Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition 38 (2005) 1275–1288
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Enhanced neural gas network for prototype-based clustering

A.K. Qin, P.N. Suganthan*

School of Electrical & Electronic Engineering, Nanyang Technological University, Nanyang Avenue, Block S2, Singapore 639798, Singapore

Received 15 June 2004; received in revised form 13 October 2004; accepted 7 December 2004

Abstract

In practical cluster analysis tasks, an efficient clustering algorithm should be less sensitive to parameter configurations and tolerate the existence of outliers. Based on the neural gas (NG) network framework, we propose an efficient prototype-based clustering (PBC) algorithm called enhanced neural gas (ENG) network. Several problems associated with the traditional PBC algorithms and original NG algorithm such as sensitivity to initialization, sensitivity to input sequence ordering and the adverse influence from outliers can be effectively tackled in our new scheme. In addition, our new algorithm can establish the topology relationships among the prototypes and all topology-wise badly located prototypes can be relocated to represent more meaningful regions. Experimental results[1] on synthetic and UCI datasets show that our algorithm possesses superior performance in comparison to several PBC algorithms and their improved variants, such as hard *c*-means, fuzzy *c*-means, NG, fuzzy possibilistic *c*-means, credibilistic fuzzy c-means, hard/fuzzy robust clustering and alternative hard/fuzzy *c*-means, in static data clustering tasks with a fixed number of prototypes.

Keywords: Prototype-based clustering; *c*-means; Neural gas; Outliers; Minimum description length (MDL); Topology formation; Relocation; Survey

1. Introduction

Cluster analysis [1] is a crucial and powerful tool for exploring and discovering the underlying structures in data by partitioning a set of N input vectors $\mathbf{X}=\{\mathbf{x_1}, \mathbf{x_2}, \ldots \mathbf{x_N} | \forall \mathbf{x}_i \in R^d,\ i=1,2\ldots,N\}$ into c, $2 \leqslant c \leqslant N$ natural groups, called clusters, such that each input vector can be assigned to each cluster with a certain degree of belongingness. It has found applications in diverse fields [2–7] ranging from pattern recognition, data mining, computer vision and communication to information retrieval and bioinformatics.

There exist no definite categorization standards for different clustering techniques. Traditionally, clustering methods can be divided into hierarchical clustering and partitioning clustering [1]. Hierarchical schemes sequentially build nested clusters with a graphical representation known as dendrogram. Partitioning methods directly assign all the data points according to some appropriate criteria, such as similarity or density, into different groups. Our paper concentrates on the prototype-based clustering (PBC) algorithm, which is the most popular class of partitioning clustering methods and can also be generalized into hierarchical methods through superimposition. Hard clustering and fuzzy clustering [8–10] are often regarded as the two main branches of PBC algorithms. Hard clustering assigns each data point to exactly one cluster while fuzzy clustering assigns each data point to several clusters with varying degrees of membership. Based on the learning strategies, PBC methods can also be subdivided into batch and sequential versions.

In real engineering and scientific applications, outliers or noisy points are likely to be brought forth in every operating

* Corresponding author. Tel.: 65 6790 5404; fax: 65 6792 0415.
E-mail addresses: qinkai@pmail.ntu.edu.sg (A.K. Qin), epnsugan@ntu.edu.sg (P.N. Suganthan).

[1] Codes available at http://www.ntu.edu.sg/home/EPNSugan.

0031-3203/$30.00
doi:10.1016/j.patcog.2004.12.007

step. Further, the ordering of the input vectors may vary significantly. If the performance of a clustering algorithm varies significantly due to the change in the initial states, input sequence orderings or the existence of outliers, the clustering results will be of little value. Hence, it is desirable for the devised clustering algorithm to effectively solve these problems.

However, many existing PBC clustering algorithms are sensitive to the initialization, input ordering or the presence of outliers. In recent years, many PBC algorithms have been proposed to address the above-mentioned problems. In this paper, we present an improved version of the neural gas (NG) method [11], called enhanced neural gas (ENG) algorithm. Although the NG algorithm has several promising characteristics suitable for many real-world clustering applications [12–15], its performance is not satisfactory with respect to sensitivity to input sequence ordering and presence of outliers. These facts hamper the suitability of the NG algorithm in real applications. By employing heuristic strategies, which take into account both the historical and current movement information of prototypes in the updating rule of the original NG algorithm, our ENG algorithm can effectively reduce the sensitiveness to initialization, input sequence ordering and the influence of outliers located at different positions in the input sequence. Moreover, by combining the competitive Hebbian learning (CHL) scheme [16] and the minimum description length (MDL) framework [17] into our ENG algorithm, the topological relations among the reference vectors can be established and all these prototypes can find meaningful regions in the dataset.

The organization of this paper is as follows. In Section 2, we review the traditional PBC algorithms and many of their improved versions. In Section 3, after a brief review of the NG algorithm we present our novel ENG algorithm. Experimental results on artificial and UCI datasets are presented in Section 4 to demonstrate the superior performance of our ENG algorithm over several existing improved PBC algorithms. Finally, in Section 5 we conclude our paper with discussions.

2. PBC algorithms

PBC algorithms have some common properties: (1) the number of clusters can be specified in advance or obtained automatically [18] during the training procedure. (2) Each cluster is represented by a prototype. Measured by some distance metric, input vectors are assigned, with a certain degree of membership, to different clusters. (3) The final positions of prototypes are usually obtained by minimizing a cost function.

Hard c-means (HCM), also called k-means [19], is the most well-known hard clustering algorithm. This method is suitable for clustering of compact and well-separated groups of data. However, in real applications, data clusters usually overlap to some extent. Therefore, some data vectors cannot be assigned to exactly one cluster with certainty and it seems more reasonable to make them partially belong to several clusters. The famous fuzzy c-means (FCM) algorithm [9,10], devised by Bezdek [20], describes this case well, where the final cluster centers can be obtained by the alternative optimization (AO) method [21]. The sequential learning version with stochastic gradient descent method is known as fuzzy competitive learning [22].

Due to a common fact that the performance of the traditional clustering algorithms, such as HCM and FCM, may be unstable under different initializations and input sequence orderings or in the presence of outliers, many improved PBC versions have been proposed in the past few decades to tackle these problems, respectively. We describe several of these methods below.

2.1. Initialization and input sequence ordering problems

Due to the use of winner-take-all (WTA) learning strategy [16], traditional HCM algorithms often suffer from initialization problem. If initial positions of the prototype vectors are not properly placed, the clustering algorithm may yield unsatisfactory results. This is also called prototype-underutilization or "dead nodes" problem because some nodes may never win the competition during the clustering process. Comparatively speaking, FCM can naturally weaken the dead node problem due to the use of fuzzy partitioning matrix $\mathbf{U} = \{u_{ij} | 2 \leqslant i \leqslant c, 1 \leqslant j \leqslant N\}$. In other words, any input vector will, to some degree, contribute to the updating procedure of all prototypes. However, for a batch learning version, different initialization may still generate different results because the use of AO scheme makes the finally obtained local minimal states heavily dependent on the initial conditions. In sequential learning schemes, initialization problems are implicitly converted into the input sequential ordering problems. For a fixed initial state, randomly chosen input sequence orderings may generate different results.

Many approaches [23–26] have been devised to tackle this problem. Frequency sensitive competitive learning algorithm [23] is a typical method that can eliminate the dead nodes by gradually increasing the wining chance for those infrequently winning nodes. Observing that sub-sampling can provide information of the joint probability density function that generates the input datasets, Bradley [24] refined the initial points for k-means clustering based on an effective approach to estimate the modes of the distribution. The k-harmonic means algorithm [25] proposed by Zhang, is an effective extension of k-means. It modified the cost function in the k-means algorithm by substituting the minimum distance from each data point to its winning node in k-means with the harmonic average of the distances from each data point to all prototypes. In this method, each input vector can influence the updating process of all prototypes to effectively cope with the initialization problem. The hard robust clustering algorithm (HRC) [26] by Yang et al. employs the sequential learning strategy. With the presence of an input