FROM NUMBERS TO INFORMATION GRANULES: A STUDY IN UNSUPERVISED LEARNING AND FEATURE ANALYSIS

ANDRZEJ BARGIELA
The Nottingham Trent University, Burton Street, Nottingham NG1 4BU, England
E-mail: andre@doc.ntu.ac.uk

WITOLD PEDRYCZ
University of Alberta, Edmonton, Canada,T6G 2G7
E-mail: pedrycz@ee.ualberta.ca

and

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

This chapter focuses on granular clustering: a way of finding structure in heterogeneous data and representing the data in the form of information granules. The main features of the proposed granular clustering approach are: (a) a noninvasive exploration of data carried out under weak assumptions made as to the nature of the data, (b) transparency of the constructed information granules which assume the form of hyperboxes in the problem space. We introduce a compatibility measure that expresses a degree of "similarity" between two information granules and takes into account both a distance between the granules and their size. We show how to "grow" clusters through a process of merging existing data points that exhibit high values of the compatibility measure. The clustering algorithm is discussed along with a comprehensive validation mechanism for the resulting structures (collections of information granules). We formulate a problem of feature analysis in the setting of the information granules and introduce some quantitative measures describing each feature. Numerical experiments use two-dimensional synthetic data and the multivariable Boston housing data.

1 Introductory comments

Combining patterns into some form of structure is the fundamental underpinning of Pattern Recognition (PR). Any in-depth analysis of patterns leads to optimal and interpretable classifiers. Interestingly, with the increasing heterogeneity of available data (patterns) and the steadily growing complexity of real-life classification problems, one has to look at a uniform and general treatment of various PR scenarios. In one way or another, a need arises for the formulation of classification problems in the language of information granules – conceptual entities that capture the essence of the overall data set while retaining the character of individual patterns. It is worth stressing that information granulation can be seen as a vehicle of abstraction supporting transition from clouds of numeric data and small information granules to larger and more general information granules [2, 3, 5, 12, 13, 16, 17, 18].

The area of clustering (unsupervised learning) with its long history has been an important endeavor in PR. Various algorithms for finding structures in data and representing the essence of such structures in terms of prototypes, dendrograms, self-organizing maps [8, 9] and alike [1, 4] have been used for a long time within the research community and industry. Commonly, if not exclusively, the direct aspect of granulation has not been tackled. The intent of this study is to address this important problem by introducing an idea of granular clustering. The simplest scenario looks like this: we start from a collection of numeric data (points in $\mathbf{R}^n$) and form information granules whose distribution and size reflects the essence of the data and reveals its structure. Forming the clusters (information granules) may be treated as a process of *growing* information granules. As the clustering progresses, we expand the clusters, enhancing the descriptive power of the granules while gradually reducing the amount of detail available to us. The information granules of interest in this study are represented as hyperboxes positioned in a high dimensional data space. The mathematical formalism of interval analysis provides a robust framework for the analysis of the information density of these granular structures. The study's intuitive objective is to match the granularity of data items used to describe physical systems to the structure of these systems. In this sense, the granulation process is attempting to achieve the highest possible generalization while maintaining the character of individual data structures.

Hybrid pattern classifiers are used in the context of this study differently to what is commonly encountered in the literature. While most hybrid systems allude to some sort of neurofuzzy architectures sometimes augmented by evolutionary mechanisms, in this study we are concerned with hybridization occurring at the level of information granules. In other words, a hybrid system operates on a spectrum of information granules ranging from numeric data through to more general (less specific) information granules.

The chapter is organized into 7 sections. In Section 2, we introduce information granules and the rationale behind them. As we are concerned with information granulation carried out in terms of sets (hyperboxes), we also provide all pertinent sets notation in this section. The principle of granular clustering is covered in Section 3 and the granular clustering algorithm is presented in Section 4. Feature analysis completed in the framework of information granules is studied in Section 5. Experimental studies are discussed within Section 6 and the conclusions are given in Section 7.

2 Information granules and information granulation

Information granulation is a process of data organization and data comprehension. Interestingly, humans granulate information almost in a subconscious manner. This makes the ensuing cognitive processes so effective and far superior to machine intelligence. Two representative categories of problems in which information granulation plays a prominent role involve processing of one and two-dimensional

signals. The first case is primarily temporal signals. The latter case pertains to image processing and image analysis. In processing, analysis and interpretation of signals, information granules arise as a result of temporal sampling and aggregation. Several samples in the same time window can be represented as a single information granule. In the simplest case, such an interval can be formed by taking the minimal and maximal value of the signal occurring in this window of granulation (refer to Figure 1). 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 builds 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 itself). Similarly, in image processing one combines pixels within some spatial neighborhood. Again, various features of an image can be granulated, such as brightness, texture, color, etc.

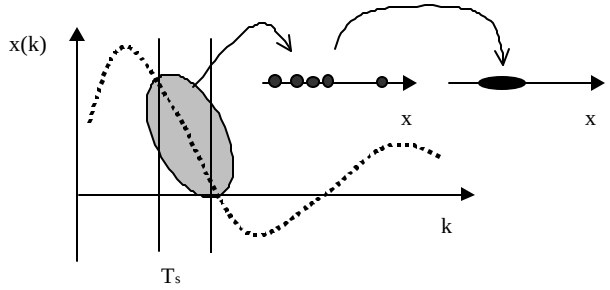


Figure 1. A fragment of a time series and its granulation through sampling (T_s denotes a sampling interval)

Information granulation has been studied in [2, 3, 10, 12], in terms of the concept itself, its computational aspects and the resulting structures. While this chapter is written as a self-contained unit the reader may be interested in a broader discussion of the information granulation issues that can be found in the above publications.

2.1 Set-based framework of information granules: the language of hyperboxes

In the overall presentation we adhere to a standard notation. A hyperbox defined in $\mathbf{R}^n$ denoted by B is fully described by its lower ($\mathbf{l}^\mathbf{B}$) and upper corner ($\mathbf{u}^\mathbf{B}$). To use explicit notation, we use B($\mathbf{l}^\mathbf{B}$, $\mathbf{u}^\mathbf{B}$) where $\mathbf{l}^\mathbf{B},\mathbf{u}^\mathbf{B} \in \mathbf{R}^n$. An evident inclusion relationship holds true. We can express it as: $l^B_i \leq u^B_i$ for i=1,…,n where $\mathbf{l}^\mathbf{B}=[l^B_1, l^B_2, \ldots, l^B_n]$ and $\mathbf{u}^\mathbf{B}=[u^B_1, u^B_2, \ldots, u^B_n]$). If $\mathbf{l}^\mathbf{B} = \mathbf{u}^\mathbf{B}$ then the hyperbox reduces to a single point (numeric datum) B($\mathbf{l}^\mathbf{B}$, $\mathbf{l}^\mathbf{B}$) = {$\mathbf{l}^\mathbf{B}$}. Hyperboxes are elements of a family of sets defined in $\mathbf{R}^n$. More specifically, we state that B $\in$ P($\mathbf{R}^n$) with P(.) being the power set of $\mathbf{R}^n$. The volume of B, denoted by V(B), is viewed as a measure of

specificity of the information granule. The point, B($\mathbf{l}^B$, $\mathbf{l}^B$) has the highest specificity. As the volume increases the specificity of information granule decreases correspondingly. Computationally, it is advantageous to consider the expression exp(-V(B)) which captures this aspect of granularity and is normalized, i.e. it attains 1 for the numeric datum and tends to zero once the hyperbox starts growing.

It is instructive to elaborate on the use of such information granules in the realm of PR (in which we are quite commonly confined to the language of probability and probabilistic granules, articulated as probability functions or probability density functions). Transparency of results is a key factor here. To illustrate this point Figure 2 shows two granular constructs. In the first case, a two-dimensional box captures the essence of the data: we may state that the Cartesian product of [a, b] and [c,d], B = [a,b] $\times$ [c,d] "covers" the data. Moreover, both features (intervals) maintain their identity. In contrast, ellipsoidal information granules (which can be otherwise quite expressive) do not provide the same transparency as hyperboxes. Obviously one can project the ellipsoid on the corresponding features. Note however that the reconstruction (the Cartesian product C =[e, f] $\times$ [g, h]) could be quite different from the original granule $\Omega \neq$ C.

