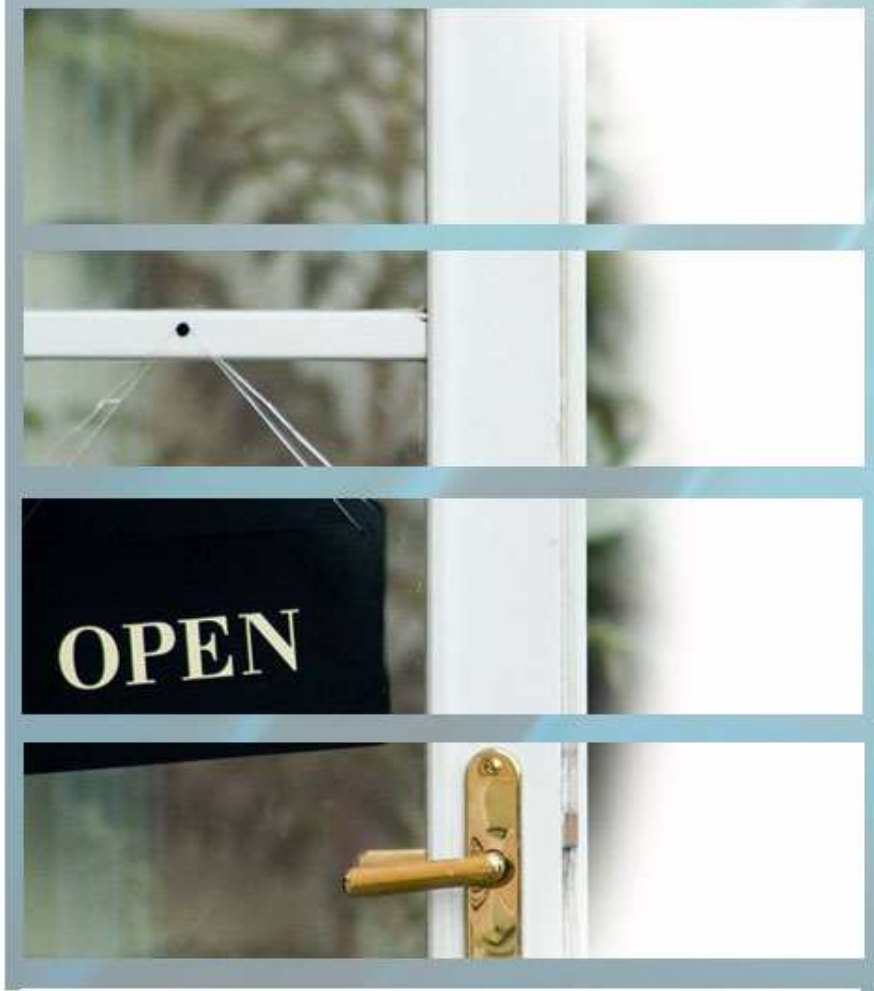
Chapter 13

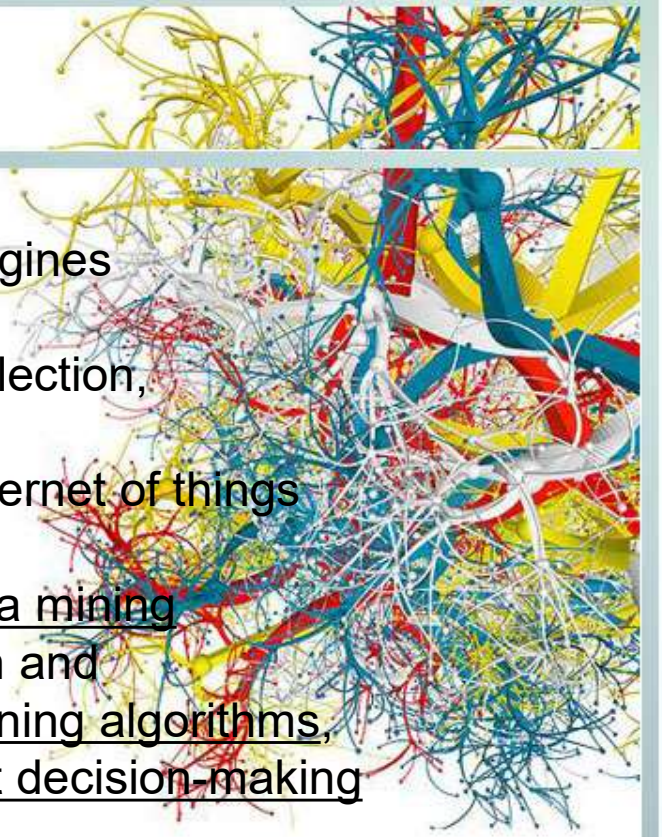
Intelligent decision



Intelligent decision is the source of the "intelligence" of the Internet of Things.

This chapter introduces the basic flow of data mining, basic types and typical algorithms.

Review



Chapter 12 introduces the knowledge of search engines

- The basic components of a search engine
- Search engine architecture (information collection, indexing technology, search services)
- The challenges of search engines in the Internet of things

This chapter introduces the basic processes of data mining (preprocessing, data mining, knowledge evaluation and representation), focuses on several typical data mining algorithms, and finally discusses the new features of intelligent decision-making in the Internet of things.


Content

13.1 Data mining overview

13.2 Basic types and algorithms of data mining*

13.3 Intelligent Decision and Internet of Things

What is data mining?
What are the three steps in data mining?

13.1 Data mining overview

Data Mining

- The process of extracting potentially useful and understandable patterns from large amounts of data
- It is an iterative process of human-computer interaction and processing, going through multiple steps, and in some steps the user needs to provide the decision

The process of data mining:

- Data preprocessing, data mining and evaluation and representation of mining results
- The output of each phase becomes the input to the next

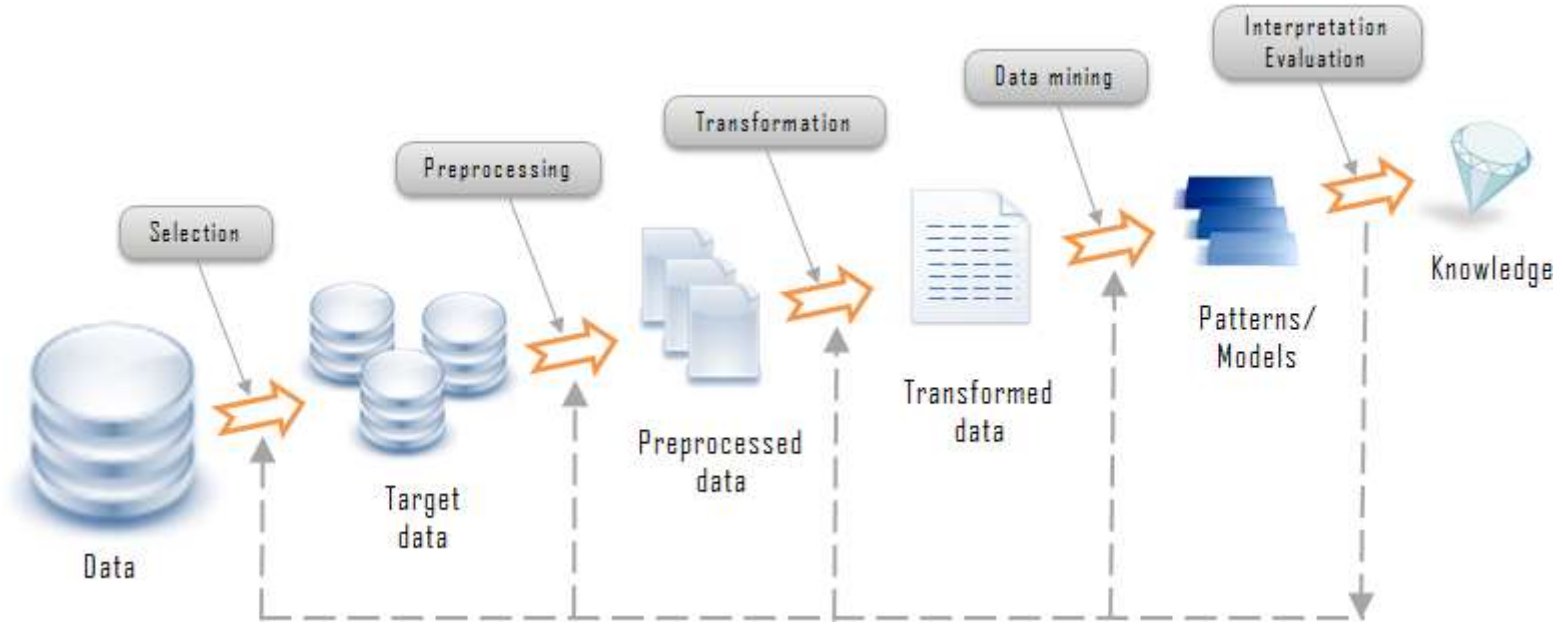
13.1 Data mining overview

The process of data mining

- **Data preprocessing stage**
  - ✓ Data preparation: understand domain characteristics and determine user requirements
  - ✓ Data selection: relevant data or samples are selected from the original database
  - ✓ Data preprocessing: check data integrity and consistency, eliminate noise, etc
  - ✓ Data transformation: reduction of data volume by projection or other manipulation
- **Data mining phase**
  - ✓ Identify mining objectives: determine the type of knowledge to be discovered
  - ✓ Selection algorithm: select the appropriate data mining algorithm according to the determined target
  - ✓ Data mining: using the selected algorithm, extract the relevant knowledge and express it in a certain way
- **Knowledge evaluation and presentation stage**
