Granular Clustering with partial supervision – the development of abstract models for simulation

*Andrzej Bargiela and **Witold Pedrycz

*Department of Computing

The Nottingham Trent University, Nottingham NG1 4BU
United Kingdom
(andre@doc.ntu.ac.uk)

**Department of Electrical & Computer Engineering

University of Alberta, Edmonton, Canada
(pedrycz@ee.ualberta.ca)

and Systems Research Institute, Polish Academy of Sciences
01-447 Warsaw, Poland

KEYWORDS
Abstract models for simulation, information granules, information abstraction, granular clustering, cluster compatibility measure.

ABSTRACT

The paper is concerned with a problem of explicit granulation of data in presence of some labeled patterns [17, 22]. The process of granulation of data is a specific demonstration of a more general activity that of deriving abstract data models which underpin any computer simulation of real-life systems. It is demonstrated here that the success of simulations depends to a large extent on the adoption of an appropriate level of abstraction for a given problem. The paper proposes an approach to combining prior knowledge with the exploratory unsupervised clustering to arrive at more representative simulation models.

The granulation process is realized as an organic growth of multi-dimensional hyperboxes guided by the compatibility measure. The organic growth signifies here that there are no prior assumptions about the number and shape of information granules. Instead, only the relative position and size of patterns in the pattern space determine the progression of the granulation process. The rationale for a specific form of the compatibility measure is explained using some illustrative examples. The inclusion of a small number of labeled patterns in the input data is shown to provide a very effective way of coping with complex decision hyperplanes in multi-dimensional pattern spaces. The method is illustrated using several synthetic data sets as well as the Iris data set that is widely regarded as a reference for the comparison of classification and clustering algorithms.

INTRODUCTION

Granular clustering, similarly to general clustering, is concerned with organizing and revealing structures in data. However, the emphasis of granular clustering is on abstracting the essential features of data so that the emerging information granules may be used as a blueprint for the description of systems at the higher level of abstraction [1, 9, 20, 21, 22].

Most experimental data, available in a raw form, are numeric. Granulation of information happens through a process of data organization and data comprehension. Interestingly; humans granulate information almost in a subconscious manner. This eventually makes the ensuing cognitive processes so effective and far superior over processes occurring under the auspices of machine intelligence. Two representative categories of problems in which information granulation emerges in a profound way involve processing of one and two-dimensional signals. Incidentally, these two categories correspond to two major ways in which humans perceive information about the environment; aural and visual. In the first case, we are concerned primarily with temporal signals. The latter case pertains to image processing and image analysis. In signal processing, granulation arise as a result of temporal sampling and aggregation. Several samples in the same time window can be represented as an information granule. In the simplest case, such interval can be formed by taking a minimal and maximal value of the signal occurring in this window of granulation. Some other ways of forming information granules may rely on statistical analysis: one determines a mean or median as a representative of the numeric data points and then build a confidence interval around it (obviously, the use of this mechanism requires assumptions about the statistical properties of the population contained in the window as well as the numeric representative under discussion). Similarly, in image processing one combines pixels exhibiting some spatial neighborhood. Again, various features of an image can be granulated, say brightness, texture, RGB, etc.

The input data (patterns) and the information granules discussed in this paper are represented as hyperboxes in a multi-dimensional pattern space. The mathematical formalism of the interval analysis provides a robust framework for the analysis of the granular structures that emerge in the process of clustering.

UNSUPERVISED CLUSTERING

Before we proceed with the details of partially supervised granular clustering, it is instructive to provide a qualitative description of the unsupervised clustering process. The full details are provided in a separate publication [3].

The granular clustering is carried out as the following iterative process:

1. Identify a pair of information granules for which the *compatibility measure* (defined below) attains a maximum and build a new granule that includes both of the identified granules; this means that a number of granules is reduced by one in each iteration step.
2. Evaluate the termination criterion to assess whether the process of condensing the original data does not distort its essential features.

Figure 1 illustrates how the clustering works. We start from a collection of patterns, which can be both data points and hyperboxes in a multi-dimensional space, and grow progressively larger information granules. It is clear that, up to the point of forming three information granules, the granulation process preserves the essential characteristics of data (grouping of data in three separate areas of the pattern space). Granulation beyond this stage (forming two or one granule) is counterproductive since the essence of data is being lost. An important feature of our unsupervised clustering technique is that the termination criterion is implicit in the *compatibility measure* itself.

It is important to note that while the above approach resembles techniques of aggregative hierarchical clustering there is a striking difference between the two approaches. In hierarchical clustering we deal with numeric objects and the clusters are sets of the same objects. No conceptually new entities are formed. By contrast, here we "grow" clusters that change "shape" (defined as a ratio of sizes measured along individual coordinates). From iteration to iteration the hyperboxes are evolved to capture the information about the spatial distribution of patterns.

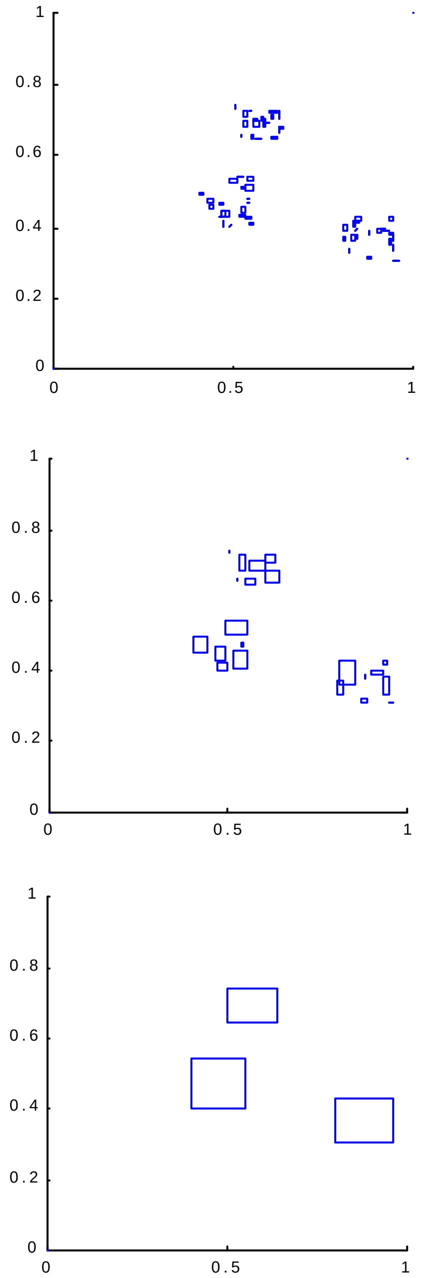


Figure 1. Snapshots of the granulation process; initial sets of patterns (which are themselves represented as hyperboxes) are grouped into coherent regions and eventually into 3 large information granules.

The above technique is also significantly different from the standard min-max clustering discussed by Simpson [18, 19]. First, Simpson's method deals with point-size data while we consider data that is represented, in general, as hyperboxes in the pattern space. Second, the compatibility measure adopted in our algorithm has a very different character to the fuzzy membership function used by Simpson.

The hyperbox *compatibility measure* is introduced here by considering two information granules (hyperboxes) A and B. More explicitly we can express the granules as A($\mathbf{l}_a$, $\mathbf{u}_a$) and B($\mathbf{l}_b$, $\mathbf{u}_b$) to point the location of the minimum and maximum vertices of the hyperboxes in the space. The expression of compatibility, compat(A, B) involves two components that is a distance between A and B, d(A,B), and a size of a newly formed information granule that comes when merging A and B. The distance d(A,B) between A and B is defined on a basis of the distance between its extreme vertices, that is

$$d(A, B) = (\|\mathbf{l}_b-\mathbf{l}_a\| + \|\mathbf{u}_b-\mathbf{u}_a\|)/2 \quad (1)$$

$\|.\|$ is a distance defined between the two numeric vectors. To make the framework general enough, we treat $\|.\|$ as an L_p distance , p>1. By chaging the value of “p” we sweep across a spectrum of well known distances that depend upon a particular value of “p”. For instance, p = 1 yields a Hamming distance, L_1. The value p = 2 produces a well – known Euclidean distance, L_2. For p =∞ we refer to a Tchebyschev distance, L_∞ .

