机器学习与结构化数据挖掘

Machine Learning and Structured Data Mining

张华平 博士 副教授

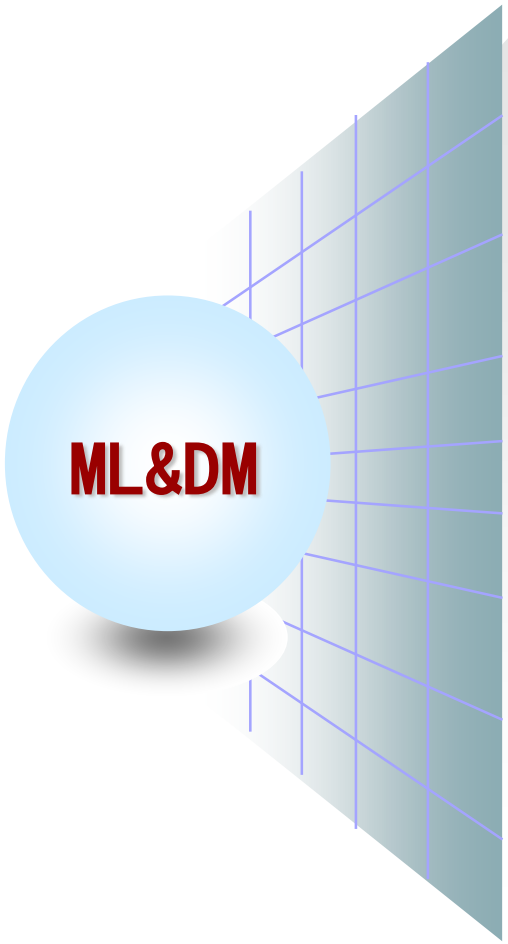
大数据搜索与挖掘实验室

kevinzhang@bit.edu.cn

@ICTCLAS张华平博士

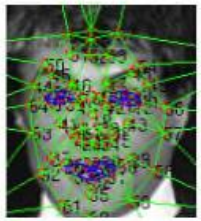
2016.11

纲要

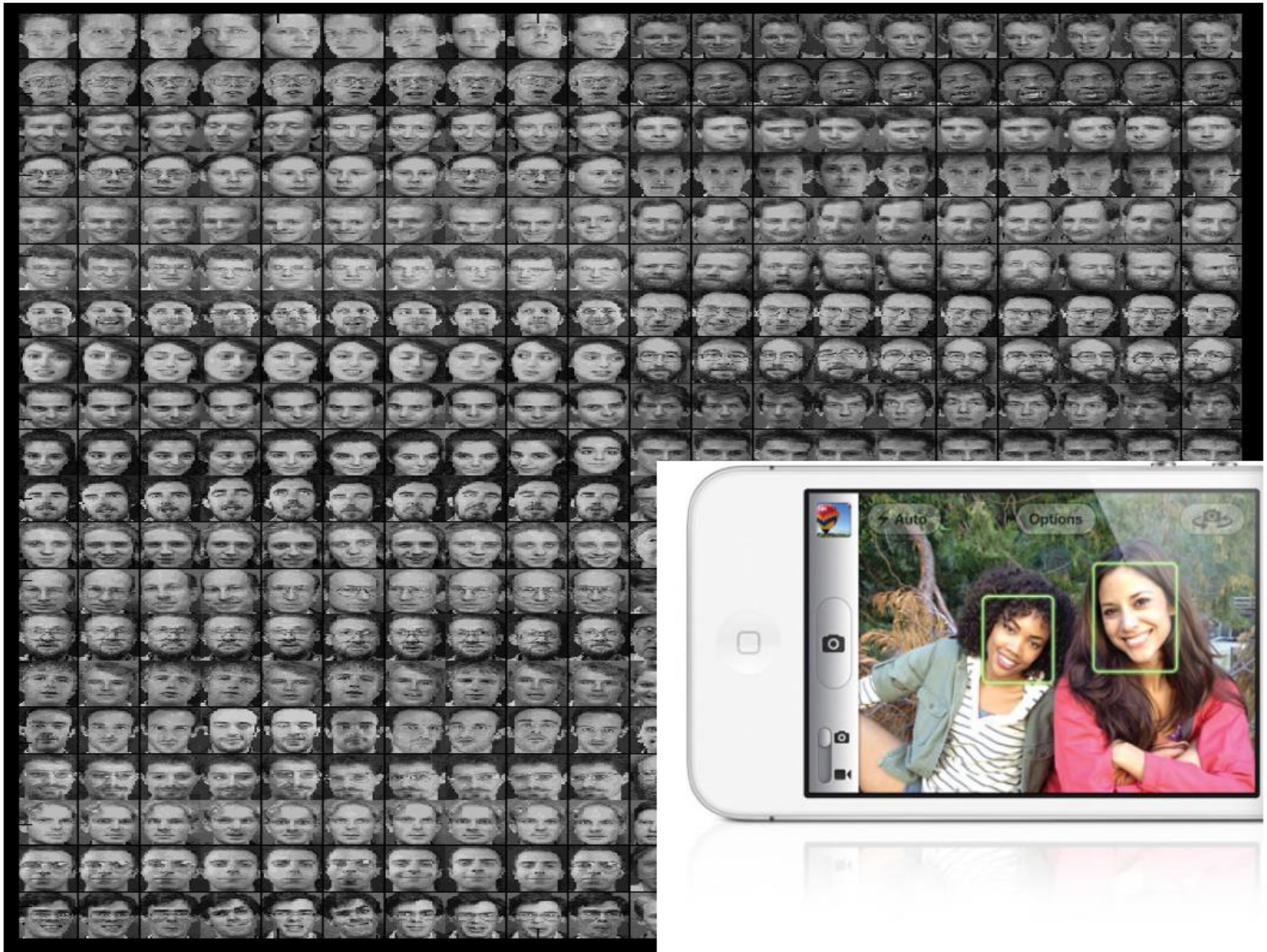


