Density based fuzzy membership functions in the context of geocomputation

Victor Lobo[1,2], Fernando Bação[1], Miguel Loureiro[1]

[1]ISEGI - UNL, [2]Portuguese Naval Academy
{vlobo, bacao, mloureiro}@isegi.unl.pt

Abstract. Geocomputation has a long tradition in dealing with fuzzyness in different contexts, most notably in the challenges created by the representation of geographic space in digital form. Geocomputation tools should be able to address the imminent continuous nature of geo phenomena, and its accompanying fuzzyness. Fuzzy Set Theory allows partial memberships of entities to concepts with non-crisp boundaries. In general, the application of fuzzy methods is distance-based and for that reason is insensible to changes in density. In this paper a new method for defining density-based fuzzy membership functions is proposed. The method automatically determines fuzzy membership coefficients based on the distribution density of data. The density estimation is done using a Self-Organizing Map (SOM). The proposed method can be used to accurately describe clusters of data which are not well characterized using distance methods. We show the advantage of the proposed method over traditional distance-based membership functions.

Keywords: fuzzy membership, fuzzy set theory, density based clustering, SOM

1 Introduction

One of the most challenging tasks in geocomputation has been the need to provide an adequate digital representation to continuous phenomena such as those typically captured in geographic information. The need to define crisp boundaries between objects in geographic space leads to data representations that, while apparently providing a rigours description, in reality have serious limitations as far as fuzziness and accuracy are concerned. “There is an inherent inexactness built into spatial, temporal and spatio-temporal databases” largely due to the “artificial discretization of what are most often continuous phenomena” [1]. The subtleties that characterize space and time changes in geo-phenomena constitute a problem as they carry large levels of fuzziness and uncertainty. While fuzziness might be characterized as inherent imprecision which affects indistinct boundaries between geographical features, uncertainty is related to the lack of information [1]. Note that these characteristics are not limited to the spatial representation but also include categorization, and attribute data. All these facts lead to the eminently fuzzy nature of the data used in geocomputation, or as [2] puts it, “uncertainty is endemic in geographic data”. Rather then ignoring these problems and dismissing them as

irrelevant, the geocomputation field should be able to devise ways of dealing with them. In many instances this will translate into attributing uncertainty levels to the representations. This way the user will be aware of the limitations involved in the use of the data, and thus be able to intuitively attribute some level of reliability.

Fuzzy Set Theory constitutes a valuable framework to deal with these problems when reasoning and modeling in geocomputation [2]. Fuzzy Set Theory allows partial memberships of entities to concepts with non-crisp boundaries. The fundamental idea is that while it is not possible to assign a particular pattern to a specific class it is possible to define a membership value. In general, the application of automatic fuzzy methods is distance-based. Thus, the (geographic or attribute) distance between a prototype and a pattern defines the membership value of the pattern to the set defined by the prototype. This approach is not only intuitive but also adequate to deal with many different applications. Nevertheless, there is yet another way of approaching the problem, which trades distance for density. In this case, it is not the distance of the pattern to the prototype that governs the membership value, but the pattern density variations between them. This perspective will emphasize discontinuous zones assuming them has potential boundaries. Membership will be a function of the changes in the density in the path between the pattern and the prototype. Thus, if density is constant then the membership value will be high. There are several classical examples in clustering where the relevance of density is quite obvious [3]. In this paper a new method for defining density-based fuzzy membership functions is proposed. The method automatically determines fuzzy membership coefficients based on the distribution density of data. The density estimation is done using a Self-Organizing Map (SOM). The proposed method can be used to accurately describe data which are not well characterized using distance methods. We show the advantage of the proposed method over traditional distance-based membership functions.