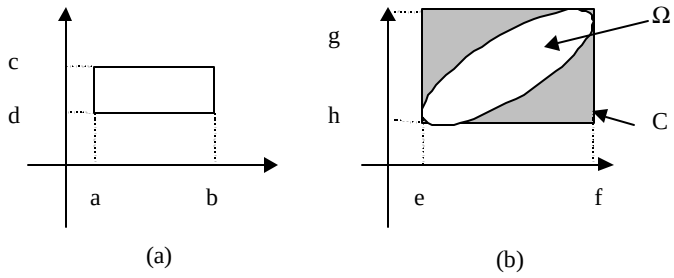


Figure 2. Expressing data as hyperboxes (a) and ellipsoidal information granules (b); note a reconstruction deficiency caused by the dependency between the features

3 The principle of granular clustering

Before we proceed with the details of the granular clustering technique, it is instructive to discuss the underlying principle, learn how the process proceeds and concentrate on the interpretation of some results generated by the proposed clustering mechanism. As emphasized in the literature [1, 4], the essence of clustering (unsupervised learning) is to discover a structure in data. It is generally true that almost all existing clustering techniques operate on numeric objects (vectors in $\mathbf{R}^n$) and produce representatives (prototypes) that are again entirely numeric. In this sense, their form does not reflect how many data points they

represent or what is the distribution of these data points. In the design of the clustering method, we add an extra dimension of granularity that helps sense the structure in the data as it becomes unveiled during the formation of the clusters.

Without loss of generality we focus our attention on the subspace of $\mathbf{R}^n$ and concern ourselves with the granular clustering algorithms defined on the unit hyperbox $[0\ 1]^n$. Consequently, as a pre-processing step, we normalize all input data to such a hyperbox. This pre-processing ensures that the granular clustering algorithm has a simpler mathematical formulation while retaining generality for all data in $\mathbf{R}^n$.

3.1 The design

The approach introduced here differs in many ways from other approaches [1, 5, 14, 19, 20, 21]. The leitmotiv is the following:

> An abstraction (no matter whether dealing with numeric or granular elements) is achieved through the *condensation* of original data elements into granules, whose location and granularity reflects the essence of the structure of data. The more condensation, the larger the sizes of the information granules that realize this aggregation.

The granular clustering is carried out as the following iterative process

- Find the two most compatible information granules (where the idea of compatibility guiding this search will be quantified later on) and on this basis build a new granule embracing them. In this way, one condenses the data while reducing the size of the data set.
- Repeat the first step until enough data condensation has been accomplished (here one has to come up with a termination criterion or introduce a sound validation mechanism).

Figure 3 illustrates how the clustering algorithm works. We start from a collection of small information granules (the original data) and start growing larger information granules. Noticeably, through their growth they tend to reflect the essential characteristics of the original data. The size of the granules reflects how much they have incorporated the original data and conveys extra information about their distribution.

This approach resembles techniques of aggregative hierarchical clustering. There is a striking difference though: in hierarchical clustering we deal with point-size data and the clusters are sets of *homogeneous* objects. No conceptually new entities are formed. Here we deal with a *heterogeneous* mix of data items and "grow" larger information granules from the smaller granules and/or the individual point-size data. It should be stressed that the nature of an information granule is significantly richer compared to that of point data. It involves additional attributes such as "shape" ($\mathbf{u}^B$-$\mathbf{l}^B$) and "size" ($\|\mathbf{u}^B$-$\mathbf{l}^B\|$) in addition to the "position" ($\mathbf{u}^B$)

attribute that is associated with both granules and point data. By monitoring the attributes of the hyperboxes we can oversee the clustering process more effectively. Essentially, once we find that the attributes of individual boxes indicate that they are incompatible with each other (the notion explained in Section 4), the process of clustering is terminated.

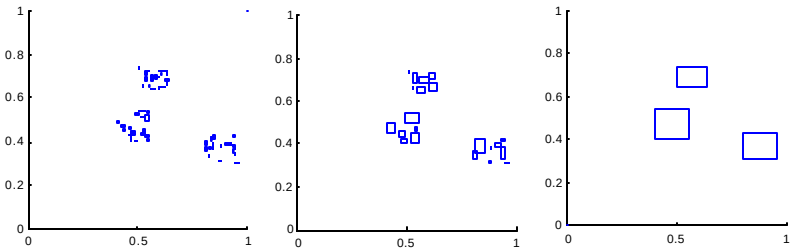


Figure 3. Several snapshots of cluster growing over the clustering process; observe the small information granules forming at the initial stage (first iteration) that are grouped in some well-confined regions and give rise to three apparent large information granules at the later stage of clustering

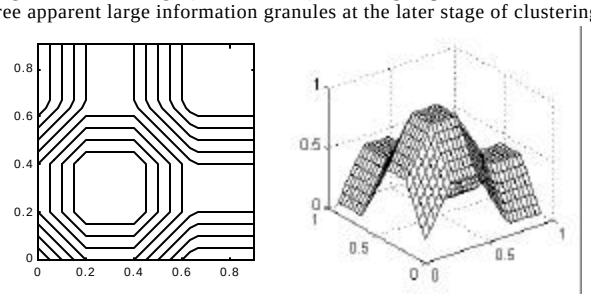


Figure 4. Simpson's membership function (as presented in [15]) for the hyperbox defined by the min point V=[0.2 0.2] and max point W=[0.4 0.4]. Sensitivity parameter γ is equal to 4.

By the same token, this concept should be contrasted with the idea of min-max clustering discussed by Simpson [14, 15] as this technique seems to bear some resemblance with the method studied here. The similarity is only superficial though. First, Simpson's method deals with point-size data while we consider data that is represented by either points or hyperboxes in pattern space. Second, the fuzzy membership functions of the information granules proposed by Simpson promote the formation of clusters whose size varies greatly in various dimensions. This is exactly the opposite to what we are trying to promote through the "compatibility measure" (discussed in Section 4). To emphasise the latter point, we present in Figure 4 a representative of a class of membership functions proposed by Simpson and refer the reader to Figure 10 for comparison with the functions that have been utilised in the proposed clustering algorithm.

3.2 Interpretation and validation of granular clustering

In the literature, there are a number of cluster validity indexes, whose role is to assess the "goodness" of clusters and, as a consequence, to identify the most "plausible" number of clusters. Validity indexes help guide the clustering process by implying what the number of clusters should be. Commonly, their behavior does not lead to clear conclusions though. Even worse, they may generate conflicting suggestions as to the proper number of clusters. In granular clustering we take another position. As the clusters capture the core of the data (and obviously, this is regarded as an important benefit of the method), our conjecture is that such a core should help establish a sound platform for assessing clusters.

When growing the information granules, a criterion worth investigating is the volume of the smallest granule (V_{min}) that needs to be constructed, at this particular step, in order to cluster two component granules (more specifically, we determine $e^{-V_{min}}$; the details will be covered in Section 4.1). If that minimal volume grows quickly then it can be deduced that the compatibility of the component granules is low and the clustering process can be terminated.

Again, it is worth emphasizing that the granularity of data adds an extra dimension to any processing. Not only the location of the information granule is essential but also its size and shape play a crucial role in the process of clustering and afterwards during the validation of the clusters.

4 The computational aspects of granular computing

There are two essential functional elements of granular clustering that need to be described prior to presenting the detailed algorithm. These concern how distance between two information granules is determined and how we compute an inclusion relation between granules. While the definitions generalize to a multidimensional case, we focus here on a two-dimensional case. Note also that these two concepts work for *heterogeneous* data, that is, granules and numeric entities in the same feature space.

4.1 Defining compatibility between information granules

In this section, we discuss how compatibility and inclusion between two information granules are computed. The issue is more complicated than in a numeric case as these notions are granular and therefore the definitions of compatibility and inclusion should reflect this fact.

Consider two information granules A and B. More explicitly, we follow the full notation A($\mathbf{l}^{\mathbf{A}}$, $\mathbf{u}^{\mathbf{A}}$) and B($\mathbf{l}^{\mathbf{B}}$, $\mathbf{u}^{\mathbf{B}}$) to point at their location in the space. The compatibility, compat(A, B), involves two components: a distance between A and B, d(A,B), and the size of the information granule that would be formed by merging

A and B. The distance d(A,B) between A and B is defined on the basis of the distance between their extreme vertices.

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

Obviously $\|.\|$ 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 the 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, the granularity of C can be captured by a volume $V(C)$ computed in the standard fashion

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

where

$$length_i(C) = \max(u_i^B, u_i^A) - \min(l_i^B, l_i^A) \quad (3)$$

i=1, 2, …, n. (Refer to Figure 5.)

