International Journal of Bifurcation and Chaos, Vol. 16, No. 7 (2006) 2073–2079
© World Scientific Publishing Company

UNSUPERVISED CLUSTERING USING FRACTAL DIMENSION

D. K. TASOULIS* and M. N. VRAHATIS†
Computational Intelligence Laboratory,
Department of Mathematics, University of Patras,
University of Patras Artificial Intelligence Research Center (UPAIRC),
GR-26110 Patras, Greece
**dtas@math.upatras.gr*
†*vrahatis@math.upatras.gr*

Received November 15, 2004; Revised April 10, 2005

Clustering can be defined as the process of "grouping" a collection of objects into subsets or clusters. The clustering problem has been addressed in numerous contexts and by researchers in different disciplines. This reflects its broad appeal and usefulness as an exploratory data analysis approach. Unsupervised clustering algorithms have been developed to address real world problems in which the number of clusters present in the dataset is unknown. These algorithms approximate the number of clusters while performing the clustering procedure. This paper is a first step towards the development of unsupervised clustering algorithms capable of identifying *clusters within clusters*. To this end, an unsupervised clustering algorithm is modified so as to take into consideration the fractal dimension of the data. The experimental results indicate that this approach can provide further qualitative information compared to the unsupervised clustering algorithm.

Keywords: Unsupervised clustering; fractal dimension; clustering within clusters; cluster analysis.

1. Introduction

Data analysis, in general, aims at extracting knowledge in large collections of data. Clustering algorithms, in this context, are employed as means of discovering subsets (clusters) in the dataset such that objects belonging to one cluster are more similar to each other than objects in different clusters.

The first references to clustering date back to Aristotle and Theophrastos in the fourth century B.C. and Linnaeus [1736] in the 18th century. It was not until 1939, however, that one of the first comprehensive foundations of these methods was published [Tryon, 1939].

The application domain of clustering techniques is very wide including data mining [Fayyad *et al.*, 1996], text mining [Boley, 1998; Dhillon & Modha, 2001], statistical data analysis [Aldenderfer & Blashfield, 1984], compression and vector quantization [Ramasubramanian & Paliwal, 1992], global optimization [Becker & Lago, 1970; Törn & Zilinskas, 1989] and web personalization [Rigou *et al.*, 2004].

A plain examination of the objects that surround us reveals that most of them are very complex and erratic in nature [Grassberg & Procaccia, 1983; Pentland, 1984; Sarkar & Chaudhuri, 1992]. Although most man-made objects, can be described by primitive geometric objects such as cubes and cones, most objects in nature involve such high complexity that classical geometry fails to describe them. The need for a model that has the ability to describe such erratic behavior was first handled by

Mandelbrot [1983] who introduced the concept of "fractals". A set is called fractal if its Hausdorff–Besicovitch dimension is strictly greater than its topological dimension [Sarkar & Chaudhuri, 1992]. Fractal sets can be characterized by their *fractal dimension*, that measures their complexity. Mandelbrot first described an approach to calculate the fractal dimension while estimating the length of a coastline.

An established approach to compute the fractal dimension of a set is the box-counting method [Falconer, 1990; Liebovitch & Toth, 1989]. In detail, for a set of N points in $\mathbb{R}^d$, and a partition of the space in grid cells of length l, the fractal dimension D can be derived from:

$$D = -\lim_{l \to 0} \frac{\log_{10} N(l)}{\log_{10} l},$$

where $N(l)$ represents the number of cells occupied by at least one point [Falconer, 1990].

Clustering algorithms that employ the fractal dimension have been proposed in the past. Barbarä and Chen [2000], proposed a grid based clustering algorithm, that uses fractal dimension to cluster datasets. The algorithm, uses a heuristic based algorithm at the initialization stage to form the initial clusters and then it incrementally adds points to a cluster, as long as, the fractal dimension remains constant. Prasad *et al.* [2003], also proposed a fractal based clustering method for two dimensions. Both these algorithms are supervised, that is, they require from the user to provide an *a priori* estimation of the number of clusters present in the dataset.

