Interpretable Information Granules with Minkowski FCM

C. Mencar
Dept. of Informatics
University of Bari
v. E. Orabona, 4
70126 –Bari, Italy
mencar@di.uniba.it

A. Bargiela
Dept. of Computing,
The Nottingham Trent
University, Nottingham,
NG1 4BU, UK
andre@doc.ntu.ac.uk

G. Castellano
Dept. of Informatics
University of Bari
v. E. Orabona, 4
70126 –Bari, Italy
castellano@di.uniba.it

A. M. Fanelli
Dept. of Informatics
University of Bari
v. E. Orabona, 4
70126 –Bari, Italy
fanelli@di.uniba.it

***Abstract* – In this study, we investigate the interpretability of information granules that arise through the application of a Fuzzy C-Means algorithm equipped with general Minkowski metric. The paper offers a link between the classical use of Euclidean norm and the more recently reported Tchebychev norm in the context of FCM-based data granulation. In particular, we focus our attention on the topology of information granules that are derived for various alpha-cuts of the resulting fuzzy sets. We quantify deformation of the granules caused by interaction between the FCM prototypes by relating their actual shape to the ideal hyper-boxes. The analysis leads to a two level characterization of information granules: the core part that has a hyper-box shape and the residual part that has complex topology and does not convey any pattern regularity.**

***Keywords* – Minkowski FCM, Information Granulation, Hyper-boxes, Deformation Quantification**

I. Introduction

Information granulation is the task of finding structures in heterogeneous data by noninvasive exploration that does not make any assumption on the statistical nature of data. As a result of granulation, we form a collection of information granules that generalize (abstract) the essential structure in data [1].

A key feature of information granules is their transparency to end users. This transparency can be accomplished by a suitable representation mechanism that involves all stages from the generation of granules through to their representation for end-users. Interpretable information granules can be effectively employed in data mining tasks, whose main objective is to find regular and ultimately understandable patterns from data [2].

The approach adopted in this paper for information granulation is based on that presented in [3] and consists in decomposing complex topologies of data structures through hierarchical partitioning of the structures into core and residual parts. The cores are represented as granular prototypes characterized by a transparent boxlike shape that can be interpreted as a decomposable relation among data features.

To generate the granular prototypes, the Minkowski FCM has been investigated as a tool that allows a degree of control over the geometry of the information granules identified through clustering. The transition from hyper-ellipsoids to hyper-boxes, prompted by the change of the order of the Minkowski distance from 2 to ∞, is somewhat distorted by two concurrent factors subject of our study: the Minkowski order and the interactions among clusters.

The first deformation factor is due to the estimation error introduced when approximating the Tchebychev distance – which characterizes boxlike granules – with Minkowski distances. To investigate such phenomenon, a first distortion measure is formalized to quantify the approximation error in relation to the Minkowski order. Experimental results show that such distortion measure quickly decreases as the Minkowski order increases, but calculi with too high orders are hampered by perturbation errors involved by finite precision arithmetic. As a consequence the distortion measure can be used as a tool to estimate the best Minkowski order that allows the approximation to the Tchebychev on a given machine.

The second deformation effect is due to interactions among clusters, which distort the boxlike shape of clusters to meet the constraints required by FCM. To deal with such problem, a second deformation measure has been formalized. This measure evaluates the distortion of each cluster alpha-cut with an associated "ideal" hyper-box. With the aid of such measure, boxlike granular prototypes can be generated within an accepted deformation threshold.

The paper is organized as follows. In the next section, the Minkowski FCM is described in detail. Then, the approach for extracting interpretable information granules is formalized, and a study on the deformation effects is presented. Finally, some illustrative material is portrayed so as to exemplify the key results of our study from a practical viewpoint.

