knowledge and data, fuzzy clustering, guidance mechanisms, proximity, inclusion, partial supervision, uncertainty, entropy

Witold PEDRYCZ*, Adam GACEK*

KNOWLEDGE-BASED CLUSTERING AS A CONCEPTUAL AND ALGORITHMIC ENVIRONMENT OF BIOMEDICAL DATA ANALYSIS

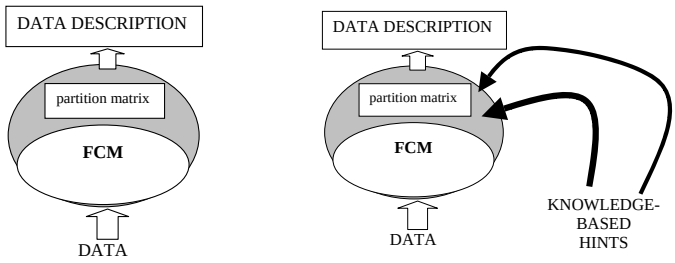
While a genuine abundance of biomedical data available nowadays becomes a genuine blessing, it also posses a lot of challenges. The two fundamental and commonly occurring directions in data analysis deal with its supervised or unsupervised pursuits. Our conjecture is that in the area of biomedical data processing and understanding where we encounter a genuine diversity of patterns, problem descriptions and design objectives, this type of dichotomy is neither ideal nor the most productive. In particular, the limitations of such taxonomy become profoundly evident in the context of unsupervised learning. Clustering (being usually regarded as a synonym of unsupervised data analysis) is aimed at determining a structure in a data set by optimizing a given partition criterion. In this sense, a structure emerges (becomes formed) without a direct intervention of the user. While the underlying concept looks appealing, there are numerous sources of domain knowledge that could be effectively incorporated into clustering mechanisms and subsequently help navigate throughout large data spaces. In unsupervised learning, this unified treatment of data and domain knowledge leads to the general concept of what could be coined as *knowledge-based clustering*. In this study, we discuss the underlying principles of this paradigm and present its various methodological and algorithmic facets. In particular, we elaborate on the main issues of incorporating domain knowledge into the clustering environment such as (a) partial labelling, (b) referential labelling (including proximity and entropy constraints), (c) usage of conditional (navigational) variables, (d) exploitation of external structure. Presented are also concepts of stepwise clustering in which the structure of data is revealed via a series of refinements of existing domain granular information.

1. INTRODUCTION

Unsupervised learning and fuzzy clustering in particular is omnipresent in our ongoing quest to understand data and produce its meaningful, concise, and user-oriented interpretation [1][2][6][7][8][9][10][11][16]. What forms a predominant trend in almost each clustering technique is its reliance on some optimization criterion. When accepted, this criterion guides the process of forming information granules – clusters [3]. Fuzzy C-Means (FCM) [2] is no exception to this paradigm. Its essence can be portrayed as visualized in Figure 1(a). The graph visualizes possible communication links with the user. This communication is primarily unidirectional where the results are communicated in the form of the resulting partition matrix (or equivalently a set of prototypes (centroids); obviously the representations in terms of prototypes or partition matrices are equivalent meaning that given one construct we could easily infer the other. Noticeably the resulting partition matrix

* Institute of Medical Technology and Equipment (ITAM), 118 Roosevelt st., Zabrze 41-800, Poland

becomes a direct consequence of the computing process being carried out by the FCM algorithm.



a) b)

Fig. 1. Fuzzy clustering being exclusively concerned with data-based processing (a) and a knowledge-based paradigm shift resulting in accepting various knowledge-based hints that augment the performance of the generic algorithm (b)

The numeric data are the exclusive source of processing and guidance of the search process. There are cases where there is some domain knowledge that would be highly beneficial to incorporate to support clustering mechanisms. In a nutshell, we are aimed at building a hybrid clustering environment based on the simultaneous usage of numeric data and domain knowledge. This paradigm shift has to do with a way in which some auxiliary knowledge can be incorporated into the clustering mechanisms. As visualized in Figure 1(b), some knowledge hints inserted by the user/analyst are accommodated at the level of results and start interacting with the FCM in an attempt to reconcile the data-driven optimization (the FCM itself). In this way they form an additional source of directing the mechanisms of clustering. The notion of knowledge hints requires more attention. So far, we have not defined them in a detailed manner (this will be dealt with later on). In a nutshell, by knowledge hints we mean some auxiliary pieces of information being available at the time of data clustering and reflecting some additional sources of problem domain knowledge. They could be very diversified. In general, they do not associate with all patterns (but their small fraction). The hints may deal with a single pattern or pairs of patterns. The partition matrix is a reflection of information granules and in this way any guidance is quantified and expressed in the language of fuzzy sets or fuzzy relations constructed at this level of generality.