Once A and B have been combined giving rise to a new information granule C, its granularity can be captured by a volume, $V(C)$ computed in a standard way

$$V(C) = \prod_{i=1}^{n} length_i(C) \quad (2)$$

where

$$length_i(C) = \max(u_{\mathbf{b}}(i), u_{\mathbf{a}}(i)) - \min(l_{\mathbf{b}}(i), l_{\mathbf{a}}(i)) \quad (3)$$

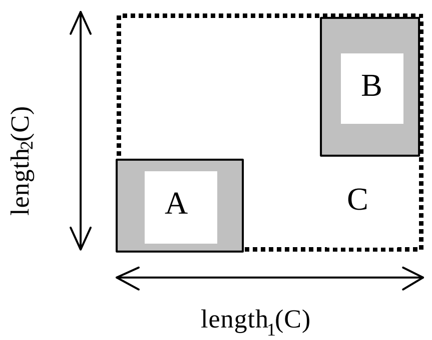


Figure 2. Information granule C as a result of combining A and B

The two expressions (1)-(2) are the contributing factors to the compatibility measure, compat(A, B) to be defined now in the form

$$\text{compat}(A, B) = 1 - d(A,B)e^{-\alpha V(C)} \quad (4)$$

The rationale behind the above form of the compatibility measure is as follows. The candidate granules to be clustered should not only be “close” enough (which is reflected by the distance component) but the resulting granule should be “compact” (meaning that the size of the granule in every dimension is approximately equal). The second requirement favors such A and B that give rise to a maximum volume for a given d(A, B). The particular exponential form of this expression has to do with the normalization criterion so that all values are kept in the unit interval. In particular, the volume of a point produces $e^{-0} = 1$ While the volume increases, its exponential function goes down to zero. To retain the values of the compatibility measure to the unit interval, all data is normalized to the unit hypercube $[0,1]^n \subset \mathbf{R}^n$ The parameter α balances the two concerns in the compatibility measure and is chosen so as to control an extent to which the volume impacts the compatibility measure.

The compactness factor ($e^{-\alpha V(C)}$) introduced in the compatibility measure is critical to the granular clustering. By contrast, it is not essential and would not play any role if we proceeded in a standard way and did not attempt to develop granules but retained a cluster of numeric data.

As the clustering proceeds (refer to Figure 1) the process of merging the progressively less closely associated patterns finds its reflection in the gradual reduction of the compatibility measure (4). A typical plot of the evolution of the compatibility measure over the complete clustering cycle is shown in Figure 3.

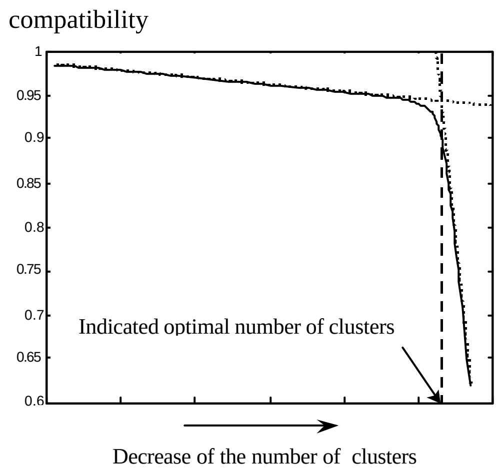


Figure 3. An example of the evolution of the compatibiliy measure over the full cycle of the clustering process.

It is self evident that the proximity of patterns that are being merged into granules at the early stages of the clustering process, is reflected in the relatively small gradient of the compatibility measure curve. By contrast, a large gradient of the curve, at the final stages of the clustering, indicates merging of incompatible clusters. The compatibility measure curve provides therefore a convenient reference for identifying which number of clusters captures the essential characteristics of the input data while providing the best generalization of them. The intersection of the two gradient lines (as indicated in Figure 3) can be used as an approximation to the optimal number of clusters.

PARTIALLY SUPERVISED CLUSTERING

The main idea behind the partially supervised clustering is to make active use of the available classification information in the process of maximization of the compatibility measure. To distinguish between labeled and unlabelled patterns we introduce a two-valued (Boolean) pattern label vector, $\mathbf{p}=[p_k]$, k=1, 2, …N with 0-1 entries

$p_k = 1$, if pattern is labeled
$p_k = 0$, otherwise

In a typical case of partially supervised clustering the number of labeled patterns is significantly lower then the total number of patterns e.i. $\sum_k p_k << N$.