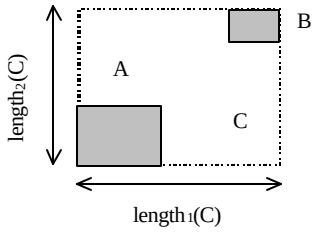


Figure 5. Information granule C as a result of combining A and B

The two expressions (1)-(2) are the elements of the compatibility measure, compat(A, B) defined as

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

The rationale behind the above form of the compatibility measure is as follows. In clustering we aggregate two information granules that are closest together, i.e. their compatibility measure is highest. In light of the above criterion, 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), in other words it stipulates formation of hyperboxes that are as similar to hypercubes as possible. The exponential term in this expression normalizes all values to the unit interval. In particular, the volume of a point produces $e^0 = 1$ When the volume increases, its exponential function goes to zero. The parameter α balances the two concerns in the compatibility measure and is chosen so as to control the 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 overall processing of the information granules. By contrast, it is not essential and does not play any role when we cluster point-size data instead of granules. To constrain the values of the compatibility measure to the unit interval, we consider the data to lie in the unit hypercube $[0,1]^n \subset \mathbf{R}^n$ (in other words we normalize the data before computing the value of (4)) and consider a normalized distance assuming values in the unit interval.

To gain a better insight of what really is accomplished when using the above compatibility measure, let us study two points (numeric values) A and B situated in $\mathbf{R}^2$. Furthermore let A be fixed and located at the origin of the coordinates while we allow B some flexibility. d(A, B) is just a standard Euclidean distance. It becomes obvious that all elements (Bs) located on a circle of a fixed radius exhibit the same distance value. Restrict now the choice of B to be from this pool. If we connect A and any such B, the resulting volume changes its value depending upon the location of B. Interestingly, out of all Bs, there are four locations on the circle for which the volume of the resulting granule attains its maximum. This happens if that box (the information granule formed by clustering A and B) is a square. In other words, the compatibility measure attains a maximal value when C is a hypercube.

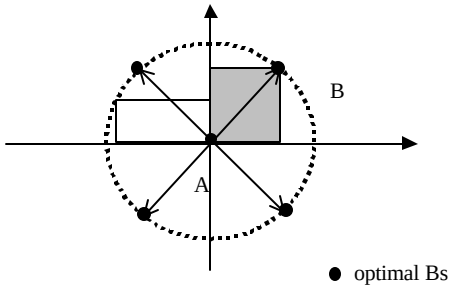


Figure 6 The calculations of the compatibility measure; note that there are four possible candidates (Bs) on the circle that maximize this measure

If we plot the compatibility measure as a function of τ (where τ is an angular position of B), we can easily see that the value of the compatibility measure is

modulated by the angle (or equivalently the shape of the resulting information granule C), see Figure 7.

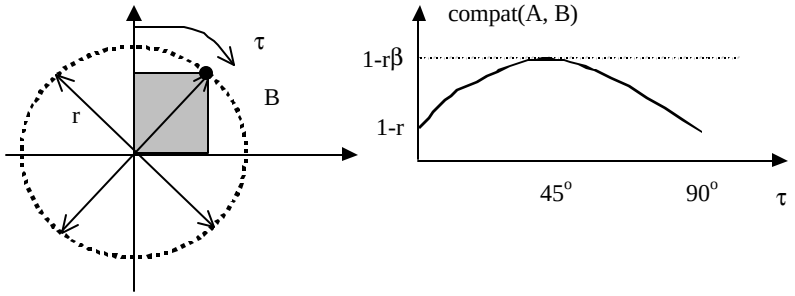


Figure 7 The compatibility measure expressed as a function of τ (the plot above is restricted to the first 90^0 degrees); $\beta = e^{-\alpha r.r/2}$

More importantly, the above graphical considerations shed light on the geometry of the information granules that are preferred by the introduced compatibility measure. This preference reflects a principle that may be termed *the principle of balanced information granularity*. In a nutshell, in building new information granules, we prefer to have entities whose granularity is balanced along all dimensions (variables) rather than constructing granules that are highly unbalanced. A number of selected examples of varying granularity are portrayed in Figure 8.

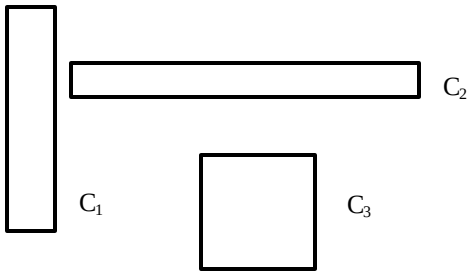


Figure 8. Examples of information granules characterized by various degrees of balance of information granularity; note that C_1 and C_2 are highly unbalanced (have high levels of information specificity along one of the dimensions only) while C_3 is well-balanced.

When we change the distance function to the Hamming (p = 1) or Tchebyschev distances ($p=\infty$), we have a number of Bs to choose from yet this selection can be made from different geometrical figures (that is a diamond and a square), Figure 9.

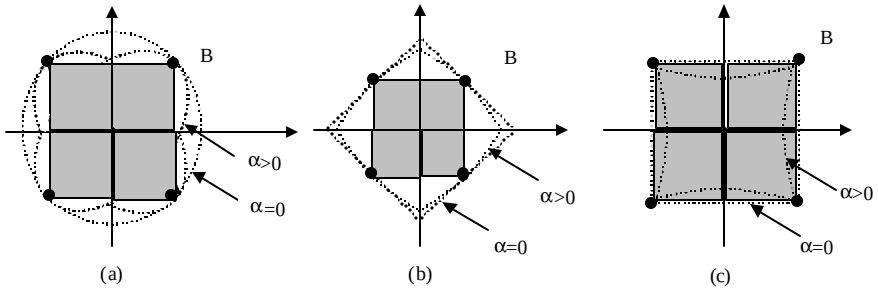
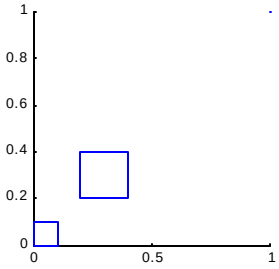
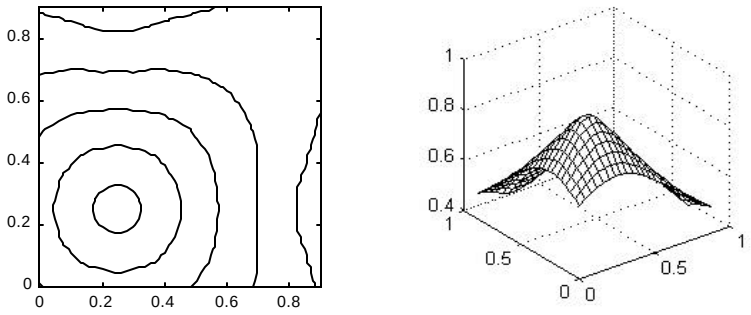


Figure 9. Identification of Bs leading to the highest value of the compatibility measure calculated with Euclidean distance (a), Hamming distance (b) and Tchebyschev distance (c).

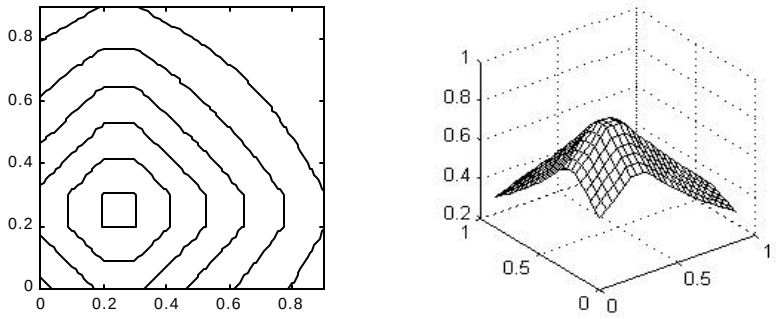
Moving on to the case where both A and B are two information granules, the resulting plots visualizing the compatibility measure are collected in Figure 10.



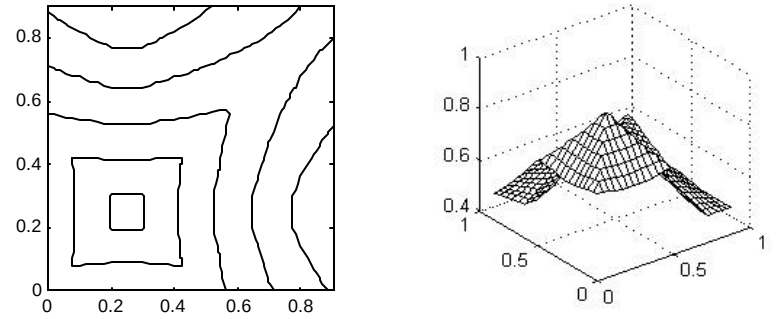
(a) Two hyperboxes representing information granules in a unit box in $\mathbf{R}^2$



(b) Compatibility measure with L_2 distance measure



(c) Compatibility measure with L_4 distance measure



(d) Compatibility measure with L_∞ distance measure

Figure 10. Comparison of compatibility measures obtained with various distance measures. Note the preference that the compatibility measure gives to hyperboxes that are well balanced in all dimensions. This contrasts with the membership function proposed in [15] and illustrated in Figure 4.

As the clustering proceeds (refer to Figure 3), the process of merging progressively less closely associated patterns is reflected 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 11.

