第1章：绪论

蒋良孝 |  中国地质大学（武汉）

机器学习与数据挖掘团队

ljiang@cug.edu.cn

http://grzy.cug.edu.cn/jlx/

本章内容

一、机器学习的定义

二、与数据挖掘的区别与联系

三、本课程的授课思路与内容安排

四、教材及参考书

一、机器学习的定义

- 机器学习是人工智能的核心研究领域之一， 其研究动机是为了让计算机系统具有人的学习能力以便实现人工智能。
- 目前被广泛采用的机器学习的定义是“利用经验来改善计算机系统自身的性能”。由于“经验”在计算机系统中主要是以数据的形式存在的，因此机器学习需要运用机器学习技术对数据进行分析，这就使得它逐渐成为智能数据分析技术的创新源之一，并且为此而受到越来越多的关注。

一、机器学习的定义

机器学习, Tom M. Mitchell著, 曾华军等译, 1997年。

- 第1章：引言
- 第2章：概念学习
- 第3章：决策树学习
- 第4章：人工神经网络学习
- 第5章：评估假设
- 第6章：贝叶斯学习
- 第7章：计算学习理论
- 第8章：基于实例的学习
- 第9章：遗传算法
- 第10章：规则学习
- 第11章：分析学习
- 第12章：归纳学习
- 第13章：增强学习

一、机器学习的定义

机器学习，周志华著，2016年。

- 第1章：绪论
- 第2章：模型评估与选择
- 第3章：线性模型
- 第4章：决策树
- 第5章：神经网络
- 第6章：支持向量机
- 第7章：贝叶斯分类器
- 第8章：集成学习
- 第9章：聚类
- 第10章：降维与度量学习
- 第11章：特征选择与稀疏学习
- 第12章：计算学习理论
- 第13章：半监督学习
- 第14章：概率图模型
- 第15章：规则学习
- 第16章：强化学习

一、机器学习的定义

- 从上述两本最具代表性的机器学习教材可以看出：机器学习的教材和课程主要讲解各种不同的机器学习技术。比如：线性学习、支持向量机学习、神经网络学习、决策树学习、贝叶斯学习、最近邻学习等等。
- 当人们在讨论机器学习的时候，经常会想到另一种智能数据分析技术：数据挖掘。
- 那到底什么是数据挖掘？

二、与数据挖掘的区别与联系

- 所谓数据挖掘就是：“识别出巨量数据中有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程”。顾名思义，数据挖掘就是试图从海量数据中找出有用的知识。
- 至于机器学习与数据挖掘的区别与联系，我们同样先来看看下面两本最具代表性的数据挖掘教材的目录。

二、与数据挖掘的区别与联系

数据挖掘概念与技术(第3版) , Jiawei Han等著, 范明等译, 2012。

- 第1章：引论
- 第2章：认识数据
- 第3章：数据预处理
- 第4章：数据仓库与联机分析处理
- 第5章：数据立方体技术
- 第6章：挖掘频繁模式、关联和相关性：基本概念和方法
- 第7章：高级模式挖掘
- 第8章：分类：基本概念
- 第9章：分类：高级方法
- 第10章：聚类分析：基本概念和方法
- 第11章：高级聚类分析
- 第12章：离群点检测
- 第13章：数据挖掘的发展趋势和研究前沿

二、与数据挖掘的区别与联系

数据挖掘导论, Pang-Ning Tan等著, 范明等译, 2006年。

- 第1章：绪论
- 第2章：数据
- 第3章：探索数据
- 第4章：分类：基本概念、决策树与模型评估
- 第5章：分类：其他技术
- 第6章：关联分析：基本概念与算法
- 第7章：关联分析：高级概念
- 第8章：聚类分析：基本概念与算法
- 第9章：聚类分析：附加问题与算法
- 第10章：异常检测

二、与数据挖掘的区别与联系

- 可见，数据挖掘的教材和课程主要讲解各种不同的数据挖掘任务。比如：分类、回归、聚类、关联分析、异常分析、演变分析等等。
- 数据挖掘可以视为机器学习和数据库的交叉，它主要利用机器学习界提供的技术来分析海量数据，利用数据库界提供的技术来管理海量数据。
- 二者既有区别又有联系，整体来说，机器学习偏理论，数据挖掘偏应用。