  - ✓ Pattern assessment: the assessment of the patterns (knowledge) discovered during the data mining steps
  - ✓ Knowledge representation: use visualization and knowledge representation techniques to present the knowledge mined

13.1 Data mining overview

Data mining process

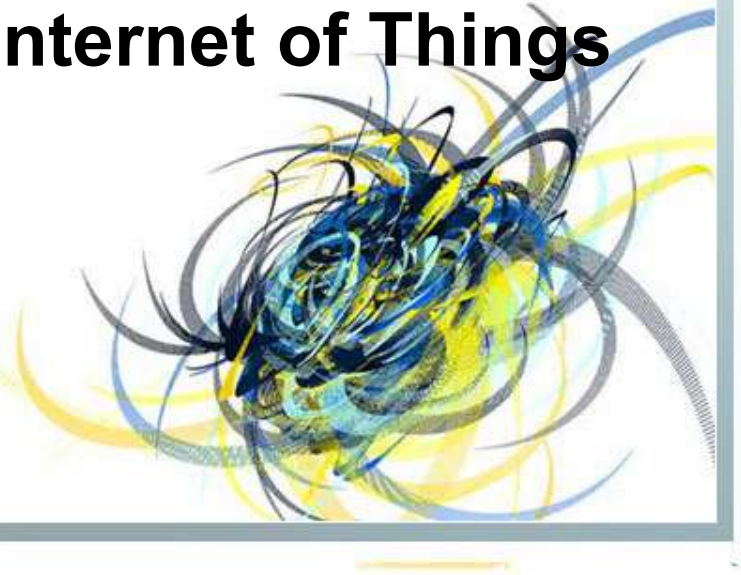


Content

13.1 Data mining overview

13.2 Basic types and algorithms of data mining*

13.3 Intelligent Decision and Internet of Things



What are the basic types and algorithms of data mining?

13.2 Basic types and algorithms of data mining

Basic types of data mining

Association Analysis

Clustering Analysis

Descriptive mining tasks: characterizing the general characteristics of data in a database

Outlier Analysis

Classification and prediction

Evolution Analysis

Predictive mining tasks: making inferences and predictions on current data

Association Analysis

The goal of association analysis is to discover frequent patterns, or association rules, from a given piece of data

Association rules are usually expressed in terms of X and Y, indicating that "records (tuples) in the database that satisfy condition X may also satisfy condition Y."

Take the sales record of an electrical appliance store as an example:

Age(Custom,"20-29") ∧ Income(Custom,"3000-5000") ⇒ Buy(Custom," laptop")

[Support=4%, Confidence=65%]

Meaning: 4% (support) of the customers are 20 to 29 years old and have a monthly income of 3,000 to 5,000 yuan, and 65% (confidence) of such customers have bought a laptop computer

Association Analysis

Mining association rules requires as much confidence and support as possible

Basic concept

Itemset: a set of data items that satisfy several conditions. If the number of conditions is k, it is called k-itemset

- ✓ Item sets that satisfy age (customer, "20~29") are 1-item sets
- ✓ Item sets that satisfy age (customer, "20~29") income (customer, "3000~5000") are 2-item sets

Calculation steps

- First, find the item set with sufficient support, namely frequent item set
- The association rules are then composed of frequent item sets and the confidence is calculated

Association Analysis

How to find frequent item sets?

