CHAPTER 1

Data Mining: Concepts

This chapter gives an overall introduction of the data mining field, its various research trends, and its applications.

1.1 CONCEPTS

Data mining is a term that usually comes in conjunction with databases. Data mining is the collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful, and understandable patterns in large databases. It usually deals with history databases. (History databases are usually huge collections of data.) As an example, the task of finding all the customers coming from London is just a database query whereas understanding that customers who buy milk tend to buy biscuits also is hidden information. The latter is a typical result of data mining. The patterns must be actionable so that they can be used in an enterprise's decision making process. Data mining is usually used by business intelligence organizations and financial analysts, but it is increasingly being used in science to extract information from the enormous data sets generated by modern experimental and observational methods [1].

Note: Data mining is a technique to find hidden patterns in a huge history database to help top level managers in decision making. It has close relation with statistics and is a hot research area nowadays.

1.2 DATA MINING APPLICATIONS

Following are the different types of applications where data mining can be directly applied:

1. In *classification*, the goal is to classify a new data record into one of the many possible classes, which are already known. For example, an applicant has to be classified as a prospective applicant or a defaulter in a loan database, given his various personal and other demographic features along with previous purchase characteristics.
2. In *estimation*, unlike classification, we predict the attribute of a data instance—usually a numeric value rather than a categorical class. An example can be "Estimate the percentage of marks of a student whose previous marks are already available".
3. The border line between *prediction*, classification, and estimation is too narrow. The main difference is that the predictive model predicts a future outcome rather than the current behaviour. The output attribute can be categorical or numeric. An example can be "Predict next week's closing price for the Google share price per unit."
4. *Market basket analysis* or *association rule mining analyses* hidden rules called association rules in a large transactional database [2]. For example, the rule {pen, pencil → book} provides the information that whenever pen and pencil are purchased together, book is also purchased; so these items can be placed together for sales or supplied as a complementary product with one another to increase the overall sales of each item.
5. In *clustering*, we use unsupervised learning technique, where target classes are unknown. For example, given 1000 applicants have to be classified based on certain similarity criteria and it is not predefined which are those classes to which the applicants should finally be grouped into.
6. Other categories of data available nowadays are scientific data collected by satellites using sensors, data collected by telescopes scanning the skies, scientific simulations generating terabytes of data, etc. Data mining can be applied to analyze these types of data too. In sequential pattern discovery, first we identify patterns and then we try to analyze whether there is any sort of sequential relationship between them. Consider the following example of buying textbooks and sports goods.
   (Introduction to Java) (Object oriented modelling) → (UML guide)
   (Shoes) (Racket, shuttle) → (Sports jacket)

From these examples, it is clear that when customers buy the first set of products on the left side of the implication, sequentially they tend to buy the products on the right side of the implication. Business intelligence, business data analytics, bioinformatics, Web mining, text mining, social network data analysis, etc., are some of the areas where data mining can be applied. It can also be applied

to problems related to social science. Some challenges in front of the data mining researchers are handling big data analysis, mining over a cloud, and optimizing searches for the big data over the cloud. In the coming sections, various stages occurring in a typical data mining problem are explained. Different data mining models that are commonly applied to various problem domains are also discussed in detail in the coming sections.

1.3 DATA MINING STAGES

Any data mining work may involve the various stages shown in Fig. 1.1. Business understanding involves understanding the domain for which data mining has to be performed. The domains can be financial domain, educational data domain, etc. Once the domain is understood properly, the domain data has to be understood next. Here relevant data in the needed format will be collected and understood.

