大数据导论
Introduction to Big Data

第1.2讲: 数据的结构化表示

叶允明

计算机科学与技术学院

哈尔滨工业大学（深圳）

目录

认识数据

不同类型的数据

- 记录数据
  - 关系表数据
  - 事务数据（Transaction Data）
- 多媒体数据：声、图、文
- 时空数据
  - 空间数据（Spatial Data）
  - 时间数据（Temporal Data）
- （关系）图数据

记录数据

- 数据是记录的汇集，每个记录包含固定的属性集

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

事务数据

- 一种特殊类型的记录数据，其中
  - 每条记录（事务）涉及一系列的项
  - 考虑一个杂货店，顾客一次购物所购买的商品的集合构成一个事务，而购买的商品是项。

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

文档数据

ABSTRACT

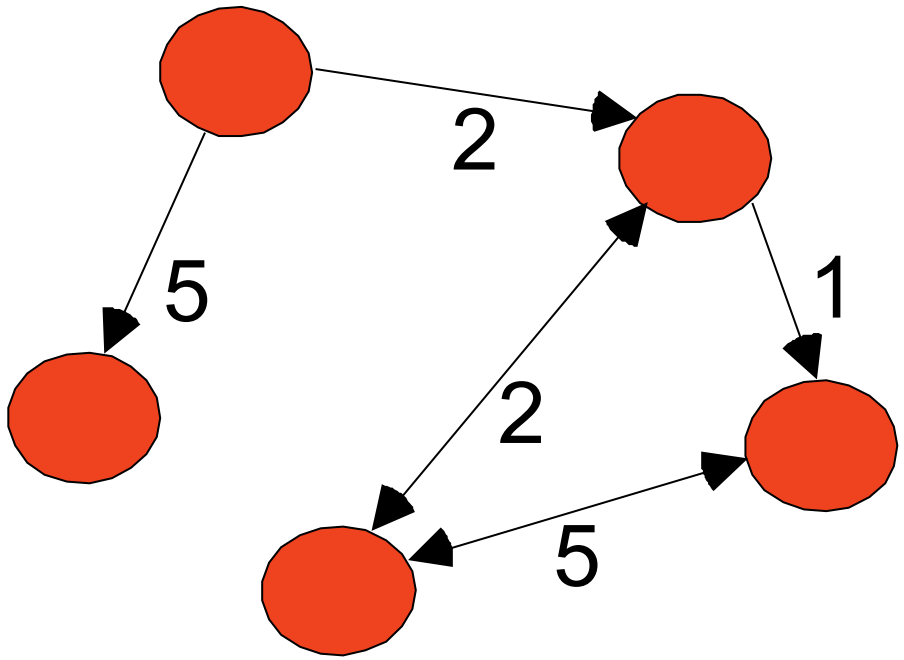
Partial differential equations (PDEs) play a prominent role in many disciplines such as applied mathematics, physics, chemistry, material science, computer science, etc. PDEs are commonly derived based on physical laws or empirical observations. However, the governing equations for many complex systems in modern applications are still not fully known. With the rapid development of sensors, computational power, and data storage in the past decade, huge quantities of data can be easily collected and efficiently stored. Such vast quantity of data offers new opportunities for data-driven discovery of hidden physical laws. Inspired by the latest development of neural network designs in deep learning, we propose a new feed-forward deep network, called PDE-Net, to fulfill two objectives at the same time: to accurately predict dynamics of complex systems and to uncover the underlying hidden PDE models. The basic idea of the proposed PDE-Net is to learn differential operators by learning convolution kernels (filters), and apply neural networks or other machine learning methods to approximate the unknown nonlinear responses. Comparing with existing approaches, which either assume the form of the nonlinear response is known or fix certain finite difference approximations of differential operators, our approach has the most flexibility by learning both differential operators and the nonlinear responses. A special feature of the proposed PDE-Net is that all filters are properly constrained, which enables us to easily identify the governing PDE models while still maintaining the expressive and predictive power of the network. These constrains are carefully designed by fully exploiting the relation between the orders of differential operators and the orders of sum rules of filters (an important concept originated from wavelet theory). We also discuss relations of the PDE-Net with some existing networks in computer vision such as Network-In-Network (NIN) and Residual Neural Network (ResNet). Numerical experiments show that the PDE-Net has the potential to uncover the hidden PDE of the observed dynamics, and predict the dynamical behavior for a relatively long time, even in a noisy environment.