2 Problem statement

Fuzzy Set Theory was introduced in 1965 by Zadeh [4]. A fuzzy set may be regarded as a set with non-crisp boundaries. This approach provides a tool for representing vague concepts, by allowing partial memberships of entities to concepts. In the context of data analysis, entities are usually represented by data patterns and concepts are also represented by data patterns that are used as prototypes. Fuzzy membership may be defined using the following formalism. Given:

- a set of n input patterns $\mathbf{X} = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_i}, \ldots, \boldsymbol{x_n}\}$, where $\boldsymbol{x_i} = (\boldsymbol{x_{i1}}, \ldots, \boldsymbol{x_{ij}}, \ldots, \boldsymbol{x_{id}})^T \in \Re^d$, and each measure $\boldsymbol{x_{ij}}$ is a feature or attribute of pattern $\boldsymbol{x_i}$
- a set of k concepts defined by prototypes $\mathbf{C}=\{\boldsymbol{c_1}, \ldots, \boldsymbol{c_c}, \ldots, \boldsymbol{c_k}\}$, where $\boldsymbol{c_c} = (c_{c1}, \ldots, c_{cj}, \ldots, c_{cd})^T \in \Re^d$, with $k<n$,

the fuzzy membership u_{ic} of pattern $\boldsymbol{x_i}$ to a prototypes c_c is defined so that:

$$u_{ic} \in [0,1],\ i=1, \ldots, n \text{ and } c=1, \ldots, k \qquad (1)$$

There are many ways of defining memberships u_{ic}. These may be divided into two major groups *e.g.* [5, 6] :

1- The probabilistic methods, where the sum of memberships of a pattern to all concepts has to add to 1:

$$u_{ic} = 1 \text{ , } i=1, \ldots, n \quad (\mathbf{2})$$

Some authors state that these memberships can be interpreted as a relative perspective *e.g.* [5], since the membership of a pattern to a given concept depends on the membership of that pattern to all other concepts.

2- The possibilistic methods, where it is only required that:

$$u_{ic} \geq 0, \; i=1, \ldots, n \quad (\mathbf{3})$$

While still maintaining the need to satisfy (1), the possibilistic approach relaxes the constraint imposed by (2). This perspective may be interpreted as an absolute way of determining the membership of a pattern to a concept, since its computation does not depend on the membership of that pattern to other concepts.

Different methods of determining fuzzy membership coefficients have been proposed *e.g.* [7]. Most of these methods are based on the distance between patterns and prototypes. However, in some cases distance based methods may not achieve the best results. A classical example of this is given by Ultsch [3] in the form of two rings, in which membership of a data pattern to one of the rings can not be defined as distance to any single point.

In this paper a new method for determining fuzzy membership coefficients based on variations of data pattern density is proposed. In order to compute the density distribution of data $\mathbf{X}$ in the $\Re^d$ input space, the use of a Self-Organizing Map (SOM) is proposed.

3 – Density estimation using a Self-Organizing Map

A Self-Organizing Map (SOM) is an unsupervised neural network. It was introduced in 1982 by Kohonen [8], and has been used as visualization tool for high dimensional data, as well as in many different tasks such as clustering, dimension reduction, classification, sampling, vector quantization and data mining [9].

The basic idea of a SOM is to map a set of data patterns onto a (usually) *2*-dimensional grid of neurons or units. That grid forms what is known as the output space, as opposed to the input space that is the original space where the data patterns are. When a SOM is trained with a given dataset, its units will tend to spread themselves in the input space in a way that is proportional to some function of the density of the data patterns [10]. This means that where data density is high, the SOM units are close to each other in the input space. In places where data density is low, SOM units, even if neighbors in the grid (output space), will be far apart from each other in the input space.

The SOM may be regarded as a graph [11], where the SOM units are the nodes,

and edges connect the units that are neighbors in the output space, *i.e.*, the edges form the regular grid of the SOM. We may associate to each of these edges a value that corresponds to the distance, in the input space, between its end nodes. Given two neighboring SOM units $\boldsymbol{w}_a$ and $\boldsymbol{w}_b$, the value associated with the edge E that joins them is:

$$\mathrm{E}(\boldsymbol{w}_a, \boldsymbol{w}_b) = \| \boldsymbol{w}_a - \boldsymbol{w}_b \| \quad (4)$$

In this paper, this graph is named a U-Graph. This U-graph is usually the first step in the computation of a U-Matrix [12], which is widely used in cluster analysis [13].