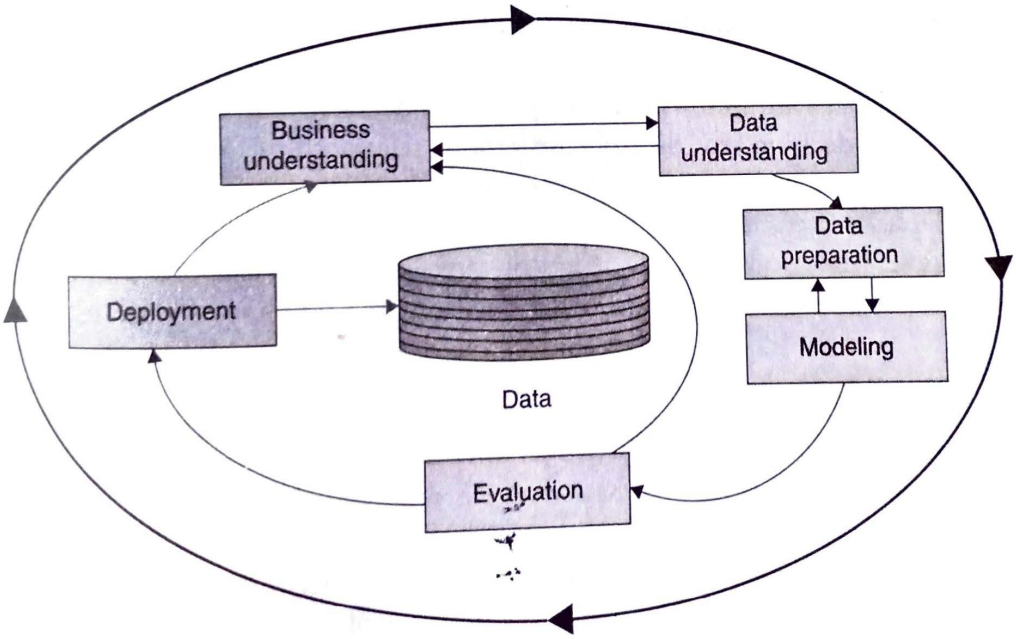


Fig. 1.1 Data mining stages

Data pre-processing is an important step in the sense that the data is to be made suitable for further processing and mining. This involves cleaning the data, transforming the data, selecting subsets of records that are of interest, and so on. When data are prepared, there are two stages, namely selection and transformation. Data selection involves choosing those data which will be useful for the data mining purpose. There are many statistical data analysis techniques that can be

adopted at this stage. Data transformations involve changing the range of data such that it can be directly used. For example, the data range can be converted from 0–100 to 0–10.

Data modelling involves building models such as decision tree, support vector machine (SVM), and neural network from the pre-processed data. Some models are described as follows:

1.4 DATA MINING MODELS

There are many popular models that can be effectively used in different data mining problems. Decision trees, neural networks, naive Bayes classifier, lazy learners, support vector machines, and regression-based classifiers are a few among them. Depending upon the type of application, nature of data, and attributes, one can decide which will be the most suited model. Still there is no clear-cut answer to the question of which is the best data mining model. One can only say, "For a particular application, one model is better than the other."

Decision trees: Decision tree is one of the most popular classification models. It is similar to a tree-like structure, where each internal node denotes a decision on the value of an attribute. A branch represents a decision, and leaves represent target classes. A decision tree displays the various relationships found in the training data by executing a classification algorithm.

Decision trees have become very popular and powerful tool for classification purpose. Their leaves usually represent target classes whereas branches represent attribute decisions. There are many decision tree building algorithms such as CART and C 4.5 which dynamically build a decision tree depending on the underlying huge history data.

Neural networks: Neural networks offer a mathematical model that attempts to mimic the human brain. Knowledge is represented as a layered set of interconnected processors called neurons. Each node has a weighted connection with other nodes in adjacent layers. Learning in neural networks is accomplished by network connection weight changes while a set of input instances is repeatedly passed through the network. Once trained, an unknown instance passing through the network is classified according to the values seen at the output layer. There are many ongoing works as shown in [4], which deal with neural network construction, to identify the important issues involved. It also explains the various domains to which neural networks can be applied.

Naive Bayes classifier: This classifier offers a simple yet powerful supervised classification technique. The model assumes all input attributes to be of equal importance and independent of one another. Naive Bayes classifier is based on the classical Bayes theorem presented in 1763 which works on the probability theory. In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other

feature. Even though these assumptions are likely to be false, the Bayes classifier still works quite well in practice.

Bayes' classifiers can be trained well in a supervised learning setting. In many applications, parameter estimation for naive Bayes model uses the method of maximum likelihood. It can often outperform the more sophisticated classification methods, even being simple conceptually. An advantage of this classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.

1.5 SUCCESS STORIES

1. **Prediction of loan defaulters:** It will be very useful for banks to predict whether a customer can be granted loan or not based on the previous experiences with the customers of the same type. The American Express in UK has been able to successfully implement this application for loan processing purpose.
2. **Bioinformatics and medicine:** Bioinformatics is the field of combining computer science with biology for various types of analysis of biological data. There are huge repositories of genetic sequence information (including the sequence of whole genomes). There is a tool called Glimmer which can identify the genes in a genome in an efficient way.

