Membership-based clustering of heterogeneous fuzzy data

Gernot Herbst[1] **Arne-Jens Hempel**[1] **Rainer Fletling**[2] **Steffen F. Bocklisch**[1]

[1]Chemnitz University of Technology, Germany
[2]University of Kassel, Germany

Abstract

This article contributes to clustering and fuzzy modelling of data such that specific characteristics of each datum can be incorporated. Particularly, each object may exhibit an individual area of influence in its feature space, for which it is representative. For such objects, a similarity measure is introduced, which is used to modify common clustering algorithms to take each object's extent into account when finding clusters. A real-world example demonstrates the practical usability of the presented methods, which deliver results in accordance to findings of experts in that field.

Keywords: Fuzzy classification, clustering, pattern recognition, engineering geodesy.

1. Introduction

1.1. Motivation, context and goal

Clustering algorithms provide powerful means to gain data-based insights into structures of a set of objects (e. g. measurements) [1]. Most algorithms and their users, however, do not account for the individual characteristics of single data points (objects) in the clustering process, which might be present for different reasons.

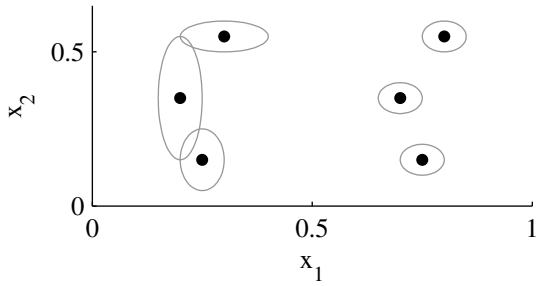


Figure 1: Two geometrically identical groups of objects with individual areas of coverage or impact.

In Fig. 1, we can identify two groups of objects with an identical geometrical layout in the feature space. Each object, however, covers a different area of the feature space. One object may be representative for a certain phenomenon in a larger area of the feature space (more on the interpretation in section 2.2). Therefore even very few objects might already constitute one cluster on their own right.

Provided the dataset of Fig. 1 is representative, one could therefore argue that the left-hand side objects should form a cluster while the other three objects should, as long as no other data become available, remain separate "clusters" on their own.

It appears reasonable that the individual extent of an object should be considered in a clustering process. A suitable clustering algorithm should therefore deliver groups of objects that cover a compact region of the feature space. For this, it would be required that information about the objects' extent is available. Given that even the objects' position will be subject to uncertainties in real-world datasets, it furthermore seems appropriate not to treat their extent in the feature space in a precise manner, but assume soft boundaries, which calls for a fuzzy representation of an object.

While there is a variety of existing fuzzy clustering algorithms [2], our approach specifically focuses on a fuzzy description of each object, rather than the cluster as a whole. In contrast to works like [3], we will also be able to effortlessly work with fuzzy objects in a higher-dimensional feature space. Commonly used clustering algorithms do not take the distinct characteristics of each object into account. In this regard our paper also aims to advertise for a paradigm shift towards an individual modelling and treatment of uncertain data.

In this article we will provide means to model the individual extent of objects in their feature space in a fuzzy manner by means of multivariate membership functions, and present ideas how to use these fuzzy descriptions in common existing hierarchical clustering algorithms in order to incorporate the individual characteristics of objects.

1.2. Structure of this article

As the basis for the ideas of this article, we will provide means to model the extent of objects in a feature space in a fuzzy manner using multivariate membership functions, which will be introduced in section 2.1. In order to be compatible with existing hierarchical clustering algorithms, we will discuss the replacement of distance measures by a fuzzy dissimilarity in section 3. An academic as well as a real-world example will be presented in sections 4.1 and 4.2, respectively.

2. Fuzzy description of objects

As already touched upon in section 1.1, an object might—apart from its position in the feature space—be characterised by a certain extent in the feature space. This extent may be used to model an object's representativeness for a certain phenomenon, spanning a region in the feature space, cf. section 2.2.

In practical applications, it appears reasonable that precise information about the borders of this region will rarely be available. We therefore propose to individually model each object and its extent in a fuzzy manner. While it has to be noted that the approach presented in section 3 is independent of the type of membership function chosen to model such objects, we will introduce (and employ throughout this article) a multivariate membership function in the following. Main advantages of this function are its adjustability by interpretable parameters, its versatile and consistent use in a uni- and multivariate form.

