Information Granulation: a Search for Data Structures

Witold PEDRYCZ * and Andrzej BARGIELA **

** University of Alberta, Edmonton, Canada*
** Systems Research Institute, Polish Academy of Sciences, Poland*
(pedrycz@ee.ualberta.ca)

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

Abstract Revealing a structure in data is of paramount importance in a broad range of problems of information processing. In spite of the specificity of the problem in which such analysis is realized, there is an evident commonality of all these pursuits worth emphasizing. One can distinguish between a core and a residual of the data structure. In this study, we propose a formal environment supporting these concepts and develop its algorithmic fabric. The algorithms leading to the development of information granules lend themselves to the Fuzzy C-Means (FCM) equipped with the Tchebyschev (l_∞) metric. The paper offers a novel contribution of a gradient-based learning of the prototypes developed in this form of clustering. The l_∞ metric promotes a design of easily interpretable hyperboxes. In this setting, we quantify the notion of the core and residual part of the data. An interaction between information granules is discussed as well.

1. Introduction and problem statement

When making sense of data, no matter if engaged in classification, modeling or data mining problems, there is a uniform perspective the data sets can be placed and processed. Denote the data set by D (it does not matter what a specific format these data points assume). We can distinguish between a core and residual structure of the data and write down this observation in the concise form as follows

$$D = \mathbf{B}_1 \cup \mathbf{B}_2 \cup \ldots \cup \mathbf{B}_c \cup R$$

By a core part of data we mean a subset of data that is quite dense (so the points are confined to a relatively small volume in the data space), exhibits a visible geometry and is significant in size 9cardinality). All these features of the core allow us to express it as a clearly defined and relatively easily describable structure – a collection of information granules [8][9], namely $\mathbf{B}_1 \cup \mathbf{B}_2 \cup \ldots \cup \mathbf{B}_c$ In this study we subscribe to the language of hyperboxes treated as basic geometric constructs capturing the core. The residual portion of D (denoted here by R) is quite loose, significantly distributed which does not quantify quite easily.The distinction between the core and residual structure can be interpreted in many different ways depending on the specific area. For instance, in pattern recognition the core part comes in the form of patterns belonging to the same category (class) while the residual is a substantial mix of patterns belonging to various classes that are far more difficult to classify.

2. Fuzzy clustering as a vehicle of information granulation

In what follows, we set up all necessary notations. The set of data (patterns) is denoted by $\mathbf{X}$, $\mathbf{X}= \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ while each pattern is an element in the n-dimensional unit hypercube, that is $[0,1]^n$. The objective is to cluster $\mathbf{X}$ into "c" clusters and the problem is cast as an optimization task (objective function based optimization)

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

where $U = [u_{ik}]$, i=1,2, ..,c, k=1, 2, …,N is a partition matrix describing clusters in data. The distance function (metric) between the k-th pattern and i-th prototype is denoted by d_{ik}. $d_{ik} = \text{dist}(\mathbf{x}_k, \mathbf{v}_i)$ while $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c$ are the prototypes characterizing the clusters. The type of the distance implies certain geometry of the clusters one is interested in exploiting when analyzing the data. For instance, it is well known that a commonly used Euclidean distance promotes an ellipsoidal shape of the clusters. Concentrating on the parameters to be optimized, the above objective function reads now as

Min Q with respect to $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c$ and U

with its minimization carried out for the partition matrix as well as the prototypes. With regard to the prototypes (centroids), we end up with a constraint-free optimization while the other one calls for the constrained optimization. The constraints assure that U is a partition matrix meaning that the following well-known conditions are met

$$\sum_{i=1}^{c} u_{ik} = 1 \quad \text{for all } k = 1,2,..,N \qquad 0 < \sum_{k=1}^{N} u_{ik} < N \quad \text{for all } i = 1,2,..,c$$

The choice of the distance function is critical to our primary objective of achieving the transparency of the findings. We are interested in such distances whose equidistant contours are "boxes" with the sides parallel to the coordinates. The Tchebyschev distance (l_∞ distance) is a distance satisfying this property. The boxes are decomposable that is the region within a given equidistant contour of the distance can be treated as a decomposable relation R in the feature space, viz.$R = A \times B$ where A and B are sets (or more generally information granules) in the corresponding feature spaces. It is worth noting that the Euclidean distance does not lead to the decomposable relations in the above sense (as the equidistant regions in such construct are spheres or ellipsoides). The above clustering problem known in the literature as an l_∞ FCM was introduced and discussed by Bobrowski and Bezdek [2] more than 10 years ago. Some recent generalizations can be found in [5]. This motivation behind the introduction of this type of distance was the one about handling data structures with "sharp" boundaries (clearly the Tchebyschev distance is more suitable with ths regard than the Euclidean distance). The solution proposed in [2] was obtained by applying a basis exchange algorithm.

