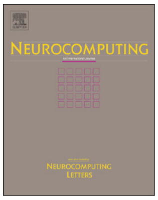
Neurocomputing 147 (2015) 47–59

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Automatic design of interpretable fuzzy predicate systems for clustering using self-organizing maps

Gustavo J. Meschino [a,*], Diego S. Comas [b,c], Virginia L. Ballarin [b], Adriana G. Scandurra [a], Lucía I. Passoni [a]

[a] Bioengineering Laboratory, Facultad de Ingeniería, Universidad Nacional de Mar del Plata, Juan B. Justo 4302, Mar del Plata B7608FDQ, Argentina
[b] Digital Image Processing Group, Facultad de Ingeniería, Universidad Nacional de Mar del Plata, Juan B. Justo 4302, Mar del Plata B7608FDQ, Argentina
[c] Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, Argentina

ARTICLE INFO

Article history:
Received 3 April 2013
Received in revised form
3 December 2013
Accepted 2 February 2014
Available online 21 May 2014

Keywords:
Self-organizing maps
Clustering
Fuzzy logic
Fuzzy predicates
Degree of truth

ABSTRACT

In the area of pattern recognition, clustering algorithms are a family of unsupervised classifiers designed with the aim to discover unrevealed structures in the data. While this is a never ending research topic, many methods have been developed with good theoretical and practical properties. One of such methods is based on self organizing maps (SOM), which have been successfully used for data clustering, using a two levels clustering approach. Newer on the field, clustering systems based on fuzzy logic improve the performance of traditional approaches. In this paper we combine both approaches. Most of the previous works on fuzzy clustering are based on fuzzy inference systems, but we propose the design of a new clustering system in which we use predicate fuzzy logic to perform the clustering task, being automatically designed based on data. Given a datum, degrees of truth of fuzzy predicates associated with each cluster are computed using continuous membership functions defined over data features. The predicate with the maximum degree of truth determines the cluster to be assigned. Knowledge is discovered from data, obtained using the SOM generalization aptitude and taking advantage of the well-known SOM abilities to discover natural data grouping when compared with direct clustering. In addition, the proposed approach adds linguistic interpretability when membership functions are analyzed by a field expert. We also present how this approach can be used to deal with partitioned data. Results show that clustering accuracy obtained is high and it outperforms other methods in the majority of datasets tested.

1. Introduction

Clustering aims to discover unrevealed structures in data and is a never-ending research topic. It is a major task in exploratory data mining and new approaches are constantly proposed, because the usage and interpretation of clustering depend on each particular application [1]. Clustering is currently applied in many fields, such as web and text mining, business and marketing, machine learning, pattern recognition, image analysis and segmentation, information retrieval, and bioinformatics (see e.g. [2–4]). In many complex problems, general clustering techniques are not able to adequately discover groups when directly applied on the data [5].

The self-organizing maps (SOM), introduced by Kohonen in 1982 [6], are widely used, unsupervised and nonparametric neural network, with remarkable abilities to remove noise, outliers, and deal with missing values. The SOM training process generates simultaneous clustering and projection of high-dimensional data. SOM have been successfully used in data clustering via two-level clustering approaches. The first level consists in training a SOM with a dataset. In the second level various techniques have been used, such as a second SOM [7], crisp-clustering methods [5,8,9] or fuzzy clustering techniques [10–12]. The second level links cells of the first-level SOM to form clusters. Then each datum is typically associated to the cluster assigned to its Best Matching Unit (BMU). Vectors of the codebook can be interpreted as "protoclusters," which are combined to form the actual clusters.

In the two-level clustering approach, the SOM in the first level generates a projection of the original data which makes a general clustering technique suitable in the second level. One advantage of SOM in the two-level approach is the reduction in the computational cost [5]. Even considering a small data package, some clustering algorithms become intractable. Grouping prototypes instead of grouping data directly is a solution for this problem.

[*] Corresponding author. Tel.: +54 223 4816600; fax: +54 223 481 0046.
E-mail addresses: gmeschin@fi.mdp.edu.ar (G.J. Meschino), diego.comas@fi.mdp.edu.ar (D.S. Comas), vballari@fi.mdp.edu.ar (V.L. Ballarin), scandu@fi.mdp.edu.ar (A.G. Scandurra), lpassoni@fi.mdp.edu.ar (L.I. Passoni).

http://dx.doi.org/10.1016/j.neucom.2014.02.059
0925-2312/© 2014 Elsevier B.V. All rights reserved.

Another additional benefit is the reduction in data noise effects, since prototypes are local averages of the data and, therefore, they are less sensitive to random variations than the original data [5].

Clustering systems based on Fuzzy Logic (FL) [13–15] have been used in a wide range of clustering problems, improving the performance of traditional approaches. Based on the fuzzy set theory, FL was proposed by Zadeh [16], who stated that a complex system will be better represented by descriptive variables of linguistic types [17]. Most of the previous works are based on Fuzzy Inference Systems (FIS). The main advantages of these models are (a) they use simple IF–THEN rules to determine the conditions a datum must satisfy to belong to each cluster and (b) FIS allow modeling the data imprecision by way of membership functions. However, in FIS, aggregation and defuzzification operations must be defined, so these models do not constitute a Boolean Logic generalization. Given that defuzzification is a pragmatic combination of operators, it lacks an axiomatic link that justifies the "logic" denomination [18].