据中央气象台消息，今年第11号台风“轩岚诺”（超强台风级）的中心今天（8月31日）早晨5点钟位于日本冲绳县那霸市偏东方向约340公里的西北太平洋洋面上，就是北纬26.1度、东经131.1度，中心附近最大风力17级以上（62米/秒），中心最低气压为915百帕，七级风圈半径220~230公里，十级风圈半径70公里，十二级风圈半径40公里。

预计，“轩岚诺”将以每小时30公里左右的速度向西偏南转西南方向移动，9月1~2日将在琉球群岛以东洋面停滞或回旋，而后转向北偏西方向移动，3日夜间移入东海东南部海面。未来4~5天“轩岚诺”将维持超强台风级的强度，最大强度可达17级以上（62~70米/秒）。

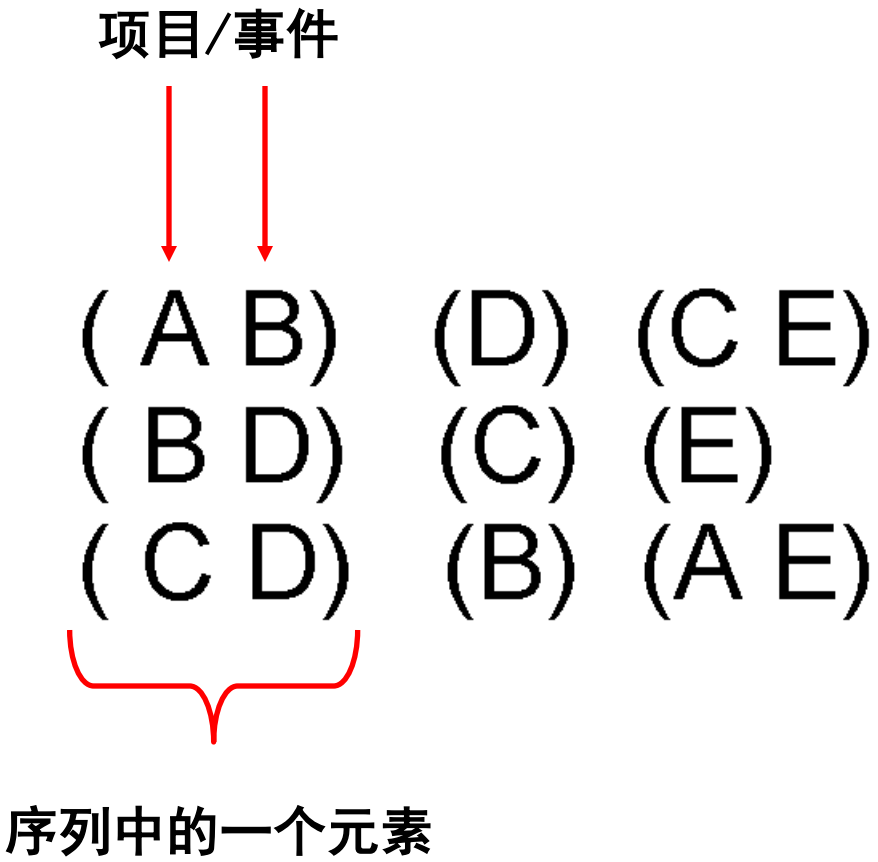
关系图数据

- 网页链接图
- 社交网络
- 文献引用图
- ……



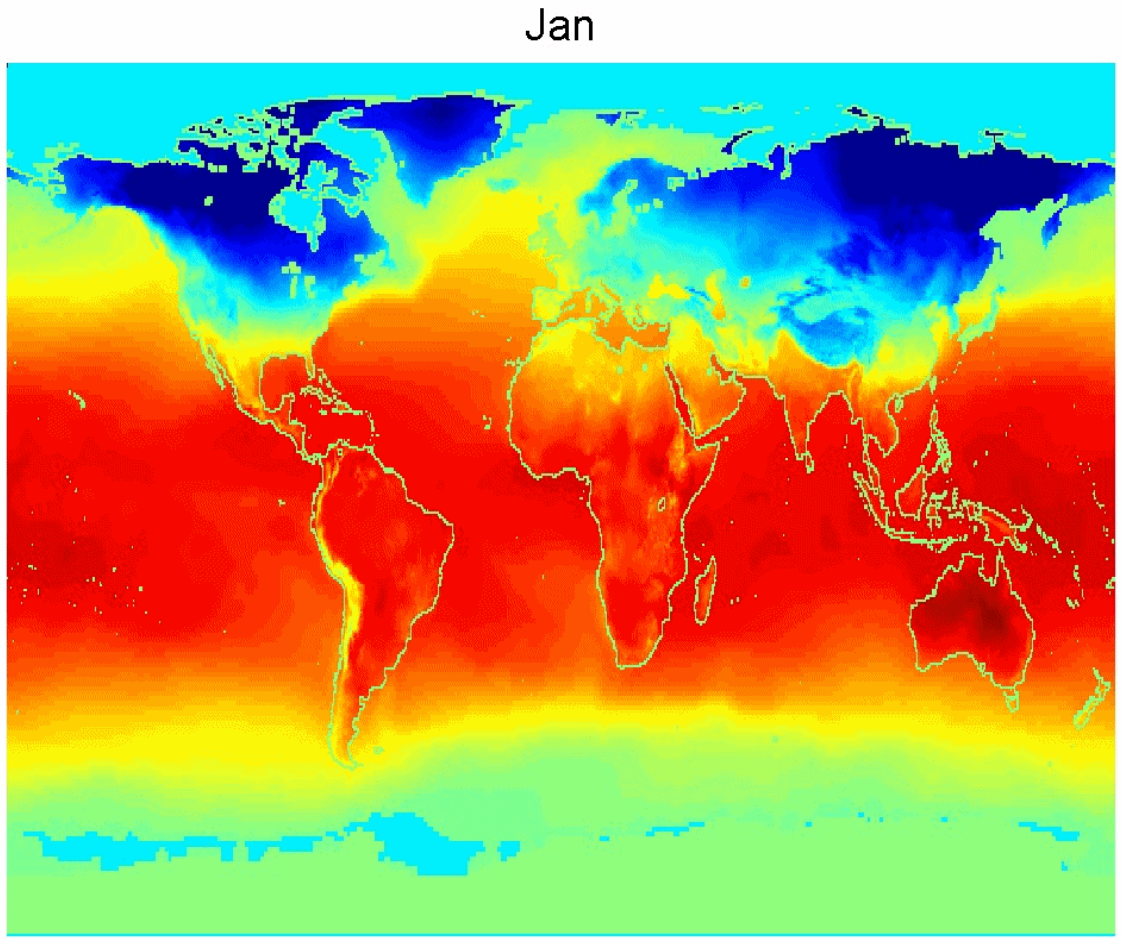
事件序列数据

- 事务序列



时空数据

陆地和海洋的
月平均温度