- **A priori algorithm**

  Basic idea: calculate (k+1)- item set with k- item set

  - ✓ First, the frequent 1- item set is evaluated
  - ✓ Then according to the two frequent k-item sets $\{p_1, p_2...p_k\}$, $\{q_1, q_2..., q_k\}$, calculate (k + 1) - frequent itemset, which $p_i = q_i$, 1 < = i < = k - 1, and the (k + 1) - item sets $\{p_1, p_2,...p_k, q_k\}$
  - ✓ Finally, determine whether the (k+1)- item set is frequent or not

  Disadvantages: A large number of candidate sets can be generated and the database needs to be scanned repeatedly

- **FP - Growth algorithm**

  The storage space needed to compute frequent item sets is reduced by saving item sets in a tree structure

Association Analysis

How do I construct association rules from frequent itemsets and calculate confidence

Association rule A $\Rightarrow$ Confidence B

$$\text{Confidence}(A \Rightarrow B) = \text{P}(B \mid A) = \frac{\text{count}(A \text{ AND } B)}{\text{count}(A)}$$

Where count(A AND B) is the number of data items satisfying condition A AND B, and count(A) is the number of data items satisfying condition A

Calculation steps

- For each frequent itemset S, all non-empty subsets of S are computed
- For every nonempty subset F of S, if $\frac{\text{count}(S)}{\text{count}(F)}$ is larger than the given confidence threshold, an association rule $F \Rightarrow (S - F)$ is obtained.

Classification and prediction

The goal of classification and prediction is to find models or functions that describe and distinguish between different data classes or concepts so that models can be used to predict data classes or mark unknown objects

The obtained **classification model** can describe the output in various forms

- ✓ Classification rules
- ✓ Decision tree
- ✓ A mathematical formula
- ✓ The neural network
- ✓ ...

The difference between classification and prediction: classification usually refers to which category the predicted data object belongs to, while when the predicted value is numerical data, it is usually called prediction

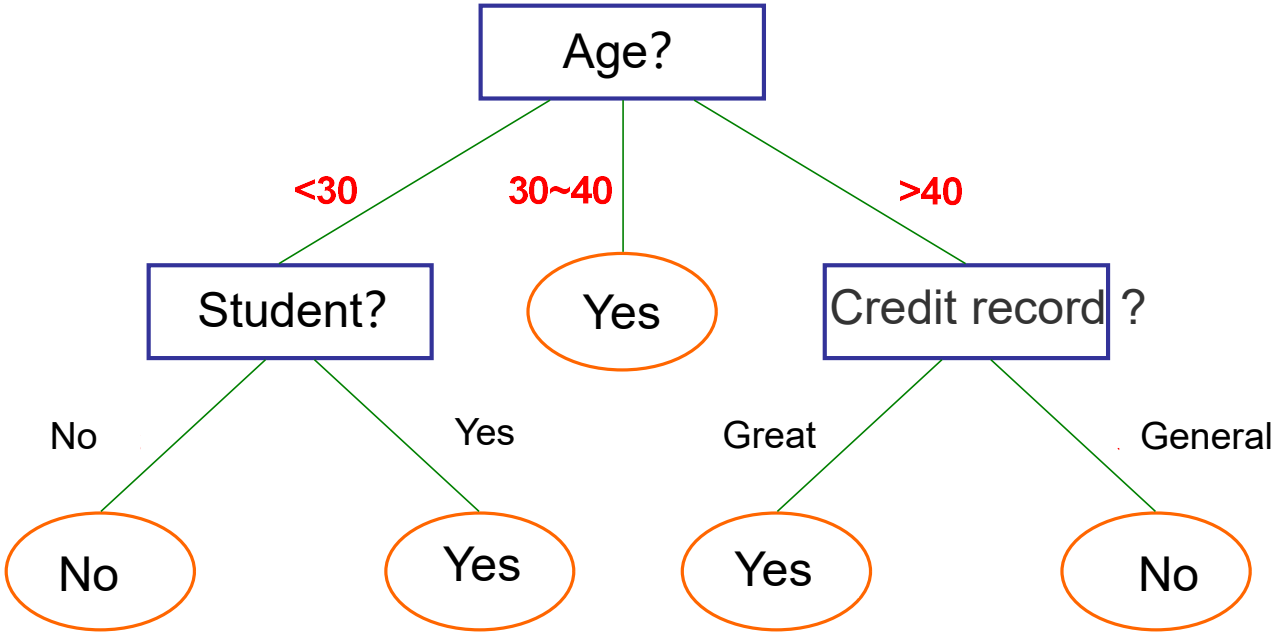
Classification and prediction

Taking the **decision tree** method as an example, the basic steps and result representation of the classification are briefly introduced.

Example of problem: suppose the store needs to mail new product information and promotional information to potential customers. The customer attributes described in the customer database include name, age, income, occupation, and credit history.
We can divide customers into two categories according to whether they will buy computers in the shopping mall, and only mail promotional materials to those customers who will buy computers, so as to reduce costs.

Classification and prediction

A decision tree used to predict whether a customer is likely to purchase a computer, where each non-leaf node represents a test on an attribute, and each leaf node represents the predicted result.