The FCM optimization procedure is standard to a high extent [1] and consists of two steps that is a determination of the partition matrix and calculations of the prototypes. The use of the Lagrange multipliers converts the constrained problem into its constraint-free version. The original objective function as shown above is transformed to the form

$$V = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^2 d_{ik} + \lambda\left(\sum_{i=1}^{c} (u_{ik} - 1)\right)$$

with λ being a Lagrange multiplier. Straightforward optimization leads to the expression

$$u_{st} = \frac{1}{\sum_{j=1}^{c} \frac{d_{st}}{d_{jt}}}$$

The determination of the prototypes is more complicated as the Tchebyschev distance does not lead to a closed-type expression (unlike the standard FCM with the Euclidean distance). Let us start with the objective in which the distance function is spelled out in an explicit manner

$$Q = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^2 \max_{j=1,2,...,n} | x_{kj} - v_{ij} |$$

The minimization of Q carried out with respect to the prototype (more specifically its t-th coordinate) follows a gradient-based scheme

$$v_{st}(iter + 1) = v_{st}(iter) - \alpha \frac{\partial Q}{\partial v_{st}}$$

where α is an adjustment rate (learning rate) assuming positive values. This update expression is iterative; we start from some initial values of the prototypes and keep modifying them following the gradient of the objective function. The detailed calculations of the gradient lead to the expression

$$\frac{\partial Q}{\partial v_{st}} = \sum_{k=1}^{N} u_{sk}^2 \frac{\partial}{\partial v_{st}} \{\max_{j=1,2,...,n} | x_{kj} - v_{sj} |\}$$

Let us introduce the following shorthand notation $A_{kst} = \max_{\substack{j=1,2,...,n \\ j \neq t}} | x_{kj} - v_{sj} |$ Evidently, A_{kst} does not depend on v_{st}. This allows us to concentrate on the term that affects the gradient. We rewrite the above expression for the gradient as follows

$$\frac{\partial Q}{\partial v_{st}} = \sum_{k=1}^{N} u_{sk}^2 \frac{\partial}{\partial v_{st}} \{\max(A_{kst}, | x_{kt} - v_{st} |)\}$$

The derivative is nonzero if A_{kst} is less or equal to the second term standing in the expression,

$$A_{kst} \leq | x_{kt} - v_{st} |$$

Next, if this condition holds we infer that the derivative is equal to either 1 or –1 depending on the relationship between x_{kt} and v_{st}, that is –1 if $x_{kt} > v_{st}$ and 1 otherwise. Putting these conditions together, we get

$$\frac{\partial Q}{\partial v_{st}} = \sum_{k=1}^{N} u_{sk}^2 \begin{cases} -1 \text{ if } A_{kst} \leq | x_{kt} - v_{st} | \text{and } x_{kt} > v_{st} \\ +1 \text{ if } A_{kst} \leq | x_{kt} - v_{st} | \text{and } x_{kt} \leq v_{st} \\ 0 \text{ otherwise} \end{cases}$$

The primary concern that arises about this learning scheme is not the one about a piecewise character of the function (absolute value) that is not (a concern that could be easily raised from the formal standpoint) but a fact that the derivative zeroes for a significant number of situations. This may result in a poor performance of the optimization method so it could be easily trapped in case the overall gradient becomes equal to zero. To enhance the method, we relax the binary character of the predicates (less or greater than) standing in the above expression. These predicates are Boolean (two-valued) as they return values equal to 0 or 1 (which translates into an expression "predicate is satisfied or it does not hold). The modification comes in the form of a degree of satisfaction of this predicate, meaning that we compute a multivalued predicate of the form "Degree(a is included in b)" that returns 1 if a is less or equal to b. Lower values of the degree arise when this predicate is not fully satisfied. This form of augmentation of the basic concept was introduced in [3, 4, 6, 7] in conjunction to studies in fuzzy neural networks and relational structures (fuzzy relational equations).

The degree of satisfaction of the inclusion relation is equal to " $\text{Degree}(a \text{ is included in } b) = a \rightarrow b$ " where a and b are in the unit interval. The implication operation $\rightarrow$ is a residuation operation, cf. [6 7]. Here we consider a certain implementation of such operation where the implication is implied by the product t-norm, namely

$$a \rightarrow b = \begin{cases} 1 & \text{if } a \le b \\ b/a & \text{otherwise} \end{cases}$$

Using this construct, we express the derivative as follows

$$\frac{\partial Q}{\partial v_{st}} = \sum_{k=1}^{N} u_{sk}^2 \begin{cases} -(A_{kst} \rightarrow | x_{kt} - v_{st} |) & \text{if } x_{kt} > v_{st} \\ (A_{kst} \rightarrow | x_{kt} - v_{st} |) & \text{if } x_{kt} \le v_{st} \end{cases}$$

In the overall scheme, this expression will be used to update the prototypes of the clusters.