In this contribution we extend the unsupervised k-windows clustering algorithm [Tasoulis & Vrahatis, 2004, 2005; Vrahatis *et al.*, 2002], by taking under consideration the qualitative information provided by the fractal dimension. The rest of the paper is organized as follows. In Sec. 2, the unsupervised k-windows algorithm is briefly described, and the proposed modification is presented. Section 3 is devoted to the presentation of the computational experiments. The paper ends with discussion and concluding remarks.

2. Unsupervised *k*-Windows Clustering Algorithm

The unsupervised k-windows algorithm uses windows to define clusters. Windows are defined as orthogonal ranges in a d-dimensional space. For a given set of points the algorithm initializes randomly a number of windows, of a user defined size, over the data. Next, it employs the procedures of *movement* and *enlargement*, to iteratively update the windows' size and center so as to capture each cluster in the dataset by one or more windows. The movement procedure updates a window's center, by setting it equal to the mean of the patterns currently included. At the enlargement stage, the size of the window is increased in order to capture as many patterns of the underlying cluster as possible. The enlargement procedure terminates when the number of patterns included in the window no longer increases. The two processes are exhibited in Fig. 1.

The final step of the algorithm, is the *merging procedure*. In this step, the resulting windows from the previous steps that are suspected to capture patterns that belong to a single cluster are considered for merging. In detail, for each pair of overlapping windows the number of points in their intersection is computed. If the ratio of this number to the number of points in each window is close to one, then one of the two windows is disregarded as the two windows are considered to be identical. If this ratio, is high the two windows are considered to capture parts of the same cluster. Finally, if this ratio is small the windows are considered to capture two different clusters. Thus by considering a sufficiently large number of initial windows, the algorithm is able to provide an approximation for the number of clusters present in the dataset. An example of this operation is exhibited in Fig. 2. In particular, in Fig. 2(a) windows W1 and W2 are considered to capture the same cluster, and W1 is deleted. On the other hand, in Fig. 2(b) windows W3 and W4 are considered to capture parts of the same cluster. Finally, in Fig. 2(c) windows W5 and W6, are considered to capture two different clusters.

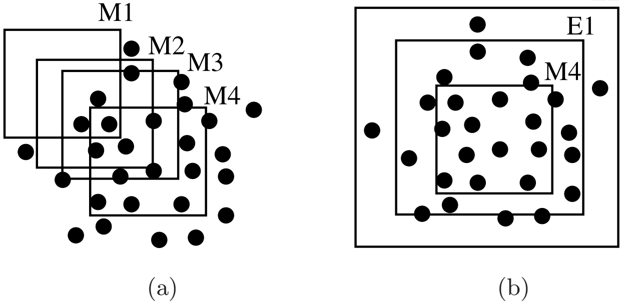

Fig. 1. (a) Sequential movements of the initial window M1 that result in the final window M4. (b) Sequential enlargements of the initial window M4 that result in the final window E2.

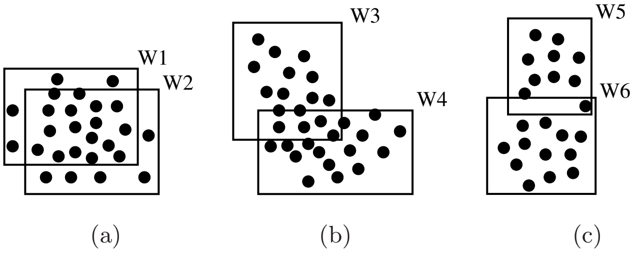

Fig. 2. The merging procedure.

