Developments in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics. Volume I: Foundations

Developments in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics Volume II: Applications

Editors

Krassimir T. Atanassov
Michał Baczyński
Józef Drewniak
Janusz Kacprzyk
Maciej Krawczak
Eulalia Szmidt
Maciej Wygralak
Sławomir Zadrożny

Editors

Krassimir T. Atanassov
Władysław Homenda
Olgierd Hryniewicz
Janusz Kacprzyk
Maciej Krawczak
Zbigniew Nahorski
Eulalia Szmidt
Sławomir Zadrożny

SRI PAS IBS PAN

Developments in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics Volume II: Applications

Systems Research Institute
Polish Academy of Sciences

Developments in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics
Volume II: Applications

Editors

Krassimir T. Atanassov
Władysław Homenda
Olgierd Hryniewicz
Janusz Kacprzyk
Maciej Krawczak
Zbigniew Nahorski
Eulalia Szmidt
Sławomir Zadrożny

IBS PAN SRI PAS

Systems Research Institute
Polish Academy of Sciences
Newelska 6, 01-447 Warsaw, Poland
www.ibspan.waw.pl

ISBN 9788389475305

Asymmetric distances and fuzzy grouping

Jan W. Owsiński
Systems Research Institute, Polish Academy of Sciences,
Newelska 6, 01-447 Warszawa, Poland
owsinski@ibspan.waw.pl

Abstract

Given asymmetric distances are a hard problem to treat explicitly in any context, including clustering. The usual way out is to somehow symmetrise hem at some point of the procedure. Even though, for quite fundamental reasons, asymmetric structures can hardly be analysed "by inspection" for even modest dimensionality of problems, it is argued in this paper that fuzzy clustering is one domain, in which asymmetric distances can be accommodated in a quite natural manner. Following the previous paper of the author, a procedure is proposed, here complemented with additional analysis, emphasis being placed on the quality criterion (objective function).

Keywords: asymmetric distances, fuzzy clustering, hierarchical grouping, k-medoids, objective function for clustering.

1 Asymmetric distances and proximities

Asymmetric distances and proximities arise in a variety of situations. In the majority of real-life cases they are given as *original data*. These are the cases of, for instance, time-wise or cost-wise road, railway or air distances, trade and financial flows, as well as traffic, migration and commuter flows. Definitely, asymmetric distances may also arise from *explicit distance calculations*, based on properties of (pairs of) objects, for which they are calculated, and additional prerequisites, allowing for asymmetry. This case, though, has to be treated separately, as open to a much wider choice and a rather specific perspective.

It is, namely, quite typical for the case, when distances are explicitly calculated, that an a priori assumption has to be made, which, actually, is responsible for

Developments in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics. Volume II: Applications *(K.T.Atanassow, W.Homenda, O.Hryniewicz, J.Kacprzyk, M.Krawczak, Z.Nahorski, E. Szmidt, S. Zadrożny, Eds.), IBS PAN - SRI PAS, Warsaw, 2009.*

the asymmetry. This assumption takes, naturally, the form of the generally valid distance definition, often stipulating quite strong properties, e.g. satisfaction of the triangle inequality. This is the case of "gravity-based asymmetric distance" and similar functions (see, e.g., [3] and [12] for the asymmetric location problems). In quite broad terms this kind of setting can be said to involve a degree of *regularity*, which can be exploited when proving relevant properties of the problems, extrema and algorithms.

Asymmetric distance or proximity definitions are frequently based on what can be referred to as *containment* or *dominance*. This is typical for the two important domains, where definitions of distances or proximities are used that can lead to asymmetry, namely in chemistry (for two compounds of particles with highly different masses, is the similarity of the smaller one to the bigger one symmetric with the reciprocal one?) and in text analysis (the same question for texts of highly different lengths), see, in particular, [6] and [7].