   Patient data can be analyzed and interesting data mining models such as probabilistic relational models can be built. The model will reveal dependencies between variables. This particular principle has been successfully implemented on analyzing the data of tuberculosis patients in the USA.
3. **Business intelligence (BI):** It is very important for businesses to clearly analyze their market to be in the cutting edge. BI technologies provide historical, current, and predictive views of business operations. Data mining forms the core of business intelligence. Online analytical processing tools in business intelligence rely on data warehousing and multidimensional data mining. Oracle, IBM, etc., have successfully implemented BI systems. Clustering plays a central role in customer relationship management, which groups customers based on their similarities.
4. **Web search engines:** Web search engines are essentially very large data mining applications. Crawling, indexing, and searching are three important aspects of these search engines. Since majority of search engines are dealing with huge amount of data, usually they have to work on clouds of computers. It is a very active research area nowadays. Also many search engines such as Google and Yahoo have successfully implemented context sensitive searching (which makes searching a quick experience), since for the users, majority of queries are auto suggested.
5. **Academic applications:** Data mining can be effectively used to predict the academic performances as well as placement chances of students [3]. Decision tree is one of the models that can be applied to data mining. There is another

important model used in data mining for prediction called "naive Bayes classifier." Students can make wise career decisions using these data mining tools. A student enters his Entrance Rank, Gender (M/F), Sector (rural/urban), and Reservation (OBC/SC/ST/GEN) category. Based on the entered information, the model will return which branch of study is Excellent, Good, Average, or Poor for him/her based on history data analysis using data mining techniques. Also in this work performances of decision trees and naive Bayes classifier on the same training and test data were compared.

6. Some other areas in which data mining can be applied are as follows:
   a. Biological data analysis
   b. Call record analysis
   c. Churn prediction for telecom subscribers, credit card users, etc.
   d. Decision support
   e. Financial forecasting
   f. Insurance fraud analysis
   g. Logistics and inventory management
   h. Trend analysis
   i. Time series analysis

1.5.1 Time Series Analysis

In time series analysis, each time series describes a phenomenon as a function of time. For example, daily stock prices can be used to describe the fluctuations in the stock market at each time during a day. In general, for a time series X with n observations, X is represented as

$$X = (v_1, t_1), (v_2, t_2), \ldots, (v_n, t_n)$$

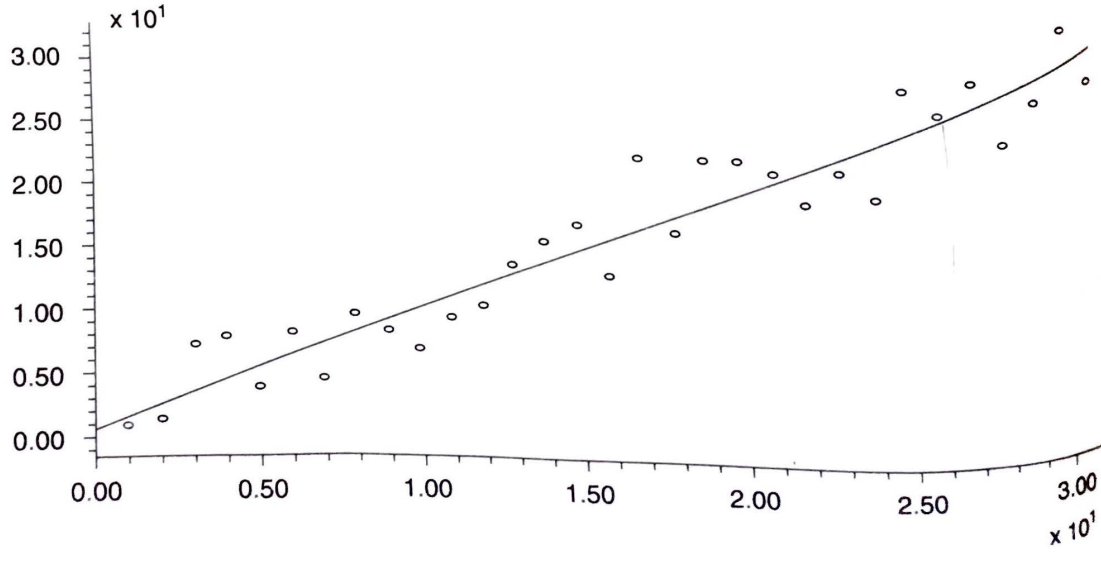


Fig. 1.2 A sample regression line used in time series analysis using XY axi

where v_i and t_i are the observation value and its time stamp, respectively. The data that is used in this work to compute the absorption rate is a time series data, where absorption rates for different years are used. Many researchers are working on applications that use time series analysis. Data mining on time series data is essential nowadays to improve business analysis and forecasting. Regression and time series analysis are very similar concepts in statistics; a graph is shown in Fig. 1.2.

