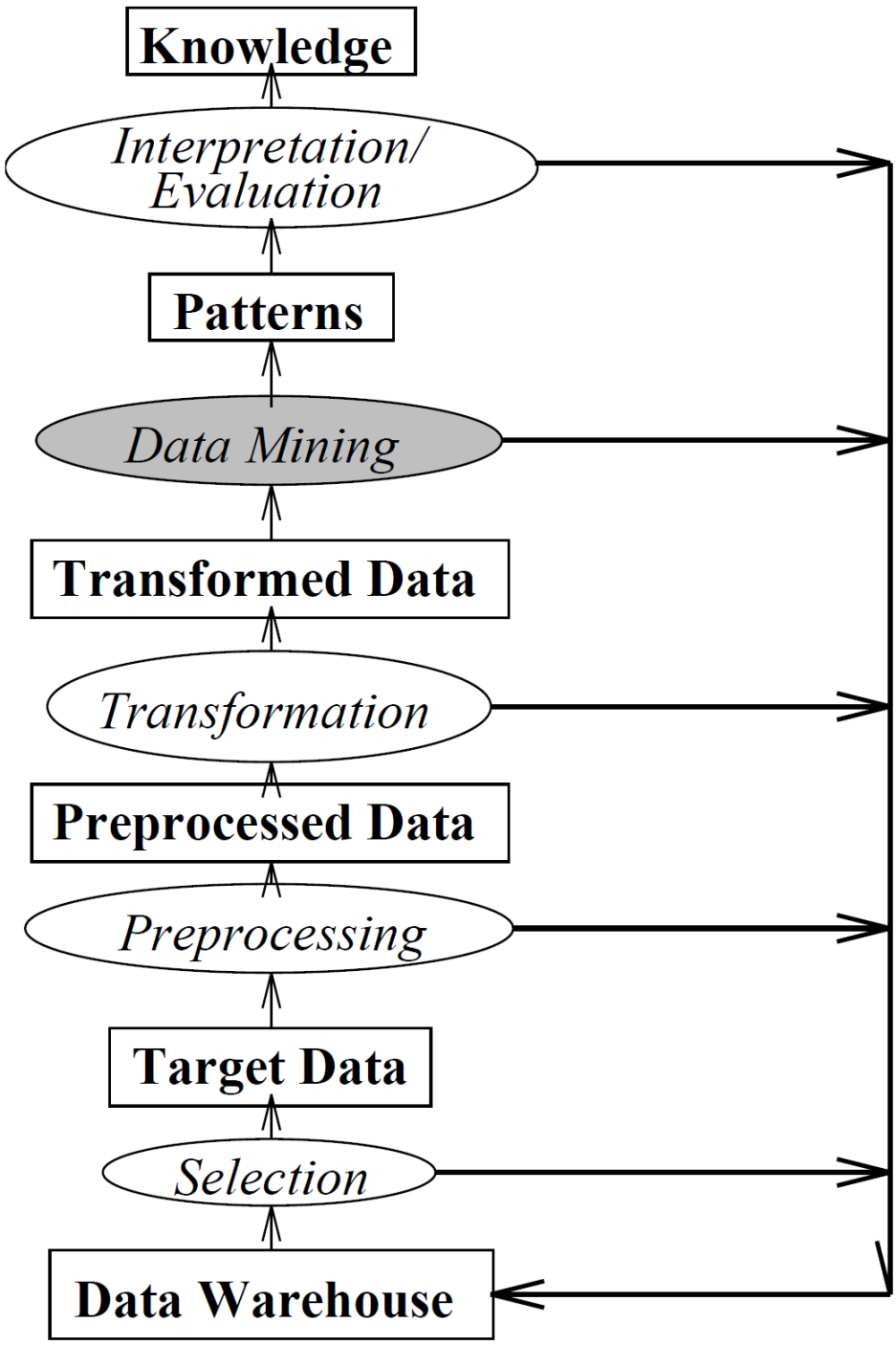
Week 10 Solutions



Analyse the situation with respect to each step of the knowledge discovery process.

1) ***Suppose that you are an analyst working for a major bank and you have been given the task of determining 'How should we decide whether to give an applicant a personal loan or not? How would you proceed.***

This is clearly a classification problem. To build the classifier we will need suitable training and test data which will have to come from the data warehouse. We will need positive examples (give loan) and negative examples (no loan).

Knowledge: A classifier that is accurate enough and makes decisions that are understandable by bank staff and customers.

