决策支持系统及其开发

主讲教师：唐晶磊

E-mail: tangjingle.cn

Tel:87091337(O)

5.5 知识发现与数据挖掘

5.6 数据挖掘的决策支持及应用

5.5 知识发现与数据挖掘

DW的兴起

（1）80年在美国召开了第一届国际机器学习研讨会；

（2）89年8月,美国底特律市召开的第一届KDD国际学术会议；

（3）95年,加拿大召开了第一届KDD和DM国际学术会议；

（4）我国于87年召开了第一届全国机器学习研讨会。

5.5.1 知识发现与数据挖掘概念

知识发现（Knowledge discovery in database）：从数据中发现有用知识的整个过程(KDD)。

KDD过程定义: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。

“模式”即是“知识”的雏形，需经过验证、完善(模式评价)后形成知识。

KDD过程概括：数据准备(data preparation)、数据挖掘(data mining)及结果的解释和评估(interpretation&evaluation)。

5.5.1 知识发现与数据挖掘概念

问题：所有企业都面临企业数据量巨大，而其中真正有价值的信息却很少。

解决方法：对大量的数据进行深层分析，获得有利于商业运作、提高竞争力的信息。

数据挖掘（DM）：KDD过程中的一个特定步骤，它用专门算法从数据中抽取模式（patterns）。

数据挖掘是一门交叉学科，涉及数据库技术、人工智能技术、数理统计、可视化技术、并行计算等方面。

5.5.1 知识发现与数据挖掘概念

（1）DM（技术角度）：从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、事先不知道的、但又是潜在有用的信息和知识的过程。即从数据源发现用户感兴趣的知识，知识要可接受、可以理解和运用；

5.5.1 知识发现与数据挖掘概念

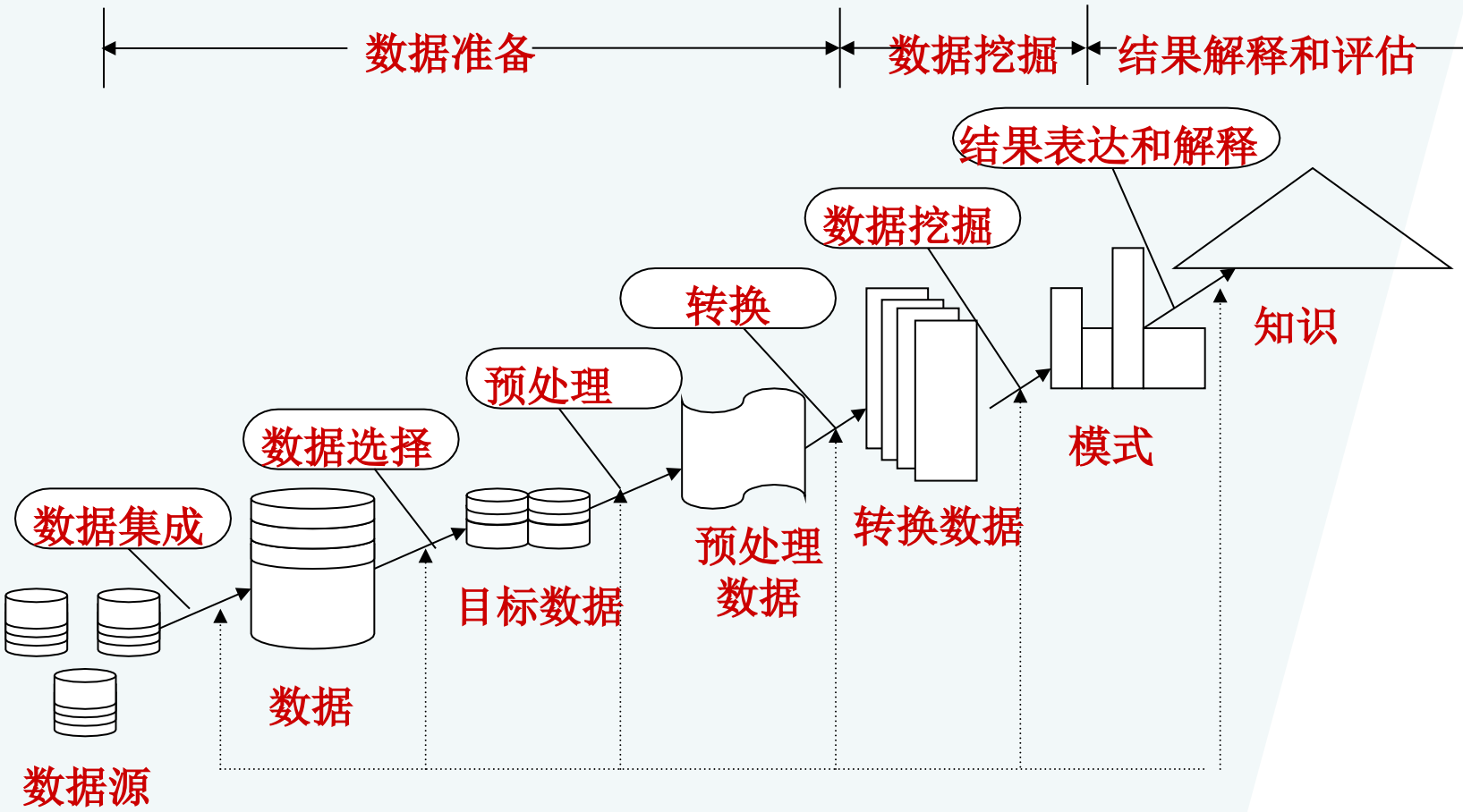
（2）（DM）商业角度：是一种新的、商业信息处理技术。对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性数据。