Classification and prediction

How to construct the above decision tree?

Basic concept: expected information of number a of n customers who have purchased a computer

$$I(a,n-a) = -\frac{a}{n}\log\frac{a}{n} - \frac{n-a}{n}\log\frac{n-a}{n}$$

When building a tree node, select the appropriate decision attribute to Maximizing the expected information gain should

- The amount of information gain on an attribute reflects the ability of the attribute to distinguish between given data.

10 customer records, 6 purchased computers and 4 did not. Three of the 10 customers are students, and two of them buy computers, while four of the non-student customers buy computers. Before choosing to distinguish the attribute, the expectation of data information $E = I(6,4) = 0.673$, after distinguishing with professional, the expectations for information $E' = \frac{3}{10}I(2,1) + \frac{7}{10}I(4,3) = 0.669$, choose a career as a distinguish between attribute of information gain $E - E' = 0.004$

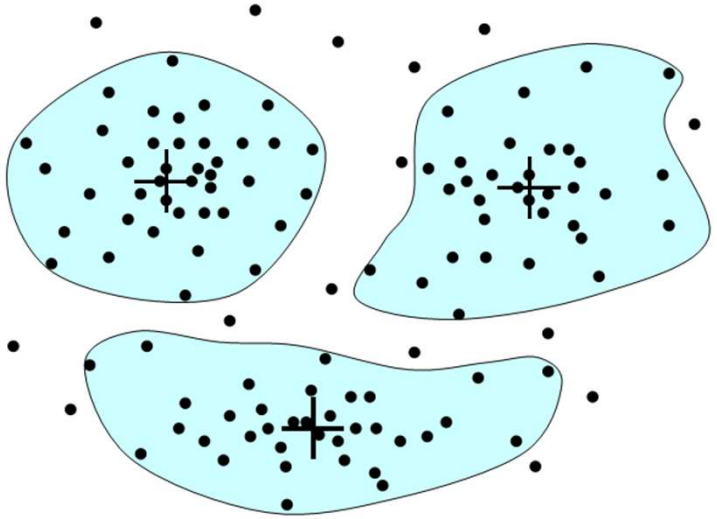
Cluster analysis

The purpose of clustering is to divide data objects into multiple classes or clusters, and the objects in the same cluster have high similarity, while the objects in different clusters differ greatly

The difference between clustering and classification: the class to be divided is unknown in advance

Application of cluster analysis

Cluster analysis of customer residence can help shopping malls adopt targeted marketing strategies for corresponding customer groups




Cluster analysis

Cluster analysis method

- **Partition method:** the number of clusters k is required in advance. An initial partition is created, and then the partition is improved by iterating over the center point of the partition. Typical algorithms include k-means algorithm and k-medoids algorithm
- **Hierarchical method:** a hierarchical recursive merge or split of a given data set, so it can be divided into merge or split methods. The merge method starts with each object as a separate class and continues to merge close classes until the termination condition is reached. The split method first places all the data objects in a class, then iterates and determines whether the current class can be further split until the termination condition is reached
- **Density-based method:** as long as the data density of a certain region exceeds the threshold, the data of that region will be clustered. Its advantage lies in the anti-interference ability of noise data and the ability to find arbitrary shape clustering

Cluster analysis

Cluster analysis method

- **Grid-based method:** the object space is quantified into regular cells to form a grid-like structure. In clustering, each cell is treated as a data. The advantage is that the processing time is fast because it is not related to the number of data objects, but only to the number of cells in the quantized space
- **Model-based approach:** if the data is known in advance to be generated according to the potential probability distribution, the model-based approach can build relevant data models for each cluster, and then find the best match of the data to a given model. There are two main categories: statistical methods and neural network methods


Outlier analysis

Outlier: a set of data objects that exist in a data set that are inconsistent with the properties or models of the vast majority of the remaining data

The meaning of finding outliers

- Discover credit card fraud. By detecting shopping location, commodity type or shopping amount and frequency, records that are different from most normal consumption can be found, which may be fraudulent use of credit CARDS
- Prevent online fraud. In online sales, fraudsters often pretend to be merchants and sell products with a price much lower than the normal price. Such behavior can also be found through outlier analysis


Outlier analysis

How to find outliers

- **Based on statistics:** a distribution or probability model of data (for example, a normal distribution) is known in advance, and outliers are determined according to the inconsistency test between data points and the model
- **Distance-based approach:** data objects that don't have enough neighbors are treated as outliers, where neighbors are defined based on the distance from a given object. The existing range-based outlier detection algorithms are divided into index-based algorithms, nested loop algorithms and cell-based algorithms, all of which aim to reduce computation and I/O overhead
- **Offset based methods:** do not use statistical tests or distance-based measures to determine the exception object. Instead, it identifies outliers by examining a set of key characteristics of the data object. Data objects that deviate from a given feature description are considered outliers

