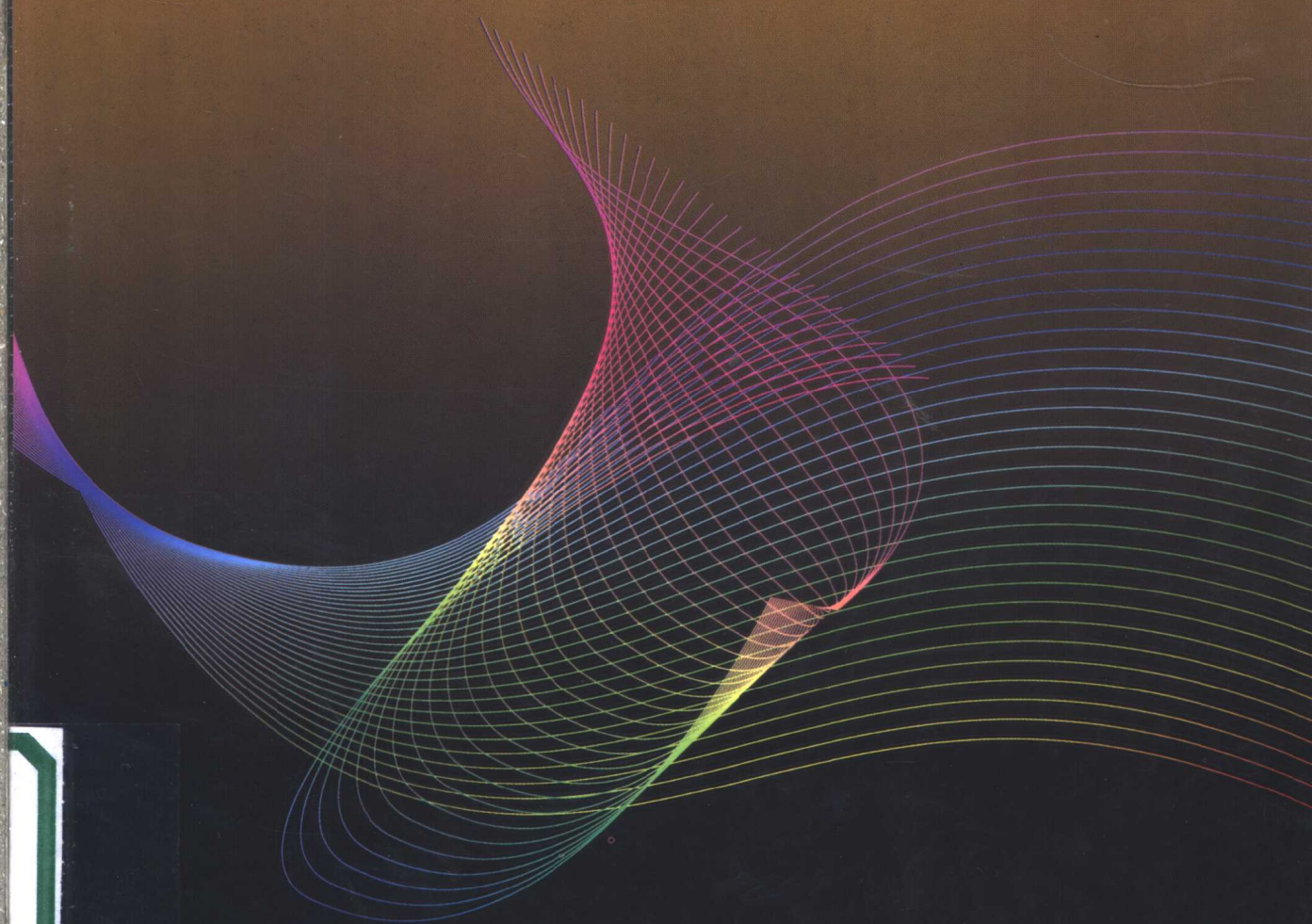
数据仓库与数据挖掘

陈文伟　黄金才　编著



人民邮电出版社
POSTS & TELECOM PRESS

数据仓库与数据挖掘

陈文伟　黄金才　编著

人民邮电出版社

图书在版编目(CIP)数据

数据仓库与数据挖掘/陈文伟,黄金才编著. —北京:人民邮电出版社,2004.1
ISBN 7-115-11902-3

Ⅰ.数... Ⅱ.①陈...②黄... Ⅲ.①数据库系统②数据采集 Ⅳ①TP311.13②TP274

中国版本图书馆 CIP 数据核字(2003)第 094315 号

内 容 提 要

数据仓库(DW)与数据挖掘(DM)是 20 世纪 90 年代中期兴起的新技术。数据仓库用于决策分析,数据挖掘用于从数据库中发现知识。数据仓库和数据挖掘的结合为决策支持系统(DSS)开辟了新方向,它们也是商业智能(BI)的主要技术。

本书主要介绍数据仓库系统、数据仓库的数据获取与管理、数据仓库的设计和开发、联机分析处理(OLAP)、数据挖掘与文本挖掘、决策树方法、粗糙集方法与关联规则挖掘、公式发现、神经网络与遗传算法、基于案例推理、决策支持系统与商业智能等内容。本书包含了作者多年来在数据仓库与数据挖掘中的研究成果。

本书可作大学计算机专业、管理科学与工程专业、系统工程专业等高年级本科生与研究生课程的教材,也可以作有关学科科技人员的参考书。

数据仓库与数据挖掘

◆ 编　　著　陈文伟　黄金才
责任编辑　邹文波

◆ 人民邮电出版社出版发行　北京市崇文区夕照寺街 14 号
邮编　100061　电子函件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
读者热线　010-67129260
北京汉魂图文设计有限公司制作
北京隆昌伟业印刷有限公司印刷
新华书店总店北京发行所经销

◆ 开本:787×1092　1/16
印张:16.5
字数:395 千字　　2004 年 1 月第 1 版
印数:1-4 000 册　　2004 年 1 月北京第 1 次印刷