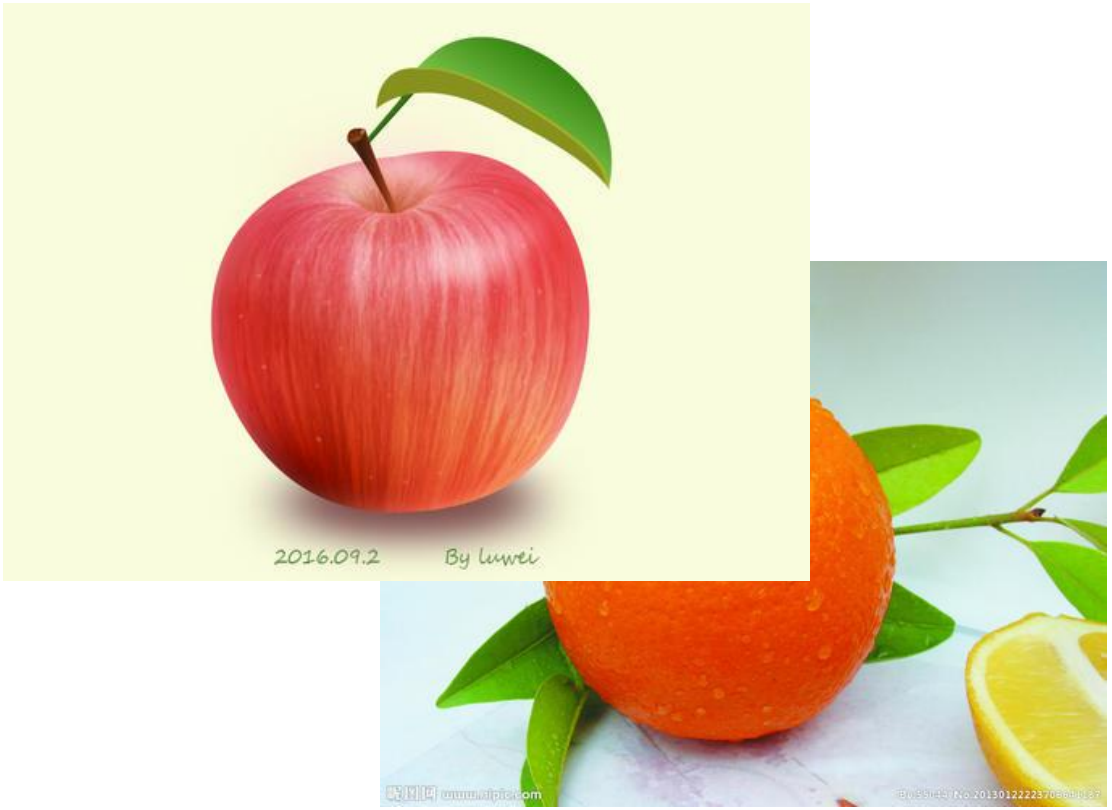


数据的定义

- 数据是对客观事物的观测（测量）或描述而得到的符号或数字集合。

序号	姓名	性别	年龄段	职业	消费收入比	剩余信贷比	历史贷款	房产	已抵押资产	家属	违约
1	张三	男	中年	个体商人	居高	高	3	2	5	2	0
2	李四	女	中年	教师	一般	低	0	1	0	1	0
3	梁五	男	青年	自由职业	超出	中	1	0	0	0	1
4	王六	男	老年	退休	正常	低	2	1	3	0	0
5	张七	男	中年	司机	一般	较高	1	0	0	0	1
6	陈八	女	中年	建筑师	一般	低	0	1	2	2	0

客户数据



图像数据

数据对象的概念

- 原始数据通常是一个包含多个数据对象（data object）的集合，每个数据对象通常对应于一个具有完整语义信息的事物，是分析事物的基本单位。

序号	姓名	性别	年龄段	职业	消费收入比	剩余信贷比	历史贷款	房产	已抵押资产	家属	违约
1	张三	男	中年	个体商人	居高	高	3	2	5	2	0
2	李四	女	中年	教师	一般	低	0	1	0	1	0
3	梁五	男	青年	自由职业	超出	中	1	0	0	0	1
4	王六	男	老年	退休	正常	低	2	1	3	0	0
5	张七	男	中年	司机	一般	较高	1	0	0	0	1
6	陈八	女	中年	建筑师	一般	低	0	1	2	2	0

结构化数据

- 数据是数据对象的集合
- 数据对象用一组刻画对象基本特征的属性描述
  - 对象也叫做记录、数据点、案例、样本、观测或实体等
- **属性（attribute）**是客观事物的性质或特性的计算机表示，而数据对象是由属性集合构成的
  - 例如：眼球颜色因人而异，物体的温度随时间而变。
  - 属性也叫做变量、字段、特征或维。

属性

对象

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Ref. this table is from Han's book slides