There exists a wide class of problems, in which asymmetry of distance- or proximity-like data does not constitute any problem whatsoever. These are the problems of maximum or minimum flow, shortest route, least cost route etc., which are usually solved via some graph-theoretic methods (see, e.g., [8] and [9]). In fact, the asymmetric distances, forming a matrix $D = \{d_{ij}\}_{ij}$, with, therefore, $d_{ij} \neq d_{ji}$ for at least one pair (i,j), are most conveniently represented by directed graphs between nodes $i,j \in I$, the latter denoting the set of nodes. In such problems the initial graph-based representation is kept throughout the solution procedure down to its successful end, consisting in specifying some ("best") subset of edges (i,j) along with their relevant characteristics.

In this paper, though, we deal with problems, in which the endpoint of the analytic procedure is constituted by a structure that differs significantly from the initial directed graph representation, even if it still can be rendered via graphs. We shall focus, namely, on clustering in the situation, where distances or proximities are given, and no assumption can a priori be made, except for nonnegativity, on their values and relations between them.

2 Clustering with asymmetric distances

2.1 Accommodating asymmetry in clustering

The issue with clustering is that the usual formulation of the clustering problem somehow implies symmetry. Let us start with the generic formulation of the clustering problem:

having objects indexed i, $i \in I = \{1,\ldots,n\}$, for which distances d_{ij} and/or proximities s_{ij} can be or are specified, partition the set I into subsets A_q, $q = 1,2,\ldots$, in such a way as to have (indices of) objects that are close to each other in the same A_q, and the (indices of) objects distant one from another in different A_q.

Even though this "definition" leaves a lot to be made more precise, it certainly appears to imply some sort of symmetry: two objects, i and j, either both belong to the same cluster A_q, or are mutually separated in two different clusters.

This, indeed, is so, if we admit (only) crisp and disjoint partitions, i.e. $A_q \cap A_{q'} = \varnothing$, $q \neq q'$, and $\mu_i(A_q) \in \{0,1\}$, where $\mu_i(A_q)$ is the membership of i in A_q. Should, by necessity, either overlapping clusters (cliques) or fuzzy clusters be allowed, as apparently admitting (some sort of) asymmetry? We shall yet return to this issue.

2.2 Why, and how cluster?

Yet, the most important issue is the one of purpose and therefore of the ultimate structure. In other words: what do we want to get? what can we get? how to get this? In this context we shall quote here some examples, for which asymmetric distances and clustering are of importance.

First, let us consider a case of importance in information retrieval. If for documents indexed i asymmetric distances are calculated (conform to the remark before, see, e.g., [6] and [7]), it may be of interest to be able to group them in such a way as to preserve (somehow) this asymmetry. This might enhance the information retrieval process, when the inherent asymmetry is of importance (for instance – for the efficiency of "directed" search in the document space).

Another example refers to the analysis of "influence areas", be it of urban centres, or companies, or countries, based on flows of goods, money and/or commuters. The resulting information may be used for planning and strategy development purposes. Here, individual objects may fall to different degrees within the influence areas of different "centres".

Yet another example is provided by the domain of social networks, of primary interest in this work. Relations in such networks are obviously asymmetric (although this is by no means easily admitted in the existing literature), and it is of utmost importance in their analysis to be able to define the "core actors", the "local nexūs", or the "spheres of influence" (see, e.g. [3]).

In these cases – and, indeed, in many others – asymmetry of distances is closely associated with asymmetry of "positions" of the objects considered (like the "big" and "small" chemical compounds) and/or the asymmetry of "meaning" of the directions of edges (distances or proximities). Indeed, although, for instance, fuzzy clustering does in no way depend upon or even refer to asymmetric distances (see, e.g., [15]), there is a problem in the reverse direction: what is the significance ("epistemological status") of the fuzzy clusters, which would try to preserve the asymmetric nature of distances? and: how to obtain them?

If we admit the asymmetry of positions or directions, then the answers to both these questions become somewhat easier.