数据挖掘是一种深层次的数据分析方法。

5.5.1 知识发现与数据挖掘概念

（3）(DM)企业角度：按企业既定业务目标，对大量的企业数据进行探索和分析，揭示隐藏的、未知的或验证已知的规律性，并进一步将其模型化的先进有效方法。

KDD过程


数据准备
数据挖掘
结果解释和评估
结果表达和解释
数据挖掘
转换
预处理
数据选择
数据集成
知识
模式
转换数据
预处理
数据
目标数据
数据
数据源

数据准备

- 数据准备：数据选择(data selection)、数据预处理(data preprocessing)和数据转换(data transformation)。
- 数据选择：确定操作对象，即目标数据(target data)，是根据用户的需要，从原始DB中选取的一组数据。
- 数据预处理：消除噪声、处理缺值数据、消除重复记录等。
- 数据转换：完成数据类型转换，进行属性约简（从初始属性中找出真正有用的属性，删除无用属性，以减少数据挖掘时要考虑的属性个数）。

数据挖掘

- 数据挖掘

  （1）首先确定挖掘的任务或目的；

  （2）确定使用何种挖掘算法。

- 选择挖掘算法需考虑2个因素：
  - 不同数据具有不同特点，需要用与之相关的算法来挖掘；
  - 考虑用户或实际运行系统的要求。如用户可能希望获取描述性的、容易理解的知识，或者希望获取预测准确度更可能高预测型知识。

