Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 127 (2018) 16–25

Procedia Computer Science

www.elsevier.com/locate/procedia

The First International Conference On Intelligent Computing in Data Sciences

Determining the Number of Clusters using Neural Network and Max Stable Set Problem

Awatif Karim[a,*], Chakir Loqman[a], Jaouad Boumhidi[a]

[a] *LIIAN Laboratory, Faculty of Sciences Dhar El Mahraz. University Sidi Mohamed Ben Abdellah, Box 30003, Fez, Morocco*

Abstract

One of the most difficult problems, in cluster analysis is the determination of the number of clusters in a data set. Solving this problem consists in detecting and finding the best number of clusters, which is an input parameter for the clustering problems. In this paper, we propose a new approach using the Maximum Stable Set Problem (MSSP) combined by Continuous Hopfield Network (CHN) to determine the number of clusters, which is a basic input parameter for K-Means method. By testing the theoretical results, the proposed approach was validated on a real application for the text mining. Some numerical examples and computational experiments assess the effectiveness of this approach as demonstrated in this paper.

Keywords: Number of cluster; Data clustering; K-Means; Maximum Stable Set Problem; Continuous Hopfield Network.

1. Introduction

Clustering problem play a vital role in data mining, where we have many applications in areas such as bioinformatics, web data analysis, text mining, and scientific data exploration. Etc.

* Corresponding author. Tel.: +212-064-124- 0191.
E-mail address: awatif.karim@usmba.ac.ma, chakirfst@yahoo.fr, jaouad.boumhidi@usmba.ac.ma

10.1016/j.procs.2018.01.093

In the literature, they are several methods of clustering classified as follows:

- Partitioning clustering decomposes a data set into a set of disjoint clusters. A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster. The partitioning methods include K-Means and its variants [1].
- In hierarchical clustering, a treelike cluster structure is created through recursive partitioning (divisive methods) or combining (agglomerative) of existing clusters. The agglomerative hierarchical clustering algorithm generates a nested group of clusters that are organized as a tree. Then, the two closest clusters are aggregated into a new combined cluster. Eventually, all records are combined into a single huge cluster. The divisive clustering methods begin with all the data in one big cluster, with the most dissimilar data being split off recursively, into a separate cluster, until each data represents its own cluster [2], [3].
- Density Based Clustering: is based on connectivity and density functions. A cluster is defined as a connected dense component which can grow in any direction that density leads. Some interesting studies include DENCLUE, CLIQUE, DBSCAN and OPTICS [4].
- Grid-Based Clustering: is based on a multiple-level granularity structure. Some typical algorithms are WaveCluster (Sheikholeslami, Chatterjee and Zhang in 1998), STING (Wang, Yang and Mutz in 1997), CLIQUE (Agrawal, Gehrke, Gunopulos, Raghavan in 1998), and GRIDCLUST (Schikuta 1997).
- Neural Network Approach: SOM (Self Organizing Maps) [5], proposed by Kohonen in 1981 is the most popular Neural Network approach for clustering data. SOM has been successfully applied for Web Document clustering [6].

This work is part of the unsupervised classification (clustering) techniques. The problem can be formulated as follows: given a set of documents in order to divide them into multiple clusters, such that documents within a cluster are similar and documents in different clusters are dissimilar. In cluster analysis, one of the most challenging and difficult problems is the determination of the number of clusters in a data set, which is considered a basic input parameter for most clustering algorithms [7]. Most of these algorithms need a user-specified number of clusters in advance. So the results of clustering obtained depend strongly on the fixed number of classes. It is, therefore, essential to choose the exact number of classes to aspire to a good classification quality. To overcome this deficiency, we will present a new approach to solve a determination of clusters number.

The proposed method is divided into four steps: the first one consists of building a graph whereby a stable set is determined, the second involves modeling the Determining the number of clusters problem as the Max stable set problem(MSSP), which can be modeled as a 0-1 quadratic programming (QP).The third step concerns applying the Continuous Hopfield Network to solve the QP problem. Therefore, the generalized energy function associated with the CHN and an appropriate parameter setting procedure about MSSP problem are given. The fourth step concerns improving the obtained solution as input parameter in K-Means.

This paper is structured as follows: in section 2, we present problem formulation, in section three Max stable set problem and an introduction of Continuous Hopfield Network (CHN) are shown. Implementation details of the proposed approach and experimental results are presented in the last section.

2. Problem formulation

The proposed system can be divided into four main steps such as:

- Preprocessing stage
- Term weighting
- The determining the number of clusters problem using the MSSP problem and CHN
- Clustering documents by K-Means