属性类型

- 类别型属性（categorical attribute）

  属性值定义域是一个固定、有限的符号或数字集合。

  Domain(“性别”)={M，F}

  Domain(“职业”)={教师，工程师，医生}

  属性值间不可以做算术运算：定性属性

属性类型

- 类别型属性（categorical attribute）的分类
  - 标称型属性（nominal attribute）：无序

身份证号、性别、职业、颜色

  - 有序型类别属性（ordinal attribute）：有大小、好坏等先后顺序的区别

学位：{“学士”、“硕士”、“博士”}

属性类型

- 独热编码（one-hot encoding）

职业-售货员	职业-教师	职业-医生
1	0	0
0	1	0
0	0	1

属性类型

- 数值型属性（numeric attribute）

  属性值定义在实数集或整数集上 。

  年龄、每月工资、债务收入比

  属性值间可以做算术运算：定量属性

属性类型

- 数值型属性（numeric attribute）
  - 区间标度型属性（interval-scaled attribute）：有序

“摄氏温度”　“30℃比10℃高20℃” ✓　属性值“0℃”并不代表无温度 ✓

“30℃的温度是10℃的三倍” ✗

  - 比率标度型属性（ratio-scaled attribute）：有序

“每月工资”　月薪1万元比月薪5000元多5000 ✓　月薪为0是有实际意义的 ✓

月薪1万是月薪5000的两倍 ✓

数据的结构化

- “**数据的结构化**”就是指将原始数据按照固定的“属性-值”序列逐行排列各个数据对象，其结果就形成了“**结构化数据**”。

- 标准结构化数据的形式化定义：

$$\boldsymbol{D} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_i}, \ldots, \boldsymbol{x_m}\},\ (i = 1 \ldots m),$$

$$\text{其中}\boldsymbol{x_i} = \{x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{in}\}\,(j = 1 \ldots n),\ x_{ij} \in \boldsymbol{R}$$

单一类型数据的结构化表示

关系表数据的结构化表示

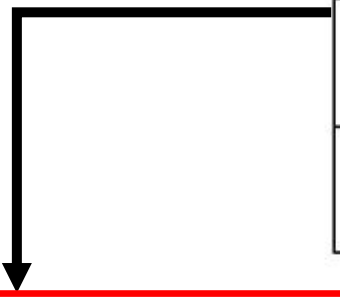
- 单个关系表数据的结构化
  - 关系表数据是指以二维表形式存储的结构化数据

用户 ID	性别	年龄	月薪	职业
1	男	45	5000	售货员
2	女	35	10000	教师
3	男	28	9000	医生

关系表数据的结构化表示

- 单个关系表数据的结构化

用户 ID	性别	年龄	月薪	职业
1	男	45	5000	售货员
2	女	35	10000	教师
3	男	28	9000	医生



用户 ID	性别	年龄	月薪	职业-售货员	职业-教师	职业- 医生
1	1	45	5000	1	0	0
2	0	35	10000	0	1	0
3	1	28	9000	0	0	1

关系表数据的结构化表示

- 多个关系表数据的结构化

客户基本信息表

用户ID	性别	年龄	月薪	职业
1	男	45	5000	售货员
2	女	35	10000	教师
3	男	28	9000	白领

客户借贷记录信息表

用户ID	累计借款金额	累计借款天数	是否存在逾期未还
1	20000	365	1
2	8000	30	0
3	15000	90	0

用户ID	性别	年龄	月薪	职业-售货员	职业-教师	职业-白领	累计借款金额	累计借款天数	是否存在逾期未还
1	1	45	5000	1	0	0	10000	365	1
2	0	35	10000	0	1	0	8000	30	0
3	1	28	9000	0	0	1	15000	90	0

文本数据的结构化表示

- 标准结构化数据——数值矩阵
  - 文本数据

    高维、稀疏
  - 结构化表示：词频法

Text_1:“背包设计的很好看，质量很好，价格便宜”

Text_2: “这是我买过性价比最高的背包，很好看”

Text_1: “背包\设计\的\很\好看\，\质量\很\好\，\价格\便宜”

Text_2: “这\是\我\买\过\性价比\最高\的\背包\，\很\好看\”

文本数据的结构化表示

- 词频法

Text_1: “背包\设计\的\很\好看\，\质量\很\好\，\价格\便宜”

Text_2: “这\是\我\买\过\性价比\最高\的\背包\，\很\好看\”

统计词频

词语 文本	很	的	好	这	是	我	买	过	背包	设计	好看	质量	价格	便宜	最高	性价比
Text_1	2	1	1	0	0	0	0	0	1	1	1	1	1	1	0	0
Text_2	1	1	0	1	1	1	1	1	1	0	1	0	0	0	1	1

