Undergraduate Topics in Computer Science

Undergraduate Topics in Computer Science (UTiCS) delivers high-quality instructional content for undergraduates studying in all areas of computing and information science. From core foundational and theoretical material to final-year topics and applications, UTiCS books take a fresh, concise, and modern approach and are ideal for self-study or for a one- or two-semester course. The texts are all authored by established experts in their fields, reviewed by an international advisory board, and contain numerous examples and problems. Many include fully worked solutions.

Also in this series

Iain Craig
Object-Oriented Programming Languages: Interpretation
978-1-84628-773-2

Hanne Riis Nielson and Flemming Nielson
Semantics with Applications: An Appetizer
978-1-84628-691-9

Max Bramer

Principles of Data Mining

Springer

Max Bramer, BSc, PhD, CEng, FBCS, FIEE, FRSA
Digital Professor of Information Technology, University of Portsmouth, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2007922358

Undergraduate Topics in Computer Science ISSN 1863-7310
ISBN-10: 1-84628-765-0 e-ISBN 10: 1-84628-766-9
ISBN-13: 978-1-84628-765-7 e-ISBN-13: 978-1-84628-766-4

Printed on acid-free paper

The use of registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant laws and regulations and therefore free for general use.

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

9 8 7 6 5 4 3 2 1

Springer Science+Business Media
springer.com

Contents

Introduction to Data Mining

The Data Explosion

Modern computer systems are accumulating data at an almost unimaginable rate and from a very wide variety of sources: from point-of-sale machines in the high street to machines logging every cheque clearance, bank cash withdrawal and credit card transaction, to Earth observation satellites in space.

Some examples will serve to give an indication of the volumes of data involved:

- The current NASA Earth observation satellites generate a terabyte (i.e. 10^9 bytes) of data *every day*. This is more than the total amount of data ever transmitted by all previous observation satellites.
- The Human Genome project is storing thousands of bytes for each of several *billion* genetic bases.
- As long ago as 1990, the US Census collected over a million million bytes of data.
- Many companies maintain large Data Warehouses of customer transactions. A fairly small data warehouse might contain more than a hundred million transactions.

There are vast amounts of data recorded every day on automatic recording devices, such as credit card transaction files and web logs, as well as non-symbolic data such as CCTV recordings.

Alongside advances in storage technology, which increasingly make it possible to store such vast amounts of data at relatively low cost whether in commercial data warehouses, scientific research laboratories or elsewhere, has come a growing realisation that such data contains buried within it knowledge that can be critical to a company's growth or decline, knowledge that could lead to important discoveries in science, knowledge that could enable us accurately to predict the weather and natural disasters, knowledge that could enable us to identify the causes of and possible cures for lethal illnesses, knowledge that could literally mean the difference between life and death. Yet the huge volumes involved mean that most of this data is merely stored—never to be examined in more than the most superficial way, if at all. It has rightly been said that the world is becoming 'data rich but knowledge poor'.

Machine learning technology, some of it very long established, has the potential to solve the problem of the tidal wave of data that is flooding around organisations, governments and individuals.

Knowledge Discovery

Knowledge Discovery has been defined as the 'non-trivial extraction of implicit, previously unknown and potentially useful information from data'. It is a process of which data mining forms just one part, albeit a central one.

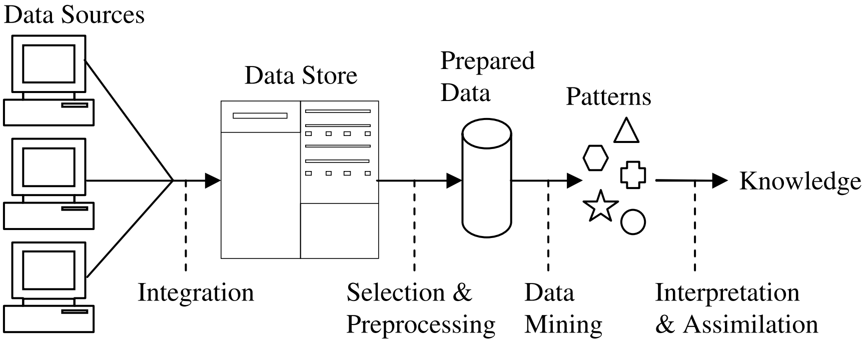

Figure 1 The Knowledge Discovery Process

Figure 1 shows a slightly idealised version of the complete knowledge discovery process.