1.6 DATA WAREHOUSING AND OLAP

Major enterprises and corporate companies have huge data to be stored and processed for various transactions and managerial analysis. The various components of these systems can be payroll, billing, HR, etc. Many of these get updated on a continuous basis. Usually these updatable data are kept on servers called OLTP (online transaction processing) servers. But the data that are mostly historic and are not frequently updated are usually stored in huge history databases called data warehouses. These data are used by top level managers for decision making process, especially for comparisons, predictions, etc. The important point to be noted is that data warehouses are not updated regularly, rather they are appended regularly. Live transaction data that are updated become the history data of the warehouse later, for analysis purpose.

OLAP technology (online analytical processing) uses data warehouses for online analysis, providing quick responses to queries. Figure 1.3 shows the structure of a data warehousing system for OLAP. Using online analysis, graphical tools, and OLAP's multidimensional data model and aggregation techniques, large

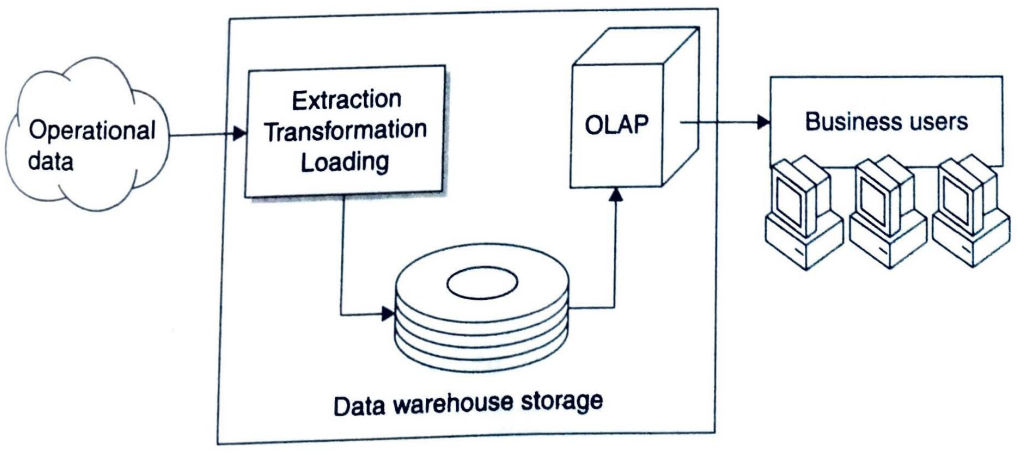


Fig. 1.3 Data warehousing and OLAP

Note: Data warehouses are huge history databases which are used by top level managers for decision making. Usually, live transactional data is not a part of this database and, mostly, the data is historic. This history data is used for data mining and OLAP activities to produce meaningful reports for managers.

amounts of data can be summarized. A query to historical data may lead to other queries also when analysis takes place. OLAP systems provide the speed and flexibility to support the analyst in real time.

A data warehouse is often implemented as a collection of data marts. Data marts are treated as a part of the data warehouse or as separate entities, the meaning can be varied depending on the context. In this section, we see various terms related to data warehouse.

Data warehouses may use relational databases themselves for their implementation. The relational database must provide rapid data transfer; update, flexible, and efficient indexing; and sophisticated and effective query capabilities to organize and retrieve data warehouse data. Locking mechanisms and high multi-table transaction throughput may be more significant in OLTP systems that are very important in data warehouse operations. A data mart is a miniature data warehouse; in others, it is just one segment of the data warehouse. Functional segments of the organization get information from data mart. Data marts for the Sales, Inventory, and Shipping department are some examples. Data marts may be used to model data warehouse data to suit a geographically distributed business. Each of the data marts may be treated as individual business unit.

Data marts are sometimes designed as complete individual data warehouses also. Usually, presentation services for clients are the main functionality of data marts. Data from data marts are loaded to a data warehouse using a batch process called ETL (extract, transform, and load).