2.1. A multivariate parametric fuzzy set

As a generalisation of Aizerman's potential function [4], (1) defines a parameterisable asymmetric fuzzy set $\mu\colon \mathbb{R} \mapsto (0, a]$:

$$\mu(x) = \begin{cases} \dfrac{a}{1 + \left(\dfrac{1}{b_\mathrm{l}} - 1\right)\left(\dfrac{r - x}{c_\mathrm{l}}\right)^{d_\mathrm{l}}}, & x < r \\ \dfrac{a}{1 + \left(\dfrac{1}{b_\mathrm{r}} - 1\right)\left(\dfrac{x - r}{c_\mathrm{r}}\right)^{d_\mathrm{r}}}, & x \geq r \end{cases} \tag{1}$$

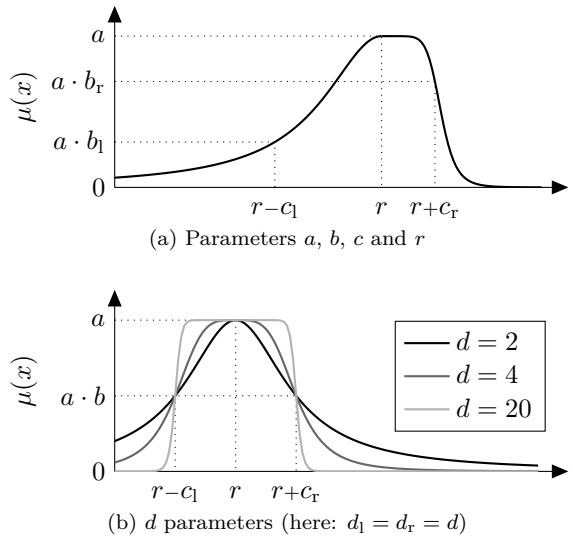
The effect of the modal point r, the maximum membership value a and the side-specific parameters b, c and d in (1) can be understood from Fig. 2a. $c_\mathrm{l}/c_\mathrm{r} > 0$ quantify the uncertainty (as in crisp sets) and $b_\mathrm{l}/b_\mathrm{r} \in (0, 1]$ and $d_\mathrm{l}/d_\mathrm{r} \in [2, \infty)$ represent the fuzziness of this uncertain information. Increasing values of d lead to sharper descents of the membership value to zero, as visible in Fig. 2b, and $d \to \infty$ will result in rectangular (crisp) sets.

To create a multivariate (N-dimensional) membership function, N normalised fuzzy sets of this type are being combined using a compensatory Hamacher intersection (2) and being assigned a joint maximum membership value a. As shown in [4] and [5], this results in an N-dimensional parametric membership function $\mu\colon \mathbb{R}^N \mapsto (0, 1]$. Visual examples for $N = 2$ are given in Fig. 3.

$$\cap_{\mathrm{Ham}}^{N} \mu_i = \left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\mu_i}\right)^{-1} \tag{2}$$

For the fuzzy description of objects with an associated area of influence in their feature space,[1]

[1] For brevity, we will refer to such objects and their fuzzy representation as "fuzzy objects" in the remainder of this article.



(a) Parameters a, b, c and r

(b) d parameters (here: $d_\mathrm{l} = d_\mathrm{r} = d$)

Figure 2: Effect of the parameters in (1).

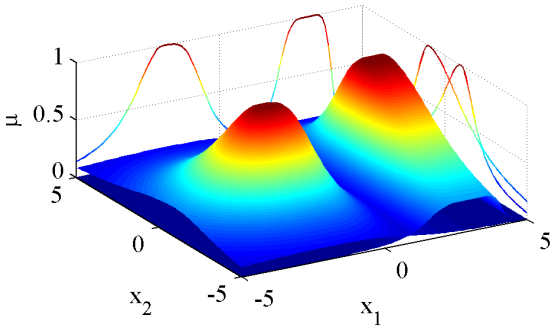


Figure 3: Two examples of two-dimensional sets.

we propose to employ the multivariate membership function based on (1). The a parameter may be used to describe objects with sub-normal significance ($a < 1$), e. g. if one is uncertain about the validity or existence of an object. More importantly, the parameters $c_\mathrm{l/r}$ can be used to quantify an object's individual extent in the feature space. The b and d parameters may furthermore be employed to modify the fuzziness of the object's extent.

