Computers and Electronics in Agriculture 150 (2018) 118–124

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Fuzzy clustering and fuzzy validity measures for knowledge discovery and decision making in agricultural engineering

Vania C. Mota, Flavio A. Damasceno, Daniel F. Leite*

Department of Engineering, Federal University of Lavras (UFLA), Lavras, Minas Gerais 37200-000, Brazil

ARTICLE INFO

Keywords:
Fuzzy clustering
Pattern recognition
Fuzzy validation measure
Decision support system
Agricultural engineering

ABSTRACT

This paper concerns the application of fuzzy clustering methods and fuzzy validity measures for decision support in agricultural environment. Data clustering methods, namely, K-Means, Fuzzy C-Means, Gustafson-Kessel, and Gath-Geva, are briefly reviewed and considered for analyses. The efficiency of the methods is determined by indices such as the Xie-Beni criterion, Partition Coefficient, and Partition and Dunn indices. In particular, fuzzy classifiers are developed to assist decision making regarding the control of variables such as bed moisture, temperature, and bed aeration in compost bedded pack barns. The idea is to identify interactive factors, promote cattle welfare, improve productivity indices, and increase property value. Data from 42 CBP barns in the state of Kentucky, US, were considered. Six classes related to the degree of efficiency of the composting process were identified. The GG method was the most accurate followed by the GK method. The main reason for the best results is the use of maximum-likelihood and Mahalanobis distance measures. A remark on the use of the Dunn validation index for different cluster geometries is given. Fuzzy models and linguistic information have shown to be useful to help decision making in cattle containment systems.

1. Introduction

Data analysis has driven knowledge discovery and decision making in a variety of application areas. Data uncertainty is quite often abstracted away or avoided in computational models and learning methods. The notion of fuzzy clustering (Bezdek, 1981) emerged as a means to summarize information and data to support various processes of comprehension and decision making. Numeric data are aggregated into kinds of information granules or sets of elements that are perceived as being similar or functionally equivalent (Pedrycz and Gomide, 2007). Agricultural engineering is a potential area of application of fuzzy methods and models. In particular, determining compost quality in agricultural environment is a problem of major importance because it is directly related to productivity and property value.

Compost Bedded Pack (CBP) barns is an alternative and modern containment system, where animals (cattle) have more freedom of movement inside the facility and are able to lie down in a more natural manner. The idea is to improve overall health and longevity of dairy cattle, and hence improve herd productivity indices (Black et al., 2013). CBP is similar to Loose Housing, except for the fact that the bed where the cows stand, walk and lie down is handled differently. In CBP, the bedded pack should be aerated twice a day to refresh the surface and stimulate microbial activity (Black et al., 2013).

According to Leso et al. (2013), the bedding material should be carefully managed to provide a dry and soft surface in which cows can stand and walk. The most commonly used materials to compose a bed are sawdust, wood shavings, and corn straw. The bedding material when mixed with cow manure produces a fertilizer with a proper amount of organic matter to improve soil fertility (Galama, 2011). An issue in CBP systems is to evaluate and maintain appropriate chemical substrate in the bed.

The success of CBP systems depends mainly on the management of the bed. When the bedded pack is handled correctly, the composting will increase the bedding temperature to reduce pathogenic microbial populations and to decrease the bedding moisture by increased drying rate (Barberg et al., 2007). A dry and soft bedding material also improves the cleaning of the animal and therefore reduces associated problems. The purpose of developing a data-driven monitoring and classification system for compost barns is to analyze variables related to the dairy cattle confinement system in order to infer about a specific index called composting process index (CPI).

Clustering algorithms with different properties can be found in the machine learning literature (Kaufman and Rousseeuw, 2008). These algorithms are useful for constructing monitoring and classification systems. Usually, algorithms give quite different solutions and in fact there is no single “best” clustering method for all possible data sets and

* Corresponding author.
E-mail addresses: esis@posgrad.ufla.br (V.C. Mota), flavio.damasceno@deg.ufla.br (F.A. Damasceno), daniel.leite@deg.ufla.br (D.F. Leite).

https://doi.org/10.1016/j.compag.2018.04.011
Received 15 January 2018; Received in revised form 8 March 2018; Accepted 13 April 2018

classes of problems. An important effort is to select a proper or the best clustering algorithm for a dataset from candidate clustering algorithms (Wang et al., 2009). The clustering methods considered in this work to construct a CPI monitoring and classification system are data-driven fuzzy clustering methods (Zadeh, 1965; Sadaaki et al., 2008).

Studies on fuzzy clustering are motivated by the need to improve results in complex issues of classification and grouping of data (Giusti and Marsili-Libelli, 2010; Pedrycz and Gomide, 2007). Existing non-fuzzy clustering methods have performance limitations and application restrictions related to linear and solid (hard) decision boundaries between classes (Leite et al., 2013; Pedrycz and Gomide, 2007; Sadaaki et al., 2008). The fundamental distinction between fuzzy and non-fuzzy clustering concerns the notion of partial membership supported by fuzzy sets. Fuzzy clusters are an important feature for describing information granules whose elements may belong only partially. Fuzzy clusters avoid specifying solid borders between full belongingness and full exclusion by means of smooth transition boundaries (Leite et al., 2013; Pedrycz and Gomide, 2007).

The goal of the CPI fuzzy classification model proposed in this article is to determine boundary functions capable of performing a mapping between data samples and their classes. The classes refer to composting grades and descriptions. In essence, the classifiers attempt to find similarities between data samples so that a distance measure between members of a cluster is minimized and the distance between members of different groups is maximized. The resulting classification model is useful to assist decision-making on farms that use CBP barns and systems. Greater composting scores can be achieved based on the monitoring of the CPI index.

The remainder of this paper is organized as follows. Section 2 describes the data clustering methods considered for evaluation, viz. K-means (KM, a non-fuzzy method), Fuzzy C-Means (FCM), Gustafson-Kessel algorithm (GK), and Gath-Geva algorithm (GG). In addition, this section addresses procedures to set initial parameters of algorithms, and indices to validate the resulting models. Section 3 describes the methodology and details the data acquisition and processing procedures. Clustering and classification results are presented and discussed in Section 4. Concluding remarks and suggestions for future work are given in Section 5.

2. Fuzzy modeling

2.1. Fundamentals

The theory of fuzzy systems provides tools that can be used to transform mathematically inaccurate or uncertain information captured in our everyday situations into computationally tractable data. The purpose is to treat natural phenomena and/or real situations as they are, in order to have computational models that are closer to what is real or true (Zadeh, 1965).

In fuzzy systems, a generalized characteristic function associated with an element of a universe of discourse gives a value in a range between 0 and 1. This value indicates the degree of relevance or membership of an element to a set (Zadeh, 1965). Let μ_A be the membership function of a fuzzy set A then $\mu_A: X \rightarrow [0,1]$ where X is the universe of discourse and [0,1] defines an infinite range of states whose endpoints mean the irrelevance of an element x of X to the set A,$\mu_A = 0$, and the overall relevance of x to A,$\mu_A = 1$. The remaining values in the range mean different degrees of membership of an element to a set.

Usually a dataset has some kind of organization that can be explained through a set of partitions (clusters). The study of these partitions is a data analysis task, which is based on some measure of similarity and an unsupervised learning algorithm. These algorithms use unlabeled data to build pattern classification systems. Often, the acquisition of labeled data requires human experts to manually classify training samples. Manual classification can be greatly influenced by subjectivity as well as not be feasible, as in when we handle large data sets. There are situations in which instances are labeled and apparently ask for fully supervised learning methods and standard procedures of classifier design. However, the labeling process may have been unreliable so that our confidence in the labels already assigned is relatively low (Pedrycz and Gomide, 2007; Leite et al., 2012). In these cases, we resort to unsupervised learning methods.

Let an input-output pair (x,y) be related through $y = f(x)$. We seek an approximation to f that allows us to predict the value of y given x. In classification problems, y is a class label, a value in a set $C_1,...,C_m \in N^m$, and the relation f specifies class boundaries. In the unsupervised case, C_k is not known and naturally cannot be used to guide learning. Patterns should be identified from input data x (Leite et al., 2012).

2.2. Fuzzy clustering

Cluster analysis refers to a broad spectrum of methods that try to subdivide a dataset X into c subsets (clusters) which are pairwise disjoint and non-empty. The subsets reproduce X via union (Bezdek, 1981). A cluster is usually formed by grouping similar data samples around a center, called centroid or cluster prototype.

Fuzzy clustering methods originated from the data analysis and pattern recognition fields (Babuska, 1998). These methods are fundamental to many classes of problems. They are intrinsically embedded in fuzzy approaches for dynamical systems modeling, time series prediction, pattern recognition, control and function approximation, where the fuzzy aspect may provide smooth nonlinear solutions (Lughofer et al., 2017; Pratama et al., 2017; Rubio and Bouchachia, 2017; Rubio, 2017). The concept of fuzzy relevance is used to represent the degree to which a given data sample is similar to other elements of a cluster. The degree of similarity between a data sample and a centroid can be calculated using a suitable distance measure.

The main similarity measures used in clustering algorithms are based on the Manhattan (1-norm), Euclidean (2-norm), Chebyshev (∞-norm), and Mahalanobis distances. In particular, the Mahalanobis distance overcomes some of the limitations of p-norms for n-dimensional (n-finite) vector spaces since it automatically takes into account the scale of the coordinated axes and correlations among variables. However, covariance matrices require additional calculations, i.e., computing time grows quadratically with the amount of variables. In practice, the higher the value of n, the greater the number of standard deviations that a data sample is away from the center of a cluster, and the lower its chance of being an element of the cluster. Moreover, for p-norms,

$$||x||_{p+a} \leqslant ||x||_p \text{ for any } x \in \Re^n, p \geqslant 1, a \geqslant 0, \quad (1)$$

that is, the p-norm $||x||_p$ of any given vector x does not grow with p; any other norm is lower bounded by the 1-norm.

The algorithms KM, FCM, GK and GG, briefly presented below, follow similar procedures: (i) setting of initial parameters and cluster centers; (ii) calculation of distances between data samples and cluster centers; and (iii) updating of a partition matrix, cluster centers, and associated parameters. The latter steps are repeated until some termination criterion is met. As these algorithms had already been extensively discussed in the literature, e.g. see (Babuska, 1998), only a brief review is given. Notice that while KM and FCM use Euclidean distance, GK and GG employ Mahalanobis and Gaussian distance functions, respectively – being this the main feature that differentiate the shape of the resulting clusters.

2.2.1. K-means clustering

The objective of the KM clustering algorithm, a crisp algorithm, is to partition a data set X into c clusters. From an $N \times n$ dimensional data set, KM allocates each data sample $x_k \in \Re^n, k = 1,...,N$, to one of the c clusters in order to minimize the within-cluster sum of squares: