Applied Soft Computing 28 (2015) 19–29

Contents lists available at ScienceDirect

Applied Soft Computing

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

A modified fuzzy min–max neural network for data clustering and its application to power quality monitoring

Manjeevan Seera [a,*], Chee Peng Lim [b], Chu Kiong Loo [a], Harapajan Singh [c]

[a] Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
[b] Centre for Intelligent Systems Research, Deakin University, Geelong, Victoria, Australia
[c] Faculty of Electrical Engineering, Universiti Teknologi MARA, Penang, Malaysia

ARTICLE INFO

Article history:
Received 28 December 2013
Received in revised form 25 July 2014
Accepted 20 September 2014
Available online 8 December 2014

Keywords:
Clustering
Fuzzy min–max neural network
Benchmark study
Power quality monitoring

ABSTRACT

When no prior knowledge is available, clustering is a useful technique for categorizing data into meaningful groups or clusters. In this paper, a modified fuzzy min–max (MFMM) clustering neural network is proposed. Its efficacy for tackling power quality monitoring tasks is demonstrated. A literature review on various clustering techniques is first presented. To evaluate the proposed MFMM model, a performance comparison study using benchmark data sets pertaining to clustering problems is conducted. The results obtained are comparable with those reported in the literature. Then, a real-world case study on power quality monitoring tasks is performed. The results are compared with those from the fuzzy c-means and *k*-means clustering methods. The experimental outcome positively indicates the potential of MFMM in undertaking data clustering tasks and its applicability to the power systems domain.

1. Introduction

Data analysis procedures can be broadly categorized as either exploratory or confirmatory, based on the models used for processing the data source [1]. Regardless the methods used in both categories, one key component is data grouping using either goodness-of-fit to a postulated model or clustering through analysis [1]. Indeed, clustering is one of the main methods in data mining [2]. It is an unsupervised method that categorizes data into groups, such that objects in a cluster are more similar to one another as compared with those in another cluster [3]. In supervised methods, data are labeled in accordance with a number of specific target classes. In clustering methods, data samples are unlabeled, and the challenge is how to categorize them into meaningful clusters. Being a fundamental data analysis method, clustering is commonly used in many applications, which include pattern recognition, image segmentation, and function approximation [4]. Unlike standard statistical methods, many clustering methods do not depend on assumptions; therefore they are useful in situations where little or no prior knowledge is available [3].

In terms of clustering methods, they can be broadly divided into two groups: hierarchical and partitional [5]. Hierarchical clustering methods recursively locate nested clusters in either (i) an agglomerative mode, where each data sample in its own cluster is merged into the most similar pair; or (ii) in a divisive mode (also known as the top-down mode), where all data samples in a single cluster are divided into smaller clusters recursively [5]. As an example, a dendrogram represents a nested grouping of patterns, where the similarity level is produced from a hierarchical algorithm [1]. On the other hand, partitional clustering methods locate all clusters at one go, as the data partition does not impose a hierarchical structure [5]. In applications with large data sets, partitional clustering methods are advantageous as construction of a hierarchical structure (e.g. a dendrogram) can be computationally prohibitive [1].

Many clustering algorithms are available in the literature. Fundamentally, clustering is accomplished with some assumptions on a distance metric, data structure, and/or the number of clusters [6]. Among different clustering methods, *k*-means clustering is one of the popular algorithms [7]. The *k*-means clustering algorithm iteratively assigns each data sample to the closest cluster center using a distance metric. Different hybrid models involving the *k*-means clustering algorithm are also available, e.g. a hybrid differential evolution and one-step *k*-means clustering model [8]. One drawback is the estimated distance metric can be inaccurate [7]. Another popular clustering method is the fuzzy c-means algorithm [9,10]. Many of the fuzzy clustering methods can only process spatial data samples and not non-spatial ones [9]. In addition, other clustering methods are available, which include fuzzy spectral clustering algorithm [11] and subspace clustering algorithms [12].

* Corresponding author. Tel.: +603 7967 6381.
E-mail address: mseera@gmail.com (M. Seera).

http://dx.doi.org/10.1016/j.asoc.2014.09.050
1568-4946/

In terms of data-based methods, incremental learning neural network models offer a number of benefits owing to their robustness in handling large scale data sets and their distributed learning capabilities [13]. Incremental learning constitutes an efficient technique in knowledge discovery, as it allows acquisition of additional knowledge/information on the fly without forgetting previously learned knowledge/information [14]. Another advantage of incremental learning is that all training data can be immediately used for learning, rather than waiting for a representative training set to be collected for learning [15]. In addition, the memory requirements tend to be smaller because a training data sample can be discarded once it has been used for learning [15]. In this domain, Simpson proposed two fuzzy min–max (FMM) networks equipped with incremental learning capabilities: one for data classification with a supervised learning model [16] and another for data clustering with an unsupervised learning model [17].

Based on both original FMM models [16,17], a number of FMM variants have been developed in the literature. In our previous work, a hybrid model consisting of supervised FMM and the classification and regression tree was proposed for fault detection and diagnosis (FDD) of induction motors [18]. The model was further enhanced with the online learning capability to tackle FDD problems [19]. A modified FMM network for tackling the phenomenon of small numbers of large hyperboxes was devised [20]. The model was then improved with the capability of rule extraction using the genetic algorithm [21]. A general FMM network with the principle of expansion and contraction combining both supervised and unsupervised learning in one model was introduced [22]. Besides that, a general reflex FMM network integrating both FMM classification and clustering algorithms, together with the concept of human reflect mechanism was proposed [23]. Based on the supervised model, the FMM network with compensatory neurons [24] that allowed online learning and, at the same time, eliminated the hyperbox contraction process was developed. In order to learn and classify data samples with multiple granularities, the granular reflex FMM network comprising hyperbox fuzzy sets to represent multi-granular data was proposed [25]. The data-core-based FMM network proposed in [26] deployed new membership functions with two types of neurons (i.e., classifying and overlapping neurons), while eliminating the contraction process. A stochastic FMM network with reinforcement learning was introduced [27]. Instead of a class label, the probability vector in a stochastic automation procedure was utilized to determine which action to take based on random selection. Motivated by the success of the aforementioned FMM-based models, we improve the clustering FMM network [17] (hereafter known as FMM) so that it is efficient for handling data clustering tasks in this study. We further demonstrate the usefulness of the modified FMM model to undertake a real-world power quality monitoring application.

In general, FMM is able to establish connection between clusters and fuzzy sets [17]. In addition, FMM possesses a number of salient features for undertaking data clustering problems, i.e., it does not requires a pre-specified number of clusters and does not limit the number of clusters (i.e., it grows incrementally); it entails a simple and efficient procedure; it has only one key parameter (i.e., the hyperbox size) that needs to be fine-tuned by users [17]. Popular clustering algorithms with batch learning procedures such as fuzzy c-means and *k*-means clustering requires a pre-defined number of clusters to begin with, which can be a difficult task for a large data set, or when the underlying data structure keeps changing, e.g. in non-stationary environments. FMM is able to circumvent this difficulty by forming a dynamic network that is able to create the number of clusters incrementally based on the characteristics of the incoming data samples.

The main contributions of this study are two-fold: a modified FMM (MFMM) model for undertaking data clustering problems and a real-world application of MFMM to power quality monitoring task. The key innovations of this research include equipping MFMM with a centroid formation procedure in online clustering as well as allowing cluster validity analysis and performance assessment using the cophenetic correlation coefficient (CCC) [28]. The original FMM network proposed by Simpson [17] forms data clusters using a hyperbox structure. The minimum and maximum vertices of each hyperbox are encoded as the network weights. However, no centroid information with respect to the data samples clustered in each hyperbox is available. As a result, a cluster centroid estimation procedure is incorporated into original FMM in this study. The centroids are constantly monitored in every hyperbox update cycle to ensure that they remain within a hyperbox during the learning stage. It should be noted that having the cluster centroid information in MFMM is useful as this allows cluster validity analysis to be conducted. In this study, the CCC metric is adopted as a quantitative measure pertaining to the generated clusters, whereby the centroids are used for performance assessment. To demonstrate the effectiveness of MFMM, a real-world power quality monitoring problem is undertaken. Power quality is an important aspect of an electrical network, as poor power quality could lead to financial loss, especially to the industrial sector [29].

The rest of the paper is structured as follows. A literature review on various clustering methods, which include commonly used fuzzy c-means and *k*-means clustering algorithms, is detailed in Section 2. The overall FMM clustering neural network and its modifications are explained in Section 3. In Sections 4 and 5, a series of experiments is presented. To evaluate the usefulness of MFMM, a number of benchmark data sets are first used in Section 4, with the results compared with those reported in the literature. Then, a real-world data set from actual measurements of a power quality monitoring task is utilized in Section 5. The results from both benchmark and real-world problems are analyzed and discussed. Conclusions and suggestions for further work are presented in Section 6.

2. Literature review

A literature review on data clustering is presented in this section. The review comprises four main clustering categories, i.e., hierarchical clustering, centroid-based clustering, distribution-based clustering, and density-based clustering.

2.1. Hierarchical clustering

Hierarchical clustering (HC), also known as connectivity-based clustering, is built upon the rationale that objects that are closer together have a tighter relationship, as compared with those far away [30]. In essence, HC connects related objects based on distance to form clusters. Numerous HC applications have been reported in the literature, which include a hierarchical agglomerative clustering algorithm for planning of distributed generator units [31]. The algorithm was validated with the weighted sum method, and evaluated using two distribution systems [31]. The simulation results demonstrated the capability of the proposed HC algorithm [31]. In [32], three Trollius species of trollflowers from various regions in China were distinguished using HC analysis. The proposed method was useful for routine analysis and quality control of trollflowers [32].

A study on the asymptotic behavior of HC in conditions where both sample size and dimension grow to infinity was conducted [33]. Explicit signal versus noise boundaries among different types of clustering behaviors were derived. The analysis showed that clustering behavior inside the boundaries was similar within a wide spectrum of asymptotic settings [33]. In analyzing microarray