Throughout the examples of this article, however, we will use normalised ($a = 1$), symmetric membership functions with $b_\mathrm{l/r} = 0.5$, $d_\mathrm{l/r} = 2$ to model fuzzy objects.[2] With these assumptions, the multivariate (N-dimensional) membership function is being reduced to (3). α-cuts of these fuzzy sets accordingly have the form of hyper-ellipses.

$$\mu(\boldsymbol{x}) = \frac{1}{1 + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\left|\frac{x_i - r_i}{c_i}\right|^2} \tag{3}$$

[2] This choice was taken only for clarity and does not affect the methods presented in this article. Each fuzzy object may be described in a completely individual manner, even with different types of membership functions.

2.2. Interpretation of fuzzy objects

As already mentioned in section 1.1, each object may be representative for a certain area of the feature space. A natural question that arises concerns the origin or determination of such an area.

One possible interpretation relates to uncertainties stemming from the underlying measurement process. For example, a nonlinear sensor characteristic could lead to non-uniform inaccuracies among several measured objects, which have to be treated individually for each object. In this case, each datum is considered as a single point known to lie somewhere in this area.

In our article, however, we consider each object as representative for all of its associated area. For example, such an object may be an abstract description for a phenomenon characterised by varying feature values within a certain area of the feature space. Therefore the aggregation of repeated measurements (which may, of course, be afflicted with uncertainties as described above) provides one possible access to obtain an object's area of influence or its fuzzy representation, respectively. For this case, [6] presents an aggregation procedure that directly results in membership functions of the type used in this article. In the spirit of fuzzy set theory, another approach to the determination of an object's area lies in the inclusion of domain-specific expert knowledge for each object. This will also be the case in our example in section 4.2.

3. Clustering of fuzzy objects

In order to identify clusters of fuzzy objects, we have to identify groups of objects whose joint extents form a (more or less) compact region of the feature space. One approach to this lies in the determination of a fuzzy degree of inclusion, which indicates adjacent or overlapping fuzzy objects.

3.1. Fuzzy degree of inclusion

As visible in Fig. 4, one fuzzy object may be "contained" in another object, but this relation is—in general—asymmetric. In order to identify overlapping fuzzy regions in the feature space, we have to take this asymmetry into account, and examine both if one object "contains" the other, or vice versa.

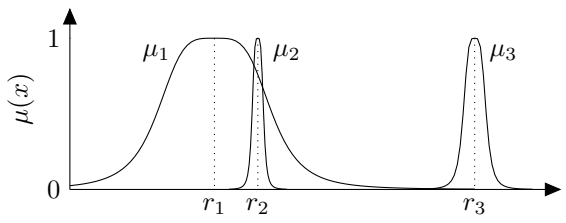


Figure 4: Examples of (non-)overlapping fuzzy objects.

To quantify the degree of inclusion of a fuzzy set $\mu_1(x)$ in a set $\mu_2(x)$, we propose the following fuzzy degree of inclusion:

$$g\left(\mu_1, \mu_2\right) = \frac{\int \mu_1(x) \ \cap \ \mu_2(x) \ \mathrm{d}x}{\int \mu_1(x) \ \mathrm{d}x} \tag{4}$$

For the operator $\cap$, we employ a standard min-conjunction, which also guarantees an interpretable range of values for $g \in [0, 1]$, with $g \to 1$ indicating high degrees of inclusion.

To identify adjacent fuzzy objects, it is sufficient if one of the degrees of inclusion $g(\mu_1, \mu_2)$ or $g(\mu_2, \mu_1)$ is high. Hence the decision whether the two objects are adjacent can be obtained by a standard disjunction (e. g. using the max-operator) of the two values:

$$g_{\max}\left(\mu_1, \mu_2\right) = \max\left(g\left(\mu_1, \mu_2\right), g\left(\mu_2, \mu_1\right)\right) \tag{5}$$

As future work, further properties and relations of (4) to existing indicators of fuzzy set inclusion as well as axioms for such indicators (cf. [7, 8]) remain to be examined.