Normally data in an organization are distributed in multiple data sources and are incompatible with each other. Consider a retail example. Let the point-of-sales data and the sales made via call-center or the Web are stored in different locations and formats. It will be a time consuming process to obtain OLAP reports such as "What are the most popular products purchased by customers between the ages 15 and 30?" OLAP process involves extracting data from the data repositories and making them compatible with each other. After this, the meaning of data across different repositories may be compatible. OLAP describes exactly what ha happened during this analysis process. Star design is the most common metho for data modelling shown in Fig. 1.4(a).

The fact table is the central table in an OLAP star data. The surrounding tabl are called the *dimensions*. Using this data model, it is possible to build reports th answer questions such as the following:

1. The supervisor that gave the most discounts.
2. The quantity shipped on a particular date, month, year, or quarter.

To obtain answers for such questions from a data model, OLAP cubes a created as shown in Fig. 1.4(b). It is the name given to the process of linking da from different dimensions.

The cubes are developed along business units such as sales or marketing. OL can be a valuable and rewarding business tool. OLAP analysis can help an orga zation evaluate targets in addition to producing reports.

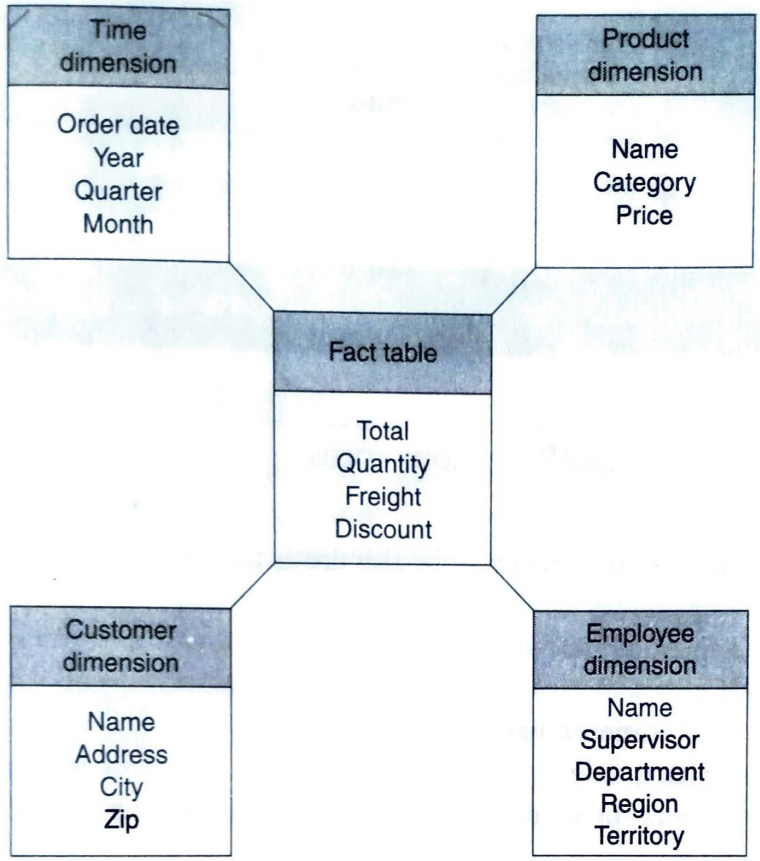


Fig. 1.4(a) Star design for an OLAP system

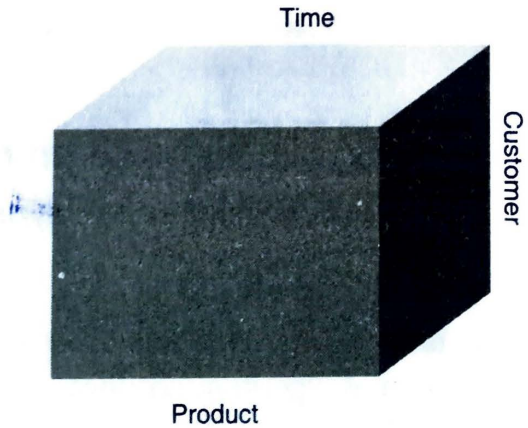


Fig. 1.4(b) Data cube for a multidimensional database

As shown in Fig. 1.5, there are many steps in obtaining reports. First the data is extracted from OLTP servers and imported to the OLAP database. Then the data is converted to multidimensional databases—new generation databases that are ready to serve queries which require lots of aggregation. From these cubes, reports are generated.

