24.75	34.85	16.67	7.58	0.5734	0.8368	0.7817		0.5332	0.8212	0.6338		0.3468	1253.10	27.13	28.69
0.10	0.00−0.15−0.40−0.60−1.00	198	7.09	136.31	49.49	21.72	2.53	0.6436	0.4713	0.6163	0.7311	0.6139	0.4724	0.5896	0.6999	0.3778	12956.34	98.64	99.66
0.10	0.00−0.25−0.50−0.75−1.00	164	6.59	96.53	51.83	15.24	3.05	0.5418	0.6029	0.8368	0.7817	0.5471	0.5701	0.8266	0.6783	0.3647	7793.59	67.19	69.29
0.10	0.00−0.40−0.60−0.80−1.00	169	6.80	111.36	47.34	16.57	2.96	0.5745	0.6163	0.8280	0.7797	0.5684	0.5757	0.7923	0.6729	0.3711	9010.98	75.19	78.79
0.20	0.00−0.33−1.00	24	3.71	13.05	16.67	8.33	20.83	0.5803	0.6046			0.4929	0.5291			0.3464	134.71	8.68	8.47
0.20	0.00−0.50−1.00	14	2.90	7.44	21.43	7.14	35.71	0.5734	0.7985			0.4724	0.7005			0.3008	44.62	4.73	4.73
0.20	0.00−0.66−1.00	18	2.81	7.04	33.33	5.56	27.78	0.6260	0.7311			0.5012	0.6035			0.2980	82.93	7.01	6.77
0.20	0.00−0.25−0.50−1.00	42	4.60	24.33	16.67	14.29	11.90	0.5418	0.6029	0.7985		0.5085	0.5334	0.7512		0.3466	440.66	15.26	15.11
0.20	0.00−0.33−0.66−1.00	52	4.85	28.81	28.85	11.54	9.62	0.5803	0.6258	0.7769		0.5425	0.5109	0.6253		0.3376	813.38	23.20	23.14
0.20	0.00−0.50−0.75−1.00	29	3.40	10.56	20.69	17.24	17.24	0.5734	0.8368	0.7817		0.5300	0.7990	0.5203		0.3178	260.15	13.49	13.97
0.20	0.00−0.15−0.40−0.60−1.00	58	4.83	28.51	36.21	13.79	8.62	0.6436	0.4713	0.6163	0.7311	0.5381	0.4733	0.5166	0.6553	0.3467	1279.90	33.64	33.46
0.20	0.00−0.25−0.50−0.75−1.00	76	4.96	31.12	35.53	21.05	6.58	0.5418	0.6029	0.8368	0.7817	0.5494	0.5736	0.8103	0.5498	0.3544	1936.47	37.93	38.36
0.20	0.00−0.40−0.60−0.80−1.00	58	4.63	24.77	34.48	12.07	8.62	0.5745	0.6163	0.8280	0.7797	0.5714	0.5465	0.7082	0.5365	0.3471	1329.24	34.43	34.56

(3) The number and percentage of singletons (%S) and doubletons (%D).

(4) The entropy of the clustering process (CE), the number of effective clusters (ENC), calculated using the following expressions[12]

$$\mathrm{CE} = -\sum_{i=1}^{\mathrm{CT}} f_i \log_2 f_i \tag{6}$$

where CT is the number of clusters and $f_i = n_i/db$ is the population frequency in each cluster, calculated as the ratio among the population of each cluster (n_i) and the number of elements of the database (db).

Knowing CE, the effective number of clusters can be calculated with the expression:[12]

$$\mathrm{ENC} = 2^{\mathrm{CE}} \tag{7}$$

(5) A representative (centroid) MCS of each cluster is selected as follows: (i) The similarity among the elements of a class is calculated (it has already been calculated previously in the preprocessing stage). (ii) The average $A_k(i)$ and variance of the similarity value obtained for each element i of the class are calculated. (iii) The average (A_k) and variance for the class are calculated. (iv) That element of the class whose difference $|A_k(i) - A_k|$ is smaller, is chosen as representative of the class.

(6) A representative (centroid) MOS from each cluster is chosen in a similar way to the representative MCS. In this case, in the calculation of the similarity, the MOS values are considered instead of the MCS. This information has been previously calculated in the preprocessing stage and will be used to analyze and compare the results of the classification process.

(7) 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 smaller.

(8) The Cartesian center (gravity) of the clusters or classes (GCL). This value is calculated in a similar way to the GCC using in its calculation the cluster centroids instead of all the database elements.

(9) The average of similarity of the clusters or classes (ASL). This value is calculated in the same way as the ASDB using the cluster centroids in the calculation instead of all the database elements.

(10) The dispersion of the representatives (RD), obtained as the sum of the Euclid distances among all the cluster centroids.

(11) The dispersion of the clustering (CD), obtained as the sum of the Euclid distances between all the centroids and the center of gravity of the clustering.

(12) The dispersion of the cluster or classes (LD), obtained as the sum of the Euclid distances between all the centroids and the center of gravity of the clusters.

These measurements allow us to observe deviations in the process of grouping of the elements of each class based on the size and number of the selected intervals of similarity as well as the dispersion of the classes a characteristic of the diversity of the database. The computational cost of this calculation is trivial since the information corresponding to

the similarity between the database elements has been calculated in the preprocessing stage.

3. APPLICATION OF THE CLUSTERING METHOD. STUDY OF THE CLASSIFICATION PARAMETERS

The tests of the clustering method proposed have been carried out on several databases with very different characteristics, such as for size and diversity of the chemical compounds that are stored. These databases are of public domain and/or they have been given freely,[13] which facilitates the reproduction of the results described in this paper for the scientific community.

The molecules of the database can be stored in files with .mol or .sdf format[16] or use of the Oracle DBMS to store a great number of molecules. On each of these databases we have carried out the preprocessing stage calculating the set of similarities among the database elements by means of the algorithm described in the bibliography.[8] This information is ordered and classified conveniently in a text file, which is analyzed for the extraction of all the parameters described in the previous section. Last comes the assignment of the database elements to their corresponding class and the selection of the centroids, information that, together with the parameters and statistical previously mentioned, is conveniently stored for their quick access.

The results shown in this paper have been obtained on a database formed by 498 natural compounds extracted from the SB2C_20T database containing synthetic organic compounds with a wide molecular diversity,[13] showing the mean value of the similarity ASDB = 0.4021, which supposes that among the database elements, on average terms, there is a maximum common substructure that corresponds to 40% of their structure.

Tables 1 and 2 show the results obtained for the two clustering methods proposed with the study of the different parameters that influence the classification process described in the previous section.

3.1. Influences of the Cell Size. According to expression 5, the cell size together with the dimensions of projection space (N) determine the maximum number of clusters that can be generated. In the tests shown in this paper different cell sizes have been used (0.2, 0.1, and 0.05) with equal size for all the dimensions in the different projection spaces.

Thus, a value of cell size of 0.2 for a space of projection of 2 dimensions ($N = 2$) supposes that the maximum number of clusters that can be formed is of 25, 100 for a cell size of 0.1, and 400 for a cell size of 0.05.

The selection of the cell size and the number of dimensions of the projection space will be chosen in function of the database size and the required average of the population of the clusters.

As Tables 1 and 2 show, when increasing the cell size the number of clusters (CT) diminishes, increasing the average of the population of the clusters (APC) and diminishing the entropy of the clustering process (CE) and the number of effective cluster (ENC), independently of the rest of the studied parameters. In general, this effect causes a decrease in the singletons and doubletons percentage.

