Cairo University
Faculty of Science
Department of Mathematics

Analysis, Investigation and Comparative Study of Some Data Mining Algorithms

Thesis Presented by
Ahmed Hassan Aly Eltawil
For the Degree of M.Sc. in
Computer Science

Supervised by

Prof. Dr. Ahmed Badr Eldin Khalil
Professor of Mathematics,
Faculty of Science, Cairo University

Prof. Dr. Laila Fahmy Abdelal
Professor of Mathematics,
Faculty of Science, Cairo University

Prof. Dr. Mohamed Nour Elsayed Ahmed
Professor of Computers
The Electronics Research Institute, Cairo

Cairo 2010

Approval Sheet for Submission

Title of M.Sc. Thesis : Analysis, Investigation and Comparative Study of Some Data Mining Algorithms

Name of the Candidate: Ahmed Hassan Aly Eltawil

This thesis has been approved for submission by the supervisors:

1- Prof. Dr. Laila Fahmy Abdelal
Signature:

2- Prof. Dr. Mohamed Nour Elsayed Ahmed
Signature:

Prof. Dr. Mohamed Zeidan Abd-allah
Head of Department of Mathematics
Faculty of Science, Cairo University

Abstract

Data clustering is an essential part of data mining and has attracted much attention recently. Data clustering is very important for several applications. Clustering methods and/or algorithms partition a set of objects into clusters such that objects in the same cluster are more similar to each other than those objects in different clusters according to some defined criteria. Most of the available algorithms that handle the clustering of categorical data are frequency-based.

This thesis proposes a technique that uses the principles of metric space isometry to extend the domain of clustering algorithms that deal with numerical datasets to cover also categorical datasets. This is done by developing a method called metric space isometry based conversion (MSIC). The MSIC method is used with the fuzzy-C-means (FCM) algorithm and compared to the traditional Fuzzy-K-Prototypes algorithm. The comparative computational results show the usefulness of this method over the Fuzzy-K-Prototypes algorithm.

A genetic algorithm (GA) for mixed numerical and categorical data is chosen, analyzed and discussed. A cost function for evaluating the fitness of clusters against datasets was presented. All genetic operators mainly, selection, mutation and crossover were applied to reach an optimal clustering solution. Five different datasets with different patterns and distributions were operated to test the capability of the algorithm. The genetic algorithm is modified by means of a newly proposed mutation operator and chromosome presentation. The modified algorithm successfully combines all the genetic operators, selection, mutation and

crossover, as well as MSIC. Moreover, the mutation and crossover probabilities are studied based on the proposed modifications. The computations show the effectiveness of the proposed modifications over the traditional GA algorithm.

Moreover, the proposed method is used to extend the domain of coverage of a chosen ant colony optimization (ACO) algorithm. The ACO algorithm and its embedded MSCI were analyzed, discussed and implemented. The algorithm was operated on the five different datasets each with two, three, four and five clusters respectively. The effect of the different algorithm parameters is profoundly studied to computationally find the parameters that provide optimal results both in terms of running time and better clustering performance.

It is concluded that the GA generally performs better than the ACO. The bio-inspired models such as genetic algorithms and ant colony optimization are promising methods to handle the fuzzy-clustering problem.

Keywords

Clustering Algorithms, Fuzzy Clustering, Genetic Algorithms, Ant Colony Optimization and Performance Analysis

Cairo University
Faculty of Science

TO WHOM IT MAY CONCERN

This is to certify that: Ahmed Hassan Ali Hassan El tawil
Has attended and passed successfully the following Postgraduate Courses as a Partial Fulfillment of the requirements of the degree of Master of Science:

1- Number Theory and Coding Theory
2- Theory of Computability
3- Parallel Processing
4- Theory of computational Complexity
5- German Language

This Certificate is issued at his own request.

Date of birth: 24/9/1976
Place of birth: Dokki – Giza

Controller **Dean**

Head of dept.
M. Z. Abdalla
23/3/2010

جامعة القاهرة
كلية العلوم
قسم الرياضيات
23 MAR 2010
Department of Mathematics

Acknowledgment

I wish to dedicate this thesis to the blessed soul of my advisor Prof. Dr. Ahmed Badr Eldin Khalil, Professor of Mathematics, Mathematics Department Faculty of Science, Cairo University. I owe my deepest gratitude to him for his motivation, enthusiasm and for giving me the space and understanding to work on this thesis, God's mercy be upon him.