The computational complexity of the algorithm depends on the complexity of determining the points that lie in a specific window. This is the well studied *orthogonal range search* problem [Preparata & Shamos, 1985]. Numerous Computational Geometry techniques have been proposed [Alevizos, 1998; Bentley & Maurer, 1980; Chazelle, 1986; Preparata & Shamos, 1985] to address this problem. All these techniques employ a preprocessing stage at which they construct a data structure that stores the patterns. This data structure allows them to answer range queries fast. For applications of very high dimensionality, data structures like the Multidimensional Binary Tree [Preparata & Shamos, 1985], and Bentley and Maurer [1980] seem more suitable. On the other hand, for low dimensional data with a large number of points the approach of Alevizos [1998] appears more attractive. The unsupervised k-windows algorithm has been successfully applied in numerous applications including bioinformatics [Tasoulis *et al.*, 2004a, 2004b], medical diagnosis [Magoulas *et al.*, 2004; Tasoulis *et al.*, 2003], time series prediction [Pavlidis *et al.*, 2003] and web personalization [Rigou *et al.*, 2004].

2.1. *Proposed modification*

In this paper we propose a modified version of the unsupervised k-windows clustering algorithm, that guides the procedures of movement, enlargement and merging using the estimation of the fractal dimension of the set of points included in a window.

In detail, the movement and enlargement of a window is considered valid only if the associated change of the fractal dimension is not significant. It is also possible to guide the merging procedure on the fractal dimension by allowing two windows to merge only if the estimated fractal dimensions are almost equal. Thus, the merging of windows that capture regions of a cluster with different fractal dimension is prevented. Such clusters appear in datasets where the density of points in the neighborhood of the cluster center is significantly higher than that of areas located further away from the center. Thus, the algorithm discovers the cluster center more efficiently and moreover, it identifies regions with qualitative differences within a single cluster. Consider for example the case exhibited in Fig. 3. The enlargement and movement procedures restrain window W3 from enclosing the right part of the cluster since the fractal dimension of this region is much higher [see Fig. 3(b)]. Similarly, window W4 is restrained from capturing the left part of the cluster. The proposed modification of the algorithm also recognizes that although the windows have many points in common [see Fig. 3(a)], the difference in the value of the fractal dimension between them is sufficiently large so as to consider them as two distinct regions of the same cluster.

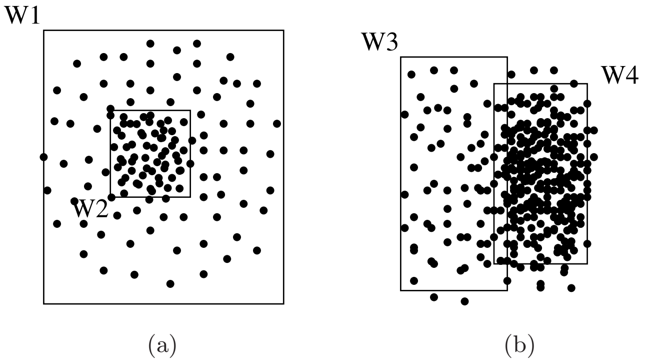

Fig. 3. Clusters with regions of different density. The proposed algorithm is able to discover the different sections of the same clusters.

3. Experimental Results

To illustrate the impact of the proposed modification on the workings of the algorithm we firstly employ an artificial two-dimensional dataset $Dset_1$ that contains 1200 points. The dataset, as well as, the outcomes of the original and the proposed algorithm are demonstrated in Fig. 4, where different colors correspond to different clusters. As it is evident both algorithms are able to discover the centers of the two main clusters. In contrast to the original algorithm, the proposed algorithm identifies the regions around the cluster centers that are characterized by higher density.

Next we consider the four-dimensional Iris dataset $Dset_{\text{iris}}$ from the UCI machine learning