二、与数据挖掘的区别与联系

中国计算机学会通讯, 2007, 3(12): 35-44. 特邀综述

机器学习与数据挖掘

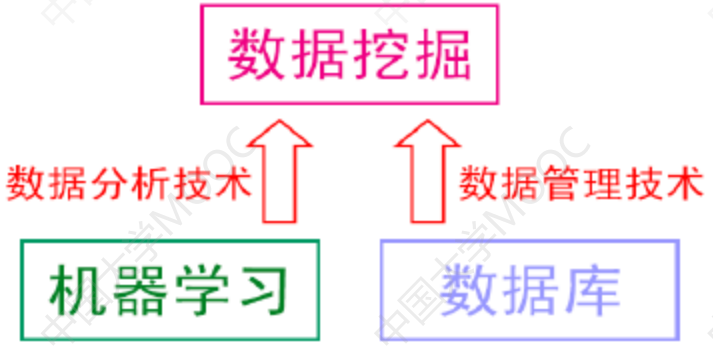
周 志 华

南京大学计算机软件新技术国家重点实验室，南京 210093

“机器学习”是人工智能的核心研究领域之一， 其最初的研究动机是为了让计算机系统具有人的学习能力以便实现人工智能，因为众所周知，没有学习能力的系统很难被认为是具有智能的。目前被广泛采用的机器学习的定义是“利用经验来改善计算机系统自身的性能”[1]。事实上，由于“经验”在计算机系统中主要是以数据的形式存在的，因此机器学习需要设法对数据进行分析，这就使得它逐渐成为智能数据分析技术的创新源之一，并且为此而受到越来越多的关注。

“数据挖掘”和“知识发现”通常被相提并论，并在许多场合被认为是可以相互替代的术语。对数据挖掘有多种文字不同但含义接近的定义，例如“识别出巨量数据中有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程”[2]。其实顾名思义，数据挖掘就是试图从海量数据中找出有用的知识。大体上看，数据挖掘可以视为机器学习和数据库的交叉，它主要利用机器学习界提供的技术来分析海量数据，利用数据库界提供的技术来管理海量数据。

因为机器学习和数据挖掘有密切的联系，受主编之邀，本文把它们放在一起做一个粗浅的介绍。



三、本课程的授课思路与内容安排

本课程的授课思路：以数据挖掘中的分类任务为例，首先讲解分类模型的评估，然后讲解各种不同的机器学习技术：

- 第1章：绪论
- 第2章：模型评估
- 第3章：线性学习（Linear learning）
- 第4章：支持向量机学习（Support vector machine learning）
- 第5章：神经网络学习（Neural network learning）
- 第6章：决策树学习（Decision tree learning）
- 第7章：贝叶斯学习（Bayesian learning）
- 第8章：最近邻学习（Nearest neighbor learning）
- 第9章：无监督学习（Unsupervised learning）
- 第10章：集成学习（Ensemble learning）
- 第11章：代价敏感学习（Cost-sensitive learning）
- 第12章：演化学习（Evolutionary learning）
- 第13章：强化学习（Reinforcement learning）

三、本课程的授课思路与内容安排

本课程的授课思路：以数据挖掘中的分类任务为例，首先讲解分类模型的评估，然后讲解各种不同的机器学习技术：

- 第1章，讲解机器学习的定义、与数据挖掘的区别与联系、本课程的授课思路与内容安排、教材及参考书
- 第2章，讲解模型评估的方法、指标和比较检验。
- 第3-9章，讲解机器学习的基础技术：以线性回归开始，讲解线性学习；以K均值聚类收尾，讲解无监督学习；中间包括支持向量机学习、神经网络学习、决策树学习、贝叶斯学习、以及最近邻学习。
- 第10-13章，讲解机器学习的进阶技术：具体包括集成学习、代价敏感学习、演化学习、以及强化学习。

三、本课程的授课思路与内容安排

本课程的授课思路：以数据挖掘中的分类任务为例，首先讲解分类模型的评估，然后讲解各种不同的机器学习技术：

- 本课程第1-11章均由蒋良孝博士主讲，第12-13章由胡成玉博士主讲。

蒋良孝博士，中国地质大学计算机学院教授，智能地学信息处理湖北省重点实验室副主任，中国计算机学会和中国人工智能学会高级会员，教育部新世纪优秀人才，湖北省杰出青年人才，主要研究方向为机器学习与数据挖掘。

胡成玉博士，中国地质大学计算机学院副教授，计算机科学系主任，中国计算机学会会员，中国仿真学会智能仿真优化与调度专委会委员。主要研究方向为进化计算、智能调度、以及大数据处理技术。