Data comes in, possibly from many sources. It is integrated and placed in some common data store. Part of it is then taken and pre-processed into a standard format. This 'prepared data' is then passed to a data mining algorithm which produces an output in the form of rules or some other kind of 'patterns'. These are then interpreted to give—and this is the Holy Grail for knowledge discovery—new and potentially useful knowledge.

This brief description makes it clear that although the data mining algorithms, which are the principal subject of this book, are central to knowledge discovery they are not the whole story. The pre-processing of the data and the interpretation (as opposed to the blind use) of the results are both of great importance. They are skilled tasks that are far more of an art (or a skill learnt from experience) than an exact science. Although they will both be touched on in this book, the algorithms of the data mining stage of knowledge discovery will be its prime concern.

Applications of Data Mining

There is a rapidly growing body of successful applications in a wide range of areas as diverse as:

– analysis of organic compounds

– automatic abstracting

– credit card fraud detection

– electric load prediction

– financial forecasting

– medical diagnosis

– predicting share of television audiences

– product design

– real estate valuation

– targeted marketing

– thermal power plant optimisation

– toxic hazard analysis

– weather forecasting

and many more. Some examples of applications (potential or actual) are:

– a supermarket chain mines its customer transactions data to optimise targeting of high value customers

– a credit card company can use its data warehouse of customer transactions for fraud detection

– a major hotel chain can use survey databases to identify attributes of a ‘high-value’ prospect

- predicting the probability of default for consumer loan applications by improving the ability to predict bad loans
- reducing fabrication flaws in VLSI chips
- data mining systems can sift through vast quantities of data collected during the semiconductor fabrication process to identify conditions that are causing yield problems
- predicting audience share for television programmes, allowing television executives to arrange show schedules to maximise market share and increase advertising revenues
- predicting the probability that a cancer patient will respond to chemotherapy, thus reducing health-care costs without affecting quality of care.

Applications can be divided into four main types: classification, numerical prediction, association and clustering. Each of these is explained briefly below. However first we need to distinguish between two types of data.

Labelled and Unlabelled Data

In general we have a dataset of examples (called *instances*), each of which comprises the values of a number of variables, which in data mining are often called *attributes*. There are two types of data, which are treated in radically different ways.

For the first type there is a specially designated attribute and the aim is to use the data given to predict the value of that attribute for instances that have not yet been seen. Data of this kind is called *labelled*. Data mining using labelled data is known as *supervised learning*. If the designated attribute is *categorical*, i.e. it must take one of a number of distinct values such as ‘very good’, ‘good’ or ‘poor’, or (in an object recognition application) ‘car’, ‘bicycle’, ‘person’, ‘bus’ or ‘taxi’ the task is called *classification*. If the designated attribute is numerical, e.g. the expected sale price of a house or the opening price of a share on tomorrow’s stock market, the task is called *regression*.

Data that does not have any specially designated attribute is called *unlabelled*. Data mining of unlabelled data is known as *unsupervised learning*. Here the aim is simply to extract the most information we can from the data available.

Supervised Learning: Classification

Classification is one of the most common applications for data mining. It corresponds to a task that occurs frequently in everyday life. For example, a hospital may want to classify medical patients into those who are at high, medium or low risk of acquiring a certain illness, an opinion polling company may wish to classify people interviewed into those who are likely to vote for each of a number of political parties or are undecided, or we may wish to classify a student project as distinction, merit, pass or fail.

This example shows a typical situation (Figure 2). We have a dataset in the form of a table containing students' grades on five subjects (the values of attributes SoftEng, ARIN, HCI, CSA and Project) and their overall degree classifications. The row of dots indicates that a number of rows have been omitted in the interests of simplicity. We want to find some way of predicting the classification for other students given only their grade 'profiles'.