4 - An algorithm to compute fuzzy membership functions

Let **X** be the set of input patterns $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_n\}$ and **W** be a U-graph with $p \times q$ nodes $\boldsymbol{w}$, obtained by training a SOM with those data patterns.

The cost of a path P with e edges between two nodes a and b is defined as:

$$C(a,b) = \sum_{j=1}^{j=e-2} \left| E(w_j, w_{j+1}) - E(w_{j+1}, w_{j+2}) \right| \quad (5)$$

Where $\boldsymbol{w}_1$, $\boldsymbol{w}_2$, ..., $\boldsymbol{w}_e$ are the nodes that form the path. The cost is obtained from the sum of absolute differences of edges, between pairs of adjacent edges on the path P from node a to node b. If the path has less than 3 nodes, then the cost is considered to be zero.

If in some area of the SOM the data patterns have a more or less constant density, the distance between units will also be more or less constant, and thus the cost will be low. On the other hand, if there are variations in the data pattern density, the SOM units will be unevenly spaced. This varying distance between nodes will lead to a high value in the cost function.

Areas where the density is constant may be regarded as a continuous cluster, and thus all data patterns in that area should have a high membership to a reference prototype. On the other hand, if there are large variations in the data density, the data patterns should be considered to form different clusters, and thus the membership of a pattern to a prototype in another cluster should be low.

To convert low cost to high membership (and vice-versa), the membership of a pattern $\boldsymbol{x}_i$ to a concept $\boldsymbol{c}_c$ is calculated using:

$$u_{ab} = \frac{1}{1 + C(a,b)} \quad (6)$$

The procedure to compute the membership of a pattern $\boldsymbol{x}_i$ to concept $\boldsymbol{c}_c$ may now be defined as follows:

1. Train a SOM with the dataset **X**
2. Compute the U-Graph of that SOM
3. Find the node $\boldsymbol{w}_x$ that maps pattern $\boldsymbol{x}_i$
4. Find the node $\boldsymbol{w}_c$ that maps concept $\boldsymbol{c}_c$
5. Find the path between $\boldsymbol{w}_c$ and $\boldsymbol{w}_x$ that minimizes the cost function (5)
6. Compute the membership of $\boldsymbol{x}_i$ to $\boldsymbol{c}_c$ using membership function (6)

The first two steps must be done only once, while the others must be repeated for each data pattern. Step 5 involves solving an optimization problem on a graph. It must be pointed out that this is not simply the shortest path along the graph, which would be a trivial problem. The fact that the absolute value of the difference between two consecutive edges is used, instead of the values of the edges themselves, renders the traditional shortest path algorithms useless. While this is a serious practical problem, it is irrelevant for the purpose of this paper. In the tests presented in the next section, this optimization was performed using a very fast heuristic based on simulated annealing, that obtains a sub-optimal but still very useful solution.

5 – Preliminary results with artificial data

From a set of problems where classical methods behave poorly in fuzzy membership determination, the one shown in Fig. 1 was chosen. It is composed of two zones, each one with approximately uniform pattern density distribution, but with numerically different pattern densities between them. The patterns were placed randomly on both zones, 1000 patterns on zone A (the left half of the square) and 100 patterns on zone B (the right half of the square).

After training the SOM, more units will end up on zone A, due to the higher density of patterns, and the opposite happens in zone B (Fig.1). It is worth noticing that each zone has a prototype density approximately constant, which is a function of the pattern density [10].

To characterize zones A and B, one data pattern from each must be selected as a prototype, and the membership of all other patterns to those two prototypes must be computed. It would be desirable that all data patterns of zone A have a high membership to it's prototype, and a low membership to the other one (and vice-versa).

Two patterns were chosen as reference prototypes, one pattern from each of the zones A and B (Fig.1). They were intentionally placed asymmetrically with respect to the border between zones A and B.