II. Minkowski FCM Algorithm

Let $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\}\subset[0,1]^n$ be a set of data patterns to be clustered into $c>1$ fuzzy partitions by minimizing the following objective function:

$$Q=\sum_{i=1}^{c}\sum_{k=1}^{N}u_{ik}^2 d_{ik}^{(p)} \quad (1)$$

where $\left[u_{ik}\right]_{i=1,2,\ldots,c}^{k=1,2,\ldots,N}$ is the partition matrix, which is subject to:

$$\begin{aligned} &(i) \quad \forall k: \sum_{i=1}^{c} u_{ik} = 1 \\ &(ii) \quad \forall i,k: 0 \le u_{ik} \le 1 \\ &(iii) \quad \forall i: 0 < \sum_{k=1}^{N} u_{ik} < N \end{aligned} \tag{2}$$

The distance between a pattern $\mathbf{x}_k$ and a cluster prototype $\mathbf{v}_i$ is denoted by $d_{ik}^{(p)}$ and is defined as:

$$d_{ik}^{(p)} = \sqrt[p]{\sum_{j=1}^{n} \left| x_{kj} - v_{ij} \right|^p} \tag{3}$$

being $p \ge 1$ the order of the Minkowski distance, $\mathbf{x}_k = \left(x_{k1}, x_{k2}, \ldots, x_{kn}\right)$ and $\mathbf{v}_i = \left(v_{i1}, v_{i2}, \ldots, v_{in}\right)$. The objective of Minkowski FCM is to find a proper choice of the partition matrix $\left[u_{ik}\right]$ and prototypes $\mathbf{v}_1, \ldots, \mathbf{v}_c$ such that (1) is minimal while satisfying constraints in (2).

Such minimization problem has been tackled specifically for Minkowski distances in [5]. There, the authors propose a gradient-free minimization method called "iterative majorization" [6]. However, such method cannot be applied for $p > 2$, hence we choose to extend the classical FCM procedure, which is standard to a high extent, and is based on alternative optimization of the partition matrix $\left[u_{ik}\right]$ and the set of prototypes $\mathbf{v}_1, \ldots, \mathbf{v}_c$ [4]. These two steps are repeated iteratively until no significant change of the objective function is registered. The update formula of the partition matrix does not depend on the chosen distance function and is derived by means of Lagrange multipliers. After some straightforward calculations, the following expression can be obtained:

$$u_{st}\left[\tau+1\right] = \frac{1}{\sum_{j=1}^{c} \frac{d_{st}^{(p)}\left[\tau\right]}{d_{jt}^{(p)}\left[\tau\right]}} \tag{4}$$

Here, we denote with τ the iteration number. Differently from the partition matrix, the derivation of the update formula for the cluster prototypes depends on the analytical form of the distance function. Specifically, the update formula is determined by solving the following equation system:

$$\frac{\partial Q}{\partial v_{st}} = 0, \quad \begin{aligned} s &= 1,2,\ldots,c \\ t &= 1,2,\ldots,n \end{aligned} \tag{5}$$

Such equation system has a simple solution for $p = 2$ while cannot be solved analytically for a generic order p. For such reason, the update formula for each prototypes follows the gradient-descent scheme:

$$v_{st}\left[\tau+1\right] = v_{st}\left[\tau\right] - \alpha \frac{\partial Q}{\partial v_{st}} \tag{6}$$

Such scheme introduces a novel hyper-parameter α whose value is critical for the convergence of the iterations. However, we found empirically that the following choice always leads to convergence:

$$\alpha \propto 1/N \tag{7}$$

The derivatives of the objective function with respect to each prototype are calculated as follows:

$$\frac{\partial Q}{\partial v_{st}} = \sum_{k=1}^{N} u_{sk}^2 \frac{\partial}{\partial v_{st}} d_{sk}^{(p)} \tag{8}$$

where:

$$\frac{\partial d_{sk}^{(p)}}{\partial v_{st}} = -\left(d_{sk}^{(p)}\right)^{1-p} \cdot \left| x_{kt} - v_{st} \right|^{p-1} \cdot \operatorname{sgn}\left(x_{kt} - v_{st}\right) \tag{9}$$

being $\operatorname{sgn}(\cdot)$ the sign function.

Although the case of $p = +\infty$ is not included in such computation, it should be remarked that the Tchebychev distance can be handled through a fuzzy logic approximation of the 'max' operation as discussed in [7]. In particular, for $p \gg 1$ the corresponding Minkowski distance quickly approaches the Tchebychev distance, hence the Minkowski FCM can be applied to derive box-shaped information granules by using a gradient descent scheme .

In summary, the clustering algorithm can be described as a sequence of the following steps:

```
1)  Randomly define prototypes v_1, ..., v_c
2)  Repeat
    a)  Compute the partition matrix according to (4)
    b)  Compute prototypes according to (6)
3)  Until a termination criterion is satisfied
```

The most commonly adopted termination criterion entails the variation of the objective function. Specifically, it is met when:

$$Q\left[\tau+1\right] - Q\left[\tau\right] < \varepsilon \tag{10}$$

being ε a user-defined threshold. Note that while the update formula of the partition matrix is the same as for the standard FCM, the update of the prototypes is more time consuming.

III. INFORMATION GRANULES GENERATION

The Minkowski order p used for clustering is user defined and is chosen so as to achieve a control over the shape of the information granules. It should be remarked that as $p \to +\infty$, the distance function approaches the Tchebychev distance defined as:

$$d_{ik}^{(\infty)} = \max_{j=1,2,\ldots,n} \left| x_{kj} - v_{ij} \right| \tag{11}$$

Such distance function is well suited to identify box-shaped information granules that can be represented as a decomposable relation $\mathbf{R}$ identified as a Cartesian product of one-dimensional information granules:

$$\mathbf{R} = R_1 \times R_2 \times \cdots \times R_n \quad (12)$$

Such decomposability leads to transparent information granules and can be effectively adopted in data mining contexts. On the other hand, for $p = 2$, the Minkowski FCM reduces to the classical FCM based on Euclidean distance, which is used to extract ellipsoidal-shaped information granules that are not decomposable in the aforementioned sense. As a consequence, an interesting study concerns the quantification of the deformation effects of information granules in relation to the Minkowski order.

Once the final partition matrix has been computed, c multidimensional fuzzy sets $\mathbf{G}_1, \ldots, \mathbf{G}_c$ can be defined by the following membership functions:

$$u_i(\mathbf{x}) = \begin{cases} \delta_{ij}, \mathbf{x} = \mathbf{v}_j \text{ for some } j \\ \left(\sum_{j=1}^{c} \frac{d^{(p)}(\mathbf{x}, \mathbf{v}_i)}{d^{(p)}(\mathbf{x}, \mathbf{v}_j)} \right)^{-1}, \text{ otherwise} \end{cases} \quad (13)$$

being:

$$d^{(p)}(\mathbf{x}, \mathbf{v}) = \sqrt[p]{\sum_{h=1}^{n} |x_h - v_h|^p} \quad (14)$$

and δ_{ij} the Kronecker's symbol. Definition (13) is a direct extension of equation (4) in a functional form.

In a small neighborhood of $\mathbf{v}_i$, the contour levels of the i-th fuzzy set have approximately the same shape of the distance function. Indeed, if for $j \neq i$ $d^{(p)}(\mathbf{x}, \mathbf{v}_i) \ll d^{(p)}(\mathbf{x}, \mathbf{v}_j)$, then:

$$u_i(\mathbf{x}) \approx 1 - \beta \cdot d^{(p)}(\mathbf{x}, \mathbf{v}_i) \quad (15)$$

being β a positive constant. As the distance between $\mathbf{x}$ and $\mathbf{v}_i$ increases, the interactions amongst the different clusters become stronger, and the shape of the contour levels will distort significantly.