Evolution analysis

The purpose of evolutionary analysis is to mine the changing rules and trends of data objects that change with time, and to model them, so as to provide references for relevant decisions

Applications of evolutionary analysis

- ✓ By analyzing the evolution of stocks, we can get the changing rules of the whole stock market and specific companies, which can help investors to make decisions
- ✓ The evolutionary analysis of ecology and climate can know the impact of human activities on nature and provide an important basis for environmental protection
- ✓ ...

Modeling methods: in addition to association analysis and classification analysis, it also includes time-related data analysis methods, including trend analysis, similar search, sequential pattern mining and periodic analysis

Evolution analysis

Time-dependent data analysis

- **Trend analysis:** the common method to determine the trend is to calculate the average value of the n-order change of the data, or use the least square method to smooth the data change curve
- **Similar search:** similar search is used to find the data sequence closest to a given sequence
- **Sequential pattern mining:** mining patterns with high frequency of occurrence of relative time or other dimensional attributes
- **Periodic analysis:** mining periodic patterns or association rules, such as "if the company leaves work more than half an hour later than usual every Saturday, the number of people who choose to take a taxi home will increase by about 20%".

Content

13.1 Data mining overview
13.2 Basic types and algorithms of data mining*
13.3 Intelligent Decision and Internet of Things

Data mining has a wide range of requirements under the background of Internet of things

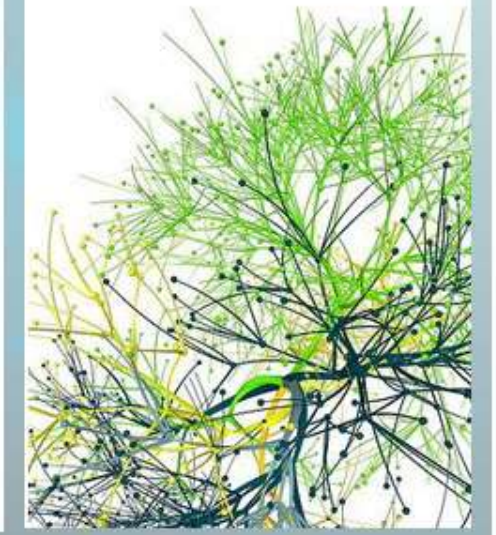
13.3 Intelligent Decision and Internet of Things

The demand of data mining technology in Internet of things

- ✓ Precision agriculture
- ✓ marketing
- ✓ Smart home
- ✓ Financial security
- ✓ Product manufacturing and quality control
- ✓ Analysis of Internet user behavior

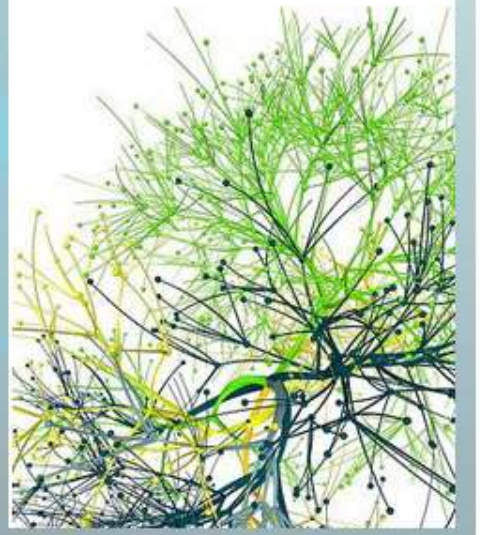
Precision agriculture

- Soil properties and environmental conditions are monitored by sensors implanted in the soil or exposed to air.
- The data is transmitted to the remote control center through the Internet of things, which can check the current growth environment and change trend of crops in time, and determine the production target of crops.
- Through the method of data mining, we can know: how do environmental temperature, humidity, soil parameters and other factors affect crop yield, and how to adjust them to maximize crop yield