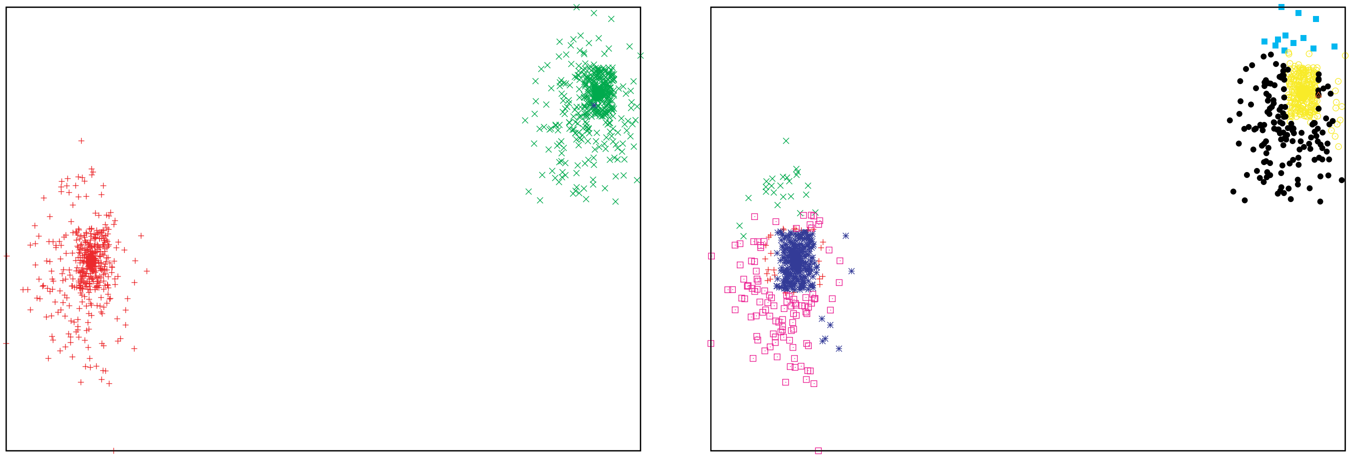

Fig. 4. *Dset*$_1$ with the results of the unsupervised k-windows algorithm (left) and the proposed version (right).

repository [Blake & Merz, 1998]. This dataset is among the best known databases to be found in the pattern recognition literature. It contains 150 records of four features. The features are measurements of the sepal and petal length and width of three different types of the iris plant (Setosa, Versicolour and Virginica). The 150 records are equally distributed in three classes, each corresponding to a different type of the plant.

The confusion matrices depicted in Table 1, report the number of members of each discovered cluster that correspond to each class. Ideally, each cluster should contain patterns that belong to a single type of the Iris plant. In terms of the confusion matrix this implies that all the nondiagonal entries are zero. Both algorithms recognize three clusters thus the physical meaning of the data is discovered by both. The only difference lies in the number of points that have different class and cluster labels, eight for the original unsupervised k-windows and six for the proposed modification.

The final benchmark problem considered, $Dset_{\mathrm{eq}}$, is a two-dimensional dataset of the longitudes and latitudes of the earthquakes with a magnitude greater than 4, in the Richter earthquake scale, that occurred in the period 1983 to 2003 in Greece. The dataset was obtained from the Institute of Geodynamics of the National Observatory of Athens [Catalogue, n.d.]. This dataset is employed not to obtain a further insight with respect to the earthquake phenomenon, but rather to study the applicability of the proposed algorithm to a real world dataset.

Figure 5 illustrates the results of the unsupervised k-windows algorithm applied on $Dset_{\mathrm{eq}}$, while Fig. 6 exhibits the results obtained by the modified algorithm. In both figures, different colors correspond to different clusters. As in the previous two cases, the modified algorithm separates regions characterized by different fractal dimensions, that were assigned to a single cluster by the original algorithm. In Fig. 6 characteristic examples of clusters that were separated by the modified algorithm are enclosed in black squares. Notice that the coastline of Greece appearing in these figures is not precise and is included to provide an approximate visualization of the relative positions of the earthquakes.

This is a preliminary investigation of the application of clustering algorithms to earthquake data. An exhaustive investigation requires the inclusion

Table 1. Confusion matrices for $Dset_{\mathrm{iris}}$.