The proximity of patterns that are merged into granules at the early stages of the clustering process is reflected in the relatively small gradient of the compatibility measure curve. In contrast, the large gradient of the curve, at the final stages of clustering, indicates the merging of highly *incompatible* clusters. The compatibility measure curve therefore provides a convenient reference for identifying how many clusters are needed to capture the essential characteristics of the input data, while providing the best generalization. The intersection of the two gradient lines (as visualized in Figure 11) can be used as an approximation to the

optimal number of clusters. This number provides a good starting point in the subsequent optimization of the overlap of the identified clusters as discussed below.

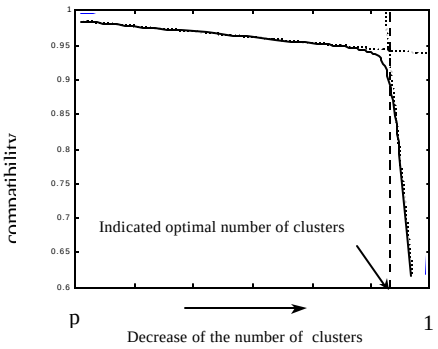


Figure 11. An example of the evolution of the compatibiliy measure over the full cycle of the clustering process (p is the initial number of patterns).

Referring to the compatibility index, we can consider a modified form where we consider a sum of the sides(edges) of the hyperboxes

$$\text{compat}(A, B) = 1 - d(A,B)e^{-\alpha L(C)} \tag{5}$$

with

$$L(C)= \sum_{i=1}^{n} \text{length}_i (C) \tag{6}$$

Considering the nature of these indexes, we refer to the first index as volume-driven (4), while the second is edge-driven (5). To compare these two forms of the compatibility index, we consider a simple two-dimensional case in which both A and X are numeric. We allow X to move on a unit circle while A is located at the origin of the coordinates (see Figure 12).

In this way the distance is always equal to 1 and the compatibility can be expressed by a single angle ϕ. For the volume driven version we have $\text{compat}(A,X) = 1 - e^{-(\sin\phi\cos\phi)}$ and for the edge-driven version $\text{compat}(A, X) = 1 - e^{-(\sin\phi+\cos\phi)}$. The plots of the compatibility measures are shown in Figure 13. It becomes obvious that the highest compatibility value is achieved for the same value of the angle, i.e., $\phi = \pi/4$.

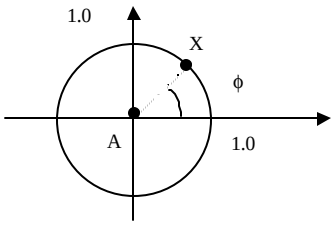


Figure 12. Computing the compatibility measure for A and X and expressed as function of ϕ.

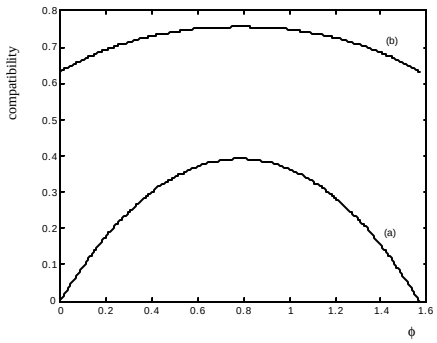


Figure 13. Compatibility measures as a function of ϕ: (a) volume-driven, (b) edge-driven; ϕ is constrained to $[0, \pi/2]$

These compatibility measures exhibit a visible difference when we look at their sensitivity defined as

$$\text{sens}(\phi) = \left| \frac{\partial \text{compat}(A, X)}{\partial \phi} \right|$$

Figure 14 reveals that the compatibility based on the volume of the information granule has higher sensitivity.

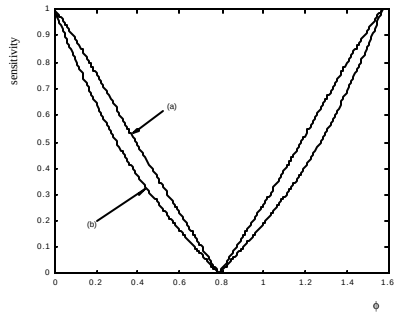


Figure 14. Sensitivity of the compatibility measures regarded as a function of ϕ: (a) volume-driven, (b) edge-driven, ϕ is constrained to $[0, \pi/2]$

4.2 Expressing inclusion and overlap of information granules

The inclusion relation expressing an extent to which A is included in B is defined as a ratio of two volumes

$$\text{incl}(A, B) = \frac{V(A \cap B)}{V(A)} \tag{7}$$

It is clear from the above that the inclusion measure is monotonic, non-commutative and satisfies the following boundary conditions: incl(A,X)=1 and incl(A,$\emptyset$)=0 where X and $\emptyset$ are the unit hyperbox and the empty set in $\mathbf{R}^n$, respectively. The calculations are straightforward; Figure 15 enumerates all cases for one-dimensional granules along with the pertinent values of this measure.

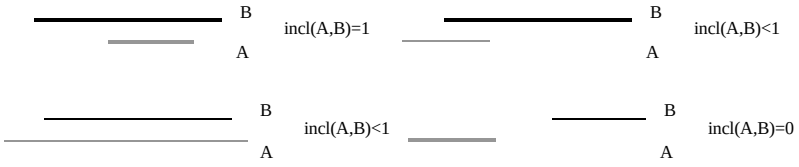


Figure 15. Computing the inclusion for two information granules A and B

It is worth mentioning that the value of the inclusion measure drops rapidly (at a rate of a^{-n} where $a \in (0, 1)$) with the increasing dimension of the feature space. For

example, if there is a 50% overlap (a=0.5) in each variable in an n-dimensional space, the inclusion level is expressed as 0.5^n. Clearly, the objective of effective information abstraction through clustering of information granules translates into identifying granules for which there is a minimum overlap. To encourage the merging of granules that have significant overlap, we calculate an average of the maximum inclusion rates of each granule in every other granule.

$$overlap(c) = \frac{1}{c-1}\sum_{i=1}^{c} \max_{\substack{j=1,..c \\ j\neq i}} (incl(A(i), A(j))) \tag{8}$$

where c is the current number of granules and A(i) and A(j) are i-th and j-th granule respectively. However, we must point out that, while the measure (7) is monotonic for any two pairs of granules (i.e. if $A \subset B$ and $C \subset D$, then $incl(A,C) \leq incl(B,D)$), the change of the number and size of granules during the clustering results in various local optima for (8). We illustrate this effect in Figure 16.

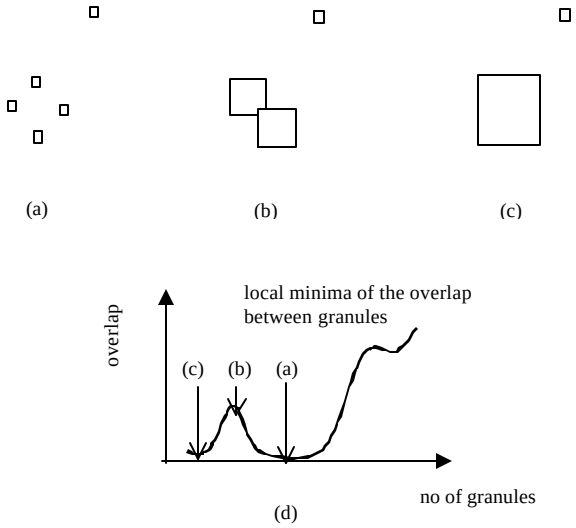


Figure 16. Progression from 5 to 2 granules involves stage (b) during which granules overlap. This is reflected in *overlap*(3) > 0 while *overlap*(5) = 0 and *overlap*(2) = 0

Because of the local minima of the *overlap*(.) function it is important to have a good initial estimate of the target number of clusters as a starting point for the local

minimization of the function. Such an estimate is provided by our earlier analysis of the compatibility measure, as discussed in the previous section.

Having completed clustering the quality of data abstraction afforded by the given set of clusters is measured using an independent validation data set. The generality of each cluster is well quantified by the sum of the inclusion rates of the validation data items in the respective cluster.

$$INCL(i) = \sum_{j=1}^{M} incl(V(j), A(i)) \qquad i = 1,..c \tag{9}$$

where c is the number of clusters, M is the cardinality of the validation data set, V(j) are the validation patterns and A(i) are the clusters. As well as indicating whether a given cluster is representative for a large proportion of data, the INCL(.) measure can be used to assess how representative the training and the validation data sets are. If the sets are representative, then INCL(.) should correlate closely with the cardinality of the individual clusters.