3.2. Approximation by membership of r

Unfortunately, an analytic or numeric computation of (4) will be costly, especially in higher-dimensional feature spaces. If a computationally efficient alternative to (4) is needed, we propose to approximate the degree of inclusion by the degree of membership of the first set's modal point to the second set. When using the membership function (1) of this article, this corresponds to the membership of the r parameter:

$$\widetilde{g}\left(\mu_1, \mu_2\right) = \mu_2(r_1) \tag{6}$$

$$\widetilde{g}_{\max} = \max\left(\mu_2(r_1), \mu_1(r_2)\right) \tag{7}$$

Equation (7) is the fuzzy measure that will be used in the following examples of section 4. While we did not obtain different results in our examples when using the original (5) instead, a thorough comparison of (7) and (5) will be left as an open issue for future work.

3.3. Application in clustering algorithms

Hierarchical clustering algorithms employ a metric to determine the distance of objects in a pairwise manner. When objects with individual extents—modelled by fuzzy sets—are to be treated, the actual distance of the object locations is less relevant, and their extents have to be considered.

With (5) or (7), possible fuzzy measures for the mutual inclusion of two fuzzy objects are available. A dissimilarity measure h, which may be used to replace the distance of two objects in a clustering procedure, can be obtained by the natural complement [9] of (5) or (7):

$$h\left(\mu_1, \mu_2\right) = 1 - g_{\max}\left(\mu_1, \mu_2\right) \tag{8}$$

In contrast to the common Euclidean distance (or other members of the Minkowski family), h is a normalised dissimilarity measure $h \in [0, 1]$. The interpretation of (8) is that two objects are considered dissimilar ($h \to 1$) if neither of these objects contains the other object in the sense of fuzzy inclusion discussed in section 3.1, and similar ($h \to 0$) if so.

Equation (8) may directly be used as a replacement for the distance measure in any hierarchical clustering algorithm, now taking the individual extent of fuzzy objects into account.

4. Examples

4.1. Artificial dataset

In a first example, we will examine an artifically generated set of two-dimensional objects with individual extents (depicted in Fig. 5) and subject these to a cluster analysis employing the dissimilarity measure of section 3.3. Each datum is being described by a membership function as given in (3) with individual uncertainties (parameter c).

On the lower edge of Fig. 5, we can observe a large fuzzy object which obviously connects the two object pairs at its sides to form one single—albeit somewhat stretched—cluster. What could furthermore be seen from Fig. 5 is that the upper-left three objects intuitively form a cluster, as these objects are significant and representative for a large and connected region of the feature space. In contrast, the three-object group in the upper-right corner does probably not form a connected region, these three fuzzy objects appear isolated. Because of the identical intra-group geometry, none of the standard clustering algorithms could identify the left-hand side group as a cluster and the right-hand side group as single objects at the same time without considering their individual characteristics.

The dataset of Fig. 5 was subjected to hierarchical single linkage clustering twice, using the Euclidean distance and the membership-based approach of this article. To illustrate the effects, Fig. 6 depicts the results for a configuration of $C = 7$ classes side by side.

Fig. 6b confirms that the membership-based clustering approach is capable of finding structures similar or identical to the intuitive judgement of a human. The five fuzzy objects on the lower edge are being connected by the middle object to form one cluster, and the overlapping three-object group in the upper-left corner is also being correctly identified as one cluster. A purely distance-based approach, as visible in Fig. 6a, must obviously fail in these aspects of this task, since the objects' individual extent or representativeness is not taken into account.

For an objective assessment of clustering results, measures comparing two partitions (such as the Rand index and its fuzzy extensions [10]) are available. The additional information contained in objects with a fuzzy extent, however, is essential to the proposed type of data representation. Thus measuring the similarity between partitions without this information will not reveal the performance of a clustering result. A prospective suitable performance measure should also incorporate these individual characteristics of the data. Throughout the remainder of this article, we will therefore confine ourselves to a visual assessment.

