



Chapter XV

Algorithms for Data Mining

Tadao Takaoka, University of Canterbury, New Zealand

Nigel K. L. Pope, Griffith University, Australia

Kevin E. Voges, University of Canterbury, New Zealand

Abstract

In this chapter, we present an overview of some common data mining algorithms. Two techniques are considered in detail. The first is association rules, a fundamental approach that is one of the oldest and most widely used techniques in data mining. It is used, for example, in supermarket basket analysis to identify relationships between purchased items. The second is the maximum sub-array problem, which is an emerging area that is yet to produce a textbook description. This area is becoming important as a new tool for data mining, particularly in the analysis of image data. For both of these techniques, algorithms are presented in pseudo-code to demonstrate the logic of the approaches. We also briefly consider decision and regression trees and clustering techniques.

Introduction

Data mining is often used to extract useful information from vast volumes of data, typically contained within large databases. In this context “useful information” usually means some interesting information that realistically can only be found by analyzing the database with a computer and identifying patterns that an unaided human eye would be unable to ascertain. Applications of data mining occur in a wide variety of disciplines — the database could contain the sales data of a supermarket, or may also be image data such as a medical x-rays. Interesting information could then be customers’ purchasing behavior in the sales database, or some abnormality in the medical image. As the size of these databases is measured in gigabytes and they are stored on disk, algorithms that deal with the data must not only be fast, but also need to access the disk as few times as possible.

One of the oldest and most widely used data mining techniques involves the identification of association rules. For example, mining an association rule in a sales database can involve finding a relationship between purchased items that can be expressed in terms such as: “A customer who buys cereal is likely to buy milk.” In the following discussion we use a simple example to illustrate a number of issues with association rule mining and to assist in the outline of data mining algorithms. Figure 1 illustrates a simple record of sales at a food supermarket, including a list of items purchased by specific customers, as well as some known attributes of the customers.

Figure 1. Example transaction and customer databases

Transactions					
Customer	Items	Total amount spent			
1	ham, cheese, cereal, milk	\$42			
2	bread, cheese, milk	\$22			
3	ham, bread, cheese, milk	\$37			
4	bread, milk	\$12			
5	bread, cereal, milk	\$24			
6	ham, bread, cheese, cereal	\$44			
Customers					
Customer	Name	Gender	Age	Annual income	Address
1	Anderson	female	33	\$20000	suburb A
2	Bell	female	45	\$35000	suburb A
3	Chen	male	28	\$25000	suburb B
4	Dickson	male	50	\$60000	suburb B
5	Elias	male	61	\$65000	suburb A
6	Foster	female	39	\$45000	suburb B

23 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/algorithms-data-mining/6030

Related Content

Classifying Inputs and Outputs in Data Envelopment Analysis Based on TOPSIS Method and a Voting Model

M. Soltanifarand S. Shahghobadi (2014). *International Journal of Business Analytics* (pp. 48-63).

www.irma-international.org/article/classifying-inputs-and-outputs-in-data-envelopment-analysis-based-on-topsis-method-and-a-voting-model/115520

Sentic-Emotion Classifier on eWallet Reviews

Tong Ming Lim, Yuen Kei Khorand Chi Wee Tan (2023). *International Journal of Business Analytics* (pp. 1-29).

www.irma-international.org/article/sentic-emotion-classifier-on-ewallet-reviews/329928

Strategies for Large-Scale Entity Resolution Based on Inverted Index Data Partitioning

Yinle Zhouand John R. Talburt (2014). *Information Quality and Governance for Business Intelligence* (pp. 329-351).

www.irma-international.org/chapter/strategies-for-large-scale-entity-resolution-based-on-inverted-index-data-partitioning/96158

Forecasting Preliminary Order Cost to Increase Order Management Performance: A Case Study in the Apparel Industry

Tüzin Akçınar Günsari, Aysegül Kayaand Yeliz Ekinci (2022). *International Journal of Business Analytics* (pp. 1-15).

www.irma-international.org/article/forecasting-preliminary-order-cost-to-increase-order-management-performance/298015

Group Processes in the Virtual Work Environment: Evidence for an Alliance-Building Dimensionality

Andrea Roofe Sattlighthightand Sungu Armagan (2012). *Managing Dynamic Technology-Oriented Businesses: High-Tech Organizations and Workplaces* (pp. 48-66).

www.irma-international.org/chapter/group-processes-virtual-work-environment/67428