5 The Granular Analysis

The hyperboxes constructed during the design phase are helpful in a thorough analysis of the data set. They shed light on the nature of data as they are perceived from the standpoint of information granularity. We will discuss two aspects of this analysis. First, we characterize the hyperboxes themselves. Second, we analyze the properties of the variables (features) forming the data space. We should emphasize that the granular analysis follows the clustering phase and does not impact it in any way. To maintain the conciseness of our presentation, we consider that each of the "c" hyperboxes located in the n-dimensional space is fully described by vectors of its lower and upper corners (coordinates), $\mathbf{B}(k) = \{\mathbf{l}^{\mathbf{B}}(k), \mathbf{u}^{\mathbf{B}}(k)\}$, k=1, 2, …, c where $\mathbf{l}^{\mathbf{B}}(k)$ and $\mathbf{u}^{\mathbf{B}}(k)$ are vectors of the corresponding coordinates, that is $\mathbf{l}^{\mathbf{B}}(k) = [l^B_1(k)\ l^B_2(k) \ldots l^B_n(k)]$ and $\mathbf{u}^{\mathbf{B}}(k) = [u^B_1(k)\ u^B_2(k) \ldots u^B_n(k)]$

5.1 Characterization of hyperboxes

The most evident characterization of the hyperboxes can be provided in their volumes, V(B(k)). The computations are obvious. First, we determine a ratio (normalized length)

$$norm\_length_i(B(k)) = \frac{u_i^B(k) - l_i^B(k)}{range_i(B(k))} \tag{10}$$

where $range_i(B(k))$ is a range of the i-th feature (variable). Since the data is normalized to a unit hypercube the $range_i(B(k)) = 1$ for all i. Second, the volume is taken as a product

$$V(B(k)) = \prod_{i=1}^{n} norm\_length_i(B(k)) \tag{11}$$

The volume quantifies the granularity of the hyperboxes. Intuitively, it states how "large" (detailed) the hyperboxes are and how much detail each captures. One can take an average of the volumes of the hyperboxes that gives a general summary of the hyperboxes

$$\bar{v} = \frac{1}{c} \sum_{k=1}^{c} V(B(k)) \tag{12}$$

If one side of the hyperbox is zero then the volume measure returns a zero value. This occurs because of the multiplicative nature of volume. To alleviate this problem, we may also introduce an additive measure. A plausible descriptor of a hyperbox could reflect a "circumference" of the hyperbox and read as follows

$$\sum_{i=1}^{n} norm\_length_i(B(k)) \tag{13}$$

5.2 Granular feature analysis

The granulation of the data space (and each feature) provides an interesting insight into the nature of the variables occurring in the problem. In what follows, we describe the variables in terms of sparsity and discriminative powers. These two descriptors are implied by the granular nature of the hyperboxes.

5.2.1 Sparsity

When looking at a certain variable in the hyperboxes, we can visualize how much of the entire range of the variable is occupied by the hyperboxes (i.e., how *sparse* the boxes are in the given space). Take the i-th feature and calculate the sum of length of the corresponding sides of the hyperboxes

$$tot\_length_i = \sum_{k=1}^{c} length_i(B(k)) \tag{14}$$

where length(B(k)) = $u^B_i(k) - l^B_i(k)$ and i=1,2,…,n. The sparsity defined in the form

$$\text{sparsity}_i = \frac{tot\_length_i}{range_i(B(k))}\frac{1}{c} \tag{15}$$

assumes values in the unit interval. If sparsity_i is less than 1 then this represents a situation when all hyperboxes (more precisely their i-th coordinates) occupy a portion of the entire range of the feature. We may state that the variable is "underutilized". In other words, we witness a highly localized usage of this feature. Sparsity value near 1 means a complete utilization of the variable. Overutilization happens when sparsity achieves values higher than 1 (in this case we have some hyperboxes overlapping along this variable).

The sparsity measure does not capture the entire picture. A situation illustrated in Figure 17 shows two cases where the distribution of the hyperbox along the given feature is very different, yet we end up having the same value of the sparsity. This leads us to another index that describes an overlap between the hyperboxes

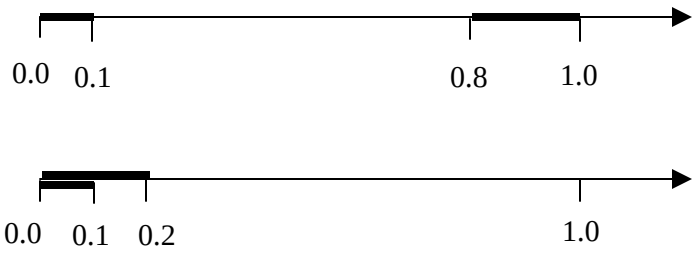


Figure 17. Two different distributions of hyperboxes (i-th feature) producing the same value of the sparsity index; in both cases the sparsity is equal to 0.3

5.2.2 Overlap index

We define the following index called coordinate overlap

$$c-\text{overlap}_i = \frac{2}{c(c-1)}\sum_{k=1}^{c-1}\sum_{l>k}^{c}\frac{\text{length}_i(I(k)\cap I(l))}{\text{length}_i(I(k)\cup I(l))} \tag{16}$$

i =1,2, …, n. In this definition, I(k) and I(l) are intervals (sides) of the hyperboxes for the i-th variable.The higher the value of this index, the more overlap between the hyperboxes projects on the given variable. When I(k) and I(l) are pairwise disjoint

then the overlap is equal to zero. This means that the feature is highly discriminative, as it separated the hyperboxes. The higher the overlap measure, the lower the discriminative power of the feature.

Each of the measures leads to a linear ordering of the features. We can easily state which of the features is highly "utilized" and which of them comes with the most significant discriminative properties. To form a comprehensive picture, one can localize each feature in the sparsity – overlap space. By doing this, one can distinguish between the variables that are essential to the PR problem. More specifically, we prefer features that exhibit low overlap (as those come with strong discriminative properties) along with low values of sparsity (localized usage of the variable). It should be stressed that these descriptors (sparsity and overlap) emerge as important quantifiers because of the existence of information granules forming the hyperboxes.

6 Experimental studies

The series of experiments is aimed at visualizing the most essential features of granular clustering. We consider both a synthetic data set and real-life data available on the WWW (Boston housing data).

6.1 Synthetic data

The synthetic data sets consist of 3 groups of information granules (hyperboxes), $A_i \in [0,1] \times [0,1]$, generated by a random number generator with a uniform distribution. Each group comprises 20 granules dispersed around pre-defined points: c_1=[0.4, 0.4]; c_2=[0.5, 0.6]; and c_3=[0.8, 0.3]. The dispersion factor σ is varied between 0.08 and 0.15 to establish the sensitivity of the clustering process to the dispersion of the data. The clustering process is governed by the compatibility measure, (4), with the distance defined according to the L_2 norm and the "compactness" factor α=0.5.

An example of the evolution of the compatibility measure throughout the clustering process is shown in Figure 18. The intersection of the two asymptotes to the compatibility measure, traced at the beginning and at the end of the clustering process, indicates that 3 clusters (iteration 57) mark a natural 'change over' point in the behavior of the system. So, the clustering process should terminate with 3 clusters, provided that the degree of overlap of clusters is also minimized for this number of clusters.

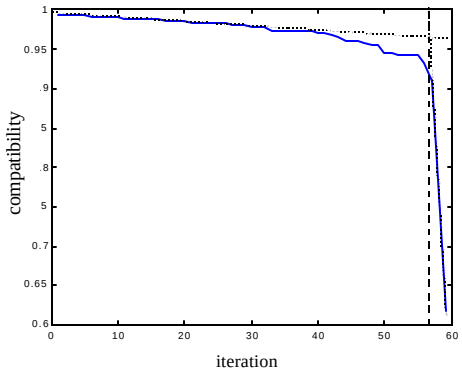


Figure 18. Compatibility measure for a single clustering process.

The degree of overlap of the clusters was evaluated at each of the 59 iterative steps of the clustering process, according to equation (16), and is depicted in Figure 19. As expected, the results of the cluster overlap analysis confirm that the test data naturally falls into 3 clusters since the overlap function assumes a local minimum for 4 or fewer clusters.

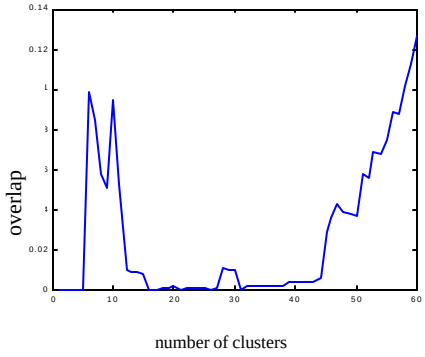


Figure 19. An average degree of overlap of clusters

The quality of data abstraction achieved through clustering is assessed by evaluating the inclusion rate, (9), of an independently generated data (with the same statistical properties) in the clusters that have been identified.