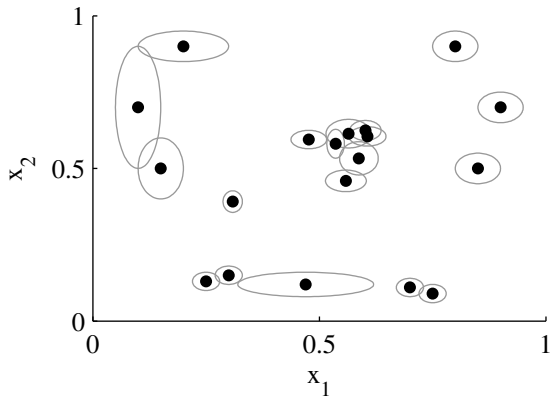
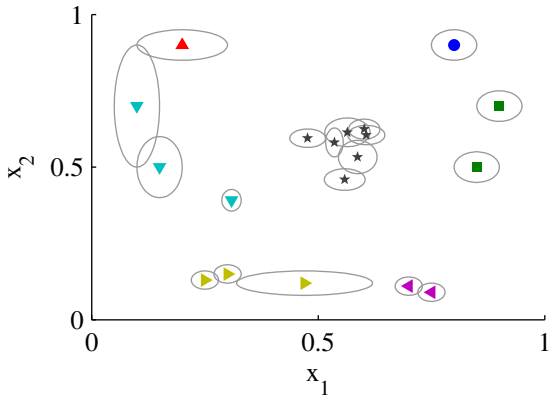
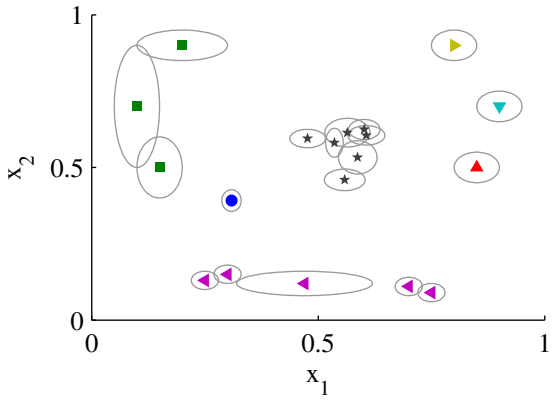


Figure 5: Objects in a two-dimensional feature space, along with α-cuts of their individual fuzzy descriptions (visible as ellipses).



(a) Distance-based

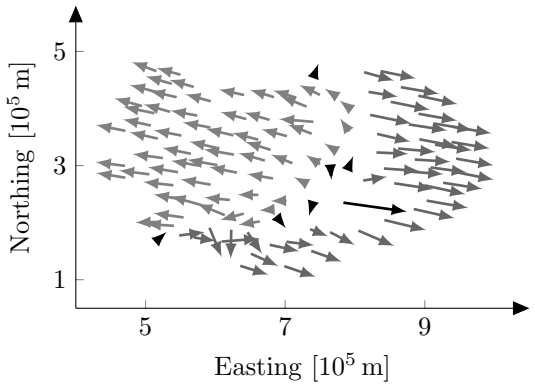


(b) Membership-based

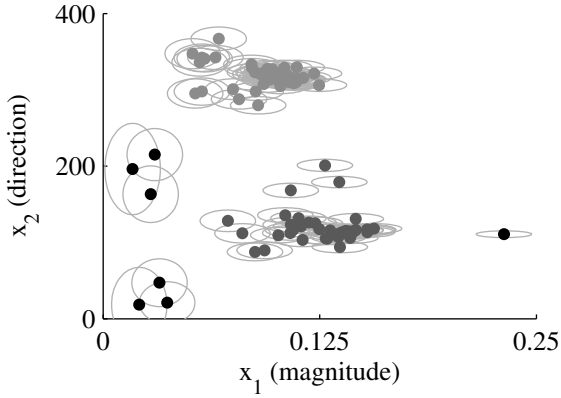
Figure 6: Results of single linkage clustering ($C = 7$ classes) of the artificial dataset from Fig. 5.