SoftEng	**ARIN**	**HCI**	**CSA**	**Project**	**Class**
A	B	A	B	B	Second
A	B	B	B	B	Second
B	A	A	B	A	Second
A	A	A	A	B	First
A	A	B	B	A	First
B	A	A	B	B	Second
....	....	....	....	....	....
A	A	B	A	B	First

Figure 2 Degree Classification Data

There are several ways we can do this, including the following.

Nearest Neighbour Matching. This method relies on identifying (say) the five examples that are 'closest' in some sense to an unclassified one. If the five 'nearest neighbours' have grades Second, First, Second, Second and Second we might reasonably conclude that the new instance should be classified as 'Second'.

Classification Rules. We look for rules that we can use to predict the classification of an unseen instance, for example:

IF SoftEng = A AND Project = A THEN Class = First

IF SoftEng = A AND Project = B AND ARIN = B THEN Class = Second
IF SoftEng = B THEN Class = Second

Classification Tree. One way of generating classification rules is via an intermediate tree-like structure called a *classification tree* or a *decision tree.*

Figure 3 shows a possible decision tree corresponding to the degree classification data.

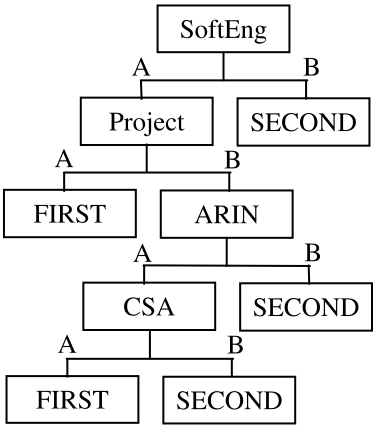

Figure 3 Decision Tree for Degree Classification Data

Supervised Learning: Numerical Prediction

Classification is one form of prediction, where the value to be predicted is a label. Numerical prediction (often called *regression*) is another. In this case we wish to predict a numerical value, such as a company's profits or a share price.

A very popular way of doing this is to use a *Neural Network* as shown in Figure 4 (often called by the simplified name *Neural Net*).

This is a complex modelling technique based on a model of a human neuron. A neural net is given a set of inputs and is used to predict one or more outputs.

Although neural networks are an important technique of data mining, they are complex enough to justify a book of their own and will not be discussed further here. There are several good textbooks on neural networks available, some of which are listed in Appendix C.

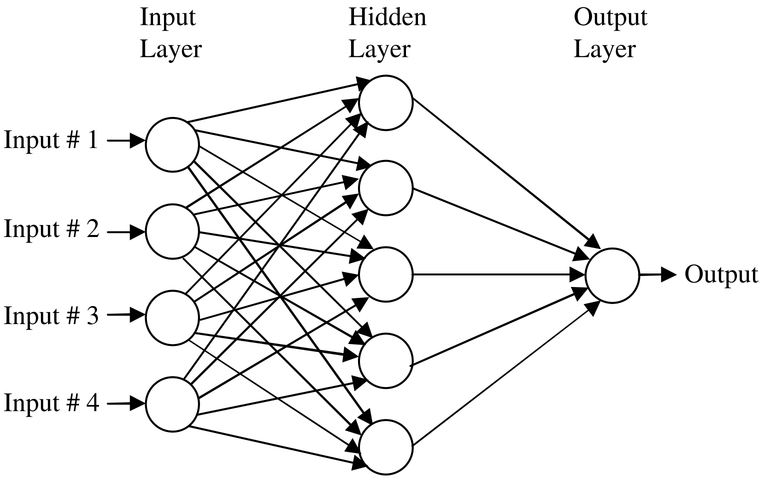

Figure 4 A Neural Network

Unsupervised Learning: Association Rules

Sometimes we wish to use a training set to find any relationship that exists amongst the values of variables, generally in the form of rules known as *association rules.* There are many possible association rules derivable from any given dataset, most of them of little or no value, so it is usual for association rules to be stated with some additional information indicating how reliable they are, for example:

IF variable_1 > 85 and switch_6 = open
THEN variable_23 < 47.5 and switch_8 = closed (probability = 0.8)

A common form of this type of application is called 'market basket analysis'. If we know the purchases made by all the customers at a store for say a week, we may be able to find relationships that will help the store market its products more effectively in the future. For example, the rule

IF cheese AND milk THEN bread (probability = 0.7)

indicates that 70% of the customers who buy cheese and milk also buy bread, so it would be sensible to move the bread closer to the cheese and milk counter, if customer convenience were the prime concern, or to separate them to encourage impulse buying of other products if profit were more important.