The change of the overall inclusion rate of the validation data throughout the clustering process is illustrated in Figure 20. It is not surprising to see that the high value of the average inclusion rate for 3 or fewer clusters confirms that 3 clusters capture the essential features of the data while the high value of the compatibility measure confirms that the clusters retain high specificity. Should the number of

clusters be reduced to 2 or 1, the inclusion rate of the validation data set would only be improved marginally while there would be a very significant reduction in the specificity of the cluster(s).

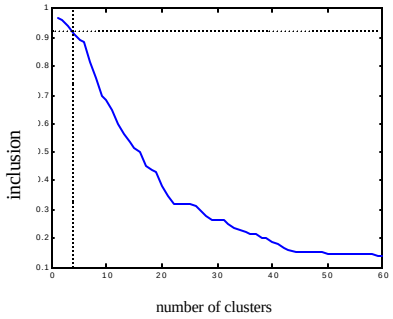


Figure 20. Average inclusion rate for the validation data set

In order to achieve a degree of independence from the statistical characteristics of the random number generator the evaluation of the inclusion of the validation data sets in the clusters was repeated 100 times for each value of $\sigma \in \{0.08, 0.09, 0.10, 0.11, 0.12, 0.13, 0.14, 0.15\}$ and the number of clusters varying from 1 to 10. A total of 8000 training sets and 8000 validation sets were processed. Figure 21 illustrates how the inclusion measure, (7), depends on the data dispersion parameter σ and the number of clusters. It is interesting to note that σ has little influence on the value of the inclusion measure. This is a very desirable characteristic of the clustering process, since it suggests that the precise statistical properties of data sets do not need to be known for the clustering to be effective.

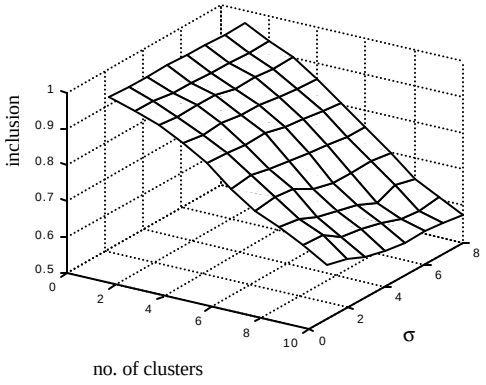


Figure 21. Average inclusion measure evaluated for 8000 training and validation sets

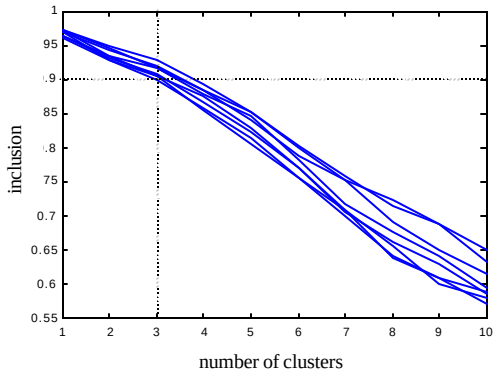


Figure 22. 2-D projection of the surface from Figure 21 resulting in a family of curves illustrating average inclusion rates of the validation data in clusters for the various values of σ

It is easy to note, from Figures 21 and 22, that the inclusion rate of 0.9 or higher is consistently attained with 3 or fewer clusters.

We now assess the progression of the clustering process using the edge-based compatibility measure defined by (5). Figure 23 illustrates a typical evolution of the compatibility measure. As expected, the asymptotic change of character of this function occurs at iteration 57, indicating that there are 3 significant clusters.

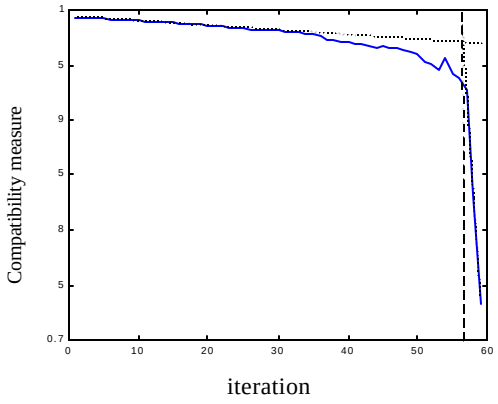


Figure 23. Edge-driven compatibility measure

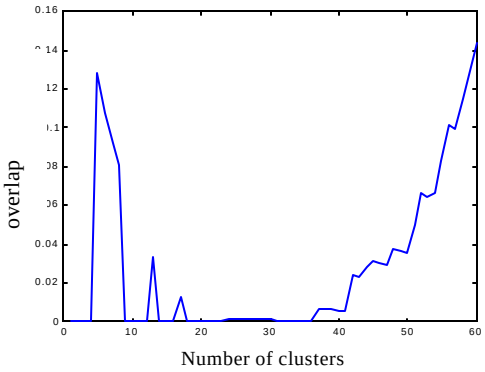


Figure 24 Average degree of overlap of clusters throughout the clustering process

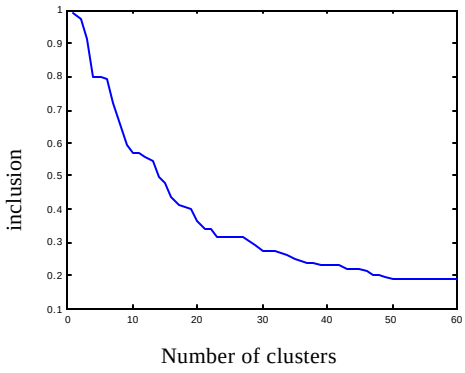


Figure 25. Average inclusion rate of the validation data in the clusters

The results illustrated above, Figures 23-25, are directly comparable to those contained in the earlier experiments. The asymptotical behaviour of the compatibility measure (Figures 18 and 23) is nearly identical and the only noticeable difference in the progression of clustering occurs at the intermediate stages.

6.2 Boston housing data

Although for 2-dimensional data sets $B \in P(\mathbf{R}^2)$ the number of clusters can be easily established by visual inspection, higher dimensional data presents a significant challenge. We have therefore applied the algorithm to a realistic 14-dimensional data set representing factors affecting house prices in Boston area

(USA). The data set was originally compiled by Harrison and Rubinfeld, [6], and is available from the Machine Learning Database at University of California at Irvine (http://www.ics.uci.edu/~mlearn/MLSummary.html). The data set comprises 506 records.

The 14 attributes of each data record are as follows:

1. CRIM per capita crime rate by town
2. ZN proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS proportion of non-retail business acres per town
4. CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX nitric oxides concentration (parts per 10 million)
6. RM average number of rooms per dwelling
7. AGE proportion of owner-occupied units built prior to 1940
8. DIS weighted distances to five Boston employment centres
9. RAD index of accessibility to radial highways
10. TAX full-value property-tax rate per $10,000
11. PTRATIO pupil-teacher ratio by town
12. B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT % lower status of the population
14. MEDV Median value of owner-occupied homes in $1000's

6.2.1 Study A

We divided the original set into two sets: the training set, comprising 253 odd-numbered records and the validation set comprising 253 even-numbered records. It should be noted that, as a pre-processing step, all data has been mapped into a 14-dimensional unit hyperbox. The compatibility measure provided direction for the clustering process and the evolution of this measure throughout the whole process is presented in Figure 26. The gradients of the compatibility measure at the beginning and the end of the process indicate that 7 clusters represent a good abstraction of the training data. In the vicinity of 7 clusters the cluster overlap indicator is minimized for 7 and 8 clusters, as shown in Figure 27. Of these two possible numbers of clusters we select the smaller number, so as to achieve greater granulation of the original data.

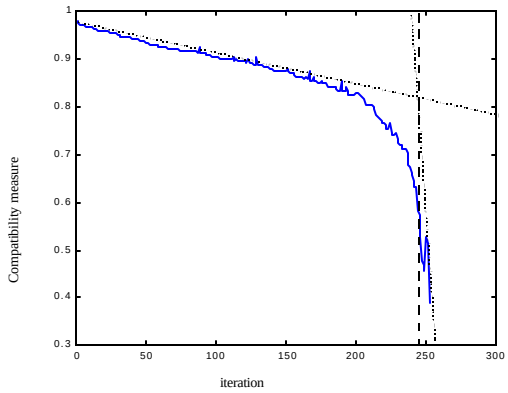


Figure 26. Compatibility measure of clusters formed from the odd-numbered records in the Boston housing data set. Iteration no. 245 corresponds to 7 clusters.