Interpretation/Evaluation of Patterns: Accuracy will be a primary criterion, but understandability is also probably important. Raw accuracy might not be enough. We might have to look at sensitivity and specificity

Data Mining: Experimentation with a range of classifiers looking for acceptable accuracy using cross validation or train/test split.

Selection, Pre-processing and Transformation.

The bank will have considerable historical records of each customer. We need to extract positive and negative customers. We might say a positive customer is one who has taken out a loan and paid it back. A negative customer is one who has taken out a loan and not paid it back. This is probably too simplistic. Somebody who has taken out a loan, is making regular payments and has not missed a payment is probably a positive customer. Somebody who has missed a few payments over a few years is probably still a positive customer. Somebody who made a few payments and then stopped is probably a bad customer. Domain knowledge from the bank will be needed. Potentially we can select any data we have about a customer – savings bank accounts and transactions, credit cards and transactions, loans past and present, changes of address.

We need data in relational form. Some fields, for example current suburb and state, can go directly into the relation. Other attributes will need to be computed, for example number of credit cards or years at current address.

Domain knowledge from the bank will be needed to determine the most relevant attributes. There may be a need for data cleaning. For example, if some of the fields in the data warehouse were collected without validation there may be letters in numeric fields or illegal values in enumerated fields.

Iteration of the process

It's very unlikely that the first pass will produce the desired classifier. What to do if the accuracy is not good enough? First look at alternative classifiers and parameter values. But probably we don't have the right data, so go back to the warehouse and look for other data and features that might give improved accuracy.

Deployment of the classifier

The classifier we generated with the above method will suit loan applicants that are already customers. But what about applicants who are customers of another bank? Some of the data will not be available for these applicants. We may have to build a separate classifier for these applicants.

2) ***A credit card company has decided to stamp out fraud. How could they proceed?***

This is primarily a classification task. Clustering and association finding might also reveal some useful patterns.

3) RMIT wants to know the numbers of students from other countries in order to do more effective marketing.

To find out the number of students from other countries, we can easily count the numbers of international or domestic students by looking the student database. This is not a data mining task.

4) A major telephone company has noticed that a steady stream of customers is terminating their contracts and suspects that they are going to a competitor. They would like to make sure that no further customers are lost. How can customers who are likely to leave be identified? (Once they are identified they can be given some inducement to stay.) This "churn" situation is a common problem faced by many companies such as banks and airlines.

This is primarily a classification task. We need to get data on customers who left and those who didn't. There should be data available in customer database. Clustering and association finding might also reveal some useful patterns.

5) Making a profit in buying and selling shares depends on the future value of the shares. Thus, if one can accurately predict the future value of the shares one can trade profitably. Data for which a new point is available on a regular cycle is called a time series. Share prices form a time series in which there is a new point every day. Can the previous history of a time series be used to predict the value 1,2...n time steps ahead?

This is a numeric prediction task. M5P or linear regression or a neural net could be used. The historical data could be used to get a prediction model which would then be used to predict future values.

6) A professional gambler wants to predict the winners of horse races.

This is a classification task in which a model would be built on historical horse racing data. However, I don't think that there is a central racing database so the data would be hard to collect. Also it's not clear how accurate the model would be.

7) A mail order company would like to 'segment' its customers into a number of groups with common characteristics. Each customer is allocated to one segment. Customers in a group should be similar enough so that promotional material can be carefully targeted.

This looks like a straightforward clustering task. The generated clusters need to be carefully examined to see if they meet customer expectations.

8) A hospital has accumulated many records over the years. They would like to identify people who have a high risk of heart attacks. How could they proceed?

This is primarily a classification task. Clustering and association finding might also reveal some useful patterns.

9) A large retailer wants to know the breakdown of sales by state and store in order to do better logistics planning.

Not a data mining task. Can be done by counting in a data base.

10) Acme computers sells computers, parts and software through a web page. As part of a sales transaction Acme would like to have a "People who bought X also bought Y." feature. How could this be implemented to maximize sales?

They can use an association rule finding algorithm like Apriori to find the pattern of $X \longrightarrow Y$.

11) The Australian tax office has been collecting access logs for their web site since 2000. They would like to know whether the data contains any indications of tax fraud. How could they proceed?

This is primarily a classification task. Examples of fraudulent and not fraudulent visits are needed. However, it might be very difficult to manually classify the visits to get a good training set. Possibly association finding might reveal something.