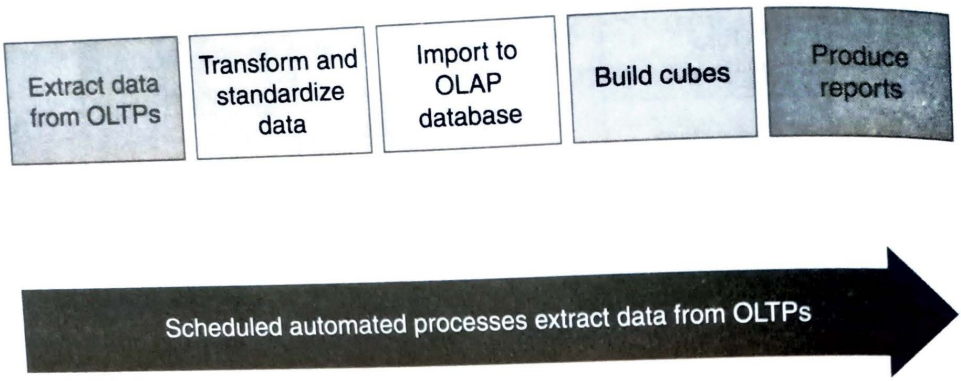


Fig. 1.5 Steps in OLAP creation process

Key points regarding data warehouses are as follows:

1. Data is appended, not updated.
2. Data is historical.
3. Data is multidimensional.
4. Uses complex queries.
5. Used mainly at the managerial level, and not at the operational level.

So, in data warehouses, the data is not updated frequently. Data warehouses are more used for reports that are based on history. Since data warehouses contain historical data, they are mostly used for meaningful data mining purposes. This information can be very useful for managers to make meaningful decisions. Chapter 8 describes some practical aspects of data warehousing using Oracle BI (Business Intelligence).

1.7 RESEARCH CHALLENGES

Currently, data mining research works proceed mainly in the following four directions:

1. Applying data mining principles to a new application domain.
2. Developing a new data mining algorithm/classifier.
3. Optimizing existing models.
4. Combining classifiers called meta-learners.

1.7.1 Application of Data Mining Principles

Majority of data mining researches belong to this category. Here data mining principles are used to predict student retention analysis, where we can understand which among the students admitted to a course will continue the course without abrupt exit [4]. Data mining principles can be applied in the banking sector to

know the most prospective customers who can respond to a new business offer. Hence, in majority of data mining researches, it is observed that first a relevant social-science problem in a domain, such as banking domain, customer domain, and student domain, is selected, where various data mining models are applied and compared.

1.7.2 Developing a New Data Mining Algorithm/Classifier

There are many popular data mining algorithms in the industry which are commercially used. For example, decision trees, neural networks, Bayes classifier, SVM, etc. But the challenge lies in developing new algorithms specially suited for a particular domain. Here the researcher has to prove its effectiveness by comparing various performance measurement factors such as accuracy and ROC value across various algorithms and prove that the advantage is statistically significant. One of the popular decision tree algorithms is C 4.5, which is included in many data mining packages [5].

1.7.3 Optimizing Existing Models

Optimization is an important activity that is being done by data mining researchers to improve data mining models. For example, there is an activity called the pruning of decision trees. It describes how the size of the decision tree can be reduced effectively. In pre-pruning, optimization is done before tree building whereas in post-pruning, the resizing of the tree is done after the tree is built.

1.7.4 Combining Classifiers

It is mere common sense to think that when we combine the efforts of many models, we get a better model. This concept has become a very hot research area now. There are two approaches. The first approach uses voting, in which that prediction is chosen which is made by maximum models. This concept is similar to voting done by humans where the majority decision is taken as final. In the second approach, the outputs of the models are given to another model called meta-level classifier that decides which decision to finalize.

There are two types of voting: bagging and boosting. In bagging, n models, usually of the same type, are constructed and for an unknown instance, for each model, predictions are recorded. The class having the maximum votes among the predictions from models is assigned.

Boosting is very similar to bagging except the model construction phase, where every time those instances that are most misclassified are trained more; here will be n classifiers which themselves will have individual weights for their ccuracies. The class having the maximum weight is assigned. An example is

AdaBoost algorithm. Bagging is better than boosting as boosting suffers from overfitting.

1.8 OTHER HOT RESEARCH AREAS IN DATA MINING

The following list shows some other hot data mining research problem areas. This list may be helpful for beginners in data mining researches for reading the corresponding literature in online journals/proceedings and then land on to a problem.