2.3 The nature of "asymmetric" clusters

Whether clusters are crisp or fuzzy, overlapping or not, the structure imposed by a partition $P = \{A_q\}_q$ is symmetric with respect to particular objects. Indeed, if, in quite a natural manner, we define by δ_{ij} the distance between objects i and j, resulting from the partition, e.g. $\delta_{ij} = \Sigma_q|\mu_i(A_q) - \mu_j(A_q)|$, then, of course, $\delta_{ij} = \delta_{ji}$. There exists, however, a structure that relates to clustering and yet is capable of implying asymmetry. It is, indeed, associated with the asymmetry of positions.

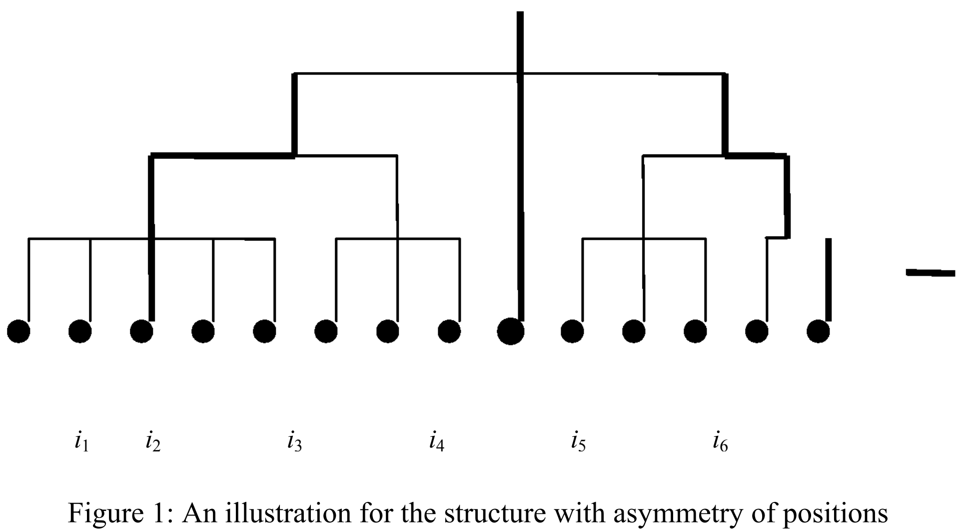

Figure 1: An illustration for the structure with asymmetry of positions

Fig. 1, taken partly from [11], will be commented upon in deeper detail here. Thus, we deal with a *hierarchical structure* of objects $i \in I$, grouped consecutively at a number of levels, in which *groups are labeled by the leading objects*, and this applies to all levels. So, the first bottom-level group to the left in Fig. 1

could be labeled $L^1(i_2)$, where the superscript "1" denotes the bottom level (leaves), while the entire set of objects in the figure would form the group $L^3(i_5)$. Formally, we will write that $L^1(i_2) = i_2$, and $L^3(i_5) = i_5$, i.e. $L^h(i) = i$, while the sets of objects, constituting respective groups, shall be denoted $A^h(i)$. If we use the general indexing of nodes, without indication of levels, say, $q \in I$, then $\boldsymbol{q}$ shall denote the cardinality of the respective group.

If a structure like this is established, we can also define a distance within it, by introducing, first, the notion of *choice-power-path* as follows: denote by $N(i,i')$ the set of nodes (corresponding to group labels) separating objects i and i' (with, of course, $N(i,i') = N(i',i)$[1]); and by $s(q,i)$, where q is a node label, the following quantity: $s(q,i) = 1$, if $q \neq i$, and $s(q,i) = 1/\boldsymbol{q}$, if $q = i$, then

$$\delta^H(i,i') = \Sigma_{q\in N(i,i')} s(q,i).$$

For the illustration of Fig. 1 we get from this definition the following values:

i → i'	***1***	***2***	***3***	***4***	***5***	***6***
1	0	1	1	3	3	4
2	0.2	0	0.2	1.325	1.325	2.325
3	1	1	0	3	3	4
4	2.33	2.33	2.33	0	2.33	3.33
5	2.07	2.07	2.07	2.07	0	2.07
6	4	4	4	4	3	0

A variant of this definition could only refer to the number of levels, required to pass from one object to another, similarly as in ultrametrics.

Let us note that while both structure of Fig. 1 and the definition of $\delta^H(i,i')$ imply crisp grouping. Yet, the definition of $\delta^H(i,i')$ can be turned into the one, involving fuzzy association of objects, by appropriately defining $\boldsymbol{q}$ and weighing $s(q,i)$ by respective membership values:

$$\delta^H(i,i') = \Sigma_{q\in N(i,i')} \mu_i(q) s(q,i),$$

with $N(i,i')$ taken, for simplicity, as the subset of I, determined by the nodes, for which $\mu_i(q)$, $\mu_i(q')$ attain maxima. (For establishment of respective principles, see, for instance, [1],[4], [5], [16], [17]).

---

[1] Minimum number of nodes, invariant with respect to object and node numbering.

It should be added that the definition, outlined here, is given mainly as an illustration for a broader family of asymmetric output structure distances, which could be applied in the context here considered.

3 A proposal

3.1 An outline for a procedure

Assume we consider whether an object i be (rather) associated in a structure of the kind shown in Fig. 1 with another object, denoted $q1$, or (rather) a different one, denoted $q2$. Actually, it does not matter whether $q1$ and $q2$ (are meant to) represent proper clusters, potentially forming a partition $P = \{A_q\}$, or "next level" nodes in a tree, representing a hierarchy, or simply other objects of the same status as i, as we consider whether to associate i with one of them.

When no "privileged direction" is defined, we deal with the following cases:

1°. $d'_{iq1} \geq d''_{iq2}$;
2°. $d''_{iq1} \geq d''_{iq2}$, and $d'_{iq1} \geq d'_{iq2}$, but $d'_{iq1} < d''_{iq2}$;
3°. $d''_{iq1} \geq d''_{iq2}$, and $d'_{iq1} \leq d'_{iq2}$,

where $d'_{ij} = \min\{d_{ij},d_{ji}\}$ and $d''_{ij} = \max\{d_{ij},d_{ji}\}$.

In the above context it appears intuitively obvious that in case 1° above, object i be associated with $q2$ in the degree $\mu(i,q2) = 1$ and with $q1$ in the degree $\mu(i,q1) = 0$, as comparison leaves no doubt as to which "distance" is smaller of the two (pairs). For the two remaining cases the values of the degrees of association can be calculated as the functions of the corresponding distance values, appearing in the above inequalities. We shall not propose any concrete form of such functions, but only will stipulate some exemplary "boundary conditions", including the one stated above, namely (assuming, in addition, quite arbitrarily, for this stage of consideration, that $\mu(i,q1) + \mu(i,q2) = 1$):

1. when $d'_{iq1} \geq d''_{iq2}$ then $\mu(i,q1) = 0$ and $\mu(i,q2) = 1$;
2. when $d''_{iq1} = d''_{iq2}$, and $d'_{iq1} = d'_{iq2}$ then $\mu(i,q1) = \mu(i,q2) = ½$;
3. when $d''_{iq1} > d''_{iq2}$ and $d'_{iq1} = d'_{iq2}$,
   then $\mu(i,q1) = (d''_{iq2} - d'_{iq2})/(d''_{iq1} - d'_{iq1} + d''_{iq2} - d'_{iq2})$, and
   $\mu(i,q2) = (d''_{iq1} - d'_{iq1})/(d''_{iq1} - d'_{iq1} + d''_{iq2} - d'_{iq2}) = 1 - \mu(i,q1)$;
4. when $d''_{iq1} = d''_{iq2}$, and $d'_{iq1} < d'_{iq2}$,

then (vice versa): $\mu(i,q1) = (d''_{iq1} - d'_{iq1})/(d''_{iq1} - d'_{iq1} + d''_{iq2} - d'_{iq2})$, and $\mu(i,q2) = (d''_{iq2} - d'_{iq2})/(d''_{iq1} - d'_{iq1} + d''_{iq2} - d'_{iq2}) = 1 - \mu(i,q1)$.