	Unsupervised k-Windows			Proposed Version		
	Iris Class			Iris Class		
	Setosa	Versicolour	Virginica	Setosa	Versicolour	Virginica
Cluster 1	50	0	0	50	0	0
Cluster 2	0	46	4	0	46	4
Cluster 3	0	4	46	0	2	48

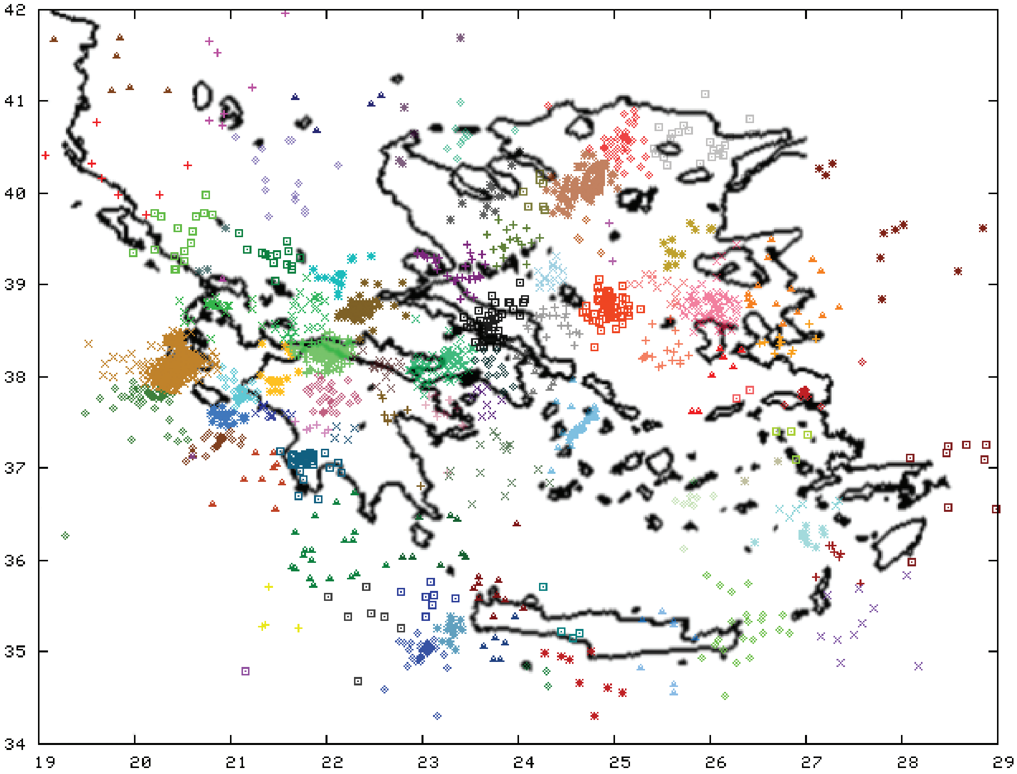

Fig. 5. $Dset_{\text{eq}}$ with the results of the unsupervised k-windows algorithm.