1. Data mining of XML data
2. Clustering of CRM (customer relationship management) applications
3. Bioinformatics (gene data analysis and finding out gene expression data combinations for a particular disease using data mining techniques)
4. Web mining over a cloud of computers
5. Intrusion detection using data mining
6. Scaling data mining algorithms, when data does not fit into memory
7. Distributed data mining
8. Email clustering for finding communication threads among mails
9. Improvements for support vector machines algorithms
10. Improvements for association rules *apriori* algorithms
11. Vitro fertilization—which embryos to use for transferring to uterus out of several embryos generated from egg–sperm fertilization

1.9 MACHINE LEARNING AND STATISTICS

Statistics and machine learning form the basis of data mining. Both these branches are inter-related too. Both have contributed to computer science too. Since inception, these two have had rather different traditions. Statistics is more concerned with testing hypotheses. Formulating the process of generalization as a search through possible hypotheses is the main concern of machine learning. Sometimes both may cover more areas than these. Statistics has been the favorite to many classification models in data mining such as naïve Bayes classifier and decision tree induction. One of the initial works in the field of decision trees [7] was conducted by Brieman et. al. They published a book on classification and regression trees in the mid-1980s. A prominent machine learning researcher, J. Ross Quinlan, was developing a system for inferring classification trees from examples before this also in a similar way. The researchers became aware of one another's work much later only. Most learning algorithms use statistical tests when constructing rules or trees and for correcting models that are "overfitted" in that they depend too strongly on the details of the particular examples used to produce them. Statistical tests are used to validate machine learning models and to evaluate machine learning algorithms. SVMs are a new breed of classification

models which is highly mathematical in nature. These models also show how close machine learning and other branches of mathematics are.

1.10 ETHICS OF DATA MINING

The use of data about people has very serious ethical implications in data mining. Practitioners of data mining techniques should always have an eye on such matters. Data mining answers questions such as who gets the loan and who gets the special offer. Sexual, religious, and other types of discriminations are not only unethical but also illegal. However, the complexity of the situation depends a lot on the application. Using sexual information for medical diagnosis is certainly ethical, but using this for mining loan payment behaviour is not. Even when sensitive information is removed, there is always a risk that models will be built that depend on variables that can be shown to substitute for such characteristics. A situation like this can be seen in real life. We know that people frequently live in areas that are associated with particular ethnic identities. It is risky to build models that are based on race, when using an area code in a data mining study. This may create a problem; it may be necessary to determine the conditions under which the data was collected and for what purposes it can be used. It is accepted that before people make a decision to provide personal information, they have the right to know its usability. Hence, individuals should be told these things in plain language that they can understand.

For example, in France, it was found that people with red cars are more likely to default on their car loans. We should consider all questions such as what is the status of such a discovery, it is based on what information, under what conditions was that information collected, and in what ways is it ethical to use it. Clearly, insurance companies are in the business of discriminating among people based on stereotypes like "young males pay heavily for automobile insurance."

But such stereotypes are solely not based on statistical correlations; they also involve common-sense knowledge about the world. Here the data scientist may have to apply common sense rather than fully depending on statistical information. When presented with data, it is needed to ask who is permitted to have access to it, for what purpose was it collected, and what kinds of conclusions are legitimate to be drawn from it. The ethical dimension raises tough questions. It is necessary to consider the norms of the community that is used to dealing with the kind of data involved, standards that may have evolved over decades but ones that may not be known to the data scientist. In addition to community standards for the use of data, logical and scientific standards must be adhered to when drawing conclusions from it.