4.2. Analysing geodetic movement patterns

The dataset for the second example stems from a geodetic observation network which monitors the impact of the continental drift in Iceland [11]. Fig. 7a shows the tectonic movement of reference points (measured by GPS over the years 1993–2004), and one can clearly observe different movement patterns caused by the Mid-Atlantic Ridge. The feature space in Fig. 7b consists of the length of movement vectors and their direction.[3] In [12], individual fuzzy representations for all measurements were determined, which will be used here and which are also depicted in Fig. 7b as ellipses around the objects in the feature space. All objects are being modelled as fuzzy sets using (3).



(a) Movement vectors (length magnified by 10^5)



(b) Movement as objects in their feature space

Figure 7: Isnet dataset. The objects are shaded only to help recognising movement vectors in the feature space. The features are: movement magnitude (in metres) and direction (in gon).

To compare the membership-based distance (i. e. the fuzzy dissimilarity measure as proposed in this article) against the common Euclidean distance, a series of experiments with hierarchical clustering algorithms was carried out. The data from Fig. 7b were subjected to single and complete linkage clustering procedures. Results for a five-, six-, and seven-class configuration (complete linkage), and an eight-class configuration (single linkage) are presented visually in Fig. 8, where the results using Euclidean distance measure are displayed on the left-hand side, and the corresponding results using the fuzzy approach on the right-hand side.

For the distance-based approach the features were scaled to fit unity intervals, whereas the membership-based approach works independently of any scaling factors, since the fuzzy dissimilarity is expressed in truth values and therefore independent of the actual distance. This is a huge benefit, since the scaling (and thus their weight in the clustering process) of the features does not have to be defined by experts anymore.

For a $C = 5$ class configuration, both approaches deliver good, albeit different results. In Fig. 8b, the two groups in the lower-left corner form one ■-cluster due to their larger extents, whereas the ●-cluster—being much tighter in its extent—is separated from the ▼-cluster and forms a cluster on its own right. Experts confirm that this does make sense also from a semantic point of view, since all ■-objects exhibit barely any movement (feature x_1), but higher variations in their measurements, while the ●-objects actually do form a movement pattern on their own [12].

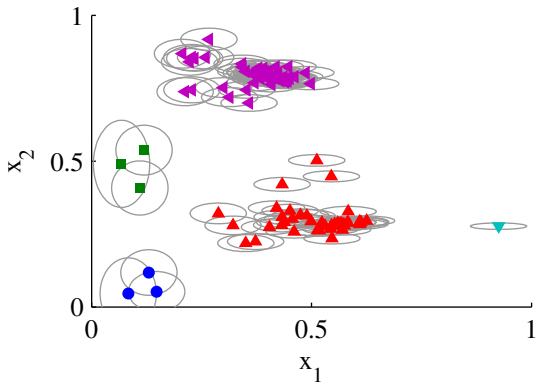
Beginning from a configuration of $C = 6$ classes, distance-based complete linkage clustering leads to results which are not acceptable both formally and from a semantic point of view, since the cluster in the centre is being split up in Fig. 8c into two clusters (● and ■). Similar objections can be raised to the results in Fig. 8e and Fig. 8g. For all class configurations $C = 6$, $C = 7$ and $C = 8$, the membership-based approach delivers well-interpretable results which are also in accordance with an expert's assessment in the field of geodesy [12]. Taking the individual extents for all objects into account—as made possible by the contributions of this paper—does very closely follow the approach of humans in this field.

As mentioned in section 4.1, an objective verification of our visual assessment would be desirable, but requires suitable performance measures incorporating individual characteristics of data, whose definition has to be left open for future work.

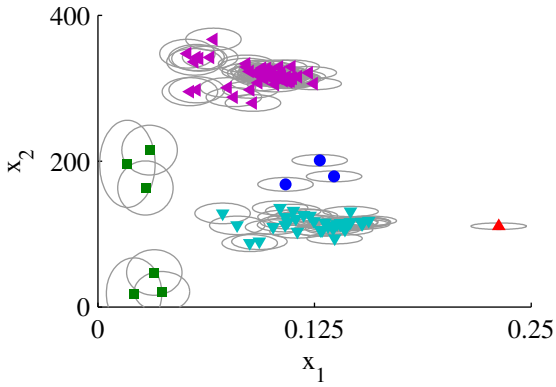
5. Conclusion

As shown in this article, multivariate membership functions are a suitable tool for a fuzzy modelling of objects in a feature space and their individual extents in a sense of representativeness. Such individual extents may stem both from data and expert knowledge. Contrary to the vast majority of data mining approaches, the distinct characteristics