Foremost, I would like to express my sincere gratitude to my advisor Prof. Dr. Mohamed Nour Professor of Computers and former head of the Informatics Research Department, the Electronics Research Institute for his continuous support throughout this research, his patience, careful revision and immense knowledge. His guidance and suggestions helped me during the research and the writing of this thesis.

I have been blessed with a great supervisor Prof. Dr. Laila Fahmy, Professor of Mathematics, Mathematics Department, Faculty of Science, Cairo University, who has encouraged me and gave me all the needed support to finalize this thesis.

Last but not least, I would like to thank Mathematics Department for providing the facilities. For everyone else that supported me in any respect during the completion of the thesis, I offer my regards and blessings.

Table of Contents

Chapter 1

Introduction and Scope of Work

Introduction and Scope of Work

1.1 Introduction

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner, [Kurt Thearling, 2007]. The relationships and summaries derived through data mining are often referred to as models or patterns. Examples include linear equations, rules, clusters, graphs, trees, structures and recurrent patterns in time series. Data mining typically deals with data that have already been collected for some purpose other than the data mining analysis. The objectives of data mining exercise play no role in the data collection strategy, [David Hand, 2001]. Data mining is often referred to as "secondary" data analysis. Data mining is the process of knowledge discovery in databases, or KDD.

Other definition of data mining includes, Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases, [Daniel T. Larose, 2005]. While the first definition emphasis the final goal of data mining this one focuses on the roots and used techniques.

A data mining algorithm can be defined as a well-defined procedure that takes data as input and produces output in the form of models or patterns [James E. Gentle and Wolfgang Härdle, 2004]. Components of the algorithm are: (a) Score Function which is used to judge the quality of a fitted models or patterns based on observed data

(e.g., misclassification error or squared error). (b) Search Method used to search over parameters and structures, i.e., computational procedures and algorithms used to find the maximum (or minimum) of the score function for particular models or patterns. (c) The data management technique to be used for storing, indexing, and retrieving data [David Hand, et.al, 2001].

The two primary goals of data mining are predictive and descriptive. A predictive task produces the model of the system described by the given data set, while a descriptive task produces new, nontrivial information based on the available data set

There are important themes of data mining tasks. Classification aims to discovery of a predictive learning function that classifies a data item into one of several predefined classes [Michael J.A. Berry and Gordon Linoff, 2000]. Regression is the discovery of a predictive learning function, which maps a data item to a real-value prediction variable. Summarization, is a descriptive task that involves methods for finding a compact description for a set of data. Dependency Modeling involves finding a local model that describes significant dependencies between variables or between the values of a feature in a data set or in a part of a data set. Change and Deviation Detection involve discovering the most significant changes in the data set [Mehmed Kantardzic, 2003]. Clustering refers to the grouping of records, observations, or cases into classes of similar objects. A cluster is a collection of records that are similar to one another and dissimilar to records in other clusters. Clustering differs from classification in that there is no target variable for clustering. The clustering task does not try to classify, estimate, or predict the value of a target variable. Instead, clustering algorithms seek to segment the entire data set into relatively

homogeneous subgroups or clusters, where the similarity of the records within the cluster is maximized, and the similarity to records outside this cluster is minimized [Daniel T. Larose, 2007].

This thesis focuses on clustering as one of the most important descriptive tasks in data mining.

Clustering algorithms can be categorized into four main categories: partition-based clustering, fuzzy clustering, hierarchical clustering and probabilistic model-based clustering.

In the first case data are grouped in an exclusive way, so that if a certain data point belongs to a definite cluster then it could not be included in another cluster. Moreover, partitioning algorithms typically represent clusters by a prototype. A well-known example of partition-based clustering is the K-means algorithm. On the contrary the second type allows overlapping, as it uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value [Matteo Matteucci, 2006]. An example of the fuzzy clustering is Fuzzy C-means or FCM. Instead, a hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every data point as a cluster. After a few iterations it reaches the final clusters wanted. The hierarchical clustering algorithms may be divisive (top-down) or agglomerative (bottom-up), [Frank Klawonn and Frank Höppner, 2003]. Finally, the last kind of clustering uses a probabilistic approach, which consists of using certain models for clusters and attempting to optimize the fit between the data and the model. In practice, each cluster can be mathematically represented by a parametric distribution, like a Gaussian (continuous) or a