In what follows, we adhere to the standard notation and terminology used throughout pattern recognition. Patterns are treated as n-dimensional vectors in $\mathbf{R}^n$, their number is denoted by "N" while the number of clusters is equal to "c". The partition matrix is denoted by U and contains all results of clustering by containing membership grades of each pattern to each cluster (so the matrix has "c" rows and "N" columns). The study uses a generic Fuzzy C- Means (FCM) as a reference algorithm. Let us recall that the minimized standard objective function (performance index) assumes a form of the weighted sum of the distances ||.|| between patterns and prototypes that is

$$Q = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^{m} \| \mathbf{x}_k - \mathbf{v}_i \|^2 \qquad (1)$$

where $U = [u_{ik}]$ is a partition matrix, $\mathbf{x}_k$ and $\mathbf{v}_i$ are patterns and prototypes, respectively. The fuzzification parameter (m) assumes values greater than 1.0.

2. SELECTED EXAMPLES OF KNOWLEDGE-ORIENTED HINTS AND THEIR TAXONOMY

There are several formats of the knowledge-based hints coming from the user or data analyst. These could be very much problem-dependent. We elaborate on their key categories that seem to be quite general and somewhat problem-independent. The first one is concerned with uncertainty of class membership using which we quantify our confidence or difficulty as to the categorization (allocation) of a certain pattern. In the second category of guidance mechanisms, we encounter proximity-oriented knowledge–based hints where we quantify knowledge about some pairs of patterns (as to their level of proximity). The third is concerned with labelling of some patterns so that their class assignment is known and the usage of the hints becomes helpful in exploring the data and discovering the structure there.

The taxonomy of the knowledge hints is governed by several criteria as to their generality, detailed knowledge or required assumptions as to the structure of the data. Table 1 elaborates on this in more coherent manner. For illustrative purposes we allude to some specific examples concerning the classification of ECG signals.

Table 1. Knowledge-based hints and their characterization

Knowledge hint	Description	Formalism	Notes
Uncertainty	Reflects *uncertainty* as to categorization of a pattern; e.g., the pattern is difficult to assign to a certain category, borderline character of the pattern, pattern's class straightforward to assess, etc.	Entropy measure of fuzziness H(.) is a basis as a suitable measure. Applies to an individual pattern. There is no requirement as to the knowledge about the number of clusters	ECG signal difficult to classify (atypical, high level of noise – poor recording quality, etc.); some hesitation exists as to its class allocation
Proximity	Reflects proximity between selected pairs of patterns and quantifies a subjective judgment as to the closeness of some pairs of patterns	Proximity measure; applies to specified pairs of patterns, does not require any fixed number of clusters to be given in advance	Some pairs of ECG signals have been compared and their proximity assessed. The process does not require any explicit label assignment but rather definition of their closeness (which could be quite easy to realize)
Labelling	Reflects the fact that some patterns are labelled (with classes assigned) and come as a part of the domain knowledge	Distance between provided membership grades and those contained in the partition matrix; requires the number of clusters to be specified in advance	Among a vast number of QRS complexes only a few patterns have been labelled. Those could have been selected as carefully investigated cases

While viewing these knowledge hints in a broader applied context, it is worth highlighting a general rationale behind them

Completeness of the feature space. The number of existing applications in which various knowledge hints become essential is directly implied by the effect of limited and incomplete feature spaces. It is obvious that a comprehensive feature space becomes a genuine asset in any pattern recognition problem and implies potentially high recognition rates. This, in particular, concerns classification tasks realized by or actively engaging clinicians. It is also quite apparent that a number of essential features may not be available or could not be easily quantified at the algorithmic end however those are the components that are implicitly used in human-based classification. The same argument holds in case of clustering: the available feature space could not involve some of the critical features. In this case any additional knowledge-based hints as to the relationships between some pairs of patterns or individual patterns start playing a pivotal role in enhancing the clustering activities. In a nutshell, these hints compensate for the reduced character of the feature space. More formally, we envision the following model, refer to Figure 2: the original feature space F ($\subset \mathbf{R}^n$) is available in its reduced version G ($\subset \mathbf{R}^m$) where usually m << n. The clustering realized in G is augmented by the logic predicates ϕ, ξ,... etc. whose active role compensates for the realization of classification activities in G. Ideally, one could anticipate a goal of achieving the close resemblance of results of clustering in F and knowledge-based clustering realized in G; refer again to Figure 2. Hopefully, our expectations are that the structures revealed in both cases are close enough that is $S \approx T$.