By virtue of (15), for $p \gg 1$ the contour levels of the fuzzy sets assume the shape of a hyper-box in the neighborhood of each prototype. To assess the distortion involved in approximating the Tchebychev distance with a Minkowski distance, the following measure can be adopted:

$$\Delta^{(p)} = \sum_{i=1}^{c} \int_{[0,1]^n} |u_i(\mathbf{x}) - \overline{u}_i(\mathbf{x})| d\mathbf{x} \quad (16)$$

being:

$$\overline{u}_i(\mathbf{x}) = \begin{cases} \delta_{ij}, \mathbf{x} = \mathbf{v}_j \text{ for some } j \\ \left(\sum_{j=1}^{c} \frac{d^{(\infty)}(\mathbf{x}, \mathbf{v}_i)}{d^{(\infty)}(\mathbf{x}, \mathbf{v}_j)} \right)^{-1}, \text{ otherwise} \end{cases} \quad (17)$$

The measure $\Delta^{(p)}$ is theoretical and cannot be evaluated in any analytical way. Nevertheless, it can be estimated as:

$$D^{(p)} = \frac{1}{N} \sum_{i=1}^{c} \sum_{k=1}^{N} |u_i(\mathbf{x}_k) - \overline{u}_i(\mathbf{x}_k)| \quad (18)$$

From a theoretical point of view, we expect that:

$$\lim_{p \to \infty} D^{(p)} = 0 \quad (19)$$

However, when the Minkowski FCM is executed on a digital computer, it is possible that for very high values of p numerical errors significantly perturb the convergence (19). As a consequence, the measure (18) can be used to evaluate the maximum value of p that allows the best approximation of the Tchebychev distance on a given digital computer.

Once the most suitable value of p has been selected, the geometric properties of the information granules can be studied. As an example, in [3] an approach is proposed to evaluate the deformation of membership grades due to cluster interactions when box-shaped information granules are required. To quantify such deformation, the γ-cuts of each fuzzy set are considered:

$$[\mathbf{G}_i]_\gamma = \{\mathbf{x} | u_i(\mathbf{x}) \geq \gamma\} \quad (20)$$

For each γ-cut, an "ideal" hyper-box is built according to:

$$[\Gamma]_\gamma = [\![\mathbf{v}_i - \mathbf{l}_i^{(\gamma)}, \mathbf{v}_i + \mathbf{s}_i^{(\gamma)}]\!] \quad (21)$$

being $\mathbf{l}_i^{(\gamma)} = \left(l_{i1}^{(\gamma)}, l_{i2}^{(\gamma)}, \ldots, l_{in}^{(\gamma)}\right)$ and $\mathbf{s}_i^{(\gamma)} = \left(s_{i1}^{(\gamma)}, s_{i2}^{(\gamma)}, \ldots, s_{in}^{(\gamma)}\right)$ such that:

$$\overline{u}_i\left(\mathbf{v}_i - l_{ij}^{(\gamma)} \cdot \mathbf{e}_j\right) = \overline{u}_i\left(\mathbf{v}_i + s_{ij}^{(\gamma)} \cdot \mathbf{e}_j\right) = \gamma \quad (22)$$

for each $j = 1, 2, \ldots, n$, and:

$$\mathbf{e}_j = \left(\underbrace{0, 0, \ldots, 0}_{j-1}, \ 1, \ \underbrace{0, 0, \ldots, 0}_{n-j} \right) \quad (23)$$

To avoid unlimited hyper-boxes, the value of γ must belong to the open interval $]1/c, 1]$. Based on the definition of $[\Gamma]_\gamma$, the deformation is measured as follows:

$$\phi_i(\gamma) = \sum_{\mathbf{u} \in \mathbf{V}_i^{(\gamma)}} |\gamma - u_i(\mathbf{u})| \quad (24)$$

being $\mathbf{V}_i^{(\gamma)}$ the set of all vertices of the hyper-interval $[\Gamma]_\gamma$:

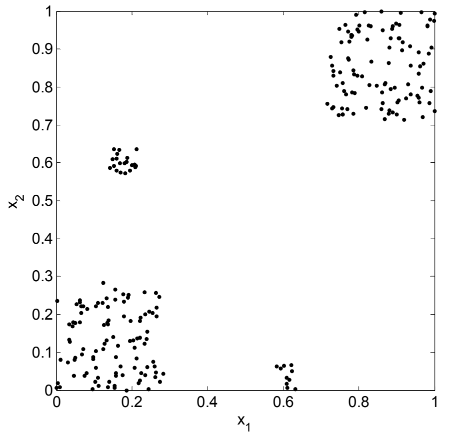

Figure 2: Two-dimensional synthetic dataset with four visible clusters of unequal size

$$\mathbf{V}_i^{(\gamma)} = \left\{ \mathbf{v}_i + (o_1, \dots, o_n) \middle| o_j = -l_{ij}^{(\gamma)} \vee o_j = s_{ij}^{(\gamma)} \right\} \quad (25)$$