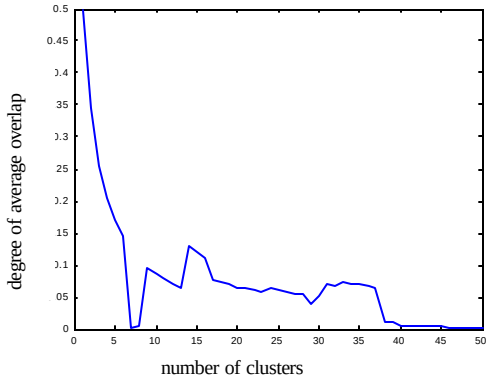


Figure 27. Degree of average overlap of clusters in the last 50 out of 252 iterations

The generality of the identified clusters was tested by evaluating the average inclusion of the validation data set (even-numbered records from the original data set) in the sets of clusters identified in the last 50 steps of the clustering process. This is illustrated in Figure 28. The value of over 90%, achieved for 7 clusters, indicates a good abstraction of the data.

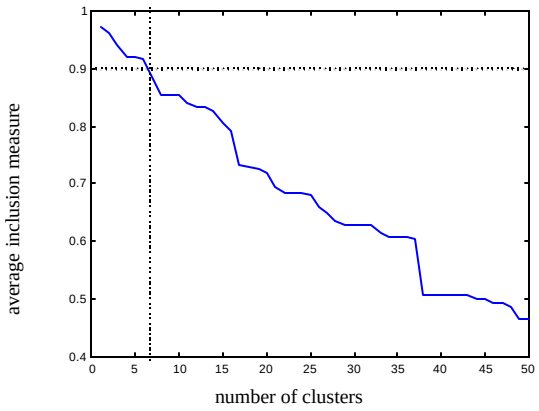


Figure 28. Average inclusion measure evaluated for 1 to 50 clusters

To gain a more detailed insight into the makeup of the 7 clusters we evaluated an aggeregate inclusion measure (9) using the validation set, and compared the results with the cardinality of each cluster. It is clear, from Figure 29, that out of 7 clusters 3 have a significant support in the two data sets, while the other 4 clusters represent data that could be described as significant exceptions. It is interesting to note, however, that the zero inclusion rates of the validation data in clusters 3, 4 and 7 indicate that the small data sample makes it difficult to do a proper evaluation of the clusters.

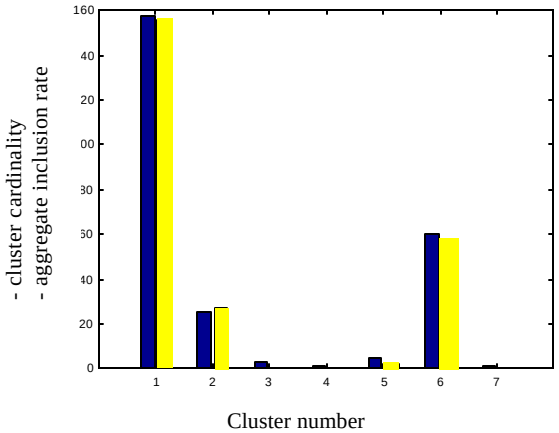


Figure 29. Cardinality (first bar) and the aggregate inclusion rate (second bar) for each of the 7 clusters

The full description of the identified clusters is given in Table 1.

TABLE 1. Description of the 7 clusters. (l_i represents minimum coordinates of the i-th hyperbox and u_i represents maximum coordinates)

Variable number	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
l_1	0.0063	0.0686	1.1265	2.0099	3.4744	2.3783	88.9762
u_1	0.0063	2.7795	3.3213	2.0099	8.9834	73.5337	88.9762
l_2	0	0	0	0	0	0	0
u_2	0	0	0	0	0	0	0
l_3	0.7399	8.1399	19.5800	19.5800	18.1001	18.1001	18.1001
u_3	0.7399	27.7400	19.5800	19.5800	18.1001	18.1001	18.1001
l_4	0	0	1	0	1	0	0
u_4	0	0	1	0	1	0	0
l_5	0.3850	0.5200	0.8710	0.6050	0.6310	0.5320	0.6710
u_5	0.3850	0.8710	0.8710	0.6050	0.7700	0.7700	0.6710
l_6	4.9730	4.9030	5.0120	7.9290	5.8750	4.1380	6.9680
u_6	4.9730	6.4580	6.1290	7.9290	8.7800	7.0610	6.9680
l_7	6.0004	69.6999	88.0004	96.2005	82.8997	41.9002	91.8999
u_7	6.0004	100.0000	100.0000	96.2005	97.4997	100.0000	91.8999
l_8	1.7984	1.3459	1.3216	2.0459	1.1296	1.1370	1.4165
u_8	10.7103	3.9900	1.7494	2.0459	2.7227	3.7240	1.4165
l_9	1.0000	2.0000	4.9999	4.9999	24.0000	24.0000	24.0000
u_9	8.0000	4.9999	4.9999	4.9999	24.0000	24.0000	24.0000
l_{10}	192.9998	188.0008	402.9980	402.9980	665.9989	665.9989	665.9989
u_{10}	469.0011	711.0000	402.9980	402.9980	665.9989	665.9989	665.9989
l_{11}	12.6000	14.7000	14.7000	14.7000	20.2000	20.2000	20.2000
u_{11}	22.0000	21.2000	14.7000	14.7000	20.2000	20.2000	20.2000
l_{12}	288.9906	70.8002	321.0184	369.2980	347.8787	0.3200	396.9000
u_{12}	396.9000	396.9000	396.9000	369.2980	395.4287	396.9000	396.9000
l_{13}	1.9199	6.4300	12.1200	3.7000	2.9600	3.2601	17.2099
u_{13}	30.8101	29.6801	26.8200	3.7000	17.5999	37.9700	17.2099
l_{14}	12.7000	8.1000	13.4002	50.0000	17.7998	5.0000	10.4000
u_{14}	50.0000	24.3000	17.0002	50.0000	50.0000	50.0000	10.4000

The results of our feature analysis are summarized in terms of sparsity and overlap values. This analysis provides an interesting observation about the discriminatory properties of the variables in the problem. The most dominant ones are: crime rate

(1), nitric oxide concentration (5), index of accessibility to radial highways (9), and proportion of non-retail business acres (3). In other words, these are the variables that discriminate between hyperboxes (we stress that these discriminatory aspects were found in the setting of the information granules, rather than classes.

TABLE 2. Characterisation of the 7 clusters in terms of sparsity and coverlap for each of the 14 variabled (dimensions)

Variable no.	sparsity	c-overlap
1	0.135	0.1826
2	0.136	0.7143
3	0.201	0.2194
4	0.143	0.3333
5	0.291	0.1933
6	0.326	0.3255
7	0.307	0.3759
8	0.210	0.3397
9	0.062	0.2109
10	0.218	0.2381
11	0.241	0.2234
12	0.344	0.4357
13	0.458	0.4399
14	0.426	0.3759

6.2.2 Study B

In order to ascertain whether the selection of records for the training and the validation data sets had significantly influenced conclusions regarding the number of clusters in the original data set, we repeated the clustering process with the training and validation sets swapped. Again, the compatibility measure directed the clustering process and the asymptotic evolution of the measure, at the initial and final stages of the process, indicated that 6 data clusters mark a 'change-over' point in the clustering process (Figure 30).

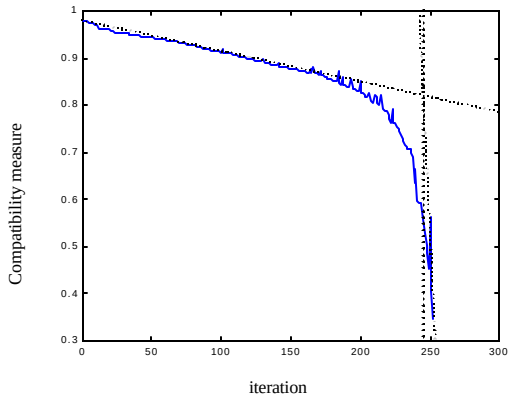


Figure 30. Compatibility measure of clusters formed from the even-numbered records in the Boston housing data set. Iteration no. 246 corresponds to 6 clusters.