An example showing this behavior can be extracted from Table 1 considering a projection space $N = 2$ with intervals of similarity equal to (0.00−0.33−1.00). We can observe the behavior previously described when the cell size increases. Thus, for cell values of (0.05, 0.10, and 0.20), CT takes the following values (88, 38, and 16), verifying there occurs a decrease in the cluster number and an increase in the averaged population of the classes (APC) with values (5.68, 13.16, and 31.25). For the entropy of the cluster process (CE) the values obtained are (5.89, 4.44, and 2.95) confirming the decrease previously described, which is also the case when we observe the values obtained for the effective number of clusters (ENC) (59.12, 21.69, and 7.74).

Generally, an increase in the cell size provides a decrease in the average of the similarity of the cluster centroids (ASL), an effect that is due to the increase in the diversity of the cluster elements when increasing the size and, therefore, to the increase of cluster population.

The increase in the cell size, maintaining the other parameters constant, causes a decrease in the dispersion measurements (RD, CD, LD). When we project over big cells, the database elements are assigned to closer cells, producing clusters with higher populations, dense and less dispersed in the classification space.

3.2. Influences of the Dimension of the Projection Space. The dimension (N) of the projection space, that is, the number of intervals of similarity on which the clustering process will be carried out, has an obvious effect on the behavior of the classification process.

As Tables 1 and 2 show, by maintaining the rest of the studied parameters constant, an increase in the value of N generates an increase in the values of CT, CE, ENC, %S, and %D. This effect is because when increasing N, the number of possible cells that can be generated considerably increases (sees expression 5), increasing the probability of dispersion of the database elements of these cells. This effect causes dispersed clusters with low populations (diminishing APC).

An example clarifying this behavior can be extracted from Table 2 considering a cell size of 0.05 and intervals of similarity of (0.00−0.50−1.00), (0.00−0.33−0.66−1.00), and (0.00−0.25−0.50−0.75−1.00). In this case the number of generated clusters (CT) is of 86, 278, and 343, the entropy of the clustering (CE) is 5.77, 7.78, and 8.14, and the effective number of clusters (ENC) is 54.75, 219.21, and 282.14, respectively, showing the tendency to increase the values of these parameters when the dimension of the projection space increases.

We can observe that the increase in the dimension of the projection space (N) produces an upward behavior of the ASL values as well as of the dispersion measurements. When increasing the projection dimension, dispersed clusters distributed in the multidimensional space of projection are generated, and these clusters are less dense and more homogeneous. This behavior is more constant in the projection method 2 than in method 1.

3.3. Influences of the Intervals of Similarity. The values of the intervals of similarity (vector V) considerably affect the classification process independently of the other parameters studied. This parameter is directly dependent on the characteristics and size of the database. Very diverse databases, where very different elements exist, produce very different values in the similarity matrix P, which produces very different values in the normalized matrix $\bar{P}$. This effect deals with database elements that are projected along the

different values of intervals of similarity, having a tendency to occupy the intervals with values close to zero.

Therefore, the selection of the values of the intervals of similarity for a dimension size (N) of the given projection space will directly determine the granularity of the classification process.

As Tables 1 and 2 show, the behavior observed is totally different for the two studied projection methods. For a given size of projection space (for example, $N = 2$), as the interval of similarity close to zero is higher ($0.33 < 0.50 < 0.66$), with method 1 an increase in the number of clusters is observed, which causes a decrease in the population of the clusters and an increase in the ASL value as well as increasing the dispersion measurements. This effect is because projection method 1 is sensitive to the characteristics of the whole database, since in the denominator of the expression (2) M_{db} is considered to represent the total number of database matching whose value of similarity is in the interval $V(k)$.

Therefore, as in the calculation of the similarities, the MCS values are taken into account, and these values in dissimilar databases are low (in our case ASDB = 0.4021), when intervals of similarity next to zero increase, a higher differentiation among the points in those where the database elements are projected, causing a higher dispersion of the clusters and less dense clusters formed by more similar elements. As is evident, this effect becomes more patent as the cell size diminishes, a greater difference in the values of the dispersion parameters and similarity being obtained.

On the other hand, projection method 2 presents a behavior, which is sensibly different to projection method 1. In method 2, an increase in the interval of similarity next to zero spreads to generate a decrease in the number of clusters (CT) and, therefore, a decrease of CE and ENC and an increase in the averaged population of the clusters (APC). But this effect is very dependent on the value of this interval (tending to be stabilized and even attain values close to, or higher than, the ASDB value), as the size of the projection space (N).

As expression 3 shows, in the denominator, the value of M_i is considered as representing the total number of matching of the i element whose value of similarity is in the interval $V(k)$, instead of M_{db}, as considered in method 1 (expression 2). That is why method 2 is more sensitive to the values of the other parameters used in the projection method, since this method is more affected by the characteristics of each element of the database in comparison with method 1 which is affected by the characteristics of the whole database.

This behavior can be appreciated in the values of the average of the similarity of the classes (ASL) and the dispersion values which spread to a maximum for a distribution of the sizes of the intervals of similarity centered in the ASL value.

3.4. Influences of the Projection Method. The behavior of the two proposed projection methods can be appreciated comparing the results of Tables 1 and 2.

Method 1 presents a slight tendency to generate a higher number of clusters than method 2. However, this behavior is affected by the values of the intervals of similarity (for the same value of the dimension of the projection space), as we commented in the previous section.

We can observe method 2 gives lower values of the average of the similarity of the classes (ASL) than method 1 in practically all cases. However, the dispersion measurements are greatly affected for the values of the other parameters in study. This behavior is due to the greater sensibility of method 2 with the values of the selected intervals of similarity.

So, method 2 generates cluster centroids that are much more different from those generated by method 1 (lesser ASL value), although the classes, generally, are distributed to each other along closer clusters in the N-dimensional Cartesian space of projection, which is shown in the values of GCC and GCL represented in Tables 1 and 2.

3.5. Analysis of the Results. The behavior described in the previous section for the clustering model proposed in function of the different classification parameters can be appreciated graphically in Figures 1−3. These figures show the distribution of the number of clusters and the population of the clusters for the two classification methods utilized and for different values of cell size, number of projection dimensions, and values of the intervals of similarity.

These figures show the influence of the parameters used in the classification process. If we observe these figures for rows we are able to appreciate the influence of the cell size; if we observe these figures for columns we are able to appreciate the influence of the intervals of similarity used in the projection; and, considering a row/column by method, we are able to appreciate the influence/behavior of the proposed projection methods or the influence of the number of dimensions of the projection space if we consider the different figures (1−3).

We can observe that for small values of cell size (0.05) the number of clusters increases excessively, producing a great number of clusters with very low populations (Figures 1−3). Conversely, for high cell sizes the number of clusters decreases, increasing the population of the clusters.

For instance, observing Figure 1 when we use the projection method 1 (or method 2) and an interval of similarity of (0.00−0.33−1.00), we can appreciate that the number of clusters generated (number of bars) for a cell size of 0.20 is smaller than when a cell size of 0.10 is used and much smaller than for a cell size of cell of 0.05. And, therefore, the population of the clusters (Y axis in the graphs) gets bigger as the cell size increases ($0.20 > 0.10 > 0.05$).

A high number of clusters with low population is not convenient because it entails a high computational cost of the screening process. On the other hand, a low number of clusters with a high population, although it increases the effectiveness of the screening process, diminishes the effectiveness of the later process of comparison of the pattern compound of search with each of the cluster elements (structural matching process).

In that case, it is necessary to consider situations of commitment depending on the objective pursued. These situations of commitment can be found by means of the consideration of the size and number of dimensions of the projection space. As Figures 1−3 show, for the same cell size, the selection of the projection space gives a variation in the number of clusters and the population of the clusters.

