A model of granular data: a design problem with the Tchebyschev –based clustering

Andrzej Bargiela * and Witold Pedrycz **

*Department of Computing and Mathematics
The Nottingham Trent University, Nottingham NG1 4BU
United Kingdom
(andre@doc.ntu.ac.uk)
and
**Department of Electrical & Computer Engineering
University of Alberta, Edmonton, Canada
(pedrycz@ee.ualberta.ca)
and
Systems Research Institute, Polish Academy of Sciences
01-447 Warsaw, Poland

Abstract -We introduce a model of granular data emerging through a summarization and processing of numeric data. It supports data analysis and casts it in the setting of data mining. The structure of data is revealed through the FCM equipped with the Tchebyschev (l_∞) metric. The study offers a novel contribution to a gradient-based learning of the prototypes developed in the l_∞ –based data space. The l_∞ metric promotes a development of easily interpretable information granules, namely hyperboxes. A detailed discussion of their geometry is provided. In particular, we discuss a deformation effect of the hyperbox-shape of granules due to an interaction between the granules. We also show how the clustering gives rise to a two-level topology of information granules. A core part of the topology comes in the form of hyperbox information granules. A residual structure is expressed through detailed, yet difficult to interpret, membership grades. Illustrative examples including synthetic data are studied.

Keywords: information granulation through clustering, FCM, l_∞ metric (distance), hyperboxes, deformation effect in clustering, geometry of data space, data mining

I. INTRODUCTION

Clustering has been widely recognized as one of the dominant techniques of data analysis. The broad spectrum of the detailed algorithms and underlying technologies (fuzzy sets, neural networks, heuristic approaches) is impressive. In spite of this diversity, the key objective remains the same which is to understand the data. In this sense, clustering becomes an integral part of data mining [4], [15]. Data mining is aimed at making the findings that are inherently *transparent* to the end user. The transparency is accomplished through suitable knowledge representation mechanisms, namely a way in which generic data elements are formed, processed and presented to the user. The notion of information granularity becomes a cornerstone concept to be discussed in this context, cf. [15] [10].

The underlying idea is that in any data set we can distinguish between a core part of a structure of the data that is easily describable and interpretable in a straightforward manner and a residual part, which does not carry any evident pattern of regularity. The core part can be described in a compact manner through several information granules while the residual part does not exhibit any visible geometry and requires some formal descriptors such as membership formulas. The approach proposed in this study dwells on the standard FCM method equipped with a Tchebyschev distance that promotes a hyperbox geometry of the information granules (hyperboxes). Starting from the results of clustering, our objective is to develop information granules forming a core structure in the data set, provide their characterization and discuss an interaction between the granules that results in their deformation.

II. PROBLEM FORMULATION

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

technique implies a 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

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

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

The choice of the distance function is critical to our primary objective of achieving the transparency of the ensuing clusters (information granules). 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 illustration of the decomposability property is illustrated in Figure 1; it becomes obvious that a hyperbox produces a collection of intervals defined for individual features.

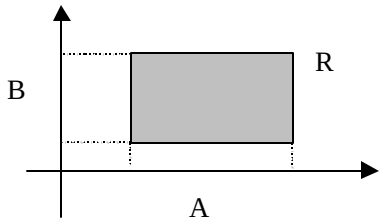


Figure 1. Decomposability property provided by the Tchebyschev distance; the region of equidistant points is represented as a Cartesian product of two sets in the corresponding feature spaces

The above clustering problem known in the literature as an l_∞ FCM was introduced and discussed by Bobrowski and Bezdek [3] more than 10 years ago. Some recent generalizations can be found in [7]. The motivation behind the introduction of this type of distance was to improve handling data structures with "sharp" boundaries (clearly the Tchebyschev distance is more suitable with this regard than the Euclidean distance). The solution proposed in [3] was obtained by applying a basis exchange algorithm.

In this study, as already mentioned, the motivation behind the use of the Tchebyschev distance is different. We are after the description of data structure and the related interpretability of the results of clustering so that the clusters can be viewed as basic models of associations existing in the data. Here, we concentrate on a gradient-based FCM technique enhanced with some additional convergence mechanism.

III. THE CLUSTERING ALGORITHM-DETAILS

The FCM optimization procedure is standard to a high extent [2] and consists of two steps: 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 introduced in the previous section is transformed to the form

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

with λ being a Lagrange multiplier. The above problem is then solved with respect to each pattern separately (k=1, 2,…,N). This leads to the equality

$$\frac{\partial V}{\partial u_{st}} = 0 \text{ and } \frac{\partial V}{\partial \lambda_t} = 0$$

s=1,2, …, c, t=1,2, …, N to be solved with respect to u_{st}. After some algebra we get

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

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 is spelled out in an explicit manner

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

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, namely

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

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. Aggregating 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 about a piecewise character of the function (absolute value) and its limited differentiability. The main concern is that the derivative zeroes for a significant range of the arguments. 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) discussed above. These predicates are Boolean (two-valued) as they return values equal to 0 or 1 (which translates into an expression "predicate is satisfied" or its negation). Instead of this predicate, we introduce a degree of satisfaction of the inclusion condition, meaning that we compute a multivalued predicate,

Degree(a is included in b)

that returns 1 if a is less or equal to b. Lower values of the degree of inclusion arise when this predicate is not fully satisfied. This form of augmentation of the basic concept was introduced in [5 6 13 14] 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. [13 14]. Here we consider a certain implementation of such operation where the implication is generated by the product t-norm, namely

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

Using this construct, we rewrite the above update formula as follows

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

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

Summarizing, the clustering algorithm comprises of a sequence of the following steps

repeat

- compute partition matrix,
- 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.