The cluster-membership values of the labeled patterns are arranged in a vector $L=[l_k]$, k=1, 2, …N, with $l_k \in \{0,1, .. ,c\}$; c represents the highest cluster label. The maximization of the compatibility measure on partially labeled patterns (p-compat(A,B)) can then proceed as follows:

1. For a given pattern A, identify pattern B for which the compatibility measure compat(A,B) attains maximimum.

2. Assess the cluster labels for the identified patterns as follows:
   - 2a. if $\neg (p_A \wedge p_B)$ then p-compat(A,B) = compat(A,B)
   - 2b. if $(p_A \wedge p_B \wedge (l_A = l_B))$ then p-compat(A,B) = compat(A,B)
   - 2c. if $(p_A \wedge p_B \wedge (l_A \neq l_B))$ then p-compat(A,B) = 0

3. If p-compat(A,B) > 0 then merge patterns A and B into granule C assigning label $l_C = \max(0, l_A, l_B)$ to the new granule.

4. Select next pattern A and return to step 1 or terminate if all patterns have been considered.

The above can be articulated as follows. Patterns (granules) that are not labeled (case 2a) or patterns (granules) that are labeled already with the same cluster label (case 2b) are considered to be eligible for granulation. Patterns (granules) that are labeled with different cluster labels (case 2c) return the value of the compatibility measure p-compat=0 and are therefore left as separate entities. Once a pattern (granule) has been found to be the nearest neighbour of the pattern (granule) labeled with different cluster label, it does not need to be considered any more in the granulation process since any attempt to merge it with other patterns would potentially increase overlap between clusters.

To illustrate the outcome of the partially supervised clustering Table 1 lists a subset of cases that are relevant for 3 patterns and 3 iterative steps of the algorithm. The full set of cases can be obtained by performing twice circular substitutions (2->1 and 3->2 and 1->3).

Table 1.
Cluster labels for 3 patterns processed by the partially supervised clustering algorithm

Step 1	Step 2	Step 3
0 , 0 , 0	0 , 0	0
0 , 0 , 1	0 , 1	1
0 , 1 , 0	1 , 0	1
0 , 1 , 1	1 , 1	1
1 , 0 , 0	1 , 0	1
1 , 0 , 1	1 , 1	1
1 , 0 , 2	1 , 2	1 , 2
1 , 1 , 0	1 , 0	1
1 , 1 , 1	1 , 1	1
1 , 1 , 2	1 , 2	1 , 2
1 , 1 , 3	1 , 3	1 , 3
1 , 2 , 0	1 , 2 , 0	1 , 2
1 , 2 , 1	1 , 2 , 1	1 , 2 , 1
1 , 2 , 2	1 , 2 , 2	1 , 2
1 , 2 , 3	1 , 2 , 3	1 , 2 , 3
1 , 3 , 0	1 , 3 , 0	1 , 3
1 , 3 , 1	1 , 3 , 1	1 , 3 , 1
1 , 3 , 2	1 , 3 , 2	1 , 3 , 2
1 , 3 , 3	1 , 3 , 3	1 , 3

The fact that the partially supervised clustering does not necessarily result in merging two patterns (granules) in each iterative step of the algorithm implies that there may be a significant number of granules left at the end of the process. A typical evolution of the compatibility measure is presented in Figure 4. It is clear that rather than identifying the number of clusters (by tracing asymptotic gradients of the curve) we identify here a number of granules for which the compatibility measure saturates before reaching the flection point. In other words, we find a set of granules, which cannot be reduced further without causing a deliberate overlap of clusters.

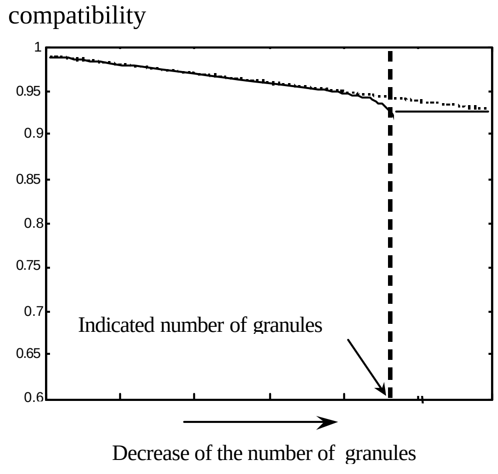


Figure 4. An example of the evolution of the compatibiliy measure in a partially supervised granular clustering.

SIMULATION STUDIES

This section summarises several numerical experiments involving synthetic data sets and the well-known IRIS data set. The patterns were clustered using partially supervised granular clustering algorithm under the control of the compatibility measure. The proportion of the labeled patterns was varied from 10% to 20% so that their relative importance of labeled patterns could be assessed.