Summarizing, the clustering algorithm arises as a sequence of the following steps

repeat

- compute partition matrix U;
- compute prototypes using the partition matrix obtained in the first phase. (It should be noted that the partition matrix does not change at this stage and all updates of the prototypes work with this matrix. This phase is more time consuming in comparison with the FCM method equipped with the Euclidean distance)

until a termination criterion satisfied

Both the termination criterion and the initialization of the method are standard. The termination takes into account changes in the partition matrices at two successive iterations that should not exceed a certain threshold level. The initialization of the partition matrix is random.

3. Granular data modeling: a classification perspective

The distinction between the core and residual structure in data implies a number of interesting findings that can be expressed in the language of pattern recognition. and cast in this framework. This somewhat brings us back to the underlying concept we demonstrated in Section 1. The structure identified in the data through the l_∞ clustering gives rise to the following classification rule

- assign $\mathbf{x}$ to class ω_i if $u_i(\mathbf{x}) > \max_{j \neq i} u_j(\mathbf{x})$

The computations of the membership grades in the above expression are done in the form

$$u_i(\mathbf{x}) = \frac{1}{\sum_{j=1}^{c} \frac{d(\mathbf{x}, \mathbf{V}_i)}{d(\mathbf{x}, \mathbf{V}_j)}}$$

It is interesting to visualize the classification boundaries as they become established in terms of the membership grades. The Tchebyschev distance results in the box-like boundaries shown in Figure 1; here c = 3 with the prototypes equal to $\mathbf{v}_1$ = (0.22, 0.21), $\mathbf{v}_2$ =(0.66, 0.63), and $\mathbf{v}_3$ =(0.65, 0.90).

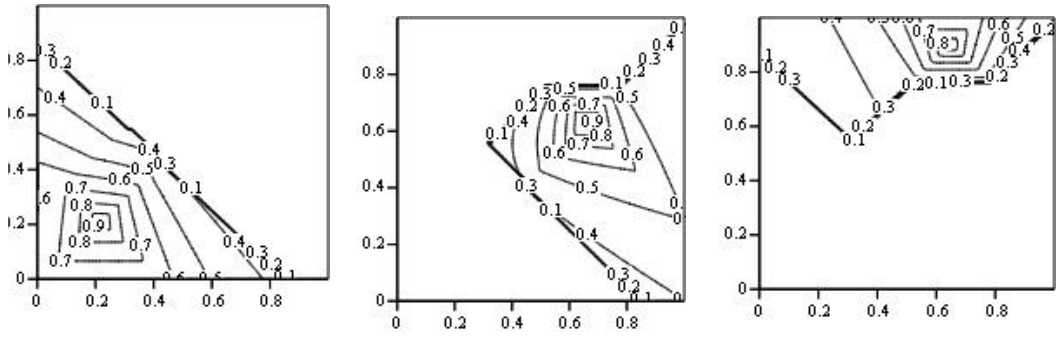

Figure 1. Classification boundaries for three classes obtained for the Tchebyschev distance

One can note that the classification boundaries become more complicated once we move apart from the core portion of the clusters. The cores themselves come with a rectangular format of the class boundaries that fully complies with the type of distance being considered there. From the classification standpoint, we may say that the core part of data requires a simple type of a classifier whereas to deal with a residual section of patterns we require a classifier of substantially higher level of complexity. Evidently, most of classification error comes associated with the residual structure of the patterns.

4. Concluding comments

In the description of data, we have developed two main components, namely cores of the data that are well-structured in the form of hyperboxes in the feature space and a far less regular structure that is described analytically through an expression for membership grades but does not carry any clear geometric interpretation. The computing backbone of this approach is based on the well-known FCM technique equipped with the Tchebyschev distance. We have introduced a new way of optimizing the prototypes in this method that uses a gradient-based technique augmented by a logic-oriented mechanisms of gradient determination. The geometry and design of the hyperbox information granules have been discussed along with an important aspect of deformation of such granules. Furthermore a quantification of this effect is discussed.

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

[2] L. Bobrowski, J.C. Bezdek , C-Means clustering with the l_1 and l_∞ norms, *IEEE Trans. on Systems Man and Cybernetics*, 21, 1991, 545-554.

[3] D. Dubois and H. Prade, Fuzzy relation equations and causal reasoning, *FSS*, 75, 1995, pp. 119-134

[4] P. J.F. Groenen, K. Jajuga, Fuzzy clustering with squared Minkowski distances, *FSS*, 120, 2001, 227-237.

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

[6] W. Pedrycz, *Fuzzy Control and Fuzzy Systems*, RSP/Wiley, 1989.

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

[8] L.A. Zadeh, From computing with numbers to computing with words-from manipulation of measurements to manipulation of perceptions, *IEEE Trans. on Circuits and Systems*, 45, 1999, 105-119.