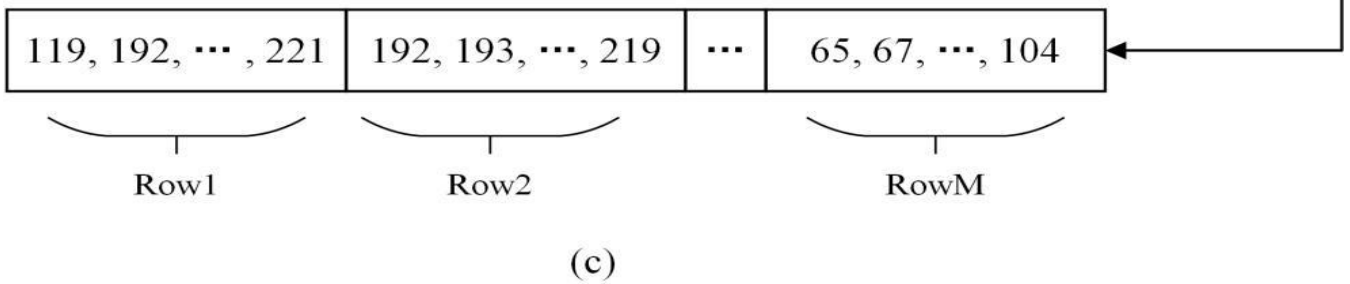
图像数据的结构化表示

- 标准结构化数据——数值矩阵
  - 数字图像与视频数据
    - ✓ 结构化表示：扁平化

(a)

$$\begin{bmatrix} 119 & 192 & 193 & \cdots & 218 & 220 & 221 \\ 192 & 193 & 194 & \cdots & 218 & 218 & 219 \\ 192 & 193 & 195 & \cdots & 219 & 218 & 218 \\ \cdots & & & & & & \\ 62 & 61 & 59 & \cdots & 59 & 60 & 63 \\ 59 & 55 & 57 & \cdots & 65 & 66 & 66 \\ 65 & 67 & 74 & \cdots & 100 & 104 & 104 \end{bmatrix}$$

(b)