ISBN 7-115-11902-3/TP・3734

定价:24.00 元

前　言

数据仓库(Data Warehouse,DW)和数据挖掘(Data Mining,DM)是20世纪90年代中期兴起的两项决策支持新技术。到了20世纪90年代末期及21世纪初,国外已经形成了研究热潮。目前我国已经跟上了国际的步伐,对数据仓库与数据挖掘技术进行了较深入的研究。

数据仓库和数据挖掘是两项不同的技术。数据仓库是区别于数据库的一种新的数据存储形式,它将数据库中的数据按决策需求(主题)进行重新组织,以多维空间结构形式存储数据。数据仓库的数据量很大,具有GB级到TB级的数据量。而一般的数据库是以二维平面结构形式存储数据,数据量一般为MB级别。

数据挖掘是从数据库中发现知识(Knowledge Discovery in Database,KDD)的核心技术,它是从人工智能的机器学习(Machine Learning,ML)中发展起来的。机器学习是让计算机通过模拟人的学习方法获取知识。机器学习中的大量学习方法已经引入到数据挖掘中。

虽然数据仓库和数据挖掘是两项不同的技术,但是它们又有共同之处,两者都是在数据库的基础上发展起来的,它们都是决策支持新技术。数据仓库利用综合数据得到宏观信息,利用历史数据进行预测;而数据挖掘是从数据库中挖掘知识,也用于决策分析。虽然数据仓库和数据挖掘支持决策分析的方式不同,但是它们完全可以结合起来,提高决策分析的能力。大量的数据仓库已经把数据挖掘技术作为它的前端分析工具,以提高数据仓库的决策分析能力。

数据仓库、数据挖掘和联机分析处理(On Line Analytical Processing,OLAP)结合起来被认为是新决策支持系统,这是以数据驱动的决策支持系统。而传统决策支持系统(Decision Support System,DSS)是以模型驱动的决策支持系统,它是由模型库系统、知识库系统、数据库系统和人机交互系统组成的。新决策支持系统并不能代替传统的决策支持系统,他们是相互补充的。新决策支持系统与传统决策支持系统结合起来形成的综合决策支持系统,将是决策支持系统发展的新方向。