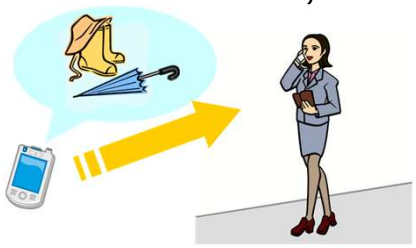
Marketing

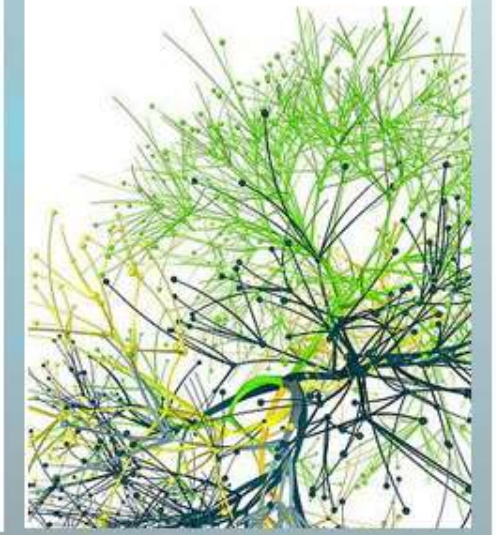
- Through the analysis of user data, information about customer's shopping orientation and interest can be obtained by using data mining technology, so as to provide basis for business decision
- Database Marketing
  - ✓ Potential customers are selected to sell to through interactive queries, data segmentation, and model prediction
  - ✓ Predict which sales channels and incentives users are most likely to be impressed with
- Basket Analysis
  - ✓ Analysis of market sales data, such as POS databases, reveals patterns of customer purchase behavior



Smart home

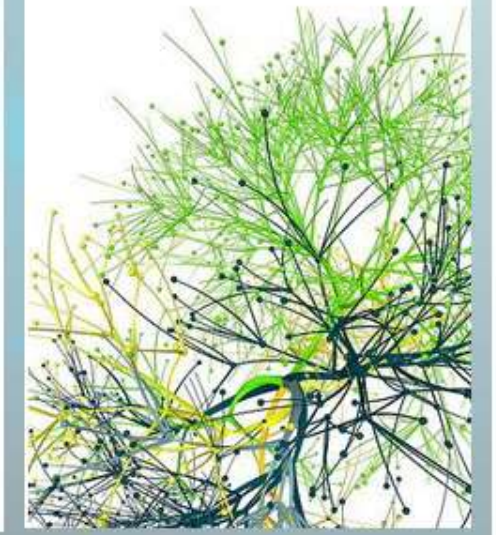
- Take weather information acquisition as an example. On the one hand, smart devices pay attention to weather information at any time and send an alarm to remind against rainy days. On the other hand, some other intelligent terminals will track the owner's whereabouts at any time, and predict his whereabouts by data mining methods based on his historical behavior characteristics
- Once the owner is expected to go out, the corresponding intelligent terminal reminds him not to forget his umbrella when appropriate. For example, smart devices installed on the door will alert the owner if he is at the door or, if he is in the car, the on-board computer





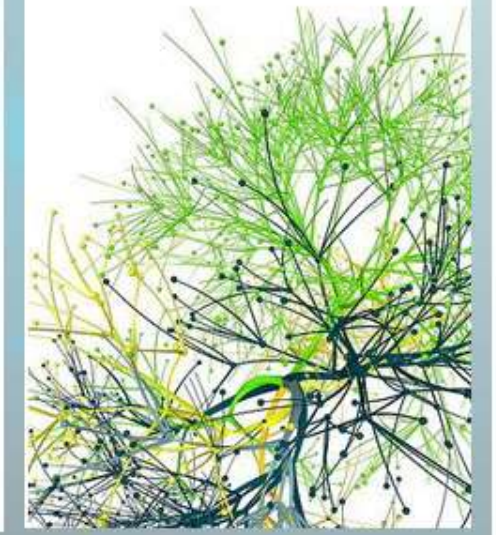
Financial security

- Due to the high risk of financial investment, it is necessary to analyze the data of various investment directions to choose the best investment direction when making investment decisions. Data mining can find the relationship between data objects through the processing of existing data, and then use the learned pattern to make reasonable prediction
- Financial fraud identification is mainly to obtain some characteristics of fraud behaviors by analyzing data and patterns of normal behaviors and fraud behaviors, so that when a certain business record conforms to such characteristics, the identification system can warn decision makers