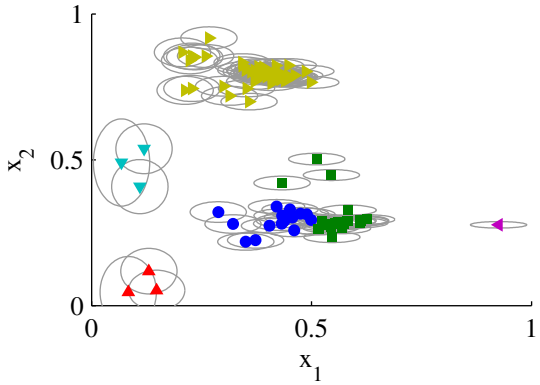
[3] The movement direction is a periodic feature, but treated as non-periodic in this example. All calculations have also been carried out considering the periodicity of this feature, but this did not lead to different results w. r. t. the goals of this article. For brevity, the necessary handling of periodic features will therefore not be covered in this article.



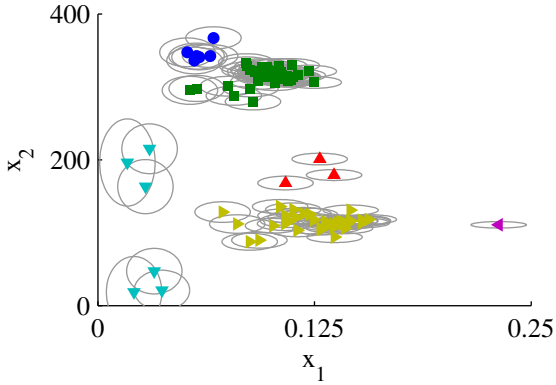
(a) Distance-based, complete linkage, $C = 5$



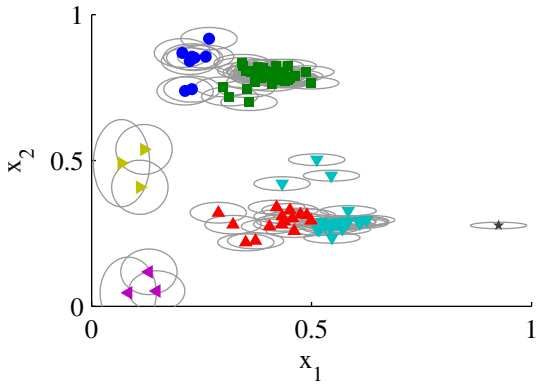
(b) Membership-based, complete linkage, $C = 5$



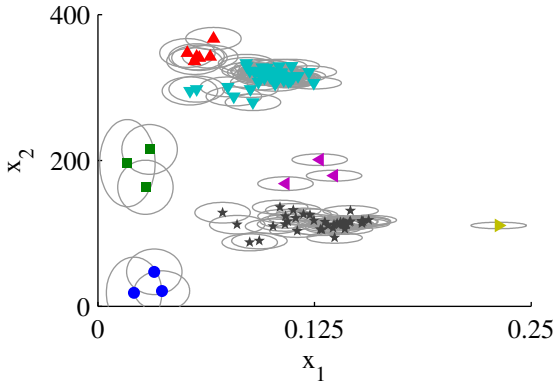
(c) Distance-based, complete linkage, $C = 6$



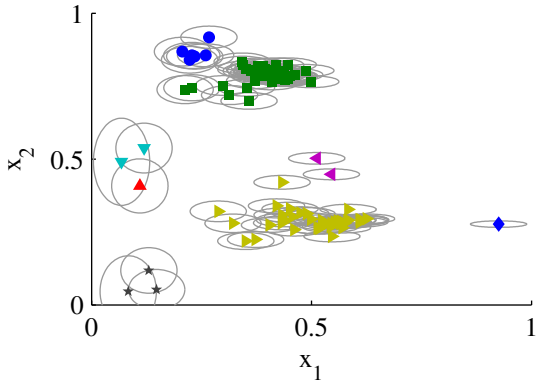
(d) Membership-based, complete linkage, $C = 6$