For example, viewing Figure 1, for a cell size of 0.20 and using method 2 is appreciated as the number of generated clusters is of 24, 14, and 18 when intervals are used of

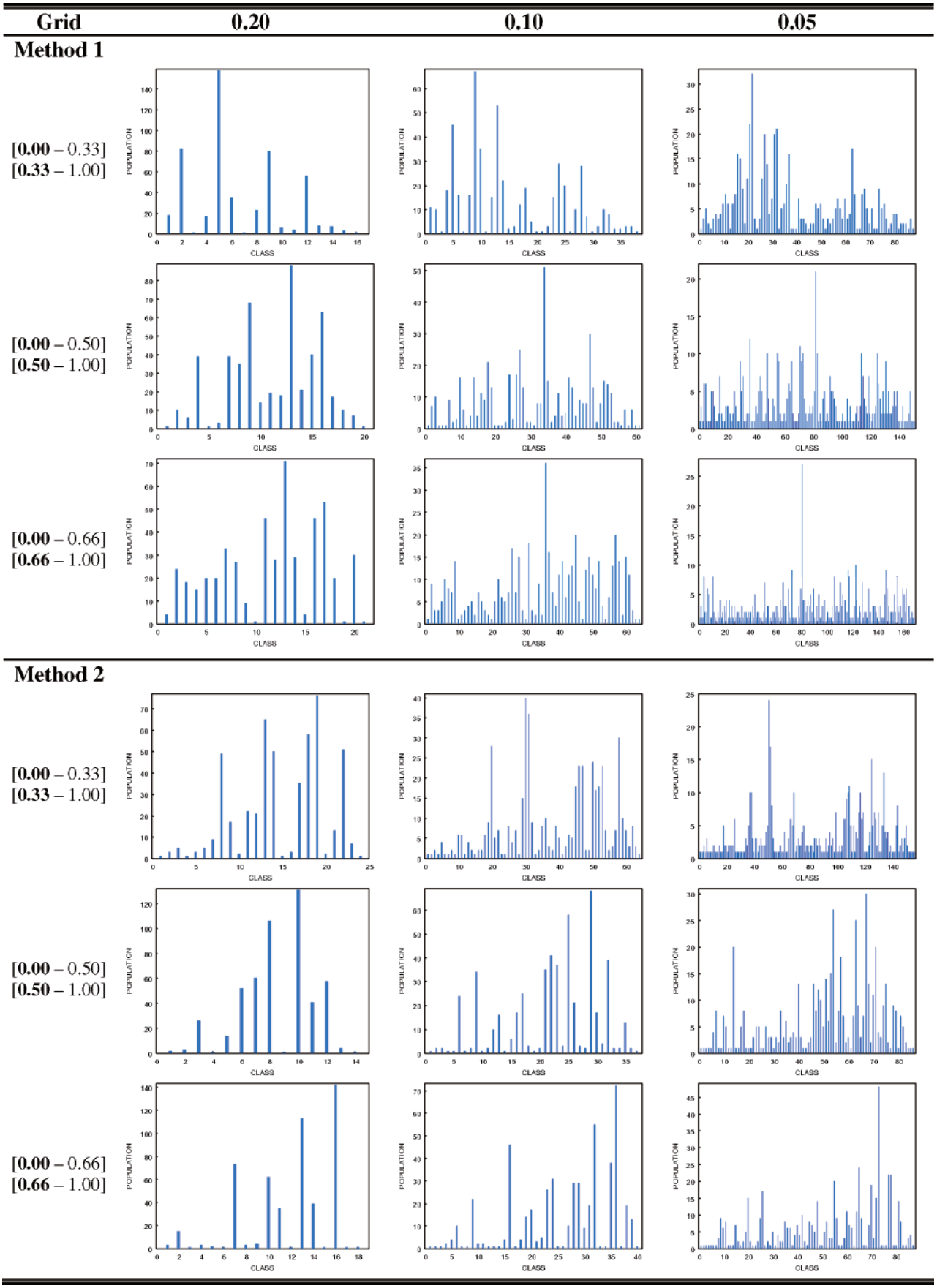

Figure 1. Behavior of the classification process using a 2D projection space for different values of cell size, intervals of similarity, and for the two proposed classification methods. The *X* axis shows the number of clusters, and the *Y* axis shows the cluster population.

(0.00−0.33−1.00), (0.00−0.50−1.00), and (0.00−0.66−1.00), respectively.

The use of both parameters (number of dimensions and cell size) allows us to find good values in the number and

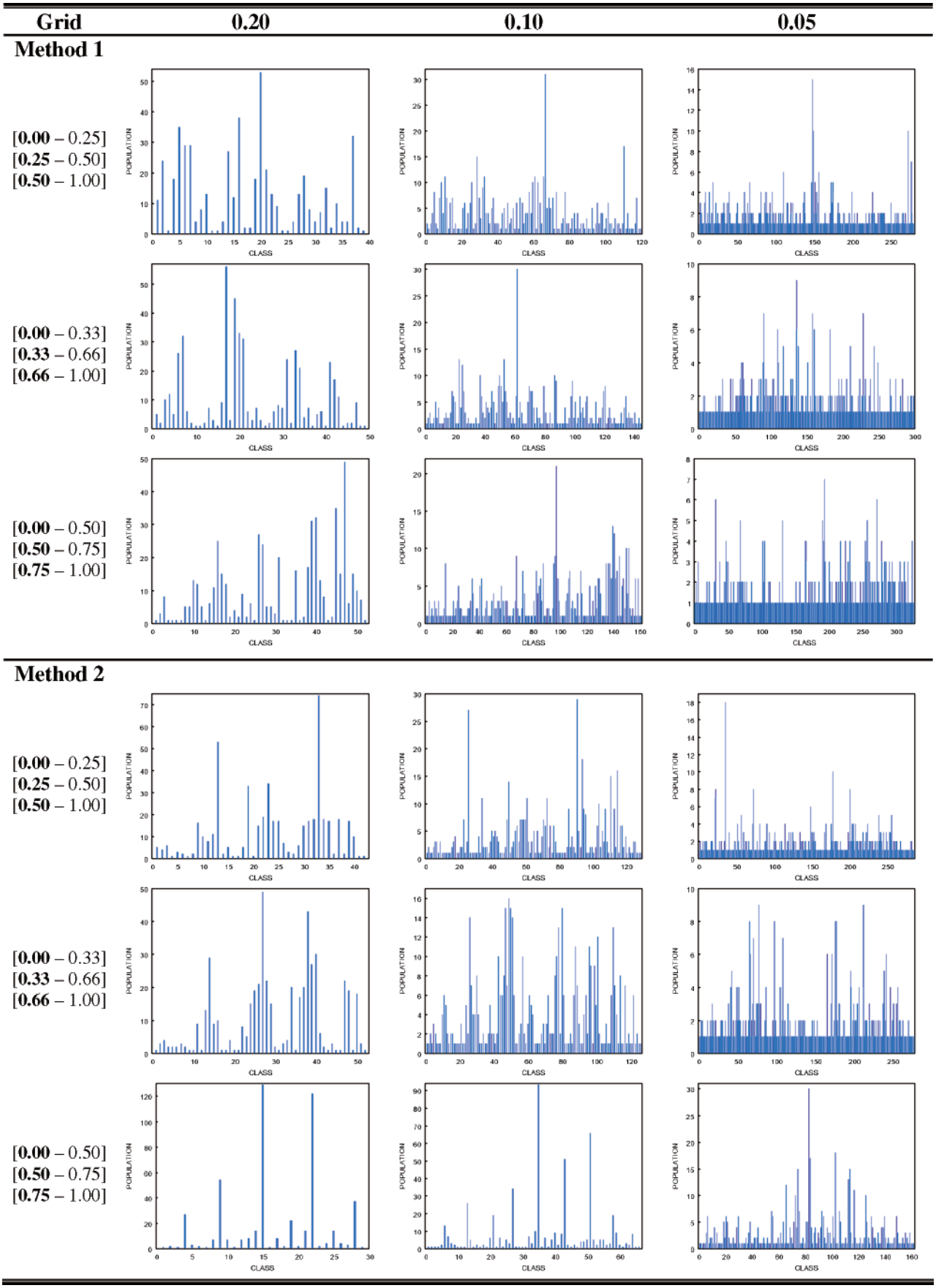

Figure 2. Behavior of the classification process using a 3D projection space for different values of cell size, intervals of similarity, and for the two proposed classification methods. The *X* axis shows the number of clusters, and the *Y* axis shows the cluster population.

population of the clusters in function of the size and database characteristics.

