ARTICLE IN PRESS
FSS:6927

Available online at www.sciencedirect.com
ScienceDirect
Fuzzy Sets and Systems ••• (••••) •••–•••
FUZZY sets and systems
www.elsevier.com/locate/fss

Rederivation of the fuzzy–possibilistic clustering objective function through Bayesian inference

Arash Abadpour

Imaging Group, Epson Edge, Epson Canada Limited, Canada

Received 10 March 2014; received in revised form 14 August 2015; accepted 8 October 2015

Abstract

Unsupervised clustering of a set of datums into homogeneous groups is a primitive operation required in many signal and image processing applications. In fact, different incarnations and hybrids of Fuzzy C-Means (FCM) and Possibilistic C-means (PCM) have been suggested which address additional requirements such as accepting weighted sets and being robust to the presence of outliers. Nevertheless, arriving at a general framework, which is independent of the datum model and the notion of homogeneity of a particular problem class, is a challenge. However, this process has not been followed organically and clustering algorithms are generally based on exogenous objective functions which are heuristically engineered and are believed to lead to the satisfaction of a required behavior. These techniques also commonly depend on regularization coefficients which are to be set "prudently" by the user or through separate processes. In contrast, in this work, we utilize Bayesian inference and derive a robustified objective function for a fuzzy–possibilistic clustering algorithm by assuming a generic datum model and a generic notion of cluster homogeneity. We utilize this model for the purpose of cluster validity assessment as well. We emphasize the epistemological importance of the theoretical basis on which the developed methodology rests. At the end of this paper, we present experimental results to exhibit the utilization of the developed framework in the context of four different problem classes.

Keywords: Fuzzy clustering; Possibility theory; Fuzzy system models

1. Introduction

Unsupervised grouping of datums into homogeneous clusters is an important process required in a vast number of signal and image processing applications. In essence, the requirement is to have a process which inputs a set of datums, of a particular model, and divides them into an appropriate number of clusters, where the clusters comply with a particular notion of homogeneity. This process is also required to satisfy additional constraints, such as being robust and allowing for datums with different priorities. Moreover, it is an important advantage for such a process to be independent of any particular datum and cluster model. Also, from the point of view of practicality, a process which depends on a few parameters with perceptual definitions is highly preferred to an alternative which requires deliberate adjustment of a larger set of incomprehensible configuration variables.

E-mail address: arash_abadpour@ea.epson.com.

http://dx.doi.org/10.1016/j.fss.2015.10.005
0165-0114/

Please cite this article in press as: A. Abadpour, Rederivation of the fuzzy–possibilistic clustering objective function through Bayesian inference, Fuzzy Sets and Systems (2015), http://dx.doi.org/10.1016/j.fss.2015.10.005

A frequently utilized approach to the data clustering problem is to adopt an objective function and to find a minimizer for it. In this context, a notion of distance between a datum and a cluster is defined and the algorithm attempts to reduce the sum of the distances between all datums and all clusters. This process takes into consideration additional factors such as datum-to-cluster membership and the presence of outliers.

In effect, clustering problems are generally translated from a verbal description into an objective function, one or more constraints, and a potential set of pruning procedures. However, it is generally the intuition of the researchers which drives the development of new clustering algorithms. In such processes, the inclusion of a new regularization term or a novel weight component in the objective function is justified through verbal descriptions which are based on heuristics. In other words, the development of fuzzy clustering algorithms is often output-based, where experimental results provide the “proof” that the heuristics of the researchers have been correct. Comparison of different fuzzy clustering algorithms, too, is often only achieved based on the outputs of the respective formulations.

In this paper, we utilize Bayesian inference and derive the formulation for loss in a fuzzy–possibilistic clustering problem. This loss model is independent of any particular datum or cluster model and utilizes a generic notion of homogeneity. This model also utilizes a robust loss function and employs the concept of the relevance of each datum to the set. Moreover, the developed loss model is independent of any parameter which might have to be trained or tuned empirically or through training or repetition of any process. It is important to emphasize that the process utilized in this work develops the objective function organically and is therefore characteristically different from previous works in the field which are based on heuristics. Nevertheless, the developed formulation bears resemblance to different works in the field.

The rest of this paper is organized as follows. First, in Section 2, we review the related literature and then, in Section 3, we present the developed method. Subsequently, we present experimental results produced by the developed method on four different problem classes in Section 4 and provide the concluding remarks in Section 5.

2. Literature review

2.1. Notion of membership

The notion of membership is a key point of distinction between different clustering schemes. Essentially, membership may be *Hard* or *Fuzzy*. Within the context of hard membership, each datum belongs to one cluster and is different from all other clusters. The fuzzy membership regime, however, maintains that each datum in fact belongs to all clusters, with the stipulation that the degree of membership to different clusters is different. K-means and Hard C-means (HCM) clustering algorithms, for example, utilize hard membership values. Iterative Self-Organizing Data Clustering (ISODATA) [1] is a hard clustering algorithm as well.

With the introduction of Fuzzy Theory [2], many researchers incorporated this more natural notion into clustering algorithms [3]. The premise for employing a fuzzy clustering algorithm is that fuzzy membership is more applicable in practical settings, where generally no distinct line of separation between clusters is present [4,5]. Additionally, from a practical perspective, it is observed that hard clustering techniques are more prone to falling into local minima [6]. The reader is referred to [7,8] for the wide array of fuzzy clustering methods developed in the past few decades. FCM, described below, is a fuzzy clustering algorithm. In this work we utilize the concept of fuzzy clustering as well.

Initial work on fuzzy clustering was done by Ruspini [9] and Dunn [10] and it was then generalized by Bezdek [7] into FCM. In FCM, datums, which are denoted as $x_1, \cdots, x_N$, belong to $\mathbb{R}^k$ and clusters, which are identified as $\psi_1, \cdots, \psi_C$, are represented as points in $\mathbb{R}^k$. FCM makes the assumption that the number of clusters, C, is known through some separate process or expert opinion and minimizes the following objective function,

$$\Delta = \sum_{c=1}^{C} \sum_{n=1}^{N} f_{nc}^{m} \|x_n - \psi_c\|^2 . \tag{1}$$

This objective function is heuristically suggested to result in appropriate clustering results and is constrained by,

$$\sum_{c=1}^{C} f_{nc} = 1, \forall n. \tag{2}$$

Here, $f_{nc} \in [0, 1]$ denotes the membership of datum n to cluster c.

Please cite this article in press as: A. Abadpour, Rederivation of the fuzzy–possibilistic clustering objective function through Bayesian inference, Fuzzy Sets and Systems (2015), http://dx.doi.org/10.1016/j.fss.2015.10.005