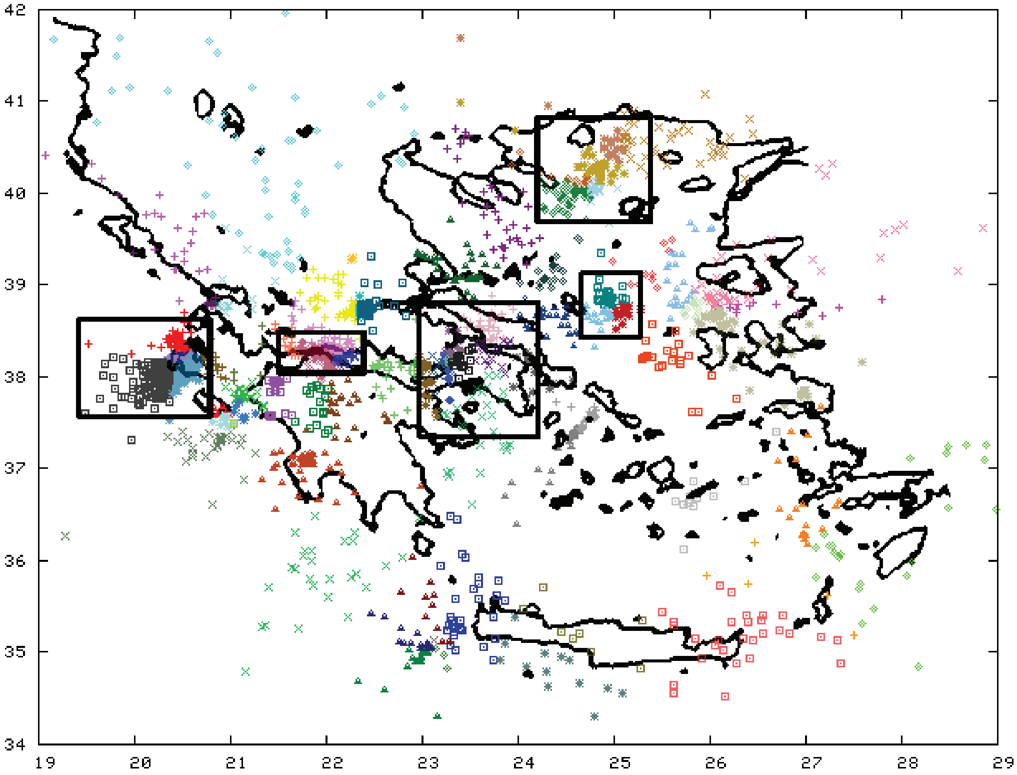

Fig. 6. $Dset_{\text{eq}}$ with the results of the proposed modification.

of additional parameters like magnitude, depth, and time.

4. Discussion and Concluding Remarks

This paper presents preliminary results of a new clustering algorithm that incorporates information about the fractal dimension of the underlying clusters in order to partition a d-dimensional dataset. To perform such *clustering within clusters* analysis it is important to use an unsupervised clustering algorithm. Unsupervised clustering algorithms have the desirable property that they do not require from the user to specify the number of clusters present in the dataset prior to their execution. As a first step, we considered the unsupervised k-windows algorithm. This algorithm uses a windowing technique to capture the clusters, and performs three procedures, movement, enlargement, and merging, to produce the final clustering result. The approximation of the fractal dimension, using the box-counting method, occurs at each step of the algorithm and provides additional, qualitative, information to determine the positioning and size of the windows, as well as, the final cluster number.

The proposed approach enables the identification of regions with different fractal dimension even within a single cluster. The design and development of algorithms that can detect clusters within clusters is particularly attractive in numerous applications where further qualitative information is valuable. Examples include time-series analysis, image analysis, medical applications, and signal processing.

Future directions will include the thorough investigation of this approach on other real life applications like the aforementioned ones.

Acknowledgments

The authors thank the Institute of Geodynamics, of the National Observatory of Athens for making the data set of the earthquake catalog available in the public domain.

References

Aldenderfer, M. S. & Blashfield, R. K. [1984] *Cluster Analysis*, Quantitative Applications in the Social Sciences, Vol. 44 (SAGE Publications, London).

Alevizos, P. [1998] "An algorithm for orthogonal range search in $d \geqslant 3$ dimensions," *Proc. 14th European Workshop on Computational Geometry*, Barcelona, Spain.

Barbarä, D. & Chen, P. [2000] "Using the fractal dimension to cluster datasets," *Knowledge Discovery in Databases* (ACM Press), pp. 260–264.

Becker, R. W. & Lago, G. V. [1970] "A global optimization algorithm," *Proc. 8th Allerton Conf. Circuits and Systems Theory*, pp. 3–12.

Bentley, J. L. & Maurer, H. A. [1980] "Efficient worst-case data structures for range searching," *Acta Inform.* **13**, 155–168.

Blake, C. L. & Merz, C. J. [1998] *UCI Repository of Machine Learning Databases.*

Boley, D. [1998] "Principal direction divisive partitioning," *Data Min. Knowl. Discov.* **2**, 325–344.

Catalogue, Earthquake, http://www.gein.noa.gr/services/cat.html, Institute of Geodynamics, National Observatory of Athens.

Chazelle, B. [1986] "Filtering search: A new approach to query-answering," *SIAM J. Comput.* **15**, 703–724.

Dhillon, I. S. & Modha, D. S. [2001] "Concept decompositions for large sparse text data using clustering," *Mach. Learn.* **42**, 143–175.