The different characteristics of the proposed projection methods can be appreciated in Figures 1−5. In Figure

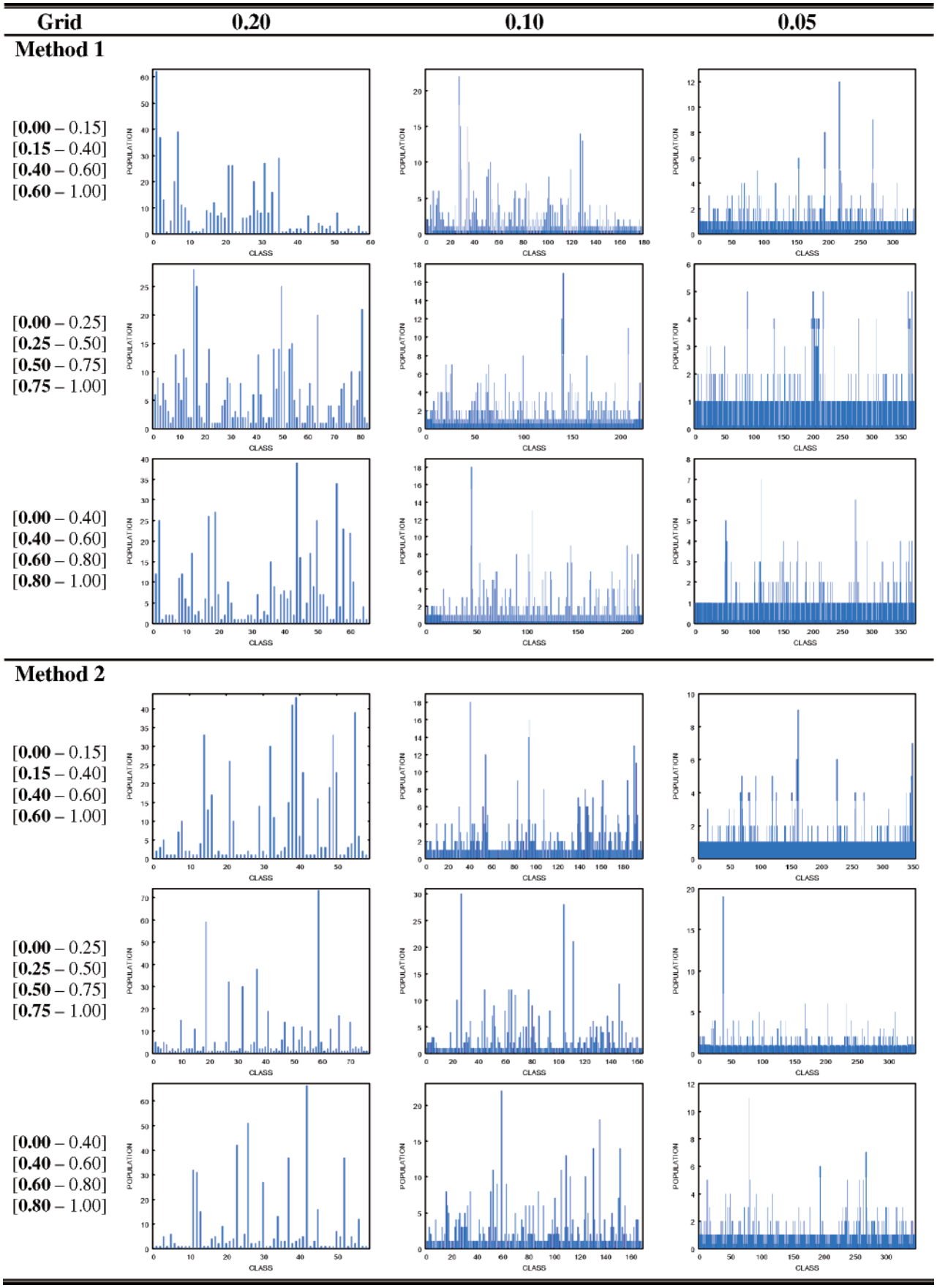

Figure 3. Behavior of the classification process using a 4D projection space for different values of cell size, intervals of similarity, and for the two proposed classification methods. The *X* axis shows the number of clusters, and the *Y* axis shows the cluster population.

4 the characteristics of methods 1 and 2 for the clustering process have been represented in a projection space 3D for different values of cell size and intervals of similarity.

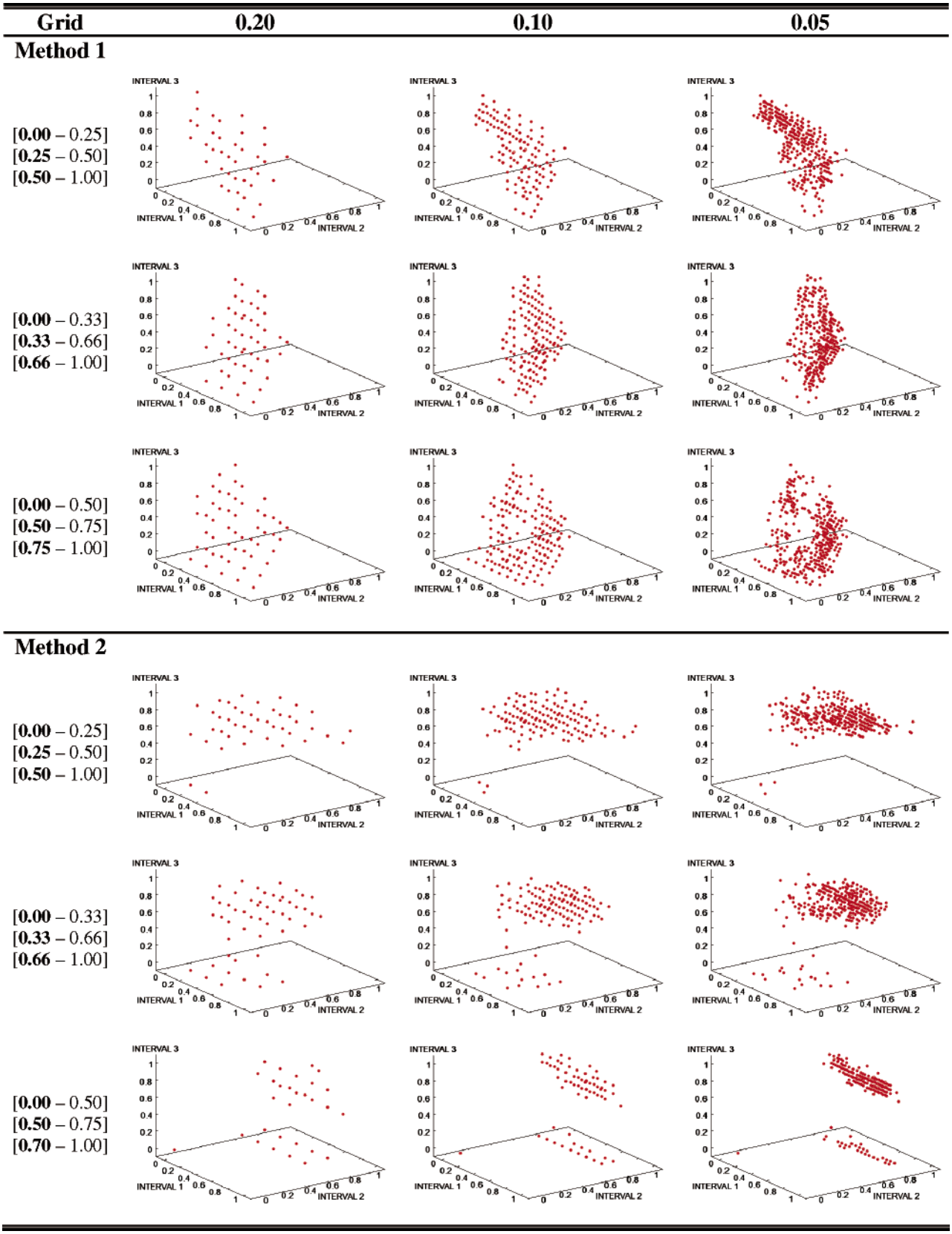

Figure 4. Cluster distribution in a 3D projection space for the two proposed projection methods and for different values of cell size and intervals of similarity.