机器学习-大数据时代眼与耳

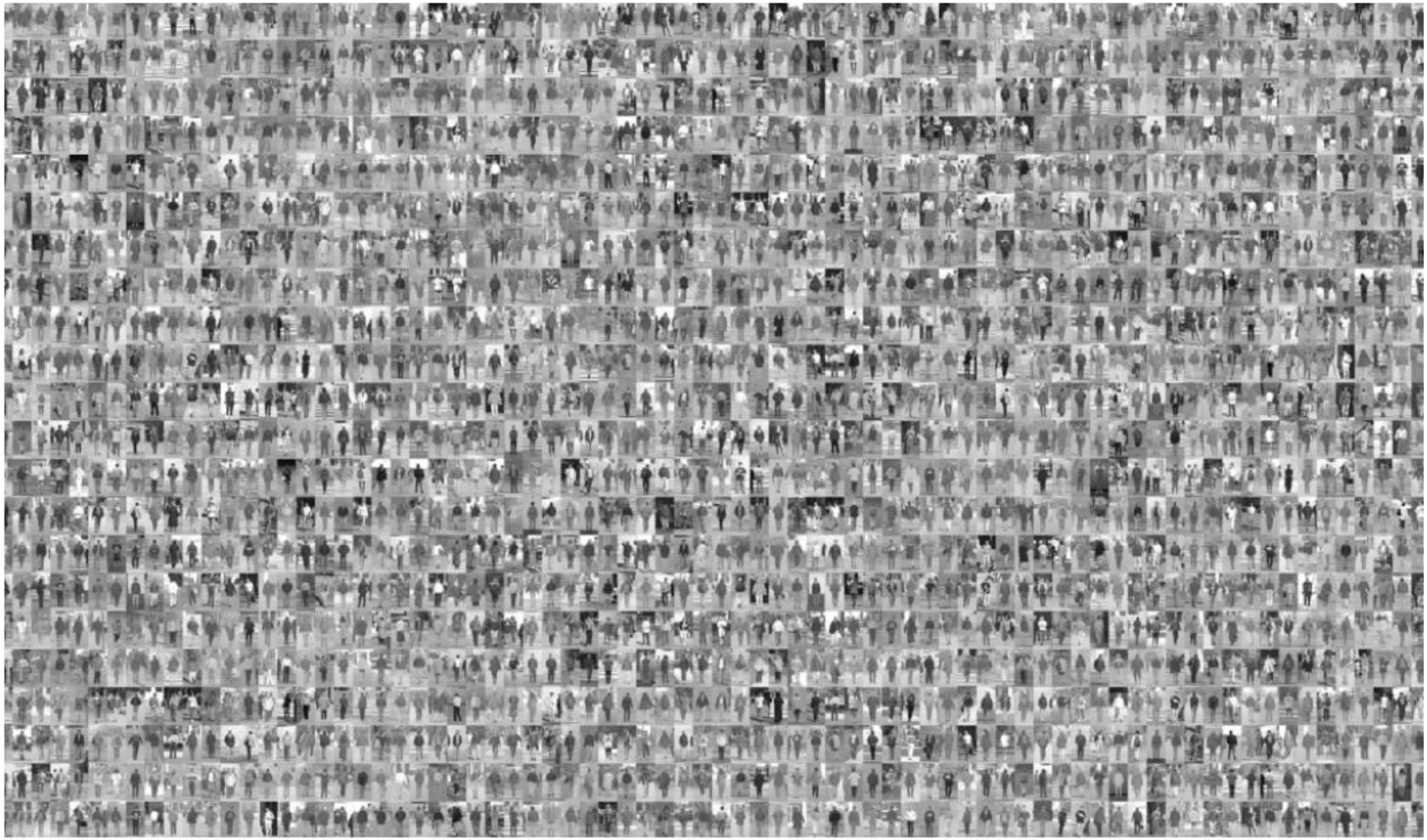
Machines that learn to recognise what they see and hear are at the heart of Apple, Google, Amazon, Facebook, Netflix, Microsoft, etc.



机器学习应用-人脸识别



机器学习应用-行人识别

