DECLARATION

I certify that all the material in this thesis that is not my own work has been identified, and that no material is included for which a degree has previously been conferred on me.

The contents of this thesis reflect my own personal views, and are not necessarily endorsed by the University.

Name: Hany Hanafy Mahmoud Said

Signature: .....

Date: 15 / 11 / 2008

We certify that we have read the present work and that, in our opinion, it is fully adequate in scope and quality as a thesis towards the partial fulfillment of the Master's Degree requirements in

Computer Engineering

From the

College of Engineering and Technology (AASTMT)

Date: 15 / 11 / 2008

Supervisor:

Name: Prof. Dr. Yasser El Sonbaty

Position: Professor – College of Computing & Information Technology – AASTMT

Signature:

Examiners:

Name: Prof. Dr. Mohamed Ismail

Position: Professor – Faculty of Engineering – Alexandria University

Signature:

Name: Prof. Dr. Amin Shokry

Position: Professor – Faculty of Engineering – Alexandria University

Signature:

ACKNOWLEDGMENTS

I owe a great gratitude to my supervisor, Prof. Dr. Yasser El Sonbaty, for his excellent advising and guidance through my thesis. I thank my mother from the bottom of my heart for her standing with me all the time. I am grateful to my father for his pushing me always to reach my goals. I also thank my colleagues in Arab Academy for their support.

ABSTRACT

Clustering is one of the important data mining techniques for extracting knowledge from datasets in various applications. Most of the clustering algorithms do not achieve the majority of the clustering requirements as scalability; discovering clusters of different shapes, dealing with noise, dealing with high dimensional data.

Density-based method is one of the most effective paradigms in cluster analysis that detects clusters with arbitrary shapes without acquire the number of clusters in advance. DBSCAN discovers the dense regions and identifies them as clusters that are separated by low density regions. Several algorithms improve DBSCAN algorithm such as fast hybrid density based algorithm “L-DBSCAN” and fast density-based clustering algorithm.

Since clustering algorithms discover clusters, which are not known a priori, the final partition of a dataset requires some sort of evaluation in most applications. *CDbw* validity index measures the average density within clusters, *Intra_dens(c)* with the separation of clusters *Sep(c)*, where *c* is the number of clusters discovered in the dataset. *CDbw* increases when clusters are highly dense, well separated and the densities in the areas among them are low.

In this thesis, an enhanced density based algorithm is proposed that improves fast density based clustering algorithm. The experimental results show that the new algorithm is superior to L-DBSCAN and fast density-based clustering and DBSCAN algorithms in more specific aspects, like running time, clusters’ compactness in terms of intra-cluster density and clusters’ separation in terms of inter-cluster density.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

KDD	Knowledge Discovery in Databases
PAM	Partitioning Around Medoids Algorithm
CLARA	Clustering LARge Applications
CLARANS	Clustering Large Applications based on RANdomized Search
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
CF	Clustering Feature
CURE	Clustering Using REpresentatives
CHAMELEON	Hierarchical Clustering Algorithm using Dynamic Modeling
RI	Relative Interconnectivity
RC	Relative Closeness
STING	Statistical Information Grid
CLIQUE	Clustering In Quest
DBSCAN	Density Based Spatial Clustering of Applications with Noise
OPTICS	Ordering Points to Identify the Clustering Structure
IDBSCAN	Improved sampling-based DBSCAN for large spatial databases
MBO	Marked Boundary Objects
FDBSCAN	Fast Density-Based Clustering Algorithm for Large Databases
CDbw	Compose Density Between and Within Clusters

LIST OF FIGURES

LIST OF TABLES