数据仓库、数据挖掘、联机分析处理等结合起来也被称为商业智能(Business Intelligence,BI)。商业智能是一种新的智能技术,它区别于人工智能(Artificial Intelligence,AI)和计算智能(Computational Intelligence,CI)。人工智能采用的技术是符号推理,符号推理过程形成了概念的推理链;计算智能采用的技术是计算推理,它模拟人和生物的模糊推理、神经网络计算和遗传进化过程;商业智能是从数据仓库和数据挖掘中获取信息和知识,对变化的商业环境提供决策支持。商业智能是目前企业界正在大力推广的知识管理(Knowledge Manage,KM)的基础。

本书在数据仓库方面介绍数据仓库概念,数据仓库系统,数据仓库的数据获取与管理,数据仓库设计、开发与应用,联机分析处理;在数据挖掘方面介绍数据挖掘概念,文本数据挖掘与Web挖掘,决策树方法,粗糙集方法与关联规则挖掘,公式发现,神经网络与遗传算法,基于案例推理。最后介绍决策支持系统与商业智能。

我们在数据仓库与数据挖掘方面的研究得到了国家自然科学基金项目的资助。

欢迎读者提出宝贵的意见，进行切磋，共同推动我国在数据仓库和数据挖掘方面的发展。

编　者

目　录

第1章　数据仓库与数据挖掘概述

1.1　数据仓库概念

数据仓库(Data Warehouse)的概念是由 W. H. Inmon 在 1992 年出版的《建立数据仓库》(Building the Data Warehouse)一书中提出的。数据仓库是以关系数据库、并行处理和分布式技术为基础的信息新技术。

从目前的形势看，数据仓库技术已紧跟 Internet 而上，成为信息社会中获得企业竞争优势的又一关键技术。

1.1.1　数据仓库的兴起

1. 数据仓库的定义

(1) W. H. Inmon 对数据仓库的定义

数据仓库是面向主题的、集成的、稳定的、不同时间的数据集合，用于支持经营管理中的决策制定过程。

(2) SAS 软件研究所的观点

数据仓库是一种管理技术，旨在通过通畅、合理、全面的信息管理，达到有效的决策支持。

从数据仓库的定义可以看出，数据仓库是明确为决策支持服务的，而数据库是为事务处理服务。

2. 从数据库到数据仓库

由数据库发展到数据仓库的主要原因如下。

(1) 数据太多，信息贫乏(Data Rich，Information Poor)

随着数据库技术的发展，企事业单位建立了大量的数据库，数据越来越多，而辅助决策信息却很贫乏，如何将大量的数据转化为辅助决策信息成了研究热点。

(2) 异构环境数据的转换和共享

由于各类数据库产品的增加，异构环境的数据也随之增加，如何实现这些异构环境数据的转换和共享也成了研究热点。

(3) 利用数据进行事务处理转变为利用数据支持决策

数据库用于事务处理，若要达到辅助决策，则需要更多的数据。例如，如何利用历史数据的分析来进行预测。对大量数据的综合得到宏观信息等均需要大量的数据。

数据仓库概念提出后，在不到几年的时间内就得到了迅速的发展。数据仓库产品也不断出现并陆续进入市场。

3. DB 数据和 DW 数据区别

传统数据库用于事务处理，也叫操作型处理，是指对数据库联机进行日常操作，即对一个或一组记录的查询和修改，主要是为企业特定的应用服务的。用户关心的是响应时间，数据的安全性和完整性。数据仓库用于决策分析，也称分析型处理，它是建立决策支持系统(DSS)的基础。

例如，银行的用户有储蓄、贷款和信用卡，这些数据是存放在不同业务处彼此独立的数据库中。现在，有了数据仓库，它把这些业务数据库集中起来，建立起对用户的整体分析，决定是否继续对用户进行贷款或发信用卡。

操作型数据(DB 数据)与分析型数据(DW 数据)之间的差别如表 1.1 所示。

表 1.1　　DB 数据和 DW 数据的对比表

DB 数据	DW 数据
细节的	综合或提炼的
在存取时准确的	代表过去的数据
可更新的	不更新
操作需求事先可知道	操作需求事先不知道
事务驱动	分析驱动
面向应用	面向分析
一次操作数据量小	一次操作数据量大
支持日常操作	支持决策需求