IV. EXPERIMENTAL STUDIES

As an illustrative example, we consider a synthetic data set involving 4 clusters, see Figure 2. The two larger data groupings consist of 100 data-points and the two smaller ones have 20 and 10 data-points respectively.

Table 1 gives a representative set of clustering results for 2 to 4 clusters. As expected, the two larger data groupings exercise dominant influence on the outcome of the FCM algorithms. Both Euclidean and Tchebyschev distance based FCM exhibit robust performance in that they find approximately the same

clusters in their successive runs (within the limits of the optimization convergence criterion). While most of the identified prototypes fall within the large data groupings, the Tchebyschev distance based FCM consistently manages to associate a prototype with one of the smaller data grouping

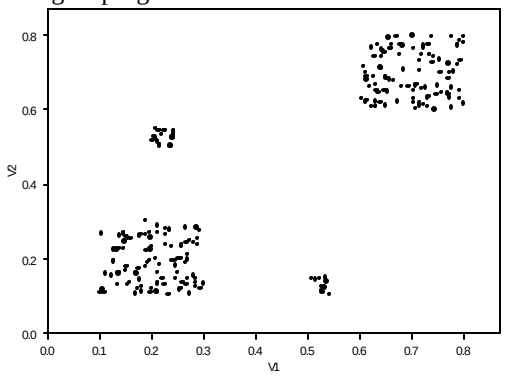


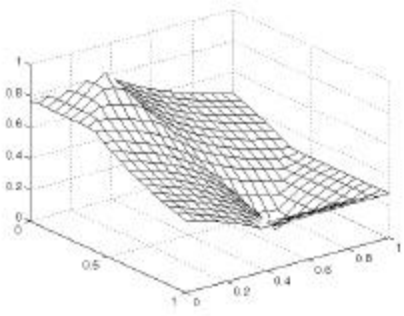
Figure 2. Two-dimensional synthetic data with four visible clusters of unequal size

Table 1. Prototypes identified by two FCM algorithms, with Euclidean and Tchebyschev distance measure respectively, for the varying number of clusters (the underlined prototypes correspond to the smaller data clusters)

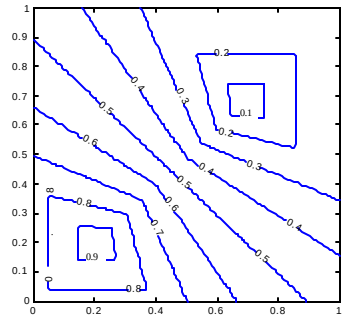
Number of clusters	Prototypes for FCM with Euclidean distance	Prototypes for FCM with Tchebyschev distance
2	0.6707 0.6706 0.2240 0.2236	0.2088 0.1998 0.6924 0.6831
3	0.2700 0.3011 0.6875 0.6841 0.2302 0.2127	0.7000 0.6847 0.2440 0.4914 0.2124 0.1852
4	0.2255 0.2035 0.2323 0.2479 0.6872 0.6814 0.6533 0.6588	0.7261 0.7377 0.2278 0.5178 0.2092 0.1846 0.6523 0.6498

(entries underlined in the table). This is clearly an advantageous feature of our modified FCM algorithm and confirms our assertion that the objective of enhancing the interpretability of data through the identification of decomposable relations is enhanced with Tchebyschev distance-based FCM.

The above results are better understood if we examine the cluster membership function over the entire pattern space. The visualization of the membership function for one of the two clusters, positioned in the vicinity of (0.2, 0.2), (c=2) is given in Figure 3.



(a)

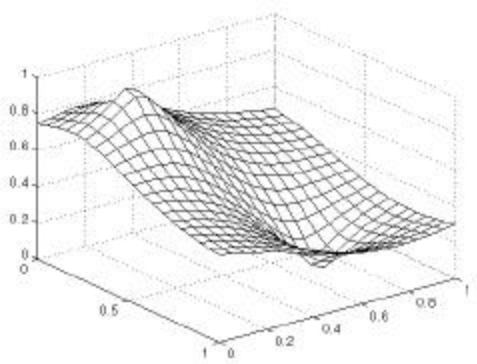


(b)

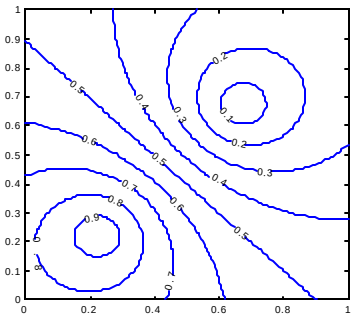
Figure 3. Visualization of the first cluster (membership function) centered around (0.2088 0.1998) : (a) 3D space and (b) contour plots.