Falconer, K. [1990] *Fractal Geometry — Mathematical Foundations and Applications* (John Wiley & Sons, Chichester).

Fayyad, U. M., Piatetsky-Shapiro, G. & Smyth, P. [1996] *Advances in Knowledge Discovery and Data Mining* (MIT Press).

Grassberg, P. & Procaccia, I. [1983] "Characterization of strange attractors," *Phys. Rev. Lett.* **50**, 346–349.

Liebovitch, L. S. & Toth, T. [1989] "A fast algorithm to determine fractal dimensions by box counting," *Phys. Lett. A* **141**, 386–390.

Linnaeus, C. [1736] *Clavis Classium in Systemate Phytologorum in Bibliotheca Botanica*, (Biblioteca Botanica, Amsterdam, The Netherlands).

Magoulas, G. D., Plagianakos, V. P., Tasoulis, D. K. & Vrahatis, M. N. [2004] "Tumor detection in colonoscopy using the unsupervised k-windows clustering algorithm and neural networks," *Fourth European Symp.* "*Biomedical Engineering.*"

Mandelbrot, B. B. [1983] *The Fractal Geometry of Nature* (Freeman, NY).

Pavlidis, N. G., Tasoulis, D. K. & Vrahatis, M. N. [2003] "Financial forecasting through unsupervised clustering and evolutionary trained neural networks," *Congress on Evolutionary Computation*, pp. 2314–2321.

Pentland, A. P. [1984] "Fractal-based description of natural scenes," *IEEE Trans. Patt. Anal. Mach. Intell.* **6**, 661–674.

Prasad, M. G. P., Dube, S. & Sridharan, K. [2003] "An efficient fractals-based algorithm for clustering," *IEEE Region 10 Conf. Convergent Technologies for The Asia-Pacific.*

Preparata, F. & Shamos, M. [1985] *Computational Geometry* (Springer Verlag, NY, Berlin).

Ramasubramanian, V. & Paliwal, K. [1992] "Fast k-dimensional tree algorithms for nearest neighbor search with application to vector quantization encoding," *IEEE Trans. Sign. Process.* **40**, 518–531.

Rigou, M., Sirmakessis, S. & Tsakalidis, A. [2004] "A computational geometry approach to web personalization," *IEEE Int. Conf. E-Commerce Technology* (*CEC'04*), pp. 377–380.

Sarkar, N. & Chaudhuri, B. B. [1992] "An efficient approach to estimate fractal dimension of textural images," *Patt. Recogn.* **25**, 1035–1041.

Tasoulis, D. K., Vladutu, L., Plagianakos, V. P., Bezerianos, A. & Vrahatis, M. N. [2003] "On-line neural network training for automatic ischemia episode detection," *Lecture Notes in Computer Science* **2070**, 1062–1068.

Tasoulis, D. K. & Vrahatis, M. N. [2004] "Unsupervised distributed clustering," *IASTED Int. Conf. Parallel and Distributed Computing and Networks* (Innsbruck, Austria), pp. 347–351.

Tasoulis, D. K., Plagianakos, V. P. & Vrahatis, M. N. [2004a] "Unsupervised cluster analysis in bioinformatics," *Fourth European Symp.* "*Biomedical Engineering*".

Tasoulis, D. K., Plagianakos, V. P. & Vrahatis, M. N. [2004b] "Unsupervised clustering of bioinformatics data," *European Symp. Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems*, *Eunite*, pp. 47–53.

Tasoulis, D. K. & Vrahatis, M. N. [2005] "Unsupervised clustering on dynamic databases," *Patt. Recogn. Lett.* **26**, 2116–2127.

Törn, A. & Žilinskas, A. [1989] *Global Optimization* (Springer-Verlag, Berlin).

Tryon, C. [1939] *Cluster Analysis* (Edward Brothers, Ann Arbor, MI).

Vrahatis, M. N., Boutsinas, B., Alevizos, P. & Pavlides, G. [2002] "The new k-windows algorithm for improving the k-means clustering algorithm," *J. Compl.* **18**, 375–391.