三、本课程的授课思路与内容安排

- 现在我们来看看分类的定义。分类就是构建一个分类模型，即分类器，然后通过分类器将数据对象映射到某个给定的类别中的过程。分类过程可以分为两步：

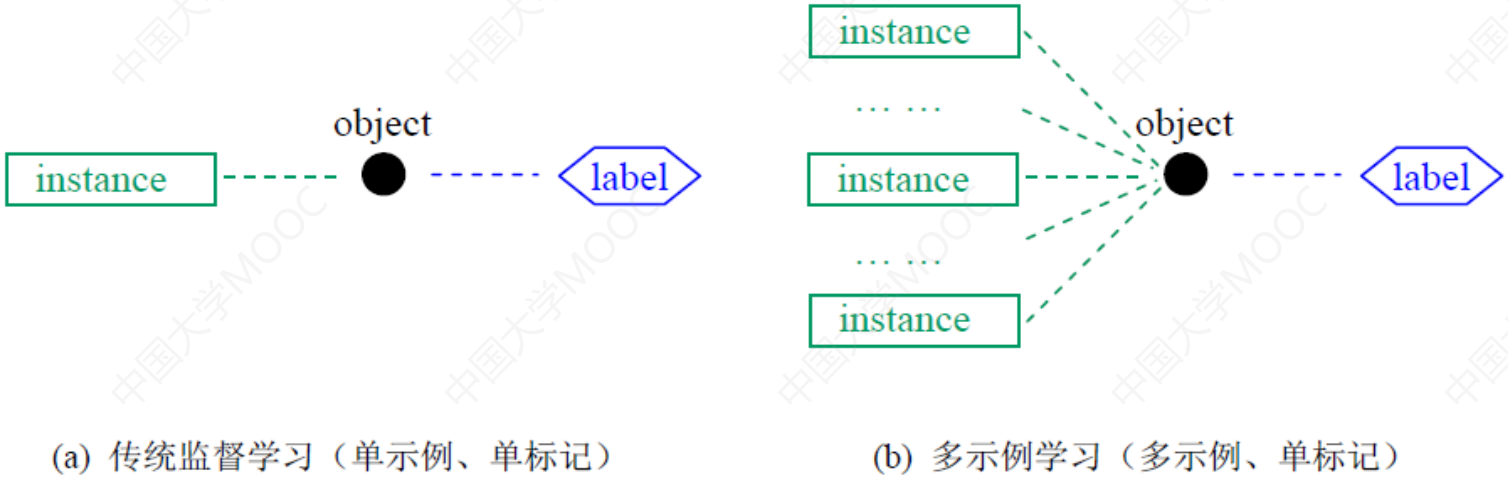
✓ 第一步使用已知类标记的训练数据集学习分类模型。这一步称为分类器的训练阶段。

✓ 第二步应用分类模型对未知类标记的对象进行分类。这一步称为分类器的工作阶段。实际上，在工作之前还应该对学到的模型进行性能测试评估（这一步称为分类器的测试阶段），如果模型的性能可以接受，才可以用它来对未知类标记的对象进行分类。

✓ **可见：分类是一个三步走的过程：训练→测试→工作。**

三、本课程的授课思路与内容安排

周志华, 张敏灵. MIML: 多示例多标记学习. 机器学习及其应用(第10章), 2009, 218-234.



(a) 传统监督学习（单示例、单标记）

(b) 多示例学习（多示例、单标记）

(c) 多标记学习（单示例、多标记）

(d) 多示例多标记学习

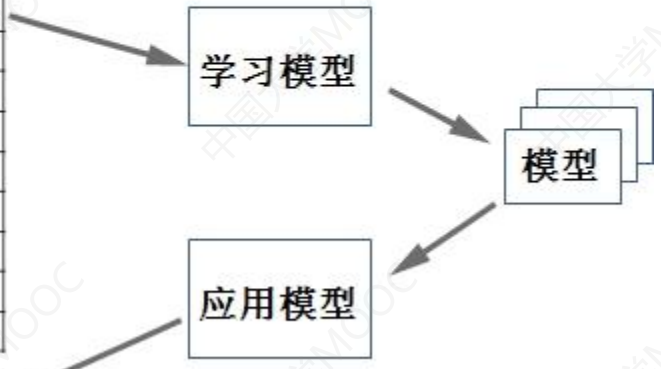
三、本课程的授课思路与内容安排

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No
15	**Sunny**	**Cool**	**High**	**Strong**	**?**