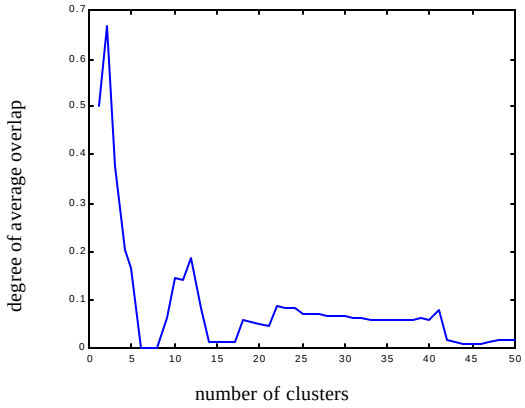


Figure 31. Degree of average overlap of clusters in the last 50 iterations

The curve showing the average degree of overlap between the clusters, illustrated in Figure 31, indicates that a minimum overlap is achieved with 6, 7 and 8 clusters. For the ease of comparison with the Study A case, we select 7 clusters for the validation stage. The average inclusion rate of the validation data set (odd-numbered records from the original data set) in the 7 clusters is slightly worse than in the previous case, averaging 86%. This is illustrated in Figure 32. The reduction of the average inclusion rate in this case suggests that the training and validation

sets contain a small number of unique patterns that do not have counterparts in the other set. The result is that although the distinctiveness of these patterns warrants their inclusion in separate clusters, the cross-comparison of these 'minority clusters' is very limited. This is further verified by the inspection of Figure 33, which shows that clusters 3, 5 and 7 are representing 1, 1 and 2 patterns respectively, and they have no corresponding patterns in the validation set. It is also interesting to note that, compared to the Study A, there is a greater discrepancy between the cardinality of the clusters and the inclusion rate. We conclude therefore that the size of the data supports only firm conclusions about 2 clusters and the characterization of further clusters requires an order of magnitude larger data sample.

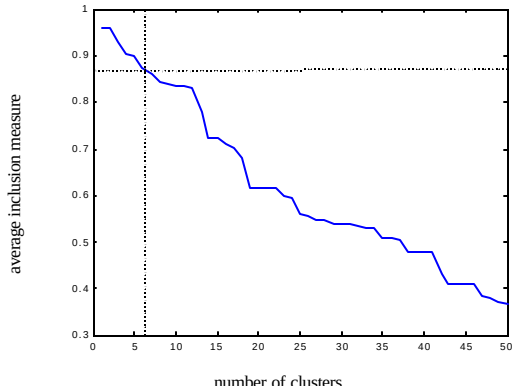


Figure 32. Inclusion measure evaluated for 1 to 50 clusters.

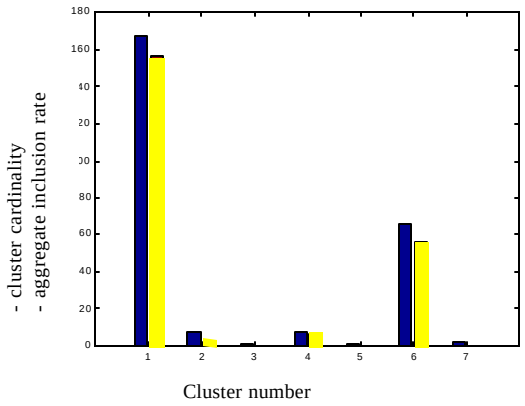


Figure 33. Cardinality (first bar) and the aggregate inclusion rate (second bar) for each of the 7 clusters

The sparsity and c-overlap of the features (variables) are very similar to Study A, meaning that some global properties discovered in the data set have been retained.

TABLE 3. Characterisation of the 7 clusters in terms of sparsity and c-overlap for each of the 14 variabled (dimensions)

Variable no.	sparsity	c-overlap
1	0.117	0.1414
2	0.229	0.2667
3	0.284	0.1432
4	0.143	0.5238
5	0.258	0.0985
6	0.348	0.3391
7	0.393	0.3144
8	0.221	0.1674
9	0.075	0.1769
10	0.155	0.1560
11	0.228	0.0760
12	0.303	0.3276
13	0.443	0.3762
14	0.412	0.3175

7 Conclusions

The study has articulated an alternative view of unsupervised pattern recognition by providing a constructive method of forming information granules that capture the essence of large collections of *heterogeneous* numeric data. In this sense, the original data are compressed down to a few information granules whose location in the data space and granularity reflect the structure in the data. The approach promotes a data-driven problem solving by emphasizing the transparency of the results (hyperboxes). Formation of information granules is guided by two aspects: distance between information granules and size (granularity) of the potential information granule formed through merging two other granules. These two aspects are encapsulated in the form of the compatibility measure. Moreover, we discussed

a number of indexes describing the hyperboxes and expressing relationships between such information granules. We show how to validate the granular structure. The resulting family of information granules is a concise descriptor of the structure of the data – we may call them a granular *signature* of the data.

Some further extensions of the hyperbox approach may deal with more detailed instruments of information granulation such as fuzzy sets [7, 11]. 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 the data to "speak" freely. This is accomplished in two main ways

- first, the hyperboxes are easily understood by a user as each dimension (variable) comes as a part of the construct.
- second, the approach finds relationships that are direction-free, meaning that we do not distinguish between input and output variables (which could be quite restrictive as we may not know in advance what implies what). Obviously, this feature is common to all clustering methods

Furthermore the granulation mechanism puts the variables (features) existing in the problem in a new perspective. The two indexes such as sparsity and overlap are useful in understanding the relevance of the variables, in particular their discriminatory abilities.

While the study was concerned with the development of information granules (hyperboxes), there are interesting inquires into their use in granular modeling. In particular, we are concerned with the fundamental inference problem

- given an input datum (information granule and numeric datum, in particular) X defined in a certain subspace of dimension n' of the original space $\mathbf{R}^{n'} \subset \mathbf{R}^{n}$ and a collection of information granules $\mathsf{B} = \{B(1)\ B(2), \ldots, B(c)\}$ determine the corresponding information granule Y

The current paper provides a basis for this investigation.

Acknowledgments

Support from the Engineering and Physical Sciences Research Council (UK), the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Alberta Consortium of Software Engineering (ASERC) is gratefully acknowledged.

References

1. A. Baraldi, P. Blonda, A survey of clustering algorithms for pattern recognition. IEEE Trans. SMC: Part B, Vol. 29, 6, 1999, pp. 778 –785.
2. A. Bargiela, Interval and ellipsoidal uncertainty models, In: W. Pedrycz (ed.) *Granular Computing,* Springer Verlag, 2001.

3. A. Bargiela, W. Pedrycz, Information granules: Aggregation and interpretation issues, submitted to *IEEE Trans. on Syst. Man and Cybernetics*.
4. J.C. Bezdek, J.M. Keller, R. Krishnapuram, N.R. Pal, *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Kluwer, 1999.
5. B. Gabrys, A. Bargiela, General fuzzy Min-Max neural network for clustering and classification, *IEEE Trans. on Neural Networks*, Vol. 11, No. 3, pp. 769-783, 2000.
6. D. Harrison, D.L. Rubinfeld, Hedonic prices and the demand for clean air, *J. Environ. Economics & Management*, vol.5, 81-102, 1978.
7. A. Kandel, *Fuzzy Mathematical Techniques with Applications*, Addison-Wesley, Reading, MA, 1986.
8. T. Kohonen, Self-organized formation of topologically correct feature maps, *Biological Cybernetics*, 43, 59-69 (1982).
9. T. Kohonen, *Self-organizing Maps*, Springer Verlag, Berlin, 1995.
10. W. Pedrycz, *Computational Intelligence: An Introduction*, CRC Press, Boca Raton, FL, 1997.
11. W. Pedrycz, F. Gomide, *An Introduction to Fuzzy Sets*, Cambridge, MIT Press, Cambridge, MA, 1998.
12. W. Pedrycz, Fuzzy equalization in the construction of fuzzy sets, *Fuzzy Sets and Systems*, vol. 119, 2, 2001, pp. 329-335.
13. W. Pedrycz, M. H. Smith, A. Bargiela, A granular signature of data, *Proc. 19th Int. (IEEE) Conf. NAFIPS'2000*, Atlanta, July 2000, pp. 69-73.
14. P. K. Simpson, Fuzzy Min-Max neural networks – Part1: Classification, *IEEE Trans. on Neural Networks*, Vol 3, No. 5, pp. 776-86, September 1992.
15. P. K. Simpson, Fuzzy Min-Max neural networks – Part2: Clustering, *IEEE Trans. on Neural Networks*, Vol 4, No. 1, pp. 32-45, February 1993.
16. L. A Zadeh, 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, 1979, 3-18.
17. L. A. Zadeh, Fuzzy logic = Computing with words, *IEEE Trans. on Fuzzy Systems*, vol. 4, 2, 1996, 103-111.
18. L. A. Zadeh, 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.
19. M. Meneganti, F.S. Saviello, R. Tagliaferri, Fuzzy neural networks for classification and detection of anomalies, *IEEE Trans. on Neural Networks*, Vol 9, 5, 1998, pp. 848 –861.
20. A. Joshi, N. Ramakrishman, E.N. Houstis, J.R. Rice, On neurobiological, neuro-fuzzy, machine learning, and statistical pattern recognition techniques, *IEEE Trans. on Neural Networks*, Vol 8, 1, 1997, pp. 18 –31.
21. L.I. Kuncheva, J.C. Bezdek, Presupervised and post-supervised prototype classifier design, *IEEE Trans. on Neural Networks*, Vol 10, 5 , 1999, pp. 1142 – 1152.