结果的解释和评估

- 结果的解释和评估（模式评价）

（1）经过评估，剔除冗余或无关的模式；

（2）不满足用户要求的模式，需回退到KDD过程的前面阶段。

（3）KDD是面向用户的，一般需对发现的模式进行可视化处理，或把结果转换为用户易懂的表示形式。

- DM质量好坏的2个影响因素：
  - （1）所采用的DM技术的有效性；
  - （2）用于挖掘的数据的质量和数量（数据量的大小）。

数据挖掘任务

❖DM任务有六项：

（1）关联分析

若两个或多个数据项的取值重复出现，且概率很高时，它就存在某种关联，则可建立起这些数据项的关联规则。

（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。

（3）聚类（通过聚类建立宏观概念）

有统计分析方法、机器学习方法、神经网络方法等。

数据挖掘任务

（4）分类：以聚类为基础，对已确定的类找出该类别的概念描述。它代表此类数据的整体信息（内涵描述）。

- 内涵描述分为特征描述和辨别性描述。
- 判别分类方法的3个标准：预测准确度、计算复杂度、模式的简洁度。

（5）偏差检测：寻找观察结果与DB中参照数据之间的差别。

（6）预测：利用历史数据找出变化规律，建立模型，并用此模型来预测未来数据的种类、特征等。

属性约简

- 属性约简常用于分类问题
  - 原则：保持数据库中分类关系不变。
  - 一般采用粗糙集(rough set)方法，也可采用信息论方法。
- 在DB的分类问题中，属性分为条件属性(C)和决策属性(D)。条件属性分为可省略属性和不可省略属性。
- 属性约简是在条件属性中，删除不影响对决策属性进行分类的多余的条件属性。
- 不可省略属性，实质上是对决策属性进行分类的核心属性。

补充：数据挖掘与传统分析方法的区别

- 传统的数据分析方法：查询、报表和联机分析等。
- 采用完全不同的工具，基于的技术差别也很大。

（1）查询和报表，告诉决策者数据库中都有什么。

（2）OLAP会进一步告诉决策者，下一步会怎么样，（假设）如果我采用这样的措施，又会怎么样。OLAP通过建立一系列的假设，来证实或推翻这些假设，以得到合理的结论。因此，OLAP本质上是演绎推理过程。

补充：数据挖掘与联机分析处理的区别

- DM在没有明确假设的前提下去挖掘信息、发现知识。
- DM所得到的信息是先前未知、有效的和可实用的。
- 数据挖掘不用于验证某个假定的模式，而是在数据库中自己寻找模型。本质是一个归纳的过程。
- DM和OLAP具有一定的互补性。
- 在利用DM出来的结论采取行动之前，利用OLAP验证一下，如果采取这样的行动，将会给公司带来什么样的影响。

5.5.2 数据挖掘方法和技术

- DM方法由人工智能、机器学习的方法发展而来。结合传统的统计分析方法、模糊数学方法以及计算科学可视化技术，以数据库为研究对象，形成了数据挖掘方法和技术。
- 数据挖掘方法和技术可以分为六大类。

