Data Mining For Selected Clustering Algorithms: A Comparative Study

Omar Yousef AL-shamesti
Supervisor: Dr. Ismail M.Romi
Master of Informatics
College of Graduate Studies
Palestine Polytechnic University

Introduction

Data mining is the process used to analyze a large amount of heterogeneous data to extract useful information from it. Clustering is one of the main data mining techniques used to divide the data into several groups and each group is called a cluster. As a reason of many applications that depend on clustering techniques, while there is no combined method for clustering, this study focuses on the comparison between k-mean, Fuzzy c-mean, self organizing map (SOM) and support vector clustering (SVC) to show how those algorithms solve the clustering problem, and then; comparing the new methods of clustering (SVC) with the traditional clustering methods (K-mean, fuzzy c-mean and SOM), and show how the studies improves SVC algorithm.

Objectives:

The main objective of this study is to compare between the various clustering techniques, which is essential for data mining, and to achieve this objective, this study will go further insight a cross the following comparisons:

1. Show how the clustering algorithms such as k-mean, fuzzy c-mean, self organizing map (SOM) and support vector clustering (SVC) solve the clustering problem.
2. Compare the SVC with the traditional clustering algorithms such as k-mean and fuzzy c-mean and SOM in term the way that it works, Time complexity, outliers and the number of clusters.
3. Compare the several enhancements that is proposed to improve SVC in term of time complexity.
4. Provide a suitable recommendation regarding to the suitable clustering algorithm.

Proposed study:

This study compares between the various traditional clustering techniques; mainly k-mean, fuzzy c-mean, SOM and SVC, in order to show how they solve the clustering problem. And then, comparing the SVC as a new clustering method with the traditional methods to find out the enhancements of the SVC.

Comparisons:

Table 1, table 2, and table 3 shows the comparisons between k-mean and Fuzzy c-mean in terms of time and space complexity. Where table 4 and figure 1 compares k-mean and SOM algorithms. And table 5 compares SVC and iSVC in term of time. And table 6 compares the different labeling strategy for SVC in term of time. Finally, table 7 compares the different enhancements of SVC algorithm in term of time complexity.

Table 1. Time comparison for FCM and K-mean

Number of clusters	Fuzzy c-mean Time complexity	K-mean Time complexity
1	3000	3000
2	12000	6000
3	27000	9000
4	48000	12000
20	900	8

Table 2. space comparison for FCM and K-mean

Number of clusters	Fuzzy c-mean Time complexity	K-mean Time complexity
5	450	2
10	600	4
15	700	6

Table 3. Time and space comparison for FCM and K-mean

Number of clusters	Fuzzy c-mean Time complexity	K-mean Time complexity
K-mean	k	cd
FCM	$O(ndc^2i)$	O(nd+nc)