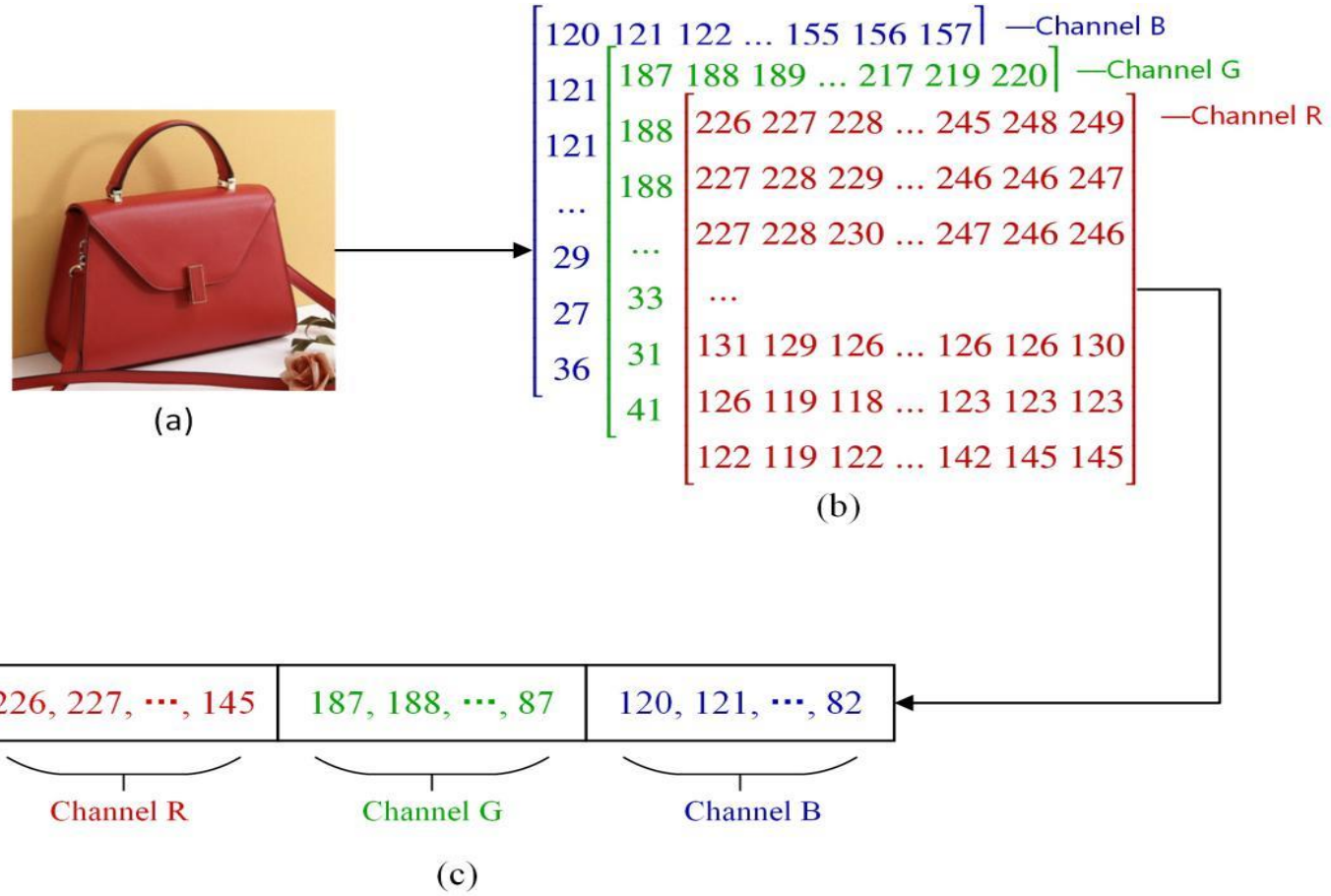


(c)

单通道灰度图像的结构化表示

图像数据的结构化表示

- 标准结构化数据——数值矩阵
  - 数字图像与视频数据
    - ✓ 结构化表示：扁平化



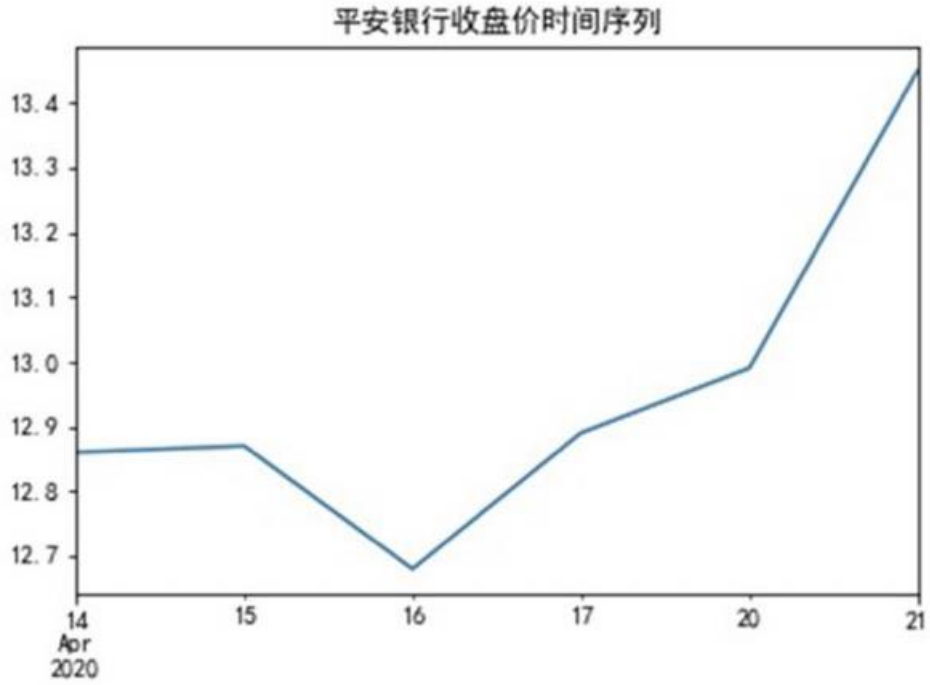
多通道彩色图像的结构化表示

时序数据的结构化表示

- 时间序列数据（time series data），是在不同的时间间隔观测同一数据对象，将其属性值相继排列所形成的序列
- 单通道时序数据

trade_date	close
2020/4/14	12.86
2020/4/15	12.87
2020/4/16	12.68
2020/4/17	12.89
2020/4/20	12.99
2020/4/21	13.45