5.5.2 数据挖掘方法和技术

（一）归纳学习方法

按采用的技术可分为信息论方法（决策树方法）和集合论方法。

1、信息论方法（决策树方法）

利用信息论的原理建立决策树或者决策规则树。

较有特色的方法有：

（1）ID3等方法（决策树方法）

（2）IBLE（决策规则树）方法。

2、集合论方法

（1）粗糙集（Rough Set）方法

对数据库中的条件属性集与决策属性集建立上下近似关系，对下近似集合建立确定性规则，对上近似集合建立不确定性规则（含可信度）。

（2）关联规则挖掘

在交易事务数据库中，挖掘出不同商品集的关联关系，即发现哪些商品频繁地被顾客同时购买。

（3）覆盖正例排斥反例方法

它是利用覆盖所有正例，排斥所有反例的思想来寻找规则。较典型的有AQ11方法、AQ15方法及AE5方法。

（二）仿生物技术

典型的仿生物技术方法是神经网络方法和遗传算法。

1、神经网络方法：包括：前馈式网络、反馈式网络、自组织网络等多个神经网络方法。

2、遗传算法：模拟生物进化过程的算法。它由三个基本算子组成：繁殖（选择）、交叉（重组）、变异（突变）

遗传算法起到产生优良后代的作用，经过若干代的遗传，将得到满足要求的后代（问题的解）。

（三）公式发现

在工程和科学数据库中对若干数据项（变量） 进行一定的数学运算，求得相应的数学公式。

1．物理定律发现系统BACON

BACON发现系统完成了物理学中大量定律的重新发现。

2．经验公式发现系统FDD

寻找由数据项的初等函数或复合函数组合成的经验公式。

（四）统计分析方法

利用统计学原理对总体中的样本数据进行分析，得出描述和推断该总体信息和知识的方法。

（五）模糊数学方法

利用模糊集合理论进行数据挖掘，如模糊聚类、模糊分类等。

（六）可视化技术

利用可视化技术分析数据库，找到潜在的有用信息。

5.5.3 数据挖掘的知识表示（一）

DM获取知识表示形式主要有六种：

规则、决策树、浓缩数据、网络权值、公式和案例。

1、规则

规则知识由前提条件和结论两部分组成

前提条件由多个谓词的合取（与）

和析取（或∨）组合而成。

结论为决策字段项（属性）的取值或者类别组成。

5.5.3 数据挖掘的知识表示（一）

	身高	头发	眼睛
第一类人	矮	金色	兰色
	高	红色	兰色
	高	金色	兰色
	矮	金色	灰色
第二类人	高	金色	黑色
	矮	黑色	兰色
	高	黑色	兰色
	高	黑色	灰色
	矮	金色	黑色