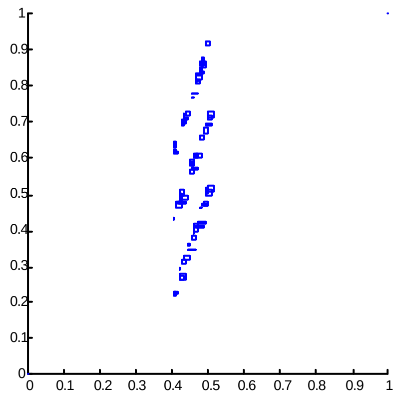


Figure 5. Synthetic 2-dimensional data.

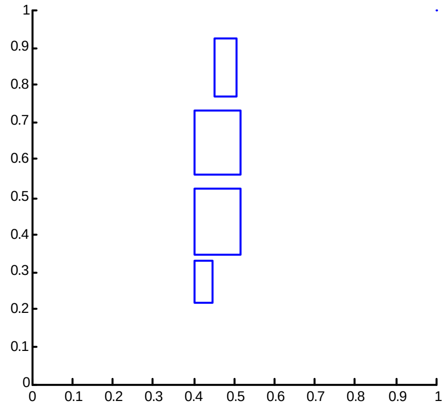


Figure 6. Unsupervised clustering results in identifying 4 clusters in the data.

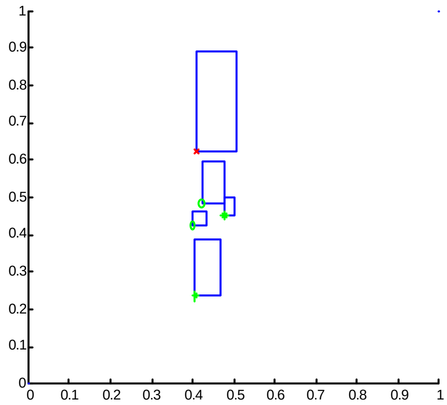


Figure 7. Partially supervised granular clustering with 10% of labeled data.

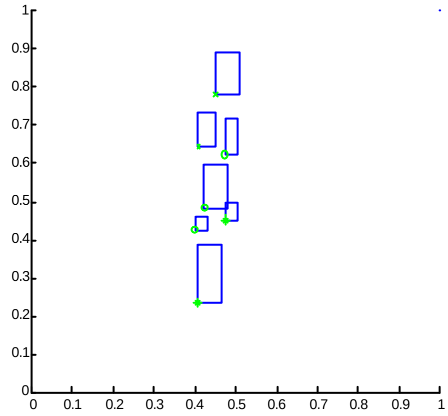


Figure 8. Partially supervised granular clustering with 20% of labeled data.

The results show that the performance of the partially supervised clustering is much superior to the unsupervised clustering counterpart. While the misclassification of patterns (the assignment of patterns from separate clusters to the same hyperbox) is over 60% in the case of unsupervised clustering, the inclusion of 10% of labeled data reduces the misclassification rate to approx. 30%. This is further reduced to less then 8% when the proportion of labeled data is increased to 20%. For 30% or more of labeled patterns the misclassification rate reduces to zero.

The proportion of labeled data that is needed to avoid misclassifications is clearly dependent on the topology of the clusters. The second numerical experiment is intended to explore this dependency by investigating a more complex topology of clusters. The two-dimensional set of patterns shown in Figure 9 consists of four overlapping clusters. Two of the clusters are ellipsoidal while the remaining two form a diagonal cross-pattern resembling the standard EX-OR problem. Overall, the diversity of forms of the clusters along with their distribution makes the problem quite challenging for unsupervised learning. The effect of partial supervision, in the context of this system, is quite remarkable significantly improving the outcomes of clustering. Obviously, to take full advantage of granular clustering with partial supervision one should ensure that the labeled patterns are representative of the respective clusters. Using all labeled patterns from the same class is not very instrumental to the improvement of the overall classification results. In a series of experiments we have attempted to quantify this effect. The percentage of labeled patterns was varied from 5% through to 20% with a 5% increment. The labeled patterns were selected on a random basis and, to provide a degree if independence from the characteristics of this random process, the selection was repeated 20 times for each percentage of the labeled patterns. The results showed significant consistency and are meaningfully reproducible.