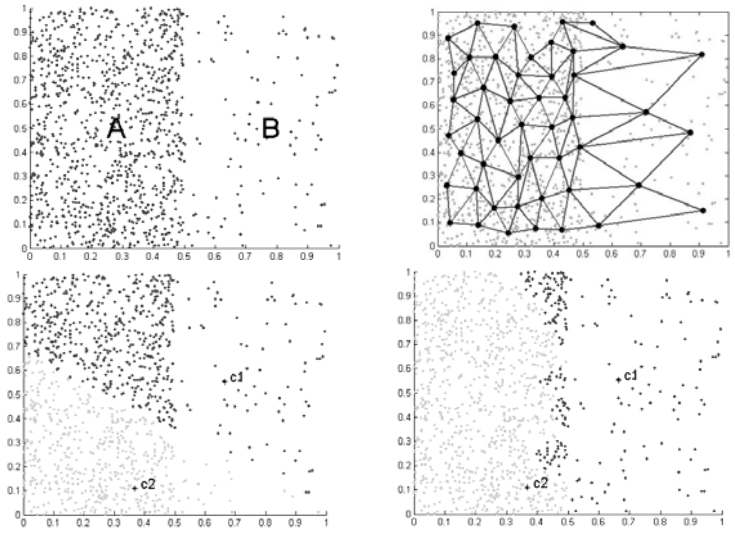

Fig. 1. Top left side: Pattern dataset and zone identification. Each zone has approximately constant pattern density. Top right side: SOM trained with the dataset. Notice the approximately constant density of SOM units in each zone. On the bottom we present the defuzzified results of using the classical membership functions (eq.7) on the left, and the proposed membership functions on the right. Darker patterns have higher membership to

To act as benchmark against the new method proposed, one of the most popular distance based methods was used [7], which computes membership of a data pattern $\boldsymbol{x}$ to concept $\boldsymbol{c}$ as:

$$u_{ij} = \frac{1}{\sum_{c=1}^{k} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_c \right\|} \right)^{\frac{2}{m-1}}} \tag{7}$$

The values of membership functions computed using this equation and using the proposed method are presented in Fig. 2. By observing these figures, we may easily see that:

1- When using probabilistic distance based methods, data patterns far away have average membership values, and these are approximately the same for all concepts. This uncertainty may be solved by using a "maximum" operator, but this defeats the purpose of fuzzy membership. On the other hand, using the approach proposed in this paper, even distant patterns are clearly identified as belonging to a cluster provided the pattern density is constant. The same effect could be obtained by certain clustering algorithms such as simple linkage, but these fail in other cases [3].

2- When using distance based methods, the borders between concepts will always be in the mid points between their respective prototypes. This means that the exact

positioning of these prototypes is critical. If the relative positions of the prototypes vary slightly, the line separating the classes may change considerably. If the borders are not straight lines, distance based methods cannot be used to identify them unless several prototypes are used for each concept (forming stepwise linear borders). In the method proposed in this paper, the borders will occur whenever there are variations in data density, regardless of how far from the reference prototype, and borders may have any shape.

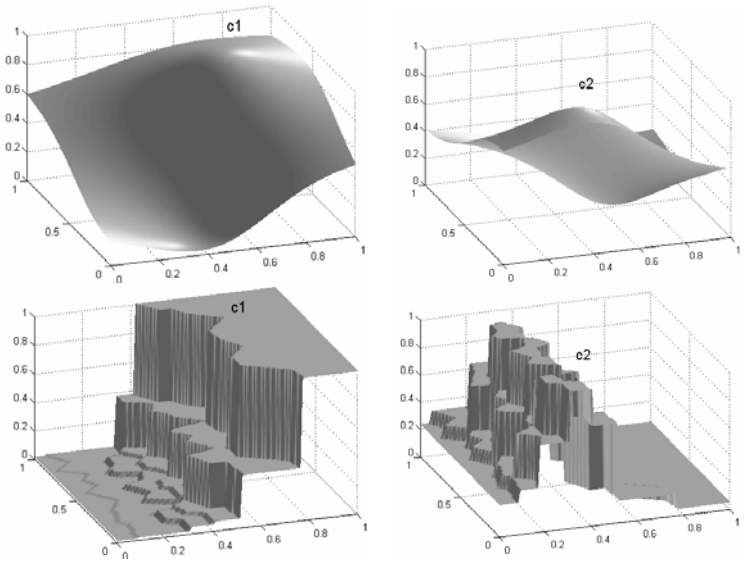

Fig. 2. Probabilistic membership functions to c_1 (right) and c_2 (left) obtained with (eq. 7) on the top row, and with the proposed method on the bottom.