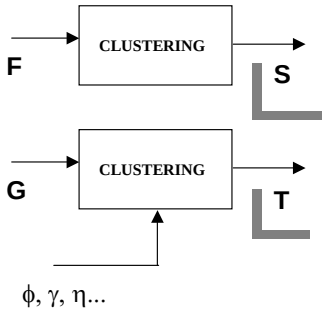


Fig. 2. Knowledge based clustering as a compensation mechanism for the use of the reduced feature space (G)

The logic predicates (knowledge-based hints) can arise in different formats. In particular, one can be provided with some referential nature of the predicates coming in the form of proximity–based information about pairs of patterns (say, the quantifications in the referential stating e.g., that two patterns are *similar* with some level of closeness, patterns are very *different*, etc). While the effect of availability of the reduced feature space is quite common, its impact is clear in the realm of various diagnostic problems where in virtue of the problem itself we are concerned with diagnostic problems. The crux of such tasks lies in its inherently heterogeneous character: in spite of the number of diagnostic tests, they still form only a certain quite limited fraction of what really becomes available in any comprehensive examination. The formation of the complete associated feature space could not be realized. This means that a possible quantification of classification results is also carried out in the format of case-based reasoning where class assignment refers to several

similar or distinct cases. The studies along this line where fuzzy clustering with proximity-based hints were involved are reported in [15]. These papers are concerned with the development of taxonomies of Web pages based on textual information and exploiting information about proximities between selected pairs of the pages and forms an alternative to other approaches existing in the literature [4][5].

Knowledge-based guidance about patterns In this general category we position a host of approaches where there is some information available about class assignment of patterns. In general, the feature space becomes available in a complete form. Depending upon the form in which categories of patterns are incorporated in the knowledge-based guidelines, two essential directions are envisioned

a) *explicit* information about class membership of selected patterns. For these patterns we are given their membership grades (those are either Boolean assignments or become represented by some membership values). This assignment requires a fixed number of classes (categories). This situation is typical in the realm of fuzzy clustering under partial supervision, cf. [14][15] where a relatively small subset of patterns has been fully labelled. This occurs e.g., in cases where labelling of all patterns is impractical while a subset of patterns can be handled quite effectively. For instance, we can deal with labelling of some OCR symbols but classifying all of them is not feasible. Likewise we may encounter a small portion of ECG signals that have been carefully labelled for the classification purposes.
b) *implicit* information about class membership. This type of information is less detailed than discussed in the previous case and concentrates on the quantification of typicality of patterns. In essence, we do not have any detailed information about allocation to classes but rather have a single numeric quantity (and thus implicit) expressing how *typical* or *relevant* a certain pattern can be sought. For instance, to express that a pattern is typical in a certain class, we envision that its uncertainty measure (entropy) [12] is close to zero, $H \approx 0$. By stating that $H \approx 1$, we express a hint about a low level of typicality of the pattern.

The essence of the implicit class allocation is visualized in Figure 3.

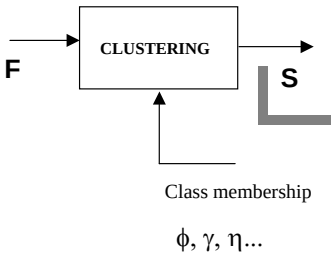


Fig. 3. The functional aspects of the knowledge-based guidance in clustering with implicit class information

3. THE OPTIMIZATION ENVIRONMENT

Following the general observations made in the previous section and following the overall architectural investigations, we can now translate these into more operational details. This will subsequently lead us to the detailed algorithmic environment. Firstly, we discuss a way in which uncertainty, proximity and labelling can be incorporated into the general scheme. In what follows (as has already been envisioned in the communication scheme between data and knowledge – oriented processing), the knowledge hints are expressed at the level of partition matrix (and specific membership grades). This becomes obvious in the ensuing notation with u_{ik} denoting a degree of membership of the k-th pattern to the i-th cluster.