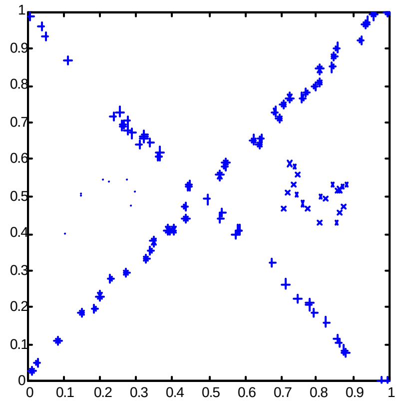


Figure 9. Synthetic 4-cluster data set

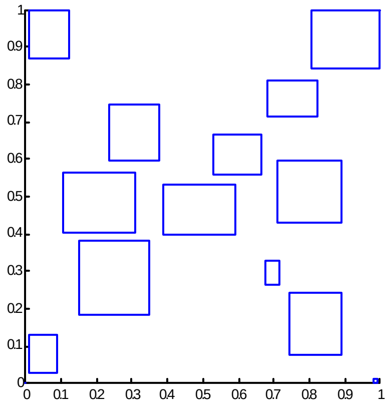
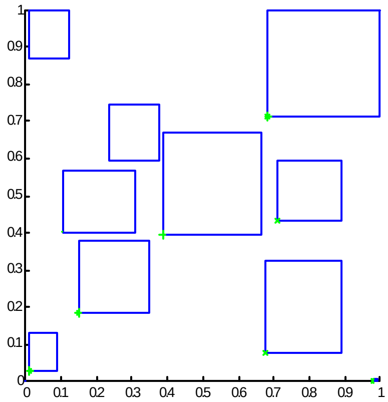
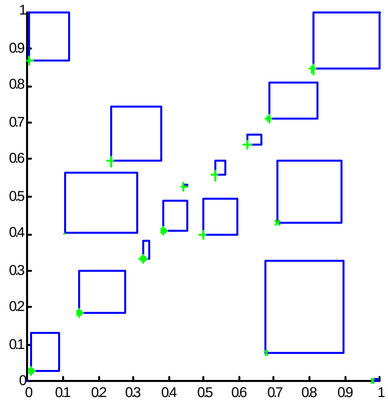
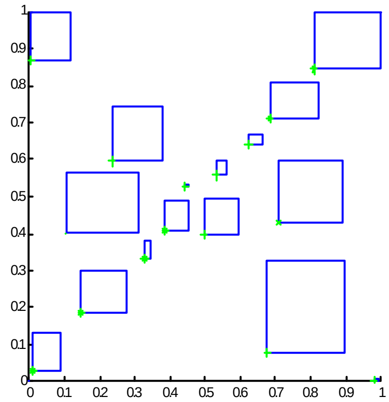


Figure 10. Unsupervised clustering







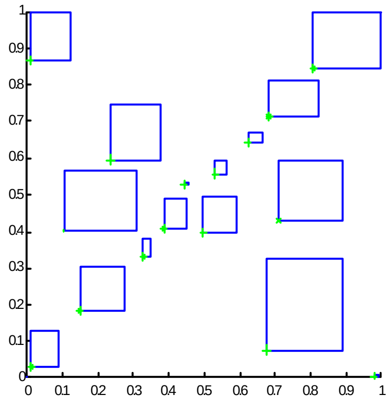


Figure 11. Partially supervised clustering with 5%; 10%; 15% and 20% of labeled patterns.

By cross-referencing Figures 9, 10 and 11 it is possible to see that even with 10% of labeled patterns the granular clustering method manages to avoid misclassifications. As could be expected, the misclassification in the area of intersection of two diagonal clusters, does occur for small proportion of labeled patterns. However, the experiments show that this is easily resolved by increasing the proportion of labeled patterns to 10% or more. Such an increase resulted in a 0% misclassification rate. Overall the method seems to have delivered highly satisfactory results.

We apply now the method to, a well known and widely regarded as the reference data set for pattern recognition, the IRIS data set. It contains three categories of species of Iris such as Iris Setosa, (C1) Iris Versicolor (C2) and Iris Virginica (C3) represented in two-dimensional space of petal length and width. The objects have been clustered using unsupervised and partially supervised clustering algorithm.

