Introduction to Data Mining

Instructor's Solution Manual

Pang-Ning Tan
Michael Steinbach
Vipin Kumar

Contents

1 Introduction 1

2 Data 5

3 Exploring Data 19

4 Classification: Basic Concepts, Decision Trees, and Model Evaluation 25

5 Classification: Alternative Techniques 45

6 Association Analysis: Basic Concepts and Algorithms 71

7 Association Analysis: Advanced Concepts 95

8 Cluster Analysis: Basic Concepts and Algorithms 125

9 Cluster Analysis: Additional Issues and Algorithms 147

10 Anomaly Detection 157

1

Introduction

1. Discuss whether or not each of the following activities is a data mining task.

   (a) Dividing the customers of a company according to their gender.
   No. This is a simple database query.

   (b) Dividing the customers of a company according to their profitability.
   No. This is an accounting calculation, followed by the application of a threshold. However, predicting the profitability of a new customer would be data mining.

   (c) Computing the total sales of a company.
   No. Again, this is simple accounting.

   (d) Sorting a student database based on student identification numbers.
   No. Again, this is a simple database query.

   (e) Predicting the outcomes of tossing a (fair) pair of dice.
   No. Since the die is fair, this is a probability calculation. If the die were not fair, and we needed to estimate the probabilities of each outcome from the data, then this is more like the problems considered by data mining. However, in this specific case, solutions to this problem were developed by mathematicians a long time ago, and thus, we wouldn't consider it to be data mining.

   (f) Predicting the future stock price of a company using historical records.
   Yes. We would attempt to create a model that can predict the continuous value of the stock price. This is an example of the

area of data mining known as predictive modelling. We could use regression for this modelling, although researchers in many fields have developed a wide variety of techniques for predicting time series.

(g) Monitoring the heart rate of a patient for abnormalities.
Yes. We would build a model of the normal behavior of heart rate and raise an alarm when an unusual heart behavior occurred. This would involve the area of data mining known as anomaly detection. This could also be considered as a classification problem if we had examples of both normal and abnormal heart behavior.

(h) Monitoring seismic waves for earthquake activities.
Yes. In this case, we would build a model of different types of seismic wave behavior associated with earthquake activities and raise an alarm when one of these different types of seismic activity was observed. This is an example of the area of data mining known as classification.

(i) Extracting the frequencies of a sound wave.
No. This is signal processing.

2. Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.

The following are examples of possible answers.

- Clustering can group results with a similar theme and present them to the user in a more concise form, e.g., by reporting the 10 most frequent words in the cluster.
- Classification can assign results to pre-defined categories such as "Sports," "Politics," etc.
- Sequential association analysis can detect that that certain queries follow certain other queries with a high probability, allowing for more efficient caching.
- Anomaly detection techniques can discover unusual patterns of user traffic, e.g., that one subject has suddenly become much more popular. Advertising strategies could be adjusted to take advantage of such developments.

3. For each of the following data sets, explain whether or not data privacy is an important issue.

   (a) Census data collected from 1900–1950. No

   (b) IP addresses and visit times of Web users who visit your Website. Yes

   (c) Images from Earth-orbiting satellites. No

   (d) Names and addresses of people from the telephone book. No

   (e) Names and email addresses collected from the Web. No