The analysis of the functions ϕ_i helps in the definition of box-shaped granular prototypes. As an example, a threshold $\phi_{\max}$ can be fixed and the corresponding value $\gamma_{\max}^{(i)}$ such that:

$$\phi_i\left(\gamma_{\max}^{(i)}\right) = \phi_{\max} \quad (26)$$

can be considered. The value $\gamma_{\max}^{(i)}$ can be then used to define the granular prototypes as:

$$\mathbf{B}_i = \left[\!\left[\mathbf{v}_i - \mathbf{l}_i^{\left(\gamma_{\max}^{(i)}\right)}, \mathbf{v}_i + \mathbf{s}_i^{\left(\gamma_{\max}^{(i)}\right)} \right]\!\right] \quad (27)$$

The derived granular prototypes serve as "focal points" of the structure underlying data and can be employed as building blocks for further information processing. Indeed, the following relation holds:

$$[0,1]^n = \Re \cup \bigcup_{i=1}^{c} \mathbf{B}_i \quad (28)$$

where $\Re$ is the residual part of the data domain that is not covered by any information granule. By virtue of relation (28), the input domain is decomposed in a "core" part, defined by transparent granular prototypes of clear semantics, and a residual part of more complex topology that does not carry an evident pattern of regularity.

IV. Illustrative Example

As an illustrative example, we consider a synthetic dataset involving four clusters, as depicted in Figure 1. The two larger data groupings consist of 100 data-points and the two smaller ones have 20 and 10 data-points respectively.

The dataset has been clustered according to four Minkowski distances corresponding to $p = 2, 4, 6, 50$. After carrying out the Minkowski FCM procedure, four multidimensional fuzzy sets have been defined according to (13). As it may be observed from Figure 5 the shape of clusters can be controlled by choosing an appropriate Minkowski distance. For $p = 2$ the fuzzy sets have the typical spherical representation, which gradually becomes boxlike for increasing values of p. For $p \gg 1$, the contour levels of each fuzzy membership function appear neatly sharp, resulting in a close approximation of the clusters attainable with Tchebychev distance.

In Figure 2, the distortion estimation $D^{(p)}$ defined in (18) is plotted for $p = 2, 3, \dots, 100$. The plot can be conveniently divided in two parts. In the first part ($p \le 62$) we notice a quick decrease of the distortion error that is coherent with the theoretical relationship between Minkowski and Tchebychev distance (19). For $p > 62$, the error perturbations due to the finite precision arithmetic[1] become increasingly significant so as to cause a random trend. As a consequence, the Minkowski distance for $p = 62$ appears as the best approximation of the Tchebychev distance for the problem at hand. It is remarkable, however, that good approximations (i.e. $D^{(p)} < 10^{-3}$) can be achieved for $p > 10$.

To evaluate the best representation of granular prototypes for different Minkowski distances, the plots of the functions $\phi_i(\gamma)$ defined in (24) are portrayed in Figure 3. Each graph consists of four plots for the same fuzzy set corresponding to $p = 2, 4, 6, 50$. As it may be easily observed, the adoption of the Euclidean distance ($p = 2$) clearly clashes with the requirement of box-shaped information granules because distortion is evident by high values of $\phi_i(\gamma)$ even for small values of γ. Conversely, higher values of p determine very similar behaviour, which is roughly characterized by a convex trend. Indeed, for $\gamma \approx 1$, interactions amongst clusters is not relevant, thus the distortion value $\phi_i(\gamma)$ is close to zero. As γ decreases, the influence of each information granule becomes significant so that the distortion value raises consequently. For