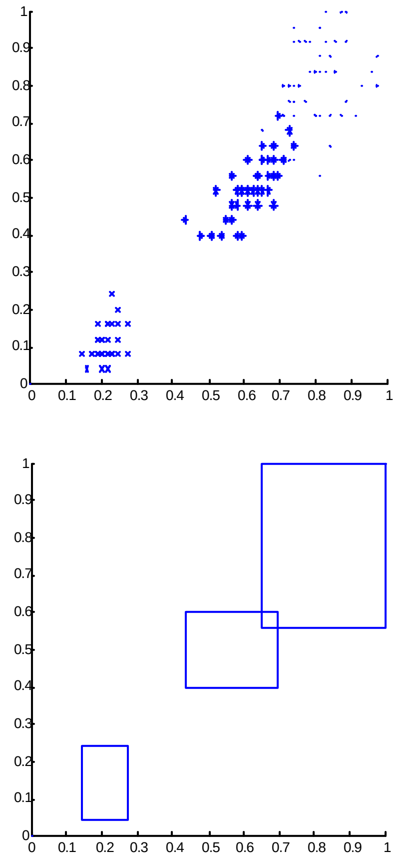


Figure 12. Unsupervised clustering

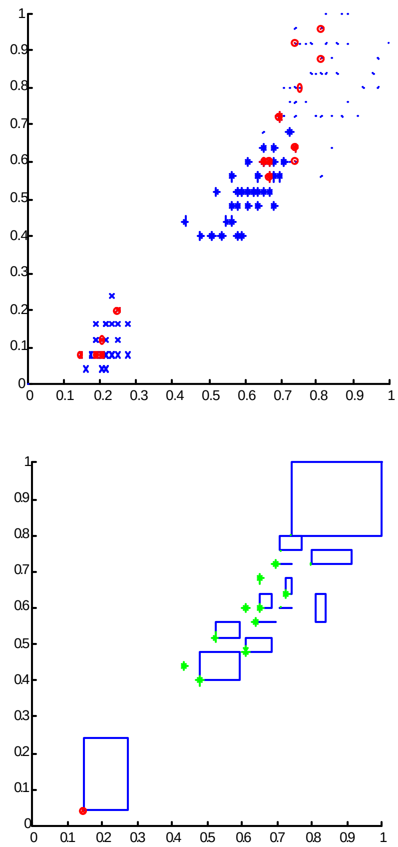


Figure 13. Partially supervised clustering (10% of labeled patterns)

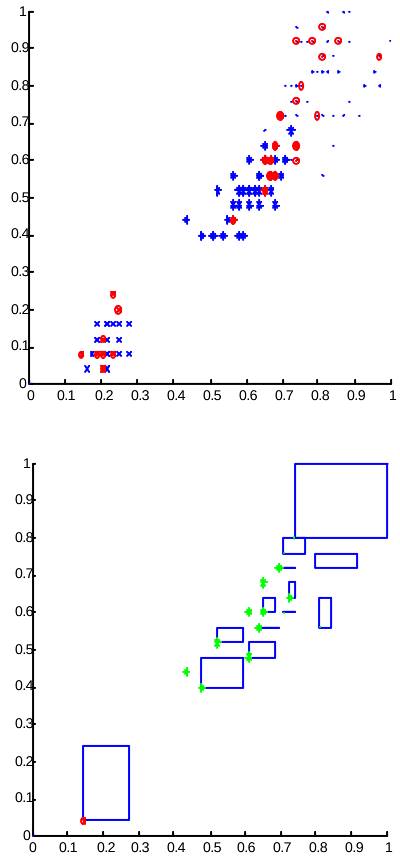


Figure 14. Partially supervised clustering (20 % of labeled patterns)

The unsupervised clustering algorithm identifies correctly that there are three types of patterns but the overlap between the clusters is not being resolved. This produces a large misclassification rate. By contrast, partially supervised clustering manages to adjust the sizes of hyperboxes so that in the area of cluster overlap there are smaller hyperboxes covering the pattern space. This is a very desirable feature because it enables formation of clusters that have complex topology using a smaller number of hyperboxes (granules).

CONCLUSION

The paper has presented a method of partially supervised granular clustering as a constructive approach to capturing the essence of large collection of numeric data. The approach emphasizes the value of explicit information granulation under the control of compatibility measure. It has been shown that while the topology of some clusters is well approximated by a single hyperbox, in general, there is a need for a collection of hyperboxes to describe a cluster. Using partially supervised granular clustering approach it has been shown that it is possible to achieve a good compromise between the accuracy of cluster representation (absence of misclassifications) and the number of hyperboxes (information granules) forming a cluster. This lays foundation for the development of simulation models that are capturing the essence of real-life systems without being distracted by the unnecessary detail.

