國立交通大學

電機與控制工程學系

碩　士　論　文

2001年資料探勘競賽研究

Study on KDD cup 2001

研 究 生：張文賢

指導教授：林心宇　教授

中華民國九十三年七月

2001年資料探勘競賽研究

Study on KDD cup 2001

研 究 生：張文賢　　Student: Wen-Hsien Chang

指導教授：林心宇　　Advisor: Shin-Yeu Lin

國 立 交 通 大 學

電機與控制工程學系

碩 士 論 文

A Thesis

Submitted to Department of Electrical and Control Engineering

College of Electrical Engineering and Computer Science

National Chiao-Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master

in

Electrical and Control Engineering

July 2004

Hsinchu, Taiwan, Republic of China

中華民國九十三年七月

2001年資料探勘競賽研究

學生：張文賢　　指導教授：林心宇 博士

國立交通大學電機與控制工程學系（研究所）碩士班

摘　　要

資料探勘是一種分析的程序，用來幫助我們發現大型資料庫中的特徵及知識。因為有關生物學的資料探勘快速的發展，2001 年資料探勘競賽聚焦在基因及藥物設計資料上。我們所熱衷的是一個分類問題，這個問題有三個有趣的特性（1）大量的遺漏值（2）大量的屬性（3）混合兩種不同型態的資料，而我們最感興趣的分類方法就是決策樹分類法，我們修改了決策樹演算法，並引入“少數服從多數”技巧來提昇分類正確性。為了結合上述兩種分類方法我們發展出“主要-輔助”分類系統。

Study on KDD cup 2001

Student: Wen-Hsien Chang Advisor: Dr.Shin-Yeu Lin

Department of Electrical and Control Engineering

National Chiao Tung University

Abstract

Data mining is an analysis process which helps discovering patterns and knowledge in large databases. Because of the rapid growth of interest in mining biological databases, KDD Cup 2001 was focused on data from genomics and drug design. We were involved in a classification problem. The problem has three interesting features: (1) the dataset contains many missing values; (2) this dataset has a lot of attributes; and (3) the dataset is a mixture of two types of data, while the classification method we interested in most is Decision Tree. We modify the Decision Tree algorithm and cite the majority vote to improve the classification accuracy. For integrating the above two classification methods we develop " Primary-Secondary " classification system.

誌　謝

首先感謝指導教授林心宇老師在學習的過程中給予協助與鼓勵，除了讓我在專業領域上有更深刻的體會外，在待人處世和為學態度方面也都獲得相當大的啟發，在此謹致上最誠摯的敬意與謝忱。另外要感謝實驗室士程學長、榮壽學長、佶興學長、志遠學長、傑愷學長提供研究上的意見與經驗，還有學弟紹興在上活上的互相幫忙。最後感謝家人的支持與關懷，讓我能專心順利完成學業，謹以此篇論文獻給你們。

張文賢　于新竹

中華民國九十三年七月

Contents

List of tables

List of figures