Table 4. The relationship between number of cluster s and algorithm performance

Performance		
Number of clusters	SOM	K-mean
8	59	63
16	67	61
32	78	84
64	85	89

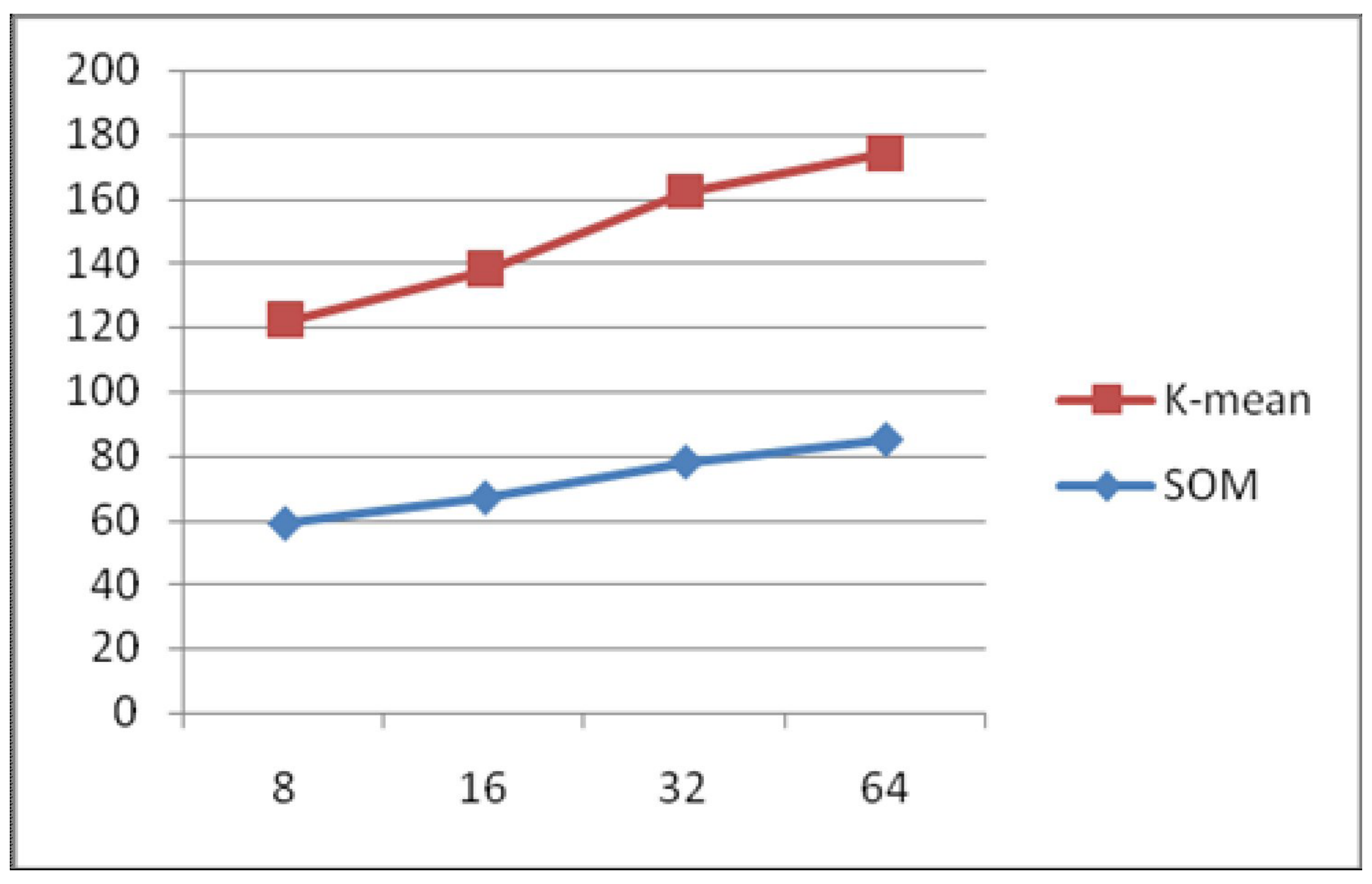

Figure 1: The relationship between number of cluster s and algorithm performance

Table 5. Time comparison for SVC and iSVC

	SVC		iSVC	
	Size	Time	Subsize	Time
Liver	354	115.1	100	0.661
Sonar	208	3.32	60.6	.093
Wine	178	2.32	52.0	0.087
Iris	150	9.09	46.0	0.138
Vote	435	126.6	125.0	0.811
Diabetes	768	261.3	219	5.687
Ionosphere	351	55.47	104	0.507

Table 6. Time comparison for different labeling approaches for SVC

	CG	SVG	PG	GD
Liver	131	109	202	657
Vote	815	286	119	89
Ionosphere	1069	301	187	205

Table 7. Time complexity analysis for the different SVC improvements

CG	SVG	PG	GD	iSVC
$O(n^2d)$	$O(n - n_{bsv})n_{sv}{}^{2}$	O(nlogn)	O(nlogn)	$O(N_{sv}{}^{3})$

Results:

The main findings of this study shows that:

•The time and space complexity for k-mean is lower the fuzzy c-mean.

•The performance of SOM algorithm becomes lower than k-mean as the number of cluster K becomes greater.

•The SVC is a better algorithm for clustering because it provides a general clustering solution which is applicable to a variety of applications, and it doesn't require any a assumption about the number or the shape of cluster, and deals with outliers.

•The proximity graph and gradient decent method are the best enhancements for SVC; because they have the best time over the other enhancements.