Unsupervised Learning: Clustering

Clustering algorithms examine data to find groups of items that are similar. For example, an insurance company might group customers according to income, age, types of policy purchased or prior claims experience. In a fault diagnosis application, electrical faults might be grouped according to the values of certain key variables (Figure 5).

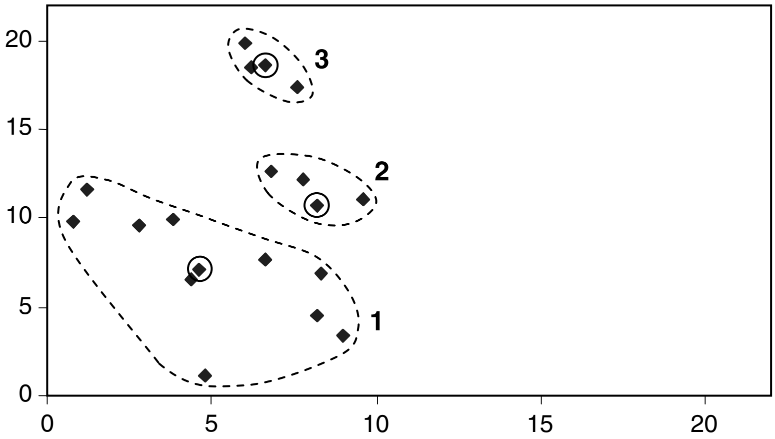

Figure 5 Clustering of Data

About This Book

This book is designed to be suitable for an introductory course at either undergraduate or masters level. It can be used as a textbook for a taught unit in a degree programme on potentially any of a wide range of subjects including Computer Science, Business Studies, Marketing, Artificial Intelligence, Bioinformatics and Forensic Science. It is also suitable for use as a self-study book for those in technical or management positions who wish to gain an understanding of the subject that goes beyond the superficial. It goes well beyond the generalities of many introductory books on Data Mining but—unlike many other books—you will not need a degree and/or considerable fluency in Mathematics to understand it.

Mathematics is a language in which it is possible to express very complex and sophisticated ideas. Unfortunately it is a language in which 99% of the human race is not fluent, although many people have some basic knowledge of it from early experiences (not always pleasant ones) at school. The author is a former Mathematician ('recovering Mathematician' might be a more accurate term) who now prefers to communicate in plain English wherever possible and believes that a good example is worth a hundred mathematical symbols.

Unfortunately it has not been possible to bury mathematical notation entirely. A 'refresher' of everything you need to know to begin studying the book is given in Appendix A. It should be quite familiar to anyone who has studied Mathematics at school level. Everything else will be explained as we come to it. If you have difficulty following the notation in some places, you can usually safely ignore it, just concentrating on the results and the detailed examples given. For those who would like to pursue the mathematical underpinnings of Data Mining in greater depth, a number of additional texts are listed in Appendix C.

No introductory book on Data Mining can take you to research level in the subject—the days for that have long passed. This book will give you a good grounding in the principal techniques without attempting to show you this year's latest fashions, which in most cases will have been superseded by the time the book gets into your hands. Once you know the basic methods, there are many sources you can use to find the latest developments in the field. Some of these are listed in Appendix C.

The other appendices include information about the main datasets used in the examples in the book, many of which are of interest in their own right and are readily available for use in your own projects if you wish, and a glossary of the technical terms used in the book.

Self-assessment Exercises are included for each chapter to enable you to check your understanding. Specimen solutions are given in Appendix E.

Acknowledgements

I would like to thank my daughter Bryony for drawing many of the more complex diagrams and for general advice on design. I would also like to thank my wife Dawn for very valuable comments on earlier versions of the book and for preparing the index. The responsibility for any errors that may have crept into the final version remains with me.

Max Bramer
Digital Professor of Information Technology
University of Portsmouth, UK
January 2007