In Figure 4 the position of each one of the clusters in the defined projection space has been represented (3D in this case). As can be observed, the increase in the cell size produces the clusters grouping and, therefore, the decrease in the clusters number. For example, for the interval of similarity (0.00−0.25−0.50−1.00) using method 1, it can be appreciated how the clusters number diminishes when the cell size increases (0.20 < 0.10 < 0.05). However, the characteristics of the projection process are not influenced by this parameter, as is appreciated in the distribution of the

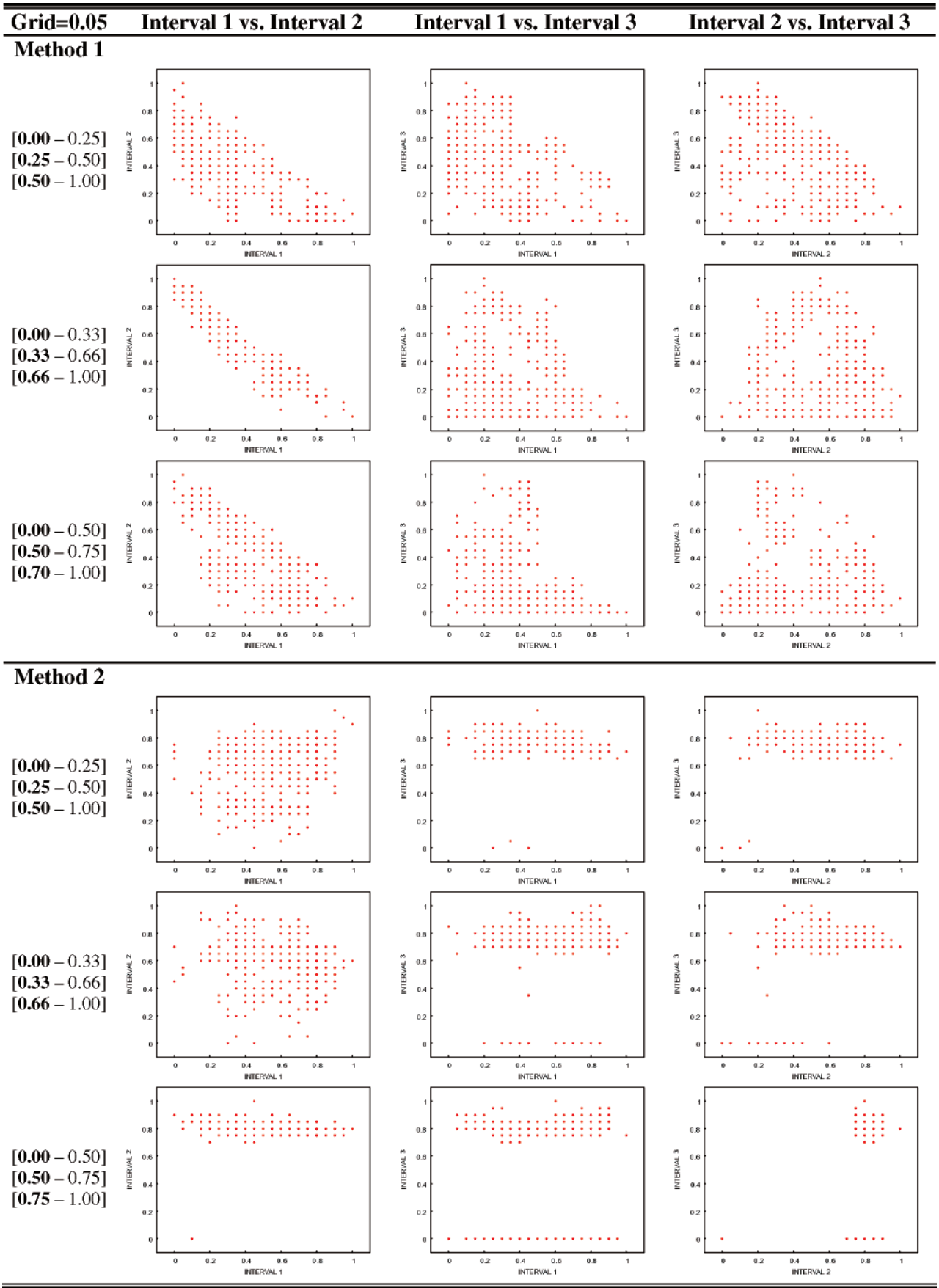

Figure 5. Behavior of the methods of classification 1 and 2 for a 3D projection space, a cell size of 0.05, and different values of intervals of similarity.

clusters (the shape/outline of the figure generated by the clusters in the space).

While distribution of the clusters for method 1 is more homogeneous in the whole multidimensional space of

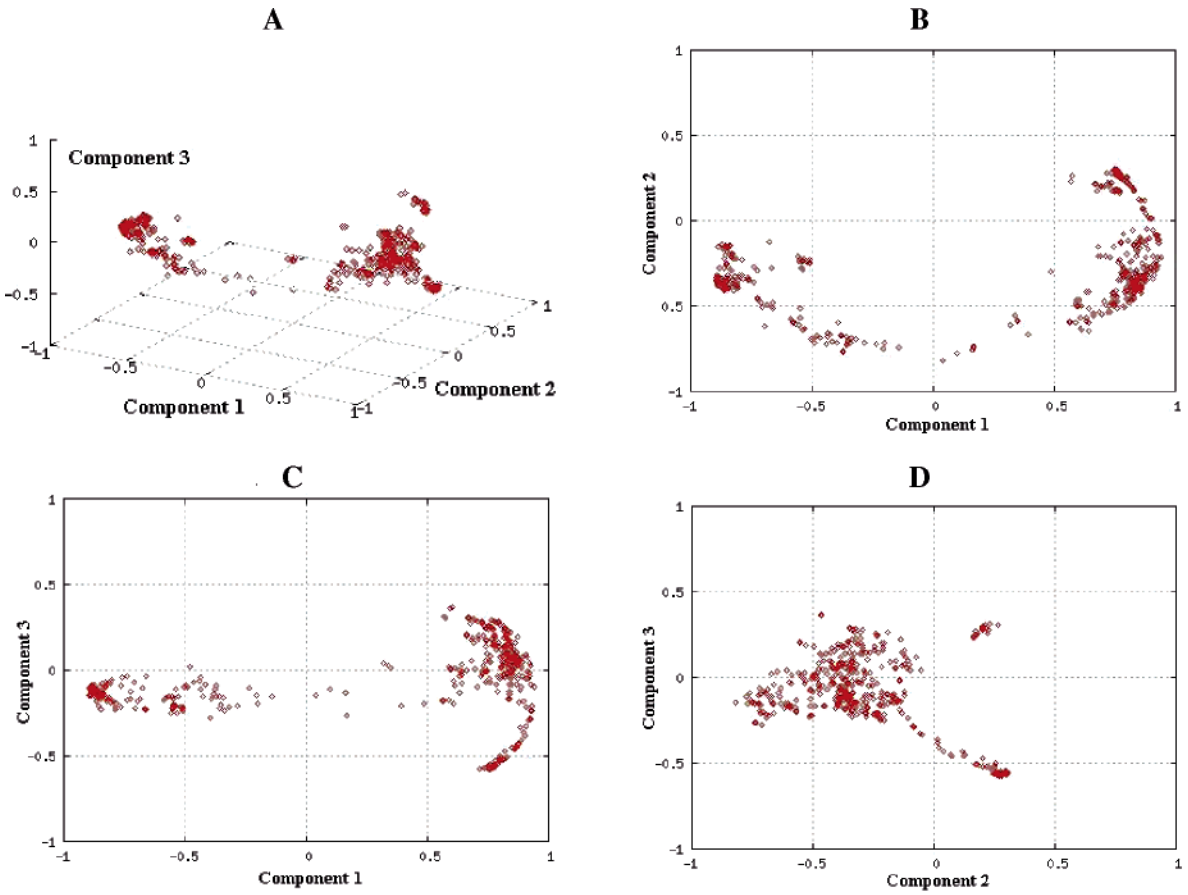

Figure 6. Results of the principal component analysis with the database in study. (A) 3D representation of the first three principal components, (B) PC1 vs PC2, (C) PC1 vs PC3, and (D) PC2 vs PC3.