三、本课程的授课思路与内容安排

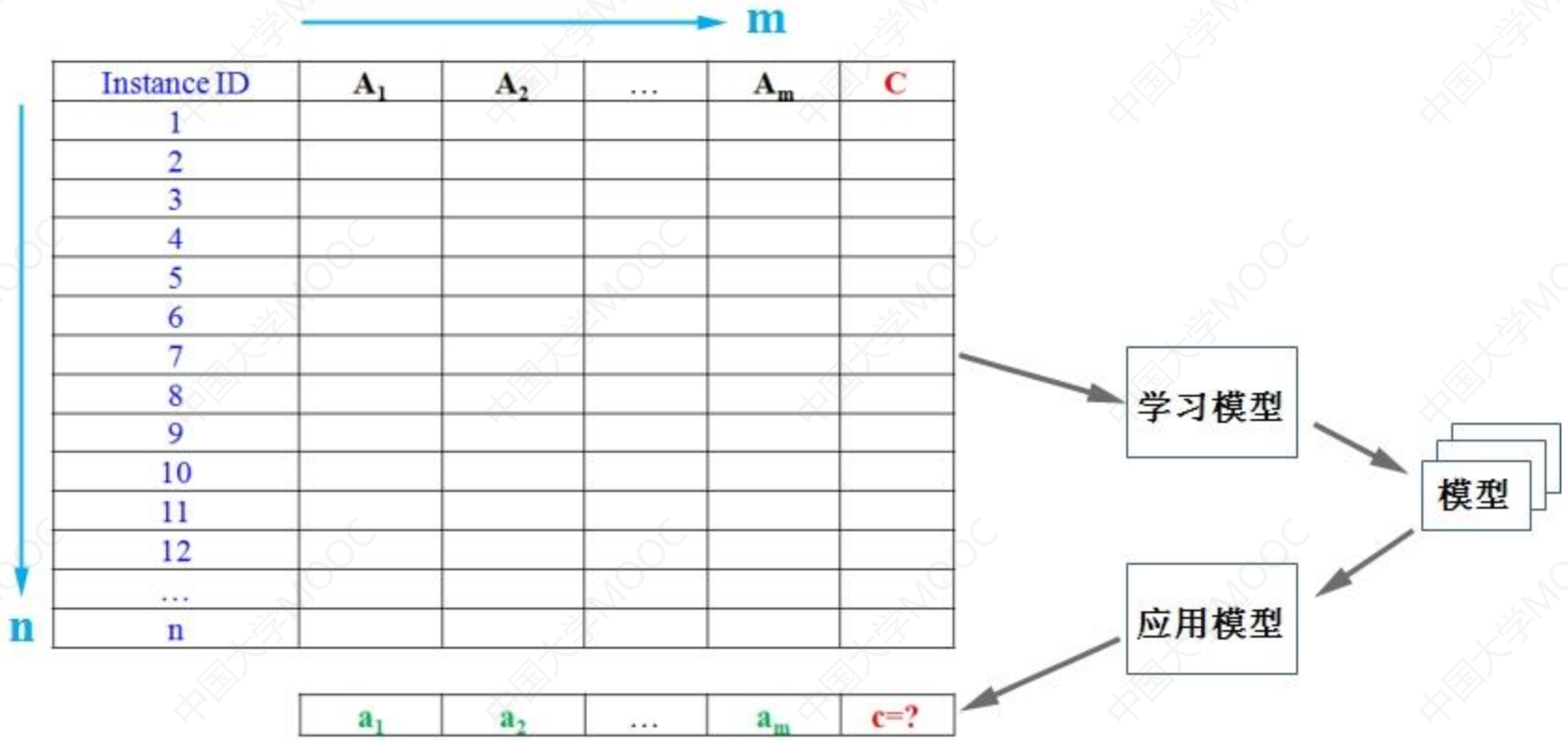
分类的基本过程

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No
15	Sunny	Cool	High	Strong	?