Uncertainty **The typical model of uncertainty and its quantification comes in the form of an entropy function [12]. Given a variable "u" which assumes values in a unit interval, an entropy function H(u) is defined as a continuous function from [0,1] to [0,1] such that (a) it is monotonically increasing in [0, ½] (b) monotonically decreasing in [½ , 1] and satisfies boundary conditions H(0)=H(1)=0 H(1/2) =1 (as intuitively expected, here the entropy function attains its maximum). Given a collection of membership grades $w = [w_1, w_2, \ldots, w_c]^T$, the entropy easily generalizes to the form**

$$H(w) = \frac{1}{c}\sum_{i=1}^{c} H(w_i) \qquad (2)$$

with $H(w_i)$ being the entropy defined for the i-th coordinate (variable). The form of the specific function coming from the class formulated above could vary. A typical example is a piecewise linear function or a quadratic function that is H(u) = 4u(1-u).

Proximity **The concept of proximity is one of the fundamental notions when expressing the mutual dependency between membership occurring two patterns. Consider two patterns with their corresponding columns in the partition matrix denoted by "k" and "l", that is u_k and u_l, respectively. The proximity between them, denoted by Prox(u_k, u_l), is defined in the form**

$$\text{Prox}(\mathbf{u}_k, \mathbf{u}_l) = \sum_{i=1}^{c} \min(u_{ik}, u_{il}) \qquad (3)$$

Note that the proximity function is symmetric and returns 1 for the same pattern (k=l) however this relationship itself is not transitive. In virtue of the properties of any partition matrix we immediately obtain

$$\text{Prox}(\mathbf{u}_k, \mathbf{u}_l) = \sum_{i=1}^{c} \min(u_{ik}, u_{il}) = \text{Prox}(\mathbf{u}_l, \mathbf{u}_k) \quad \text{Prox}(\mathbf{u}_k, \mathbf{u}_k) = \sum_{i=1}^{c} \min(u_{ik}, u_{ik}) = 1 \qquad (4)$$

In addition to the proximity or uncertainty guidance expressed in terms of specific thresholds, we can envision their relaxed versions allowing for the quantification involving

some type of containment predicates, say " *low* uncertainty level, uncertainty not exceeding *about* 0.5", *high* proximity" where the terms quantifying these constraints are regarded to be fuzzy sets. Or, equivalently, we can relax the predicates "less than", etc by allowing their truth values and regard these to be modelled by means of fuzzy sets. This makes these expressions more in par with the language being used by the user and its usage in forming the interfaces with the clustering environment contributes to the enhanced relevance and readability of the knowledge hints.

The use of navigational variables In this scenario, we are concerned with clustering where the knowledge based hints "direct" the search of the structure by using the labels of the patterns. For instance, let the class membership values of the patterns x_k be equal to f_k where f_k assumes values in [0,1]. The labels play a role of the navigational variable. The optimization task is realized through the minimization of Q but now the membership constraints are no longer standard and sum up to 1 but satisfy the constraint $\sum_{i=1}^{c} u_{ik} = f_k$

The use of the external structure in data This category of knowledge hints arises when patterns are labelled by other processes of data organization or unsupervised learning. Given that some features of the patterns are not available directly to the process of clustering but come in the form of auxiliary (external) partition matrix $T=[t_{ik}]$ we can take advantage of these hints when navigating the optimization process of the original partition matrix. More specifically we are concerned with the standard minimization of the augmented objective function Q that assumes the following format

$$Q = \sum_{i=1}^{c} \sum_{k=1}^{N} u_{ik}^{m} \| \mathbf{x}_k - \mathbf{v}_i \|^2 + \alpha \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ik} - t_{ik})^2 \| \mathbf{x}_k - \mathbf{v}_i \|^2 \qquad (7)$$

where α is a positive scaling factor capturing the impact of departure of the values of the partition matrix from the external hints T. The minimization of Q is completed under the "standard" assumptions (that is we require that U satisfies the conditions imposed in partition matrices).