projection, for method 2 clearly different groups of clusters are observed. This distribution in different groups becomes more patent in Figure 4 as we diminish the cell size, due to the increase in the number of clusters.

We can observe in Figure 4 that the projection method 2 is more sensitive than method 1 to the values of the intervals of similarity, as mentioned in the previous section. The changes in the sizes of the intervals of similarity affect the clusters groups and the dispersion among the clusters sensitively, a higher differentiation being obtained as the size of the interval of similarity next to zero ($0.50 > 0.33 > 0.25$) increases.

This behavior is appreciated more clearly if we represent the distribution of the clusters in function of the different intervals of similarity. The data shown in Figure 4 for a cell size of 0.05 are shown in Figure 5 representing the influence of the intervals of similarity.

Figure 5 shows in a 2D graph, the distribution of the clusters with regard to the different intervals of similarity, which allows us to appreciate the different behavior more clearly (distribution of the clusters) of the two clustering methods proposed, already represented in Figure 4 for a 3D projection space.

We can observe that for method 1 the distribution of the clusters is very homogeneous in the whole projection space, being little influenced by the values of the intervals of similarity. However method 2 is sensitively influenced by the values of the intervals of similarity, clearly observing groups of independent clusters whose number, size, and distribution depend on these values.

Moreover, in the case of method 1 the clusters are distributed practically for all the cases in the interval [0−1] for all the intervals of similarity. However, in the case of method 2, this behavior only takes place for close values of interval of similarity and, mainly, smaller than the ASL value, the other cases being distributed in values very close to 1.

3.6. Comparison with Other Classification Methods. The difficulty of evaluating the useful of different classification methods comparatively has been considered in the bibliography.[4,14,15] In fact, each method is based on a different variable for the clustering, which generates different cluster numbers, etc.

Usually, the principal component analysis (PCA)[2,3,15] is used as a preparatory technique for the classification process. 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*.

This analysis has demonstrated that 17 components are necessary in order to explain 95% of the variance; 79.32% is explained by the first three components. In Figure 6 the first three components (Figure 6A) and their corresponding 2D projections (Figure 6B−D) have been represented. As we can observe the PCA technique is able to find some groups of molecules that could be contained in clusters or classes, although due to the database characteristics few and populated clusters can be observed, there being a high number of molecules that are difficult to classify, as shown

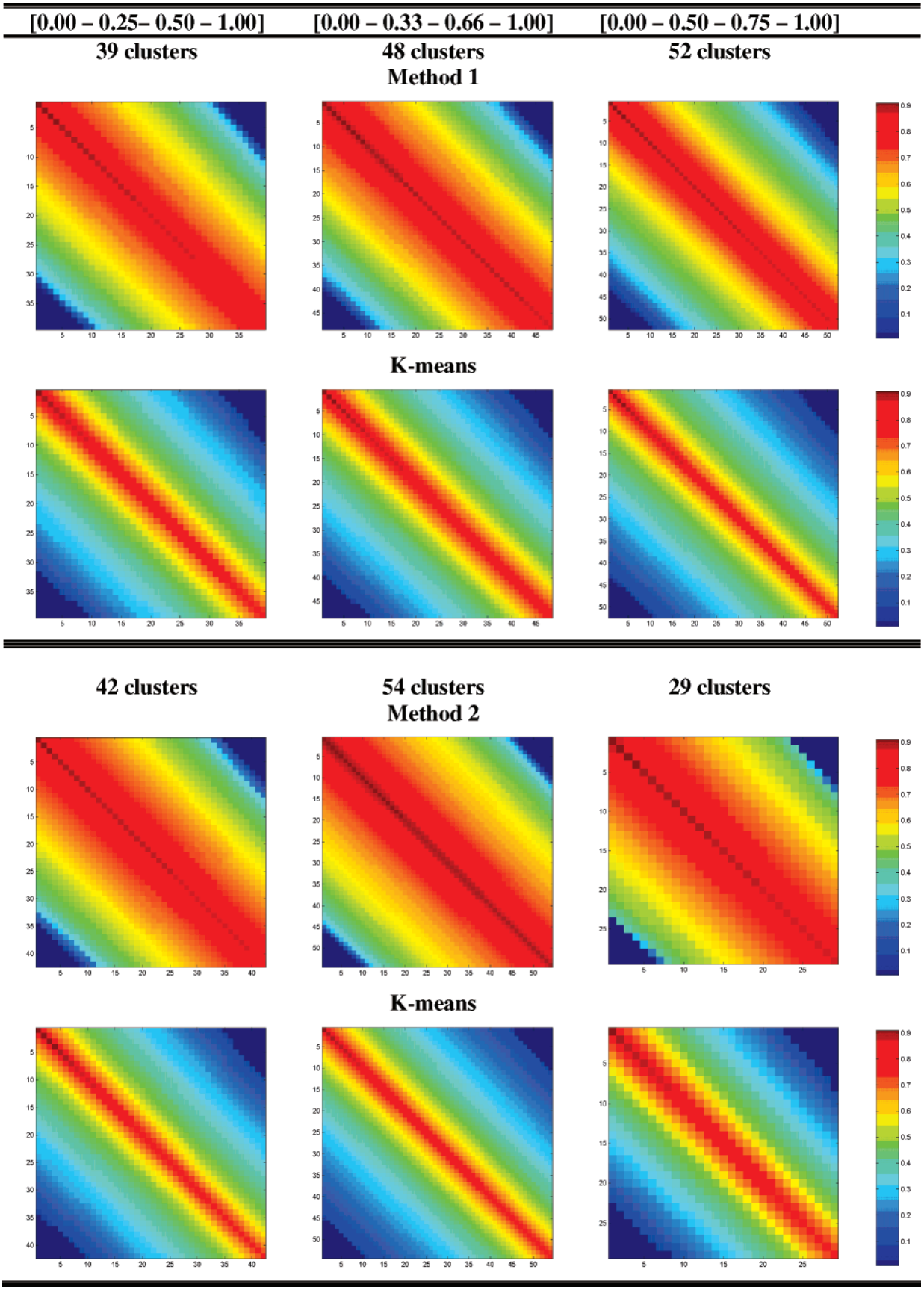

Figure 7. Representation of the distances between cluster centroids for the proposed methods (cell size = 0.2) and K-means method with the same number of clusters.

in the representations among the first three main components (see Figure 6B−D) in which that not more than 2−4 groups of molecules can be clearly appreciated.