3- When using probabilistic based methods, data patterns on the borders are considered to have 0.5 membership to each concept. In many cases this is not at all true, since the border belongs simultaneously to both concepts, with high membership values. In the method proposed in this paper, the borders have high membership to both concepts.

6 – Conclusions and future work

In this paper a new method for computing fuzzy memberships was proposed. This method allows membership to be sensibly determined in special cases where distance based methods fail.

The method requires computing a large SOM and U-Graph with the available data. This is not a problem, since the algorithm for training SOMs is quite fast and is easily

parallelized [14].

One of the bottlenecks of the proposed algorithm is the optimization of the path along the U-Graph. This is an interesting problem requiring some research, but fortunately, as shown in our tests, simple heuristics are quite effective for this kind of problem.

Preliminary results with artificial datasets indicate that the new method is indeed efficient in characterizing clusters which are adjacent but have different densities.

Although not tested experimentally, it seems clear that the new method will also characterize correctly any clusters, provided that between them there are areas where data density is low, since the variation in data density will induce high values in the cost function, and thus low values in membership. This effect will occur even if the clusters have very irregular or "long" shapes.

7 – References

1. Peuquet, D.J., *Making Space for Time: Issues in Space-Time Data Representation.* GeoInformatica, 2001. **5**(1): p. 11-32.
2. Goodchild, M., *Introduction: Special Issue on 'Uncertainty in geographic information systems'.* Fuzzy sets and systems, 2000. **113**(1): p. 3-5.
3. Ultsch, A. *Clustering with SOM: U*C*. in *WSOM 2005*. 2005. Paris.
4. Zadeh, L.A., *Fuzzy Sets.* Information and Control, 1965. **8**: p. 338-353.
5. Baraldi, A. and P.Blonda, *A survey of Fuzzy Clustering Algorithms for Pattern Recognition - Part I.* IEEE Trans. on Systems, Man and Cybernetics - Part B: Cybernetics, 1999. **29**: p. 778-785.
6. Krishnapuram, R. and J.M. Keller, *A possibilistic approach to clustering.* IEEE Transactions on Fuzzy Systems, 1993. **1**: p. 98-110.
7. Bezdek, J.C., et al., *Fuzzy models and algorithms for pattern recognition and image processing*. 1999: Kluwer Academic Publishers.
8. Kohonen, T., *Self-Organizing Maps*. 3rd ed. Information Sciences. 2001, Berlin-Heidelberg: Springer. 501.
9. Vesanto, J., *Data Mining Techniques Based on the Self-Organizing Map*, in *Department of Engineering Physics and Mathematics*. 1997, HELSINKI UNIVERSITY OF TECHNOLOGY: Helsinkia. p. 71.
10. Cottrell, M., J.C. Fort, and G. Pages, *Theoretrical Aspects of the SOM algorithm.* Neurocomputing, Elsevier, 1998. **21**: p. 119-138.
11. Bauer, H.-U. and K.R. Pawelzik, *Quantifying the Neighborhood Preservation of Self-Organizing Maps.* IEEE Transactions on Neural Networks, 1992. **3**: p. 570-579.
12. Ultsch, A. and H.P. Simeon, *Exploratory Data Analysis Using Kohonen Networks on Transputers*. 1989, Department of Computer Science, University of Dormund, FRG.
13. Vesanto, J. and E. Alhoniemi, *Clustering of the Self-Organizing Map.* IEEE Transactions on Neural Networks, 2000. **11**(3): p. 586-600.
14. Bandeira, N. and V. Lobo. *Training a Self-Organizing Map distributed on a PVM network*. in *IEEE World Conference on Computational Intelligence*. 1998. Anchorage, Alaska, USA.