4. THE ORGANIZATION OF THE INTERACTION PROCESS

As the fuzzy clustering and the incorporation of the knowledge-based hints are the two streams of cooperating and intertwined activities, we can portray the following scheme of computing

Fuzzy clustering We deal here with a general class of clustering techniques that return results of clustering arranged in a form of a certain partition matrix U. The minimization of the underlying objective function induces the clustering to become a certain minimization problem, min Q(U) where the minimization is carried out for U as well as the associated prototypes (centroids) of the clusters. The number of clusters (c) is specified in advance. The minimization of Q is a result of some iterative process as we cycle through computations of the partition matrix and the resulting prototypes.

Knowledge-based hints here we are concerned with the optimization of the logic predicates leading to the maximization of the assumed performance index. The accommodation of the knowledge hints is realized by the maximization of the truth value of the corresponding predicates realized with respect to the membership grades (entries of the partition matrix), max $\mathcal{P}$(U) where $\mathcal{P}$ stands for the general form of the predicate (inclusion or similarity) computed over the entire set of patterns and the associated maximum computed over the partition matrix.

5. CONCLUSIONS

The knowledge-based guidance augmentation of unsupervised learning opens up some new and promising avenues of exploration of data structures. By building the unified optimization framework in which we seamlessly combine data and knowledge-based computing activities, we are able to address the fundamental matter of hybrid information processing. The interface between data and knowledge-based computing exploits models of logic optimization where we develop a certain predicate and maximize its truth value by determining the underlying structure of the clusters (partition matrix). The fuzzy predicates become helpful in expressing linguistic relational constraints (such as less than, approximately equal, etc.) that are in line with the assessment made by designers and data analysts. We have presented an array of possible scenarios arising in various areas of applications ranging from Web exploration, uncertainty guidance in pattern classification, and partial supervision involving labelling of selected patterns.

BIBLIOGRAPHY

[1] M.R. ANDERBERG, *Cluster Analysis for Applications*, Academic Press, New York, NY, 1973.

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

[3] A. BARGIELA, W. PEDRYCZ, *Granular Computing: An Introduction*, Kluwer Academic Publishers, Dordrecht, 2002.

[4] D. BOLEY et al., Partitioning-based clustering for Web document categorization, *Decision Support Systems*, 27, 1999, 329-341.

[5] D. GUILLAUME, F. MURTARGH, Clustering of XML documents, *Computer Physics Communications*, 127, 2000, 215-227.

[6] R.J. HATHAWAY, J.W. DAVENPORT, J.C. BEZDEK, Relational dual of the C-means clustering algorithms, *Pattern Recognition*, 22, no. 2, 1989, 205-212.

[7] R.J. HATHAWAY, J.C. BEZDEK, NERF-c means: non-Euclidean relational fuzzy clustering, *Pattern Recognition*, 27, 1994, 429-437.

[8] R.J. HATHAWAY, J.C. BEZDEK, J.W. Davenport, On relational data versions of c-means algorithms, *Pattern Recognition Letters*, 17, 1996, 607-612.

[9] R.J. HATHAWAY, J.C. BEZDEK, Y. HU, Generalized Fuzzy C-Means clustering strategies using L_p norm distances, *IEEE Trans. on Fuzzy Systems*, 8, no. 5, 2000, 576-582.

[10] F. HOPPNER, F. KLAWONN, R. KRUSE, T. RUNKLER, *Fuzzy Cluster Analysis – Methods for Image Recognition*, J. Wiley, N. York, 1999.

[11] A. K. JAIN, R.C. DUBES, *Algorithm for Clustering Data*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[12] G.J. KLIR, T.A. FOLGER, *Fuzzy Sets, Uncertainty, and Information*, Prentice Hall, Englewood Cliffs, NJ, 1988.

[13] W. PEDRYCZ, J. WALETZKY, Fuzzy clustering with partial supervision, *IEEE Trans. on Systems, Man, and Cybernetics*, 5, 1997, 787-795.

[14] W. PEDRYCZ, J. WALETZKY, Fuzzy clustering in software reusability, *SOFTWARE: PRACTICE & EXPERIENCE*, 27, 1997, 245 - 270.

[15] W. PEDRYCZ, V. LOIA, S. SENATORE, P-FCM : A proximity-based clustering, *Fuzzy Sets & Systems*, to appear.

[16] T. A. RUNKLER, J.C. BEZDEK, Alternating cluster estimation: a new tool for clustering and function approximation, *IEEE Trans. on Fuzzy Systems*, 7, no. 4, 1999, 377-393.