The results obtained with the proposed methods have been compared with the classification method K-means[2−4,14,15] using as input data the similarity matrix *S*. For an objective

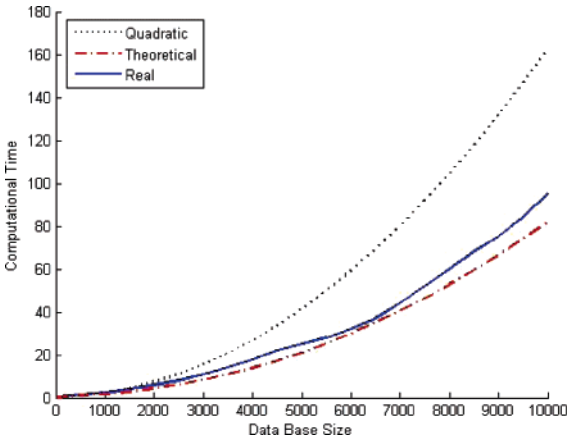

Figure 8. Analysis of the computational cost of the classification method with different database sizes.

comparison we have imposed the generation of the same number of clusters with the K-means method as the clusters generated by the proposed methods for values of classification parameters given.

K-means method generates clusters with a similar average similarity to the proposed methods. However, the proposed methods generate more dispersed clusters than the K-means method.

In Figure 7 we have represented graphically, by means of a color scale, the normalized distances among the cluster centroids for given values of classification parameters (cell size of 0.2, 3D projection space, for three different values of intervals of similarity, and for the two proposed classification methods). In each case we have applied K-means imposing the generation of the same clusters number as those generated for the proposed methods.

For both methods the distances among the centroids of the clusters have been calculated, and they have been normalized in the interval [0,1] assigning to each value in this interval a color in the range [blue-red] making use of statistics toolbox of Matlab.[17] Thus, the most separate clusters (normalized distances next to 1) have assigned a color close to red, while the very next clusters (normalized distances next to zero) have assigned a color close to blue.

As Figure 7 shows, the normalized distances among the cluster centroids for the proposed methods are, in all cases, greater than the K-means method. This fact 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 by in our proposal) improves the classification of chemical databases with regard to the use of the similarity matrix (used by other clustering models as K means).

As we commented previously, the distance among the clusters generated by method 2 is markably higher than method 1 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.

3.7. Computational Cost. Figure 8 shows the study of the computational cost of the proposed classification method using a PC Pentium II 400 MHz. For this study databases of different size have been built (from 100 to 10000 molecules) and the corresponding similarity matrices have been obtained in the preprocessing phase.

The computational cost of the processing phase supposes the following: (i) The projection of the similarity matrix on the defined projection space for the selected intervals of similarity. This stage has a computational cost $O[(1/2)n^2 - n]$, n being the database size (dimension of the similarity matrix), since the similarity matrix is symmetrical. (ii) The normalization of the new generated matrix of size nxd (d being the number of dimensions or number of selected intervals of similarity). The computational cost of this stage is $O(nd)$. (iii) The assignment of the normalized matrix entries to each of the defined cells (see expression 5) and the deleting of the empty clusters. This process is carried out with a computational cost of $O(n + s)$, s being the maximum clusters number generated in function of both the cell size and number of intervals of similarity. (iv) And, last, the extraction of the statistical parameters. This process has different orders of complexity depending on the statistical parameter.

In Figure 8, the theoretical computational cost of the process, the real computational cost compared to a computational cost $O(n^2)$ has been represented. As can be observed the real cost approaches the theoretical cost (the deviations are due to the cost of the measure of times, cost of operating system processes, etc.) being much lower than $O(n^2)$.

3.8. A Screening Example. The classification of the databases has as an objective, among others, to improve the performance of the recovery process of information. Generally, these processes are based on the selection of a pattern molecule with the aim of recovering those database molecules most "similar". So, if the database is classified based on measures of structural similarity, as is the case of the methods described in this manuscript, the screening process will be based on recovering those molecules whose structural similarity with regard to the pattern molecule satisfies a selected similarity threshold. In this screening process the pattern molecule should be compared with the centroids of each cluster, recovering those clusters that satisfy the given similarity threshold.

Figure 9 shows an example of a screening process with the database studied in this manuscript for a randomly selected pattern molecule. For this example, the database has been classified by method 2, using a cell size of 0.05 and considering different values of intervals of similarity and sizes of the projection space (2D, 3D, and 4D).

For each of the working conditions (classification parameters) we have recovered the clusters in which the pattern molecule is classified, and the statistical parameters that characterize each of these clusters have been analyzed.

Figure 9 shows some of the values of the most representative statistical extracted: the average of the similarity of the clusters (ASC), the variance (V), and the average of the similarity of the representative of the cluster (ASR). These statistical values have been calculated considering two approaches in the measures of the similarity of the clusters: (a) MCS (maximum common substructure), in which the described methods are based, and (b) MOS (maximum overlapping set), a similarity approach generally utilized to measure the structural similarity among molecules.

Figure 9 also shows the housed molecules in each cluster in which the pattern molecule is classified. For each cluster the centroid has been colored in yellow when the MCS approach is used as measure of similarity, in blue when the