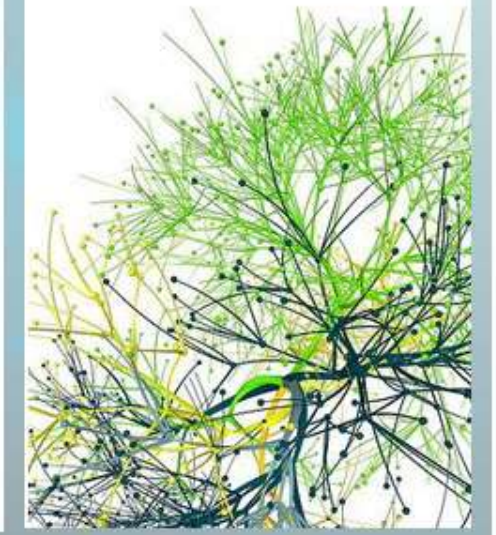
Product manufacturing and quality monitoring

- With the development of science and technology, the manufacturing industry is not a simple manual labor, but the integration of a variety of advanced technology flow. In the manufacturing process of products, there is often a large amount of data, such as various processing conditions or control parameters of products (such as time, temperature, etc.).These data collected by various monitoring instruments reflect the state of each production link and play an important role in the smooth production.
- By analyzing the data through data mining, the relationship between product quality and these parameters can be obtained, so that highly targeted Suggestions can be obtained to improve product quality, and it is possible to find new and more efficient and economical control mode, bringing rich returns for manufacturers



Internet user behavior analysis

- With the proliferation of Internet users in China, the analysis of users' behavior on the Internet has gradually attracted attention. For example, users often have to jump from one web page to another through HTTP links while surfing the web
- Access to the Internet user access model brings many benefits. First, it can assist in improving the performance of distributed network systems, such as providing fast and effective access between highly relevant sites. Second, it can help organize and design web pages better, and help improve marketing strategies (such as placing advertisements on appropriate web pages) to better attract customers' attention



Conclusion

Review

This chapter introduces the basic process of data mining, focusing on five typical data mining algorithms and steps. Finally, the extensive application of data mining technology under the background of Internet of things is discussed.

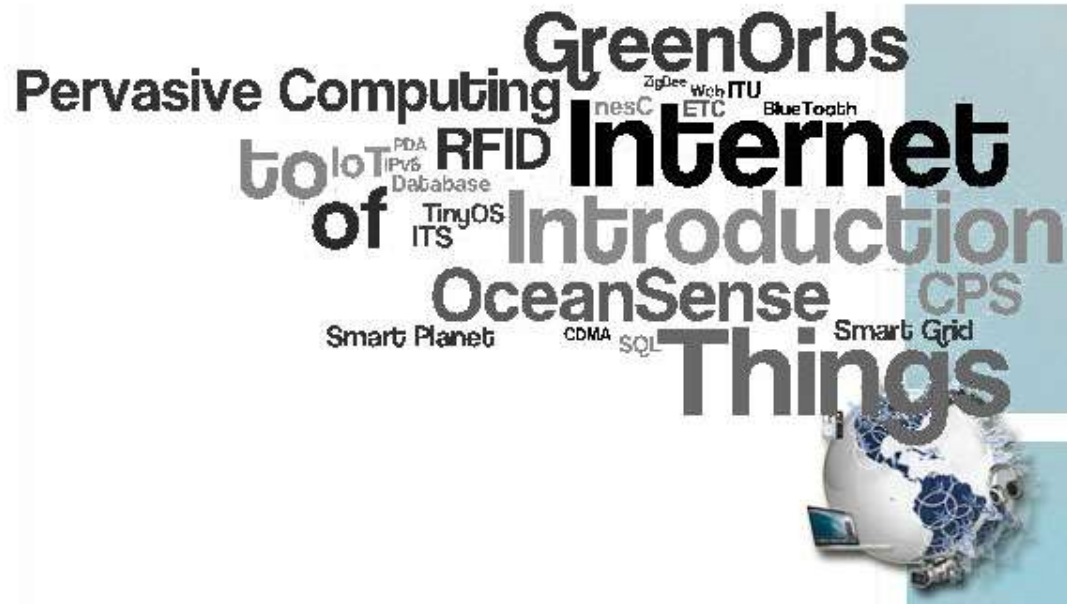
Key Points

- Understand the concepts and characteristics of data mining (iterative, human-computer interaction).
- Familiar with data mining process (data preprocessing, mining knowledge, knowledge evaluation and representation).
- Understand the related concepts of association analysis: association rules (support/confidence), Apriori algorithm, frequent item set.
- Understand the concepts of classification and prediction: decision tree, expected information, and information gain.

Conclusion

Key Points

- Understand the difference between cluster analysis and classification, and understand the k-means algorithm.
- Understand three methods of outlier analysis (based on statistics, distance offset).
- Understand the basic concepts of evolutionary analysis.
- Illustrate the extensive application of data mining technology in the Internet of things.



Thank you!

Internet of Things