1.1.2　数据仓库的特点

从数据仓库的定义中可以看出数据仓库的特点如下。

1. 数据仓库是面向主题的

主题是数据归类的标准，每一个主题基本对应一个宏观的分析领域。

例如，保险公司数据仓库的主题为：客户、政策、保险金和索赔等。

基于应用的数据库组织则完全不同，它的数据只是为处理具体应用而组织在一起的。保险公司按应用组织的数据库有：汽车保险、生命保险、健康保险和伤亡保险等。

2. 数据仓库是集成的

数据进入数据仓库之前，必须经过加工与集成。对不同来源的数据进行数据结构统一和编码。统一原始数据中的所有矛盾之处，如字段的同名异义，异名同义，单位不统一，字长不一致等。总之，将原始数据结构做一个从面向应用到面向主题的大转变。

3. 数据仓库是稳定的

数据仓库中包括了大量的历史数据。数据经集成进入数据仓库后是极少或根本不更新的。

4. 数据仓库是随时间增长的

数据仓库内的数据时限为 5～10 年，故数据的键码包含时间项，需标明数据的历史时期，这有助于 DSS 进行时间趋势分析。

而数据库只包含当前数据，即存储当前时间的正确的有效数据。

5. 数据仓库中的数据量很大

通常的数据仓库数据量为 10GB 级，相当于一般数据库 100MB 的 100 倍，大型数据仓库是一个 TB(1000GB)级数据量。

数据仓库中数据的比重为索引和综合数据占 2/3，原始数据占 1/3。

6. 数据仓库软硬件要求较高

(1) 需要一个巨大的硬件平台。

(2) 需要一个并行的数据库系统。

1.1.3　数据集市

1. 数据集市(Data Marts)的产生

数据仓库的工作范围和成本常常是巨大的。信息技术部门必须对所有的用户并以全企业的眼光对待任何一次决策分析，这就形成了代价很高的、时间较长的大项目。

这样提供更紧密集成的、拥有完整图形接口并且价格吸引人的工具——数据集市，就应运产生。

目前，全世界对数据仓库总投资的一半以上集中在数据集市上。

2. 数据集市概念

数据集市(Data Marts)是一种更小、更集中的数据仓库，是为公司提供分析商业数据的一条廉价途径。

数据集市是指具有特定应用的数据仓库，主要针对某个具有战略意义的应用或者具体部门级的应用，支持用户利用已有的数据获得重要的竞争优势或者找到进入新市场的具体解决方案。

3. 数据集市与数据仓库的关系

数据集市不等于数据仓库，多个数据集市简单合并起来并不能成为数据仓库。主要由于以下几点原因。

(1) 各数据集市之间对详细数据和历史数据的存储存在大量冗余。

(2) 同一个问题在不同的数据集市的查询结果可能不一致，甚至相互矛盾。

(3) 各数据集市之间以及与源数据库系统之间难以管理。

4. 数据集市的特性

(1) 规模很小；

(2) 特定的应用；

(3) 面向部门；

(4) 由业务部门定义、设计和开发；

(5) 由业务部门管理和维护；

(6) 快速实现；

(7) 购买较便宜；

(8) 投资快速回收；

(9) 工具集的紧密集成；

(10) 更详细的、预先存在的数据仓库的摘要子集；

(11) 可升级到完整的数据仓库。

1.2 知识发现和数据挖掘概念

从数据库中发现知识(Knowledge Discovery in Database,KDD)是20世纪80年代末开始的,KDD一词是在1989年8月于美国底特律市召开的第一届KDD国际学术会议上正式形成的。KDD研究的问题有:(1) 定性知识和定量知识的发现;(2) 知识发现方法;(3) 知识发现的应用等。

1995年在加拿大召开了第一届知识发现和数据挖掘(Data Mining,DM)国际学术会议。由于把数据库中的"数据"形象地比喻成矿床,"数据挖掘"一词很快流传开来。