(a)



(b)

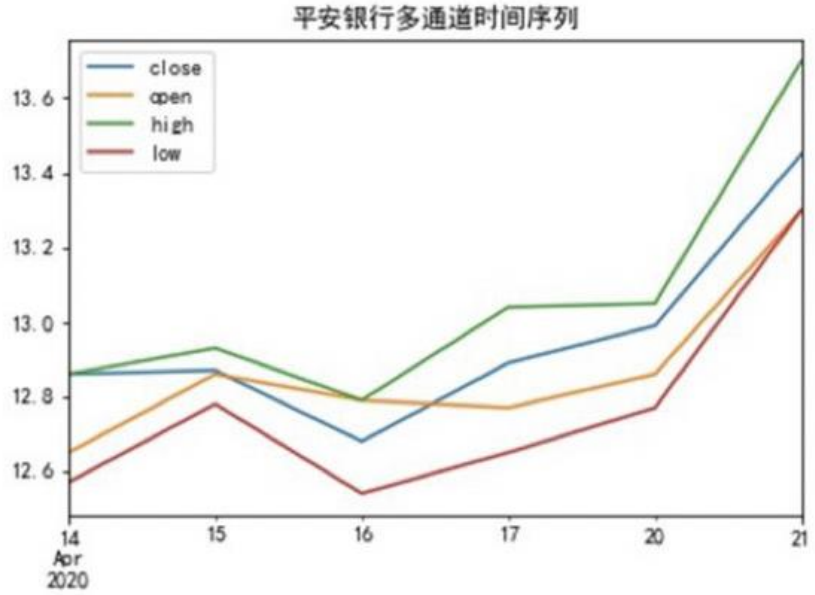
单通道：股票的"每日收盘价"的涨跌情况

时序数据的结构化表示

- 多通道时序数据

trade_date	close	open	high	low
2020/4/14	12.86	12.65	12.86	12.57
2020/4/15	12.87	12.86	12.93	12.78
2020/4/16	12.68	12.79	12.79	12.54
2020/4/17	12.89	12.77	13.04	12.65
2020/4/20	12.99	12.86	13.05	12.77
2020/4/21	13.45	13.3	13.7	13.3

(a)



(b)

多通道：股票的“每日开盘价”、“每日收盘价”、“每日最高价”、“每日最低价”涨跌情况

多源异构数据的统一结构化表示

多源异构数据的结构化表示

- 输入数据来自多个来源、且包含多种类型——多源异构
  - 电商平台商品销售额预测
    - ✓ 商品的基本属性数据：商品ID、商品的大类、商品单价、颜色、尺寸等，属于关系表类型的数据
    - ✓ 商品描述数据：对每个商品（对应于每个商品ID）的文字描述，属于文本数据；
    - ✓ 商品的展示图片：每个商品（对应于每个商品ID）的外观图，属于图像数据。

多源异构数据的结构化表示

商品的展示图片

商品描述数据

"2020新款红色时尚牛皮包，洋气百搭"

商品的基本属性数据

商品ID	商品价格	商品类别
312	99.9	服饰

广告ID	商品ID	开始时间	结束时间
12	312	2020/5/31 17:00	2020/5/31 21:00
25	312	2020/6/1 17:00	2020/6/1 21:00

扁平化图像矩阵 → [226, 227, 228, ···, 76, 80]
向量维度:360*360*3

词频法 → [0, 2, 2, 0, 0, 0, 1, 0, ···, 1, 0]
向量维度:56857(总词数)

多元关系表拼接 → [99.9, 0, 1, 0, 0, 0, 8]
向量维度:7

拼接 → [226, 227, ···, 0, 2, 2, ···, 99.9, ···, 8]
向量维度:360*360*3 + 56857 + 7

原始数据

各自应用结构化表示方法
生成子向量

子向量拼接
成异构数据统一表示向量

致谢

- 一小部分图表、文字来自互联网，仅供公益性的学习参考，在此表示感谢！如有版权要求请联系：yym@hit.edu.cn，谢谢！