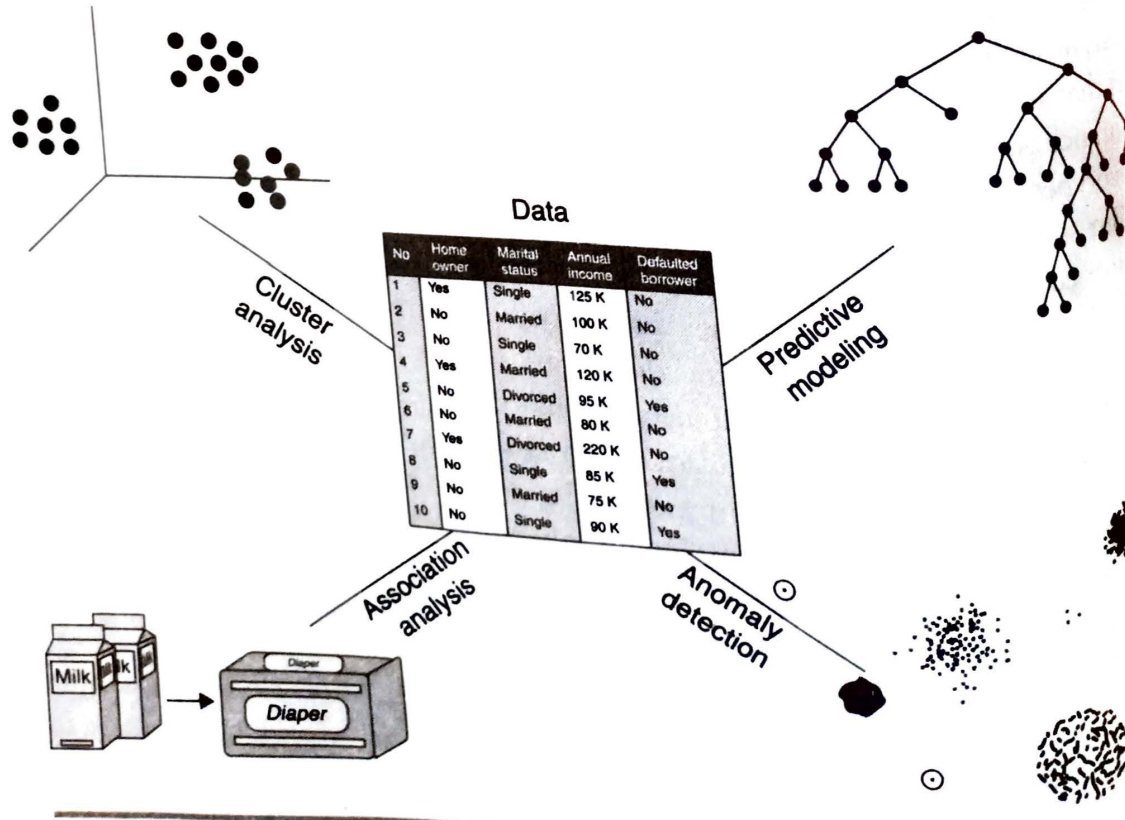
The point is that data mining is just a tool in the whole process. It is people who take the results, along with other knowledge, and decide what action is to be taken. Of course, those who use advanced technologies should consider the wisdom of what they are doing. So the conclusion that can be drawn is, "when data mining is conducted in a particular domain, one should seriously address the

questions, whether the objective of mining is useful for the mankind and is there anything non-ethical hidden behind the rules extracted from the data mining process." Hence, the conclusion is that data mining applications may involve many ethical dimensions and application of judgements based on common sense, which may vary from domain to domain [6].

1.11 POPULAR TOOLS

There are many popular data mining tools that are used all over the world. Some of them are used for research/academic purposes whereas others are used for commercial purposes. These tools can be used for data mining for different purposes and are given in Fig. 1.6. Following are some popular data mining tools:

1. WEKA
   URL: http://www.cs.waikato.ac.nz/ml/weka/
2. SPSS Clementine
   URL: http://www.spss.com/clementine/
3. Rapid Miner
   URL: http://rapid-i.com/content/blogcategory/10/69/lang,en/
4. SAS/SAS Enterprise Miner
   URL: http://www.sas.com/technologies/analytics/datamining/miner/index.html
5. MATLAB, Microsoft SQL Server, Oracle BI, etc., are also very popular data mining research/commercial tools.



No	Home owner	Marital status	Annual income	Defaulted borrower
1	Yes	Single	125 K	No
2	No	Married	100 K	No
3	No	Single	70 K	No
4	Yes	Married	120 K	No
5	No	Divorced	95 K	Yes
6	No	Married	80 K	No
7	Yes	Divorced	220 K	No
8	No	Single	85 K	Yes
9	No	Married	75 K	No
10	No	Single	90 K	Yes

Fig. 1.6 Important data mining tasks

SUMMARY

In this chapter, we studied the fundamentals of data mining and understood the real advantages of using this concept in social science problems. We saw various ways by which data mining can be applied to domains such as banking, student data, and customer data. We also saw lots of hot research problem areas and various success stories.

Data mining is usually conducted over huge history non-transactional databases called warehouses. Various analyses are carried out using online analytical processing. The architecture of a data warehouse was analyzed in this chapter.

Different types of data mining applications such as clustering and classification were analyzed. Finally, we saw various popular data mining tools used in the academia and industry.