数据挖掘是知识发现中的核心工作,主要研究发现知识的各种方法和技术。

1.2.1 知识发现和数据挖掘的定义

知识发现(KDD)被认为是从数据中发现有用知识的整个过程。数据挖掘被认为是KDD过程中的一个特定步骤,它是用专门算法从数据中抽取模式(Pattern)。

KDD过程定义如下(Fayyad,Piatetsky-Shapiror和Smyth,1996)。

KDD是从数据集中识别出有效的、新颖的、潜在有用的,以及最终可理解的模式的高级处理过程。

其中,数据集:数据库记录的集合F;模式:用语言L表示的F中部分记录的表达式E,它所描述的数据集是集合F的一个子集F_E,我们称表达式E为模式;有效、新颖、潜在有用、可理解:表示发现的模式应该是新的,将来有实用价值,能被用户所理解。

KDD过程图如图1.1所示。

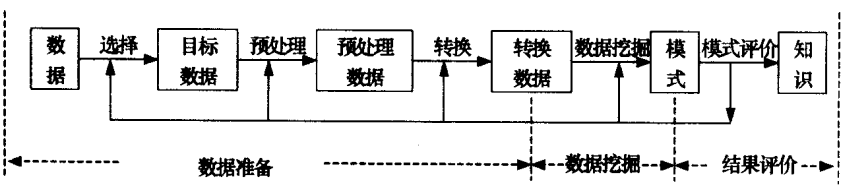


图1.1 KDD过程图

KDD过程可以概括为三部分:数据准备(Data Preparation),数据挖掘(Data Mining)及结果的解释和评估(Interpretation & Evaluation)。

1. 数据准备

数据准备又可分为3个子步骤:数据选取(Data Selection)、数据预处理(Data Preprocessing)和数据变换(Data Transformation)。

数据选取的目的是确定发现任务的操作对象,即目标数据(Target Data),是根据用户的需要从原始数据库中抽取的一组数据。数据预处理一般包括消除噪声、推导计算缺值数据、消除重复记录、完成数据类型转换(如把连续值型数据转换为离散型数据,以便于符号归纳;或是把离散型数据转换为连续值型数据,以便于神经网络计算)等。数据变换的主要目的是削减数据维数或降维(Dimension Reduction),即从初始特征中找出真正有用的特征以减少数据挖掘时要考虑的特征或变量个数。

2. 数据挖掘

数据挖掘阶段首先要确定挖掘的任务或目的，如数据分类、聚类、关联规则发现或序列模式发现等。确定了挖掘任务后，就要决定使用什么样的挖掘算法。选择实现算法有两个考虑因素：一是不同的数据有不同的特点，因此需要用与之相关的算法来挖掘；二是用户或实际运行系统的要求，有的用户可能希望获取描述型的(Descriptive)、容易理解的知识(采用规则表示的挖掘方法显然要好于神经网络之类的方法)，而有的用户只是希望获取预测准确度尽可能高的预测型(Predictive)知识。选择了挖掘算法后，就可以实施数据挖掘操作，获取有用的模式。

3. 结果的解释和评估

数据挖掘阶段发现出来的模式，经过评估，可能存在冗余或无关的模式，这时需要将其剔除；也有可能模式不满足用户要求，这时则需要退回到发现过程前面的阶段，如重新选取数据、采用新的数据变换方法、设定新的参数值，甚至换一种挖掘算法等等。另外，KDD 由于最终是面向人类用户的，因此可能要对发现的模式进行可视化，或者把结果转换为用户易懂方式，如把分类决策树转换为“if ... then...”规则。

数据挖掘仅仅是整个过程中的一个步骤。数据挖掘质量的好坏有两个影响要素：一是所采用的数据挖掘技术的有效性，二是用于挖掘的数据的质量和数量(数据量的大小)。如果选择了错误的数据或不适当的属性，或对数据进行了不适当的转换，则挖掘的结果不会成功。

整个挖掘过程是一个不断反馈的过程。比如，用户在挖掘途中发现选择的数据不太满意，或使用的挖掘技术产生不了期望的结果。这时，用户需要重复先前的过程，甚至从头重新开始。

可视化技术在数据挖掘的各个阶段都起着重要的作用。特别是在数据准备阶段，用户可能要使用散点图、直方图等统计可视化技术来显示有关数据，以期对数据有一个初步的了解，从而为更好地选取数据打下基础。在挖掘阶段，用户则要使用与领域问题有关的可视化工具。在表示结果阶段，则可能要用到可视化技术以使得发现的知识更易于理解。