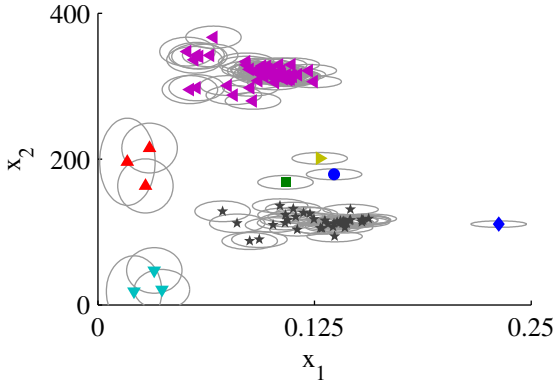
(e) Distance-based, complete linkage, $C = 7$



(f) Membership-based, complete linkage, $C = 7$



(g) Distance-based, single linkage, $C = 8$



(h) Membership-based, single linkage, $C = 8$

Figure 8: Results of single/complete linkage clustering of the Isnet dataset with C classes. The membership-based approach works independently of feature scaling and can therefore also use the original feature values.

of every object can thus be modelled and incorporated into subsequent steps of data analysis.

Based on the fuzzy description of data, we derived a fuzzy dissimilarity measure for a pair of objects, which can effortlessly be used as a plug-in replacement for distance measures in hierarchical clustering algorithms. As the main benefit of this approach, the individual extent of each object is now being considered in the clustering process. Regardless of the absolute distances of object locations, a group of objects is now considered as a cluster if the extents of its objects jointly form a (more or less) compact region of the feature space.

As a very useful side-effect of working with truth values instead of actual (e. g. Euclidean) distances for the dissimilary of objects, one can omit the scaling of features prior to a clustering analysis. The scaling is implicitely performed by an object's extent, and—even more importantly—individually for each pair of objects.

References

[1] Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, September 1999.

[2] Miin-Shen Yang. A survey of fuzzy clustering. *Mathematical and Computer Modelling*, 18(11):1–16, 1993.

[3] Wen-Liang Hung and Miin-Shen Yang. Fuzzy clustering on LR-type fuzzy numbers with an application in Taiwanese tea evaluation. *Fuzzy Sets and Systems*, 150:561–577, 2005.

[4] Steffen F. Bocklisch. *Prozeßanalyse mit unscharfen Verfahren*. Technik, Berlin, 1987.

[5] Arne-Jens Hempel and Steffen F. Bocklisch. Fuzzy pattern modelling of data inherent structures based on aggregation of data with heterogeneous fuzziness. In Gregorio Romero Rey and Luisa Martinez Muneta, editors, *Modelling Simulation and Optimization*, chapter 28, pages 637–655. INTECH, 2010.

[6] Arne-Jens Hempel, Gernot Herbst, and Steffen F. Bocklisch. Modelling and aggregation of heterogeneous fuzzy data. In *Proceedings of EUSFLAT-LFA 2011 (to appear)*, 2011.

[7] Divyendu Sinha and Edward R. Dougherty. Fuzzification of set inclusion: Theory and applications. *Fuzzy Sets and Systems*, 55:15–42, April 1993.

[8] Chris Cornelis, Carol Van der Donck, and Etienne Kerre. Sinha-Dougherty approach to the fuzzification of set inclusion revisited. *Fuzzy Sets and Systems*, 134:283–295, March 2003.

[9] Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, June 1965.

[10] Eyke Hüllermeier and Maria Rifqi. A fuzzy variant of the Rand index for comparing clustering structures. In *2009 International Fuzzy Systems Association World Congress and 2009 European Society for Fuzzy Logic and Technology Conference (IFSA-EUSFLAT 2009)*, pages 1294–1298, 2009.

[11] Guðmundur Þór Valsson, Þórarinn Sigurðsson, Christof Völksen, and Markus Rennen. ISNET2004: Niðurstöður úr endurmælingum grunnstöðvanets islands. Technical report, Landmælingar Islands, 2007.

[12] Rainer Fletling. *Methodische Ansätze zur unscharfen Mustererkennung bei Deformationsmessergebnissen*. PhD thesis, TU Braunschweig, 2010.