Unlike previous works, in this paper we propose the design of a new clustering system in which (a) we use predicate fuzzy logic [19] to perform the clustering task, which is a natural extension of predicate Boolean Logic and (b) the system is automatically designed (unsupervised) [20] using a two-level clustering approach that combines SOM and Fuzzy C-Means (FCM) as a second level clustering method. First, SOM are trained from the original data, considering a SOM with a number of cells much larger than the expected number of clusters. Then the SOM codebook (the set of protoclusters) is clustered. Next, the codebook clustering is analyzed and membership functions and predicates are defined. Thus we obtain a ranked clustering criteria represented as self-discovered fuzzy predicates using data information [20] which consider the behavior of the variables into the different clusters. Hereafter we will name the proposed method SOM-based Fuzzy Predicate Clustering (SFPC).

Given a datum, degrees of truth of fuzzy predicates associated to each cluster are computed using continuous membership functions defined over data features. Finally, the predicate with the maximum degree of truth determines the cluster to be assigned (the first of the ranking). From a linguistic point of view we obtain a ranking for each datum, computing the degree of truth of "*The datum D belongs to cluster k*", being $k=1, 2, \ldots, K$, and K the amount of clusters. This ranking could also be used to determine the group to which new similar input data (not contained in the training dataset) belongs. Besides, this system allows comparing the degree of membership to the clusters and also assessing the contribution of individual features to the final decision. If required, it allows assessing how far the datum was from being assigned to other cluster (the next one in the ranking).

The proposed approach adds value to previous knowledge-based fuzzy clustering methods [21]. In this case knowledge is discovered from data, obtained using the SOM generalization aptitude and taking advantage of the well-known SOM abilities to discover natural data groupings when compared with direct clustering [5,7,22]. Also SOM provides useful information about the features in each cluster, an interesting property that was used for different applications [23]. In addition, the proposed approach adds linguistic interpretability when the clustering obtained and the knowledge discovered from the membership functions are analyzed by a field expert. This approach can also be used to deal with partitioned data as we present later in the paper, where we explain how the predicate system can join the results obtained from different partitions.

The paper is structured as follows. Section 2 discusses the main works related to SOM-clustering and fuzzy systems. Some important concepts concerning this work on SOM and fuzzy predicates systems are presented in Section 3. Section 4 details with the proposed methodology to design the fuzzy clustering system by way of a SOM–FCM scheme and the method for assessing the clustering quality is explained. In Section 5 we show the accuracy of the results on several datasets, and then we develop two examples of interpretability of membership functions and predicates. Finally we conclude by discussing the results, limitations and future research directions in Sections 6 and 7.

2. Related works

Before presenting the SFPC method proposed in this paper, in this section we discuss relevant works related to SOM-clustering and fuzzy systems applications in clustering.

2.1. SOM-clustering

In the literature, several approaches were developed in the attempt to achieve clusters of data from a trained SOM. The general approach uses a two-level process. In the first level, a SOM is trained. The second level can involve another SOM, making a hierarchical SOM [7] or some crisp-clustering method [5,8,9] which allows linking codebook cells of the first-level SOM in bigger clusters. In other works the SOM fuzzification problem was specifically studied using a FCM algorithm [11] or other fuzzy methods [10,12] as second level clustering.

Lampinen and Oja [7] proposed a clustering method using a multilayer SOM (HSOM) to achieve complex data groupings. They argue that when the abstraction level of the classification task increases, the shapes of the regions associated together become complex, requiring very large amounts of training data to form the class boundaries. In these cases, using unsupervised learning techniques can reduce the required training. They state that as the goal of SOM-learning is not only to find the most representative code vectors for the input space in mean square sense, but at the same time to realize a topological mapping from the input space to the grid of neurons; updating the internal weights of the network tends to preserve the input topographic space. Therefore, neighboring cells are related to nearby data in the original space and a grouping of data can be achieved through the SOM, where topographically nearby cells are associated with the same cluster. In particular, the HSOM algorithm is able to achieve clustering of complex data structures.

Vesanto and Alhoniemi [5] developed a two-level SOM-clustering approach that uses agglomerative clustering in the second level. A large set of prototypes, interpreted as "protoclusters," larger than the number of clusters, is achieved by the SOM in the first level. The protoclusters are combined together to form the actual clusters in the second level. Each data vector of the original dataset belongs to the same cluster as its nearest prototype. While extra abstraction levels yield higher distortion, they also effectively reduce the complexity of the reconstruction task. They indicate some advantages of using SOM in the first level that we cited in the previous section.

On the way to finding the best emergent clustering, Murtagh [8] presented a convenient framework for clustering, applying the contiguity-constrained clustering method in the second-level clustering. A SOM is trained and the prototype vectors are interpreted as clusters defined by a minimum distance criterion, with the cluster centers located on the discretized plane in such a way that proximity reflects similarity. The algorithm searches for the minimum distance between prototype vectors belonging to two different clusters and integrates them into a single group. This recursive algorithm ends when a stop condition defined by the user is reached. This condition may be the number of defined clusters.