1.2.2　数据挖掘任务

数据挖掘任务有六项：关联分析、时序模式、聚类、分类、偏差检测以及预测。

1. 关联分析

关联分析是从数据库中发现知识的一类重要方法。若两个或多个数据项的取值之间重复出现且概率很高时，它就存在某种关联，可以建立起这些数据项的关联规则。

例如，买面包的顾客有 90%的人还买牛奶，这是一条关联规则。若商店中将面包和牛奶放在一起销售，将会提高它们的销量。

在大型数据库中，这种关联规则是很多的，需要进行筛选，一般用“支持度”和“可信度”两个阈值来淘汰那些无用的关联规则。

“支持度”表示该规则所代表的事例(元组)占全部事例(元组)的百分比，如买面包又买牛奶的顾客占全部顾客的百分比。

“可信度”表示该规则所代表事例占满足前提条件事例的百分比，如买面包又买牛奶的顾客占买面包顾客中的 90%，可信度为 90%。

2. 时序模式

通过时间序列搜索出重复发生概率较高的模式。这里强调时间序列的影响。例如，在所有购买了激光打印机的人中，半年后 80%的人再购买新硒鼓，20%的人用旧硒鼓装碳粉；在所有购买了彩色电视机的人中，有 60%的人再购买 VCD 产品。

在时序模式中，需要找出在某个最小时间内出现比率一直高于某一最小百分比（阈值）的规则。这些规则会随着形式的变化做适当的调整。

时序模式中，一个有重要影响的方法是“相似时序”。用“相似时序”的方法，要按时间顺序查看时间事件数据库，从中找出另一个或多个相似的时序事件。例如在零售市场上，找到另一个有相似销售的部门，在股市中找到有相似波动的股票。

3. 聚类

数据库中的数据可以划分为一系列有意义的子集，即类。在同一类别中，个体之间的距离较小，而不同类别上的个体之间的距离偏大。聚类增强了人们对客观现实的认识，即通过聚类建立宏观概念。例如，鸡、鸭、鹅等都属于家禽。

聚类方法包括统计分析方法、机器学习方法和神经网络方法等。

在统计分析方法中，聚类分析是基于距离的聚类，如欧氏距离，海明距离等。这种聚类分析方法是一种基于全局比较的聚类，它需要考察所有的个体才能决定类的划分。

在机器学习方法中，聚类是无导师的学习。在这里距离是根据概念的描述来确定的，故聚类也称概念聚类，当聚类对象动态增加时，概念聚类则称为概念形成。

在神经网络方法中，自组织神经网络方法用于聚类。如 ART 模型、Kohonen 模型等，这是一种无监督学习方法。当给定距离阈值后，各样本按阈值进行聚类。

4. 分类

分类是数据挖掘中应用的最多的任务。分类是找出一个类别的概念描述，它代表了这类数据的整体信息，即该类的内涵描述，一般用规则或决策树模式表示。该模式能把数据库中的元组影射到给定类别中的某一个。

一个类的内涵描述分为特征描述和辨别性描述。特征描述是对类中对象的共同特征的描述。辨别性描述是对两个或多个类之间的区别的描述。特征描述允许不同类中具有共同特征，而辨别性描述要求不同类不能有相同特征，一般情况下辨别性描述用的更多。

分类是利用训练样本集（已知数据库元组和类别所组成的样本）通过有关算法而求得。

建立分类决策树的方法，典型的有 ID3、C4.5 和 IBLE 等方法。建立分类规则的方法，典型的有 AQ 方法、粗集方法和遗传分类器等。

目前，分类方法的研究成果较多，判别方法的好坏，可从下述 3 个方面进行：(1) 预测准确度（对非样本数据的判别准确度）；(2) 计算复杂度（方法实现时对时间和空间的复杂度）；(3) 模式的简洁度（在同样效果的情况下，希望决策树小或规则少）。

在数据库中，往往存在噪声数据（错误数据）、缺损值和疏密不均匀等问题，他们对分类算法获取的知识将产生坏的影响。

5. 偏差检测

数据库中的数据存在着很多异常情况，从数据分析中发现这些异常情况也很重要，以引起人们对它更多的注意。

偏差包括很多有用的知识，如以下 4 类：