In this manner we can compare association of an i with any object in the set considered, whether this object is a "cluster representative" or not. The simple association rules, given above, can be then appropriately extended over all objects in I, leading to the matrix $\{\mu(i,i')\}_{i,i'}$, with preservation of the condition $\Sigma_{i'\neq i}\mu(i,i')=1\ \forall i \in I$. (Here, again, we refer to [1], [4], [5], [16], [17] for respective basic principles.)

The above procedure appears numerically cumbersome, even if, by definition, it consists of a "single pass", since it requires $O(n^3)$ comparisons, where n = card$I$, with yet a number of additional arithmetic operations. This, indeed, would be prohibitive, if it were not so that in the domain of main interest to us the respective matrices of distances, or rather proximities (existence of any relation) are either relatively sparse (down to below 10% or even less) or relatively small (hundreds of objects at most).

This procedure has been first presented in [11], and referred to as a simile of the k-medoids algorithm. Actually, it resembles this algorithm only slightly, not like the clustering procedures for asymmetric distances, proposed in [13] and [14], which, indeed, directly draw upon the classical k-medoids. The procedure here proposed: (i) is deterministic, (ii) does not specify directly (unambiguously) a partition of the set I, but a fuzzy hierarchy with a priori undefined number of "levels" (in principle, it can be taken to be equal card$I$-1).

3.2 An objective function

Following the principles, formulated in [10], we shall now introduce an objective function, calculated for the structure, resulting from the procedure, outlined in Section 3.1, and using the distance function, proposed in Section 2.3, defined for such a structure:

$$Q^H(\{\delta^H(i,i')\}_{i,i'}) = \Sigma_{q\in I}\Sigma_{i,i'\in I}\ \delta^H(i,i' \mid i,i' \in A^h(q)) + \Sigma_{q,q'\in I,\ q\neq q'}\Sigma_{i,i'\in I}\ \sigma^H(i,i' \mid i\in A^h(q),\ i'\in A^h(q'))$$

where $\sigma^H(i,i')$, corresponding to proximity, as opposed to distance, is defined analogously to $\delta^H(i,i')$, with $1-\mu_i(q)$ replacing $\mu_i(q)$ and $1-s(q,i)$ replacing $s(q,i)$.

This objective function is minimized (possibly small distances within groups and possibly small proximities between groups). Yet, the deterministic procedure proposed leaves, indeed, in principle, no room for choice and optimization.

Still, the values of the objective function, calculated along with the realization of the procedure, can serve the purpose similar to that of the criteria used in classical agglomerative algorithms of cluster analysis, namely: selection of the hierarchy level, at which partition is defined. It is, namely, obvious, that as cardinalities of $A^h(q)$ forming the partition get bigger, the respective increments to $Q^H(\{\delta^H(i,i')\}_{i,i'})$ get smaller, so that, in general, the optimum number of hierarchy levels is lower than the maximum. This means that the structure, constituting the output from the procedure, is cut at a certain height, according to the sequence of values of $Q^H(\{\delta^H(i,i')\}_{i,i'})$ along the cardinalities of $A^h(q)$ generated.

4 Conclusions

This paper presents outlines an approach to grouping of objects, for which distances or proximities are given that are asymmetric. The procedure proposed tries to preserve the asymmetric nature of data through the establishment of a tree-like structure, in which objects are associated one with another in fuzzy degrees. This procedure is composed of the association rules, the distance / proximity definitions and the objective function, serving to determine the optimum height of the tree-like structure. Not being numerically effective, the approach can either be used to pre-processed sparse data matrices, or to relatively small data sets. The approach proposed shall be first tested on the data sets pertaining to local networks, as observed through connections, announced via websites.

References

[1] Bezdek J. C. (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York.

[2] Carrington P.J., Scott J. and Wasserman S., eds. (2005) *Models and Methods in Social Network Analysis*. Cambridge University Press, New York.

[3] Drezner Z. and Wesolowsky G. O. (1989) The asymmetric distance location problem. *Transp. Sci.*, 23, 201-207.

[4] Dubois D. and Prade H. (1985) *Possibility Theory*. Plenum Press, New York.

[5] Dubois D. and Prade H. (1990) The three semantics of fuzzy sets. *Fuzzy Sets and Systems*, 90, 141-150.