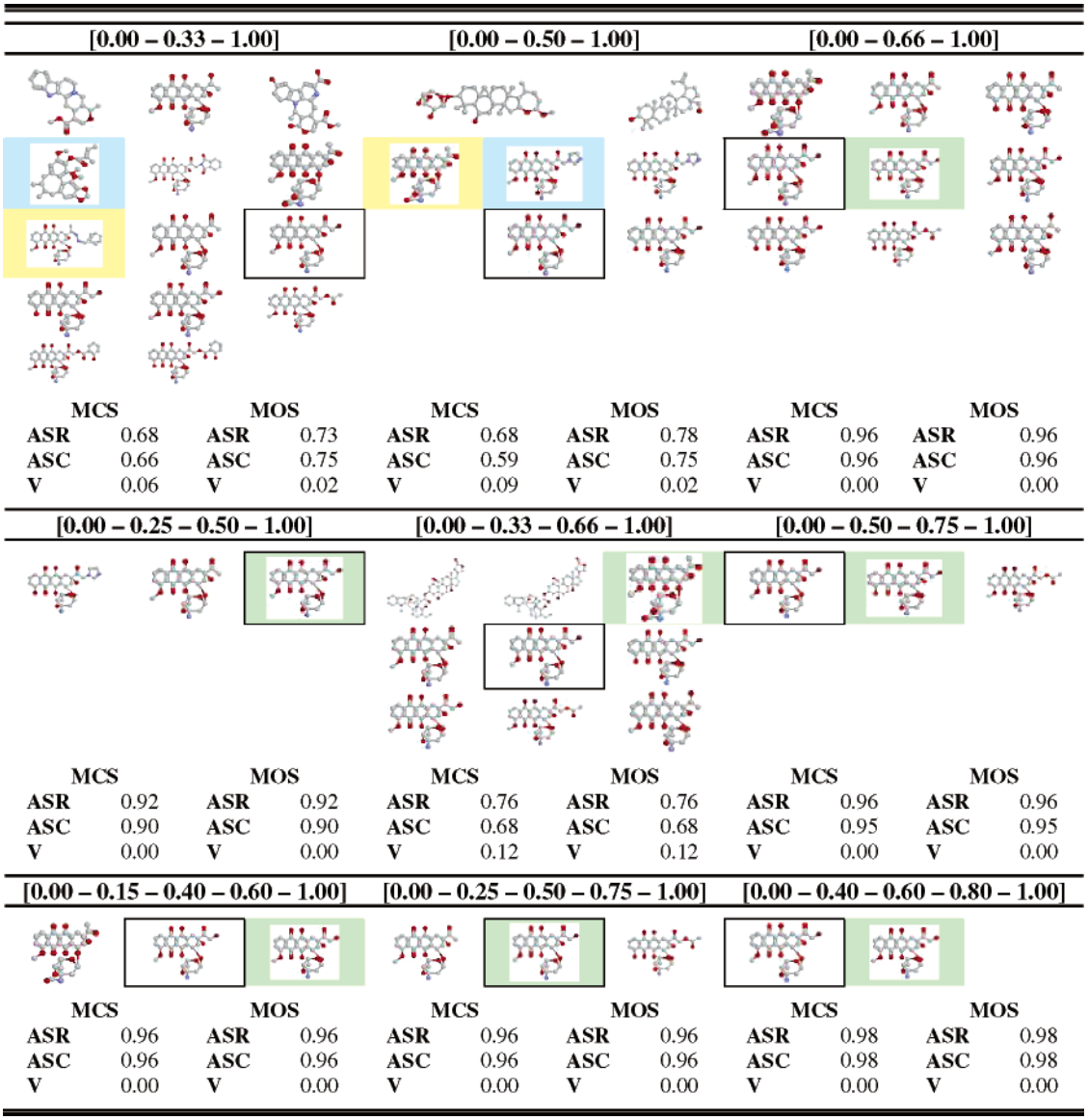

Figure 9. Screening example for different sizes of the projection space and intervals of similarity making use of method 2 in the classification process. ASR: average of the similarity of the representative of the cluster, ASC: average of the similarity of the cluster, V: variance, MCS: in the calculation of the similarity the maximum common substructure is used, MOS: in the calculation of the similarities the maximum overlapping set is used.

MOS approach is used, and in green when the centroids with both approach are the same.

The use of the similarity measurements based on the MCS for the classification process shows less sensitive behavior to the characteristics of the database than the use of the MOS.[18] Although the similarity values based on the MCS are usually smaller than those based on the MOS, they are less influenced by the size of the structures (molecules),[19] generating a more homogeneous distribution of the clusters, as can be appreciated in Figure 9. However, this characteristic should be taken into account in the screening processes, where smaller thresholds of similarity are to be chosen.

Evidently, ASR and ASC values are smaller when MCS instead of MOS is considered, which will be taken into account in the selection of the threshold of similarity in the recovery process, as commented previously. In all cases the ASR and ASC values (using the MCS) are much higher than the value of ASL, it being also very high when the MOS is used in the calculation.

4. DISCUSSION AND REMARKS

The proposal carried out in this paper for the classification of chemical databases using measurements of structural similarity based on the MCS has proven appropriate. With low computational cost (without considering the preprocessing cost) the characteristics of the classification process can be adjusted as a function of the database characteristics and of the objectives pursued.

The versatility of the method allows us to readjust the number of dimensions of the projection space and the values of the intervals of similarity, carrying out adjustments in the clustering until obtaining the desired state with low computational costs and even carrying out significant changes in the state of the obtained clustering by means of changes in the cell size, with a low computational cost.

The proposed clustering model allows it to adapt to different requirements or objectives, since the definition of different projection dimensions, cell size, and values of the intervals of similarity allow the researcher to generate more

or fewer clusters, more or less dispersed. Also, the adjustment of these parameters allows the extraction of the clusters database (groups of molecules) that present a greater diversity in the database. Thus, defining small intervals of similarity close to zero groups of molecules (clusters) which are very different to the rest of the database molecules can be found, and, on the other hand, defining intervals of similarity close to one, clusters of greatly similar molecules can be obtained.

Evidently, the effectiveness or adjustment of the classification process should be interpreted by the researcher as a function of the pursued objectives and the analysis of the great number of statistical values that the developed method provides. At the moment we are working on the idea of designing a self-adaptive method that gives support to the user, using the set of statistical and the researcher requirements/objectives.

The clustering method described in the manuscript is a nonhierarchical method since the clusters generated are based on the projection of the values of similarity among the database molecules in previously defined cells through the selection of the number and characteristics of the intervals of similarity. However the proposed method takes advantage of the hierarchical methods, since the redefinition of the cell size allows the grouping/disaggregating of the clusters.

Thus, once the database is classified for a given value of similarity intervals, we can carry out the database reclassification (with a negligible computational cost) simply by changing the cell size and generating more clusters if we diminish the cell size and, vice versa, fewer clusters if we increase the cell size. This way, the selection of the cell size has a similar effect to the selection of the level or depth of the clustering in the hierarchical methods. This characteristic allows the proposed classification method to adapt to the different necessities, researcher objectives, and databases characteristics.

ACKNOWLEDGMENT

The Comisión Interministerial de Ciencia y Tecnología (CICyT) is thanked for financial support (Project TIN2004-04114-C02-01).

REFERENCES AND NOTES

(1) Cundari, T. R.; Russo, M. Database Mining Using Soft Computing Techniques. An Integrated Neural Network-Fuzzy Logic Genetic Algorithm Approach. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 281−287.

(2) Kantardzic, M. *Data Mining: Concepts, Models, Methods, and Algorithms*; Wiley-IEEE Computer Society Pr: 2002.

(3) Kaufman, L.; Rousseeew P. J. *Finds Group in Data: An Introduction to Clustering Analysis;* John Wiley & Sons: 1990.

(4) Downs, G. M.; Barnard, J. M. Clustering and Their Uses In Computational Chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 2003; Vol. 18, pp 1−39.

(5) Gillet, V. J.; Downs, G. M.; Ling, A.; Lynch, M. F.; Venkataran, P.; Wood, J. V.; Dethlefsen, W. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 8. Reduced Chemical Graph and Their Applications in Generic Chemical Structure Retrieval. *J. Chem. Inf. Comput. Sci*. **1987**, *27*, 126−137.

(6) Butina, D. Unsupervised Database Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way to Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci*. **1999**, *39* (4), 747−750.

(7) Turner, D. B.; Tyrrel, S. M.; Willett, P. Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 18−22.

(8) Cerruela García, G.; Luque Ruiz, I.; Gómez-Nieto, M. A. Step-by-Step Calculation of All Maximum Common Substructures Through a Constraint Satisfaction based Algorithm. *J. Chem. Inf. Comput. Sci*. **2004**, *44* (1), 30−41.

(9) Raymond, J. W.; Gardiner, E. J.; Willett, P. Heuristics for Similarity Searching of Chemical Graphs Using a Maximum Common Edge Subgraph Algorithm. *J. Chem. Inf. Comput. Sci*. **2002**, *42* (2), 305−316.

(10) Rouvray, D. H.; Balaban, A. T. *Chemical Applications of Graph Theory. Applications of Graph Theory*; Wilson, R. J., Beineke, L. W., Eds.; Academic Press: 1979; pp 177−221.

(11) Willett, P.; Barnard, J. M.; Downs, G. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci*. **1998**, *38* (6), 983−996.

(12) Taraviras, S. L.; Ivanciuc, O.; Carbol-Bass, D. Identification of Groupings of Graph Theoretical Molecular Descriptors Using a Hybrid Cluster Analysis Approach. *J. Chem. Inf. Comput. Sci*. **2000**, *40* (5), 1128−1146.

(13) SPECS and BioSPECS B.V. http://www.specs.net.

(14) Everett, B. S.; Landau, S.; Leese, M. *Cluster Analysis*; Arnold Publishers: 2001.

(15) Jajuga, K.; Sokoowski, A.; Hermann Bock, A. *Classification, Clustering and Data Analysis*; Springer-Verlag: 2002.

(16) MDL Information Systems, Inc. http://www.mdli.com.

(17) The MathWorks, Inc. http://www.mathworks.com.

(18) Cerruela García, G.; Luque Ruiz, I.; Gómez-Nieto, M. A. Clustering Chemical Databases Through Projection of MOS Similarity Measures on Multidimensional Spaces. In *Lectures Series on Computer and Computational Sciences*; VSP International Science Publisher: 2004; Vol. 1, 181−184.

(19) Cerruela García, G.; Luque Ruiz, I.; Gómez-Nieto, M. A. Representation of the Molecular Topology of Cyclical Structures by means of Cycle Graphs: 2. Application to Chemical Databases *J. Chem. Inf. Comput. Sci*. **2004**, *44* (4), 1383−1393.

CI050035O