It is easily noticed that that for higher values of the membership grades (e.g. β=0.9), the shape of contours is rectangular. This changes for lower values of the membership grades when we witness a gradual departure from this geometry of the clusters. This is an effect of interaction between the clusters that manifests itself in a deformation of the original rectangles. The deformation depends on the distribution of the clusters, their number and a specific threshold β being selected. The lower the value of this threshold is, the more profound departure from the rectangular shape. For higher values of β such deformation is quite limited. This suggests that when using high values of the threshold level the rectangular (or hyperbox) form of the core part of the clusters is fully legitimate.

Let us contrast these results with the geometry of the clusters constructed when using a Euclidean distance. Again, we consider two prototypes, as identified by the Euclidean distance based FCM, refer to Figure 4. The results are significantly different: the clusters are close to the Gaussian-like form and do not approximate well by rectangular shapes.



(a)



(b)

Figure 4. Visualization of the first cluster (membership function) centered around (0.2240 0.2236) : (a) 3D space and (b) contour plots. The Euclidean distance function was used in the clustering algorithm.

V. THE GEOMETRY OF INFORMATION GRANULES

As the contour plots of the clusters reveal (Figure 3), an interaction between the clusters become responsible for the deformation of the hyperbox shape of the cores. This poses an interesting question as to the structure of the data. With the development of the granular prototypes guided by the clustering algorithm, we can concisely describe the data in the form

$$D = B_1 \cup B_2 \cup \ldots \cup B_c \cup R$$

where D is a data set under discussion, B_i are granular prototypes (viz. hyperboxes around the prototypes) and R is a residual structure of the data set. Briefly speaking, hyerboxes and a residual component of the data structure can be described as follows

Hyperboxes: Hyperboxes constructed on the basis of the Tchebyschev distance for a given threshold level. The choice of this level depends on an acceptable level of deformation of the hyperboxes allowed in the structure. The hyperboxes can be decomposed into intervals for individual features.

Residual structure: A lack of explicit description, more complex geometry of the regions. The degree of membership to the hyperbox is expressed through a standard expression encountered in the FCM computing

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

where $\mathbf{x}$ is a pattern under discussion and $u_i(\mathbf{x})$ stands for its membership grade.

VI. 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 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 proposed approach to data analysis can be exploited in many different ways. Several promising avenues are as follows

(a) Data mining. Articulating the main pursuit of data mining in the language of well-defined, semantically sound and easily interpretable constructs. The information granules discussed here appear to provide good abstraction on which to base data mining activities. They are easy to interpret and thus cope with the underlying structure of data while leaving out the residual portion of data not exhibiting strong patterns of dependencies

(b) In any modeling pursuit, the above data description helps concentrate on the design of local models assigned to the core parts. The residual part of data can be handled separately with an anticipation that these data points may not lead to a model with a strongly manifested character

(c) In classification problems, the core part of the data implies a collection of simple classifiers while the residual part invokes more demanding and conceptually advanced classifiers such as neural networks.

Acknowledgments

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

VII. REFERENCES

1. A. Bargiela, *Interval and Ellipsoidal Uncertainty Models* In: Granular Computing, W. Pedrycz (ed.), Springer Verlag, 2001
2. J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, N. York, 1981.
3. 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.
4. K. Cios, W.Pedrycz, R. Swiniarski, *Data Mining Techniques*, Kluwer Academic Publishers, Boston, 1998.
5. D. Dubois and H. Prade, Fuzzy relation equations and causal reasoning, *Fuzzy Sets and Systems*, 75, 1995, pp. 119-134
6. S. Gottwald, Approximate solutions of fuzzy relational equations and a characterization of t-norms that define metrics for fuzzy sets, *Fuzzy Sets and Systems*, 75, 1995, pp. 189-201
7. P. J.F. Groenen, K. Jajuga, Fuzzy clustering with squared Minkowski distances, *Fuzzy Sets and Systems*, 120, 2001, 227-237.
8. F. H Öppner, F. Klawonn, R. Kruse, T. Runkler, *Fuzzy Cluster Analysis*, J. Wiley, Chicester, 1999.
9. A. Kandel, *Fuzzy Mathematical Techniques with Applications*, Addison-Wesley, Reading, MA, 1986.
10. O. Maimon, A. Kandel and M. Last, Information-theoretic fuzzy approach to data reliability and data mining, *Fuzzy Sets and Systems*, 117, 2001 pp. 183-194
11. Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic, Dordrecht, 1991.
12. W. Pedrycz, *Computational Intelligence: An Introduction*, CRC Press, Boca Raton, FL, 1997.
13. W. Pedrycz, F. Gomide, *An Introduction to Fuzzy Sets*, Cambridge, MIT Press, Cambridge, MA, 1998.
14. W. Pedrycz, *Fuzzy Control and Fuzzy Systems*, RSP/Wiley, 1989.
15. 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, 111-117.
16. 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.