IF（发色=金色∨红色）∧（眼睛=兰色∨灰色）THEN 第一类人

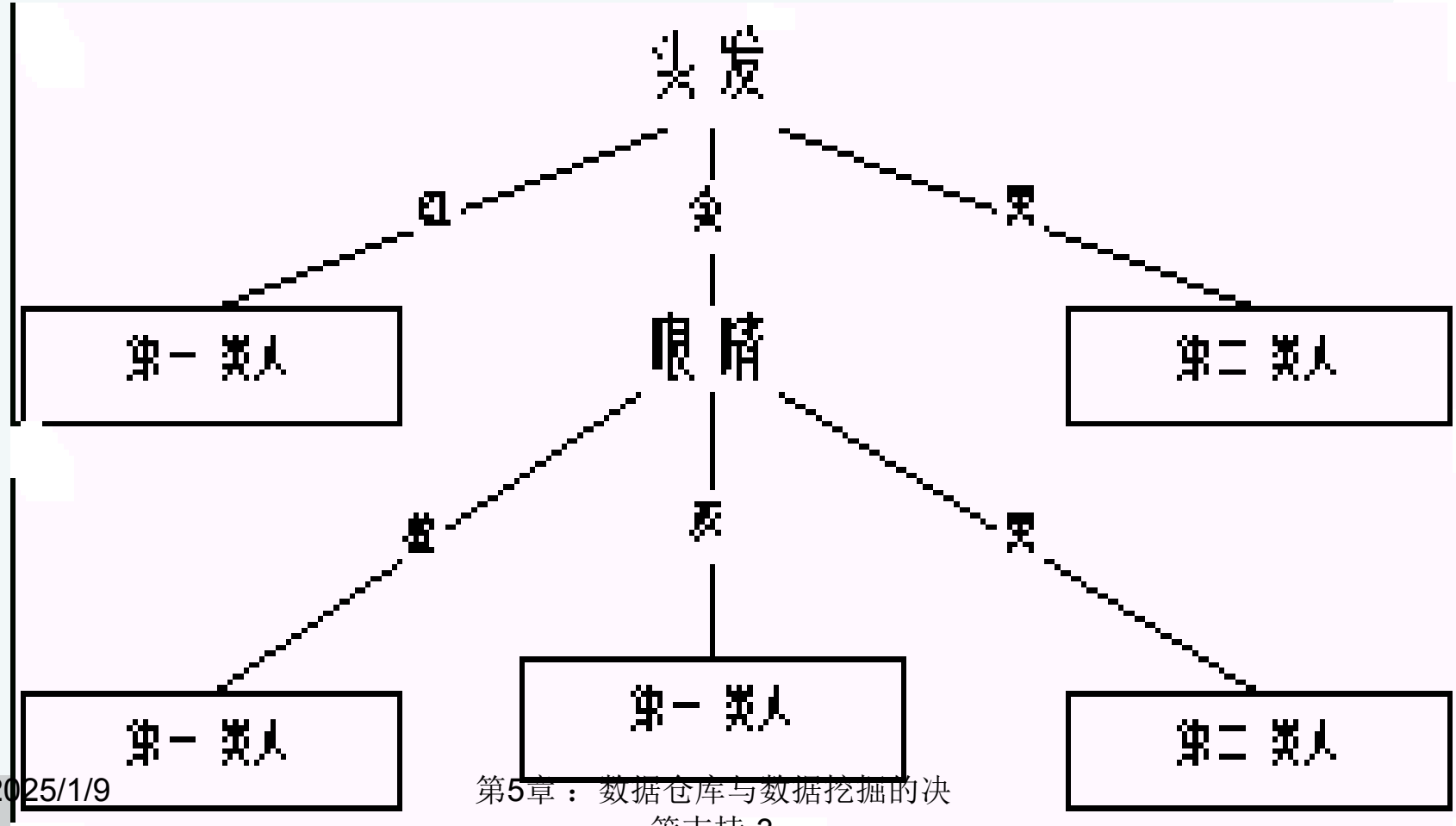
IF（发色=黑色）∨（眼睛=黑色）THEN 第二类人

即：凡是具有金色或红色的头发，并且同时具有兰色或灰色眼睛的人属于第一类人；凡是具有黑色头发或黑色眼睛的人属于第二类人。

数据挖掘的知识表示（二）

2、决策树

例如：上例的人群数据库，按ID3方法得到的决策树如下：



数据挖掘的知识表示（三）

3、知识基（浓缩数据）

例如上例的人群数据库，通过计算可得出身高是不重要的字段，删除它后，再合并相同数据元组，得到浓缩数据如下表：

	头发	眼睛
1 类人	金色	兰色
1 类人	红色	兰色
1 类人	金色	灰色
2 类人	金色	黑色
2 类人	黑色	兰色
2 类人	黑色	灰色

数据挖掘的知识表示（四）

4、网络权值

神经网络方法经过对**训练样本**的学习后，**所得到的知识**是网络**连接权值**和**结点的阈值**。

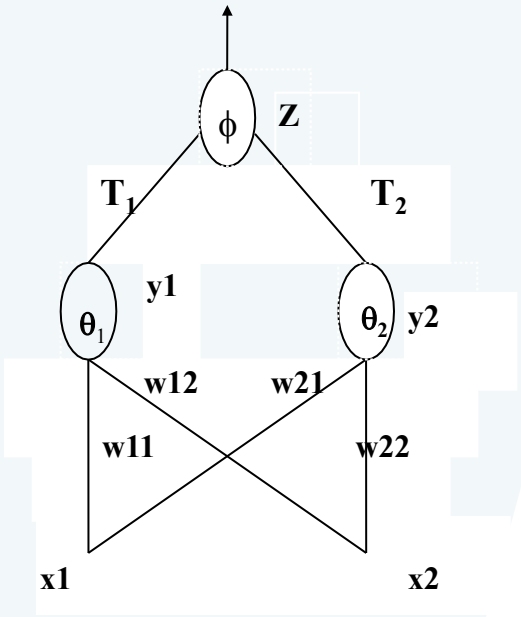
$$\begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 1.5 \end{pmatrix}$$

，

$\phi = 0.5$

$$(T_1, T_2) = (-1, 1)$$



数据挖掘的知识表示（五）

5、公式

例如，太阳系行星运动数据中包含行星运动周期（旋转一周所需时间，天），以及它与太阳的距离（围绕太阳旋转的椭圆轨道的长半轴，百万公里），数据如下表：

	水星	金星	地球	火星	木星	土星
周期 P	88	225	365	687	4343. 5	10767.5
距离 d	58	108	149	228	778	1430

通过物理定律发现系统BACON和经验公式发现系统FDD，都可得到开普勒第三定律：$d^3/p^2=25$

5.6 数据挖掘的决策支持及应用

5.6.1 决策树及其应用

1、决策树概念：

用样本的属性作为结点，用属性的取值作为分支的树结构。利用信息论原理对大量样本的属性进行分析和归纳。

决策树的根结点是所有样本中信息量最大的属性。

中间结点是以该结点为根的子树，所包含的样本子集中信息量最大的属性。

叶结点是样本的类别值。