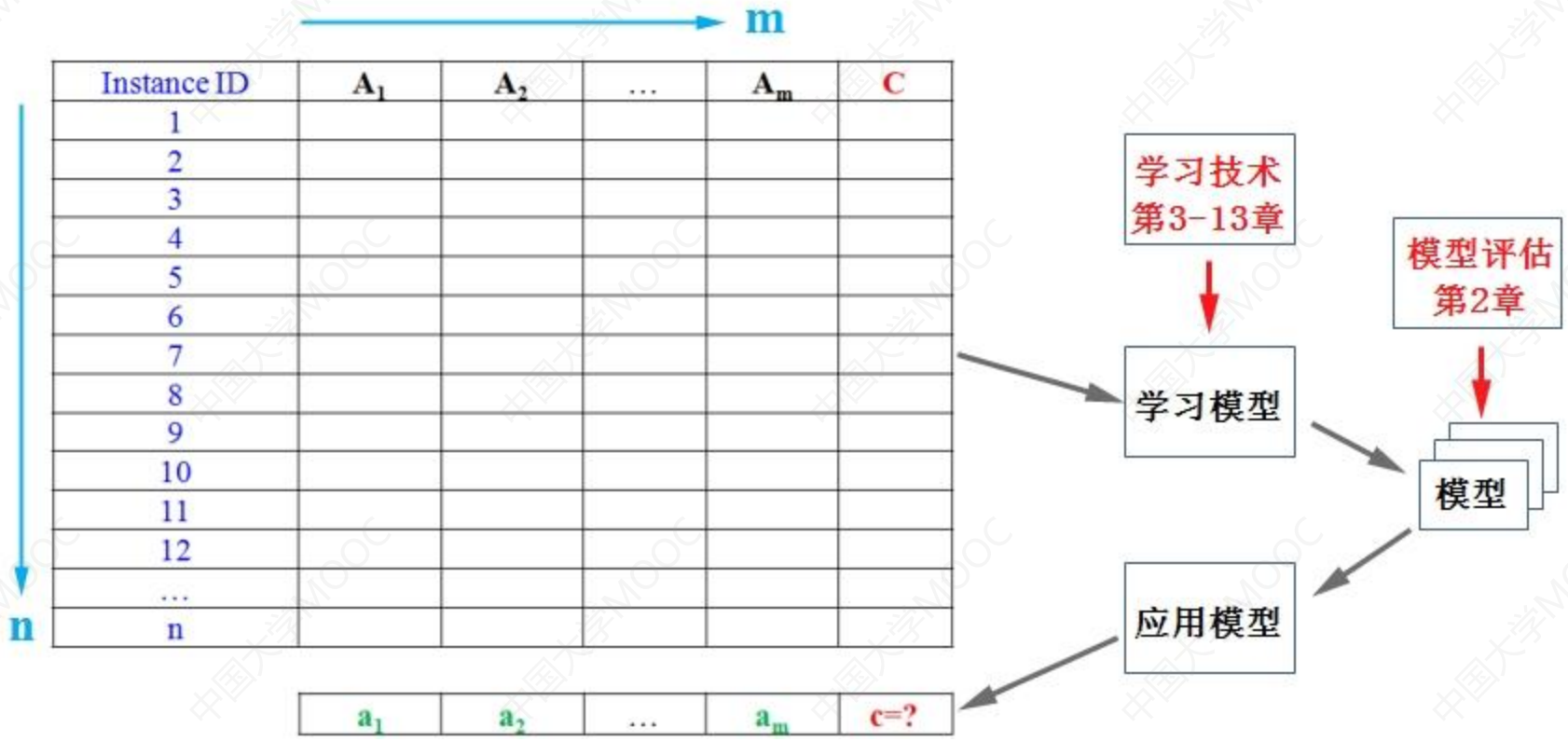
三、本课程的授课思路与内容安排

分类：定义与过程



三、本课程的授课思路与内容安排

分类：定义与过程



四、教材及参考书

- **教材：**
  - ✓ 周志华著. 机器学习. 清华大学出版社, 2016年.
- **参考书：**
  - ✓ Tom M. Mitchell著, 曾华军等译. 机器学习. 机械工业出版社, 1997年.
  - ✓ 于剑著. 机器学习：从公理到算法. 清华大学出版社, 2017年.
  - ✓ Jiawei Han等著, 范明等译. 数据挖掘：概念与技术（第3版）. 机械工业出版社, 2012年.
  - ✓ Pang-Ning Tan等著, 范明等译. 数据挖掘导论. 人民邮电出版社, 2006年.
  - ✓ I. H. Witten等著, 董琳等译. 数据挖掘：实用机器学习工具与技术（第4版）, 机械工业出版社, 2018年.