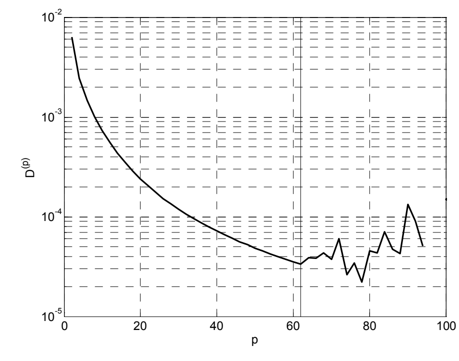

Figure 1: Distortion evaluation for Minkowski distance in approximating Tchebychev distance

[1] All simulations were run on a IEEE754 double-precision computer

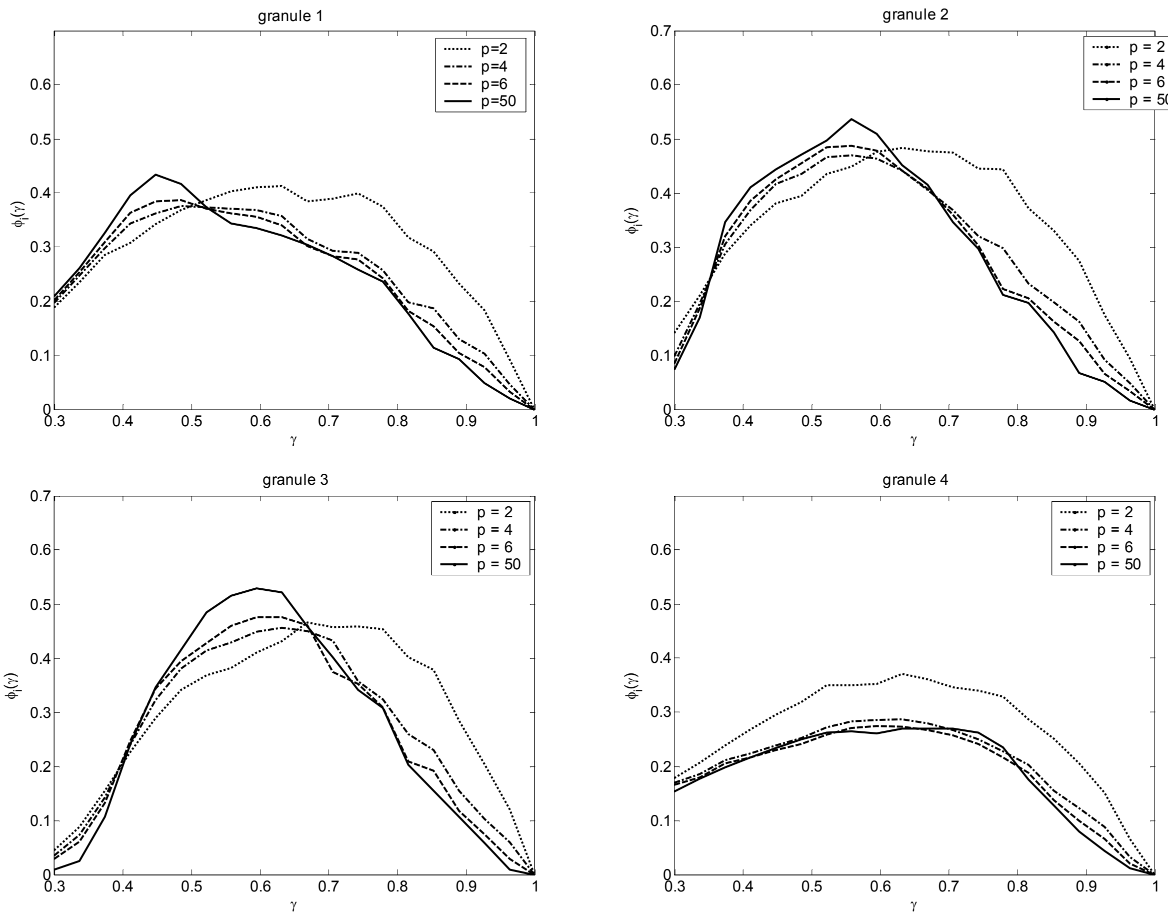

Figure 3: Evaluation of the deformation effects for each information granule with different Minkowski orders

small values of γ, i.e. $\gamma \approx 1/c$, the distortion measure becomes small since the membership value of each fuzzy set is asymptotically constant for large distances. Such behaviour suggests to select the value of γ approximately within the range $[0.7,1]$ to extract meaningful granular prototypes.

To derive granular prototypes, we select the maximum distortion measure $\phi_{max} = 0.1$. Based on such threshold, the granular prototypes corresponding to each cluster have been derived for $p = 2$ and $p = 50$ and represented in Figure 4. As it may be observed, hyper-boxes derived for $p = 2$ appear too narrow, especially for the two biggest clusters, thus excluding a significant number of pattern. Conversely, hyper-boxes derived for $p = 50$ better capture the underlying structure of data.

V. CONCLUSIONS

The approach adopted in this paper for data understanding consists in the decomposition of complex topologies described by data into a regular structure and a residual part. The regular structure is defined in terms of granular prototypes, whose transparency is determined by their representation in form of hyper-boxes.

To generate the granular prototypes, the Minkowski FCM has been investigated. Specifically, two deformation effects that perturb the desired boxlike shape are analysed. The first

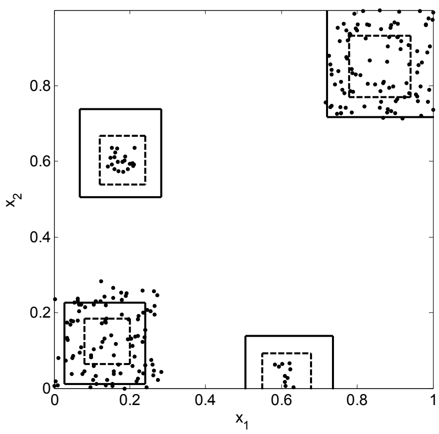

Figure 4: Granular prototypes generated for p=2 (dashed boxes) and p=50 (solid boxes) with $\phi_{max} = 0.1$

Figure 5: Contour levels of the membership function of the bottom-left information granule for p=2 (a), p=4(b), p=6 (c) and p=50 (d)

deformation effect is due to the estimation error introduced when approximating the Tchebychev distance – which characterizes boxlike granules – with Minkowski distances. Experimental results show that such approximation error quickly decreases as the Minkowski order increases, but calculi with too high orders are hampered by perturbation errors involved by finite precision arithmetic. As a consequence, we found that the Minkowski order should be chosen between 10 and 60 to avoid both rough approximation and perturbation errors.

The second deformation effect is due to interactions among clusters, that distort the boxlike shape of clusters to meet the constraints required by FCM. To deal with such problem, a deformation measure has been formalized. With the aid of such measure, boxlike granular prototypes can be generated within an accepted deformation threshold.

References

[1] Bargiela, A., Pedrycz, W., *Granular Computing. An Introduction*, Kluwer Academic Publishers, 2003

[2] Frawley W.J., Shapiro, G., Matheus, C.J., "Knowledge Discovery in Databases: An Overview", in Shapiro and Frawley (eds.), *Knowledge Discovery in Databases*, AAAI/MIT, pp. 1-27, 1991

[3] Bargiela, A., Pedrycz, W., "A model of granular data: a design problem with the Tchebyschev FCM", *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, Springer-Verlag, 2003

[4] Bezdek, J. C., *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, NY, 1981

[5] Groenen, P.J.F., Jajuga, K., "Fuzzy clustering with squared Minkowski distances", *Fuzzy Sets and Systems*, vol. 120(2), pp. 227-237, 2001

[6] Groenen, P.J.F., Mathar, R., Heiser, W.J., "The majorization approachto multidimensional scaling for Minkowski distances", *Journal of Classification*, vol.12, pp. 3–19, 1995

[7] Castellano, G., Fanelli, A.M., Mencar, C., "An Empirical Risk Functional to Improve Learning in a Neuro-Fuzzy Classifier", in *IEEE Transactions on Systems, Man and Cybernetics (part B)*, vol. 34(1), pp. 725-731, 2004