[6] MacCuish N. and MacCuish J. (2003a) Tversky Shape Clustering for Screening Data. *OpenEye Scientific Software User Group Meeting CUPIV*, Santa Fe, New Mexico, February 2003.

[7] MacCuish J. and MacCuish N. (2003b) Mesa Suite Version 1.1. Fingerprint Module.

[8] Mażbic-Kulma B., Owsiński J. W. and Sęp K. (2008) Application of selected methods of graph theory and combinatorial heuristics to minimize the number of transit nodes in an air network. *Total Logistics Management*, 1.

[9] Mażbic-Kulma B., Potrzebowski H., Stańczak J. and Sęp K. (2008) Evolutionary Approach to Solve Hub-and-Spoke Problem Using Alpha-cliques. *Prace Naukowe PW*, issue 165: *Evolutionary Computation and Global Optimization*, 121-130.

[10] Owsiński J. W. (1990) On a new naturally indexed quick clustering method with a global objective function. *Applied Stochastic Models and Data Analysis*, 6, 157-171.

[11] Owsiński J.W. (2009) Asymmetric distances – a natural case for fuzzy clustering? In: D. A. Viattchenin, ed., *Developments in Fuzzy Clustering*. Vever, Minsk (Belarus'), 36-45.

[12] Plastria F. (1992) On destination optimality in asymmetric distance Fermat-Weber problems. *Annals of Operations Research*, 40, 355-369.

[13] Saito T. and Yadohisa H. (2005) *Data Analysis of Asymmetric Structures – Recent Development of Computational Statistics*. Marcel Dekker, New York.

[14] Takeuchi A. and Yadohisa H. (2009) Evaluation of asymmetric *k*-medoids algorithms. *IFCS@GFKL. Classification as a Tool for Research. 11th Conference of the International Federation of Classification Societies*. March 13-18, 2009, University of Technology, Dresden, Germany. IFCS, 269.

[15] Viattchenin D. A. (2004) A new heuristic algorithm of fuzzy clustering. *Control & Cybernetics*, **33**, 2, 323-340.

[16] Zadeh L. A. (2002) From computing with numbers to computing with words – from manipulation of measurements to manipulation of perceptions. *Int. J. Appl. Math. Comput. Sci*. **12**, 3, 307-324.

[17] Zadeh L. A. (1975) The concept of a linguistic variable and its application to approximate reasoning. *Information Sciences*, **8**, 3, 199-249.

The papers presented in this Volume 2 constitute a collection of contributions, both of a foundational and applied type, by both well-known experts and young researchers in various fields of broadly perceived intelligent systems.
It may be viewed as a result of fruitful discussions held during the Eighth International Workshop on Intuitionistic Fuzzy Sets and Generalized Nets (IWIFSGN-2009) organized in Warsaw on October 16, 2009 by the Systems Research Institute, Polish Academy of Sciences, in Warsaw, Poland, Centre for Biomedical Engineering, Bulgarian Academy of Sciences in Sofia, Bulgaria, and WIT – Warsaw School of Information Technology in Warsaw, Poland, and co-organized by: the Matej Bel University, Banska Bistrica, Slovakia, Universidad Publica de Navarra, Pamplona, Spain, Universidade de Tras-Os-Montes e Alto Douro, Vila Real, Portugal, and the University of Westminster, Harrow, UK:

http://www.ibspan.waw.pl/ifs2009

The Eighth International Workshop on Intuitionistic Fuzzy Sets and Generalized Nets (IWIFSGN-2009) has been meant to commence a new series of scientific events primarily focused on new developments in foundations and applications of intuitionistic fuzzy sets and generalized nets pioneered by Professor Krassimir T. Atanassov. Moreover, other topics related to broadly perceived representation and processing of uncertain and imprecise information and intelligent systems are discussed.

We hope that a collection of main contributions presented at the Workshop, completed with many papers by leading experts who have not been able to participate, will provide a source of much needed information on recent trends in the topics considered.

ISBN-13 9788389475305
ISBN 838947530-8
9 788389 475305