It should be stressed that the proposed approach to data analysis is noninvasive meaning that we have not attempted to formulate specific assumptions about the distribution of the data but rather allow data to demonstrate its essential features. The granulation mechanism puts the features existing in the problem in a new perspective. They can be regarded as the composite of the basic characteristics of hyperboxes (granules) such as position, size, shape and number.

REFERENCES

1. Bargiela, A., (2001) *Interval and ellipsoidal uncertainty models,* In: W. Pedrycz (ed.) Granular Computing, Springer Verlag, to appear.

2. Bargiela, A., Pedrycz, W., *Informarion granules: Aggregation and interpretation issues*, submitted to IEEE Trans. on Syst. Man and Cybernetics.

3. Bargiela, A., Pedrycz, W., *From numbers to information granules: A study of unsupervised learning and feature analysis*, In: H. Bunke, A. Kandel (eds.) Hybrid Methods in Pattern Recognition, World Scientific, to appear.

4. Bezdek, J.C., (1981), *Pattern recognition with fuzzy objective function algorithms,* New York, Plenum.

5. Dubois, D., Jaulent, M.C., *Some techniques for extracting fuzzy regions*, Proc. 1st IFSA Congr., July 1985.

6. Gabrys, B., Bargiela, A., (2000) *General fuzzy Min-Max neural network for clustering and classification*, IEEE Trans. on Neural Networks, vol. 11, 3, pp. 769-783.

7. Hathaway, R.J., Bezdek J.C., (1988), *Recent convergence results for the fuzzy C-means clustering algorithms*, J. Classification, vol. 5, pp.237-247.

8. Hirota, K., Pedrycz, W., (1986), *Subjective entropy of probabilistic sets and fuzzy cluster analysis,* IEEE Trans. Syst. Man Cybernetics, 1, pp. 173-179.

9. Jain, A., Dubes, R., (1988), *Algorithms for clustering data*, Englewood Cliffs, NJ: Prentice Hall.

10. Kandel, A., (1986), *Fuzzy mathematical techniques with applications*, Addison-Wesley, Reading, MA

11. Pedrycz, W., (1985), *Algorithms of fuzzy clustering with partial supervision*, Pattern Recognition Lett., vol. 3, pp. 13-20.

12. Pedrycz, W., Waletzky, J., (1997) *Fuzzy clustering with partial supervision*, IEEE Trans. SMC, vol. 27, 5, pp. 787-795.

13. Pedrycz, W., Bargiela, A., *Granular clustering: a granular signature of data*, submitted to IEEE Trans. on Fuzzy Systems.

14. Pedrycz, W., (1997), *Computational Intelligence: An Introduction*, CRC Press, Boca Raton, FL.

15. Pedrycz, W., Gomide, F., *An Introduction to Fuzzy Sets*, Cambridge, MIT Press, Cambridge, MA, 1998.

16. Pedrycz, W., Fuzzy equalization in the construction of fuzzy sets, *Fuzzy Sets and Systems*, to appear.

17. Pedrycz, W., Smith, M.H., Bargiela, A (2000), *Granular signature of data*, Proc. 19th Int. (IEEE) Conf. NAFIPS'2000, Atlanta, July 2000, pp. 69-73.

18. Simpson, P.K., (1992), *Fuzzy Min-Max neural networks – Part1: Classification*, IEEE Trans. on Neural Networks, Vol 3, No. 5, pp. 776-786.

19. P. K. Simpson, (1993) *Fuzzy Min-Max neural networks – Part2: Clustering*, IEEE Trans. on Neural Networks, Vol 4, No. 1, pp. 32-45.

20. Zadeh, L.A., (1979), *Fuzzy sets and information granularity*, In: M.M. Gupta, R.K. Ragade, R.R. Yager, eds., Advances in Fuzzy Set Theory and Applications, North Holland, Amsterdam, 3-18.

21. Zadeh, L.A., (1996), *Fuzzy logic = Computing with words*, IEEE Trans. on Fuzzy Systems, vol. 4, 2, 103-111.

22. Zadeh, L.A., (1997), *Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic*, Fuzzy Sets and Systems, 90, 1997, pp. 111-117.