سامية محمد مصطفى

شبكة المعلومات الجامعية

بسم الله الرحمن الرحيم

سامية محمد مصطفى

شبكة المعلومات الجامعية

شبكة المعلومات الجامعية

التوثيق الالكتروني والميكروفيلم

سامية محمد مصطفى

شبكة المعلومات الجامعية

جامعة عين شمس

التوثيق الإلكتروني والميكروفيلم

قسم

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها علي هذه الأقراص المدمجة قد أعدت دون أية تغيرات

يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار

سامية محمد مصطفى

شبكة المعلومات الجامعية

بعض الوثائق الأصلية تالفة

سامية محمد مصطفى

شبكة المعلومات الجامعية

بالرسالة صفحات لم ترد بالأصل

Cairo University
Faculty of Economics and Political Science
Department of Statistics

Simulation Study for Cluster Analysis

Prepared by
Ahmed EL-Tabey Okasha

Supervised by

Prof. Dr. Ramadan Hamed
Professor of Statistics
Dept. of statistics

Prof. Dr. Kamal Selim
Associate Prof. of statistics
Dept. of Social-Science computing

Faculty of Economics and Political Science

A Thesis submitted to the department of Statistics, Faculty of Economics and Political Science as Partial Fulfillment of the Requirements for the M.Sc. Degree in Statistics

2005

B
15--0

TO MY PARENTS AND MY PARENTS-IN-LAW

I'm greatly indebted to Prof. Dr. Ramadan Hamed, Professor of Statistics, Dept. of Statistics, Faculty of Economics and Political Science for his valuable and generous assistance. My sincere thanks are also dedicated to him for this constructive guidance and warm encouragement throughout the preparation of this thesis.

Prof. Dr. Kamal Selim, Professor of Statistics, Dept. Of Social-Science Computing, Faculty of Economics and Political Science deserves my deepest gratitude and appreciation for his kind supervision, continuous help and active discussions during the preparation of this thesis.

My deepest gratitude and appreciation and extended also to Prof Dr. Tarek Emira, Professor of Statistics, Chief of Dept. of Statistics, Faculty of Economics and Political Science for his valuable and constructive comments, and to Prof Dr. Maged George for his generous acceptance of discussion of this thesis.

Simulation Study for Cluster Analysis

Abstract

A Monte Carlo simulation study of four clustering methods –Complete linkage , Ward's method, K-means method and AutoClass Bayesian classification-for determining the best partitioning is conducted on some artificial data sets. The results indicated that the classification obtained by the four clustering methods is affected by increasing the number of clusters in the data sets when error-free data sets are used. All clustering methods perform better than AutoClass Bayesian classification method when number of clusters in data sets is small while the performance of AutoClass Bayesian classification is better than other clustering methods when number of clusters in data sets is large. The results indicated that the classification obtained by the four methods is affected by dimension, noise, and outliers conditions. All clustering methods are affected by outliers. The performance of all clustering methods is decreased with increasing the number of clusters except for AutoClass Bayesian classification, its performance is increased with increasing the number of clusters.

Keywords

- Cluster analysis
- Bayesian classification
- Monte Carlo simulation
- Cluster validity
- Homogeneity and Separation measures

Supervised by

Prof. Dr. Ramadan Hamed
Professor of Statistics
Dept. of statistics

Prof. Dr. Kamal Selim
Associate Prof. of statistics
Dept. of Social-Science computing

A Thesis submitted to the department of Statistics, Faculty of Economics and Political Science as Partial Fulfillment of the Requirements for the M.Sc. Degree in Statistics

2005

Name: Ahmed EL-Tabey Okasha

Nationality: Egyptian

Date and Place of birth: 16/9/1972, Damietta, Egypt

Degree: Master

Specialization: Statistics

Supervisors:

Prof. Dr. Ramadan Hamed

Professor of Statistics
Dept. of statistics

Prof Dr. Kamal Selim

Associate Prof. of statistics
Dept. of Social-Science computing

Title of the Thesis:

Simulation Study for Cluster Analysis

Summary of the thesis:

Few studies which focus on the comparison of hierarchical and nonhierarchical clustering methods with Bayesian classification are available in the literature. These studies are very limited in the variation of test conditions, such as cluster size and number of clusters present in the data.

For a given data set, each clustering method may identify groups whose member objects and/or number of clusters are different. Thus a decision must be taken for choosing the clustering method that produce the best data partition for a given data collection.

The objective of the present work is to perform a comparison study between four clustering methods-Complete linkage, Ward's method, K-means method