سامية محمد مصطفى

بسم اللّه الرحمن الرحيم

بعض الوثائق الأصلية تالفة

بالرسالة صفحات لم ترد بالأصل

A Data Mining Tool Using a Modified Ordered Attribute Trees Algorithm for Handling Missing Values

By

Mona Farouk Ahmed Kamal El Deen

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
In
COMPUTER ENGINEERING

Under the Supervision of

Prof. Dr. Nevin M. Darwish
Computer Engineering Department
Faculty of Engineering
Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
September 2006

A Data Mining Tool Using a Modified Ordered Attribute Trees Algorithm for Handling Missing Values

By

Mona Farouk Ahmed Kamal El Deen

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

COMPUTER ENGINEERING

Approved by the
Examining Committee:

Prof. Dr. Nevin M. Darwish (Thesis Advisor),
Chairman Computer Engineering Department Nevin Darwish
9/2006

Prof. Dr. Mohamed Zaki Abdel-Mageid, M. Zaki

Prof. Dr. Ashraf Abdel-Wahab Ashraf

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
September 2006

Acknowledgement

First and foremost, I express my gratefulness to God the most merciful and the most gracious who gave me the strength and guidance to complete this work.

Next, I wish to thank Prof. Nevin M. Darwish for her kind supervision, precious advice and effective guidance throughout my M.Sc. study.

I would like to thank Prof Amir Attiya Souryal for his fruitful suggestions and illustrations. I also wish to express my appreciation to Eng. Mohamed Ismail for his great help all the way throughout this work.

I wish with great love to express my gratitude to my husband for his endless support, encouragement and patience.

Last but not least I wish to express my special thanks to my parents for their ceaseless love and great support all along the way, also to the greatest love and the source of happiness in my life, my children.

Abstract

Data mining techniques are becoming increasingly useful in a wide range of environments as a source of business intelligence. A wide range of companies has deployed successful applications of data mining. When attempting to discover by learning concepts embedded in data it is common to find that information is missing from the data. Such missing information can diminish the confidence on the concepts learned from the data. When missing values occur in the data, the learning algorithm fails to find an accurate representation of the concept. Properly filling missing values in data helps in reducing the error rate of the learned concepts. This work aims at solving the problem of missing values in data presented to a data mining decision tree learner. The missing values handling technique proposed in this work is a modification of the Ordered Attribute Trees method which is a machine learning approach to the missing values problem. A decision tree is constructed to determine the missing values of each attribute by using information contained in other attributes. Also, an ordering for the construction of the decision trees for the attributes is formulated.

This work is presented together with an implementation of two other missing values handling techniques namely, Unordered Attribute Trees which is also a machine learning approach and the Probabilistic method which is a good example of the statistical approach for handling missing values. The three methods are tested on the same data sets and performance results and evaluation are presented. Results show the proposed modification is at advantage in comparison to the other two techniques.

Table of Contents