I.J. Information Engineering and Electronic Business, **2013, 2, 22-27**
Published Online August 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2013.02.04

Optimal Clustering Algorithms for Data Mining

Omar Y. Alshamesti
Department of Computer science, Palestine Technical Colleges, Al-Aroub, Hebron, Palestine
oshamesti@ptca.edu.ps

Ismail M. Romi
College of Administrative sciences and Informatics, Palestine Polytechnic University, Hebron, Palestine
ismailr@ppu.edu

Abstract— Data mining is the process used to analyze a large quantity of heterogeneous data to extract useful information. Meanwhile, many data mining techniques are used; clustering classified to be an important technique, used to divide data into several groups called, clusters. Those clusters contain, objects that are homogeneous in one cluster, and different from other clusters. As a reason of the dependence of many applications on clustering techniques, while there is no combined method for clustering; this study compares k-mean, Fuzzy c-mean, self-organizing map (SOM), and support vector clustering (SVC); to show how those algorithms solve clustering problems, and then; compares the new methods of clustering (SVC) with the traditional clustering methods (K-mean, fuzzy c-mean and SOM). The main findings show that SVC is better than the k-mean, fuzzy c-mean and SOM, because; it doesn’t depend on either number or shape of clusters, and it dealing with outlier and overlapping. Finally; this paper show that; the enhancement using the gradient decent, and the proximity graph, improves the support vector clustering time by decreasing its computational complexity to O(nlogn) instead of O(n^2d), where; the practical total time for improvement support vector clustering (iSVC) labeling method is better than the other methods that improve SVC.

Index Terms **—** Data Mining, Clustering, Self-Organizing Map, Support Vector Clustering, Computational Complexity.

I. INTRODUCTION

In the early 1990's, the establishment of the internet made a huge quantity of data to be stored electronically; therefore, handling this quantity of data became to be necessary. Therefore, data mining emerged to extract useful information from a large quantity of heterogeneous data [1] using several techniques such as clustering. Where clustering divides data into several groups [2] depending on one of the proposed algorithms that have been developed by researchers [3, 4] such as K-mean, fuzzy c-mean, Self Organizing Map (SOM) and Support Vector Clustering (SVC). K-mean is a well-known partitioning method and one of the most popular clustering algorithms used in scientific and industrial applications [5]. Fuzzy c-mean [5, 6] is an iterative algorithm which is frequently used in pattern recognition, it allows one piece of data to be classified to more than one cluster. SOM algorithm can be classified as a powerful method for clustering high dimensional data [7]. SVC [8] is a nonparametric clustering process which depends on Support vector machine (SVM) concepts.

The fact that; there is no fixed method or technique, encourages researchers to keep developing algorithms and techniques to perform clustering in a variety of ways, where part of the studies focus on improving data clustering algorithms [3, 5, 9, 10, 11, 12], or develop new clustering methods [7, 8, 13], the other part focuses on comparing different data clustering algorithm using different factors [5, 12, 14, 15, 16, 17, 18].

This paper will focus on comparing k-mean, fuzzy C-mean, SOM and SVC algorithms to show how those algorithms solve clustering problems, and then compare those traditional methods with the new clustering method; mainly SVC, to find out the improvements and characteristics that reduce the computational complexity of this algorithm. Those comparisons will provide a tool for selecting the best clustering algorithm in specified area such as text mining, geographical information system, and information retrieval that depend on clustering.

This paper is organized as follow: A short description of data mining, k-mean clustering algorithm, fuzzy c-mean algorithm (FCM), self organizing map algorithms (SOM), and support vector clustering (SVC) are included in section 2. Section 3 includes the comparisons among the different data mining algorithms. Section 4, presents the conclusion of this paper, recommendations, and the required future researches.

II. BACKGROUND AND LITERATURE REVIEW

Data mining is the process of analyzing a large quantity of heterogeneous data to extract useful information [1]. This process could be performed using several techniques based on two types of learning paradigms [19]; mainly supervised and unsupervised learning. Clustering is one of those techniques which depend on unsupervised learning paradigm, and used to divide data into several groups; each of which called a cluster. Many algorithms are proposed for data clustering; where prior research's shows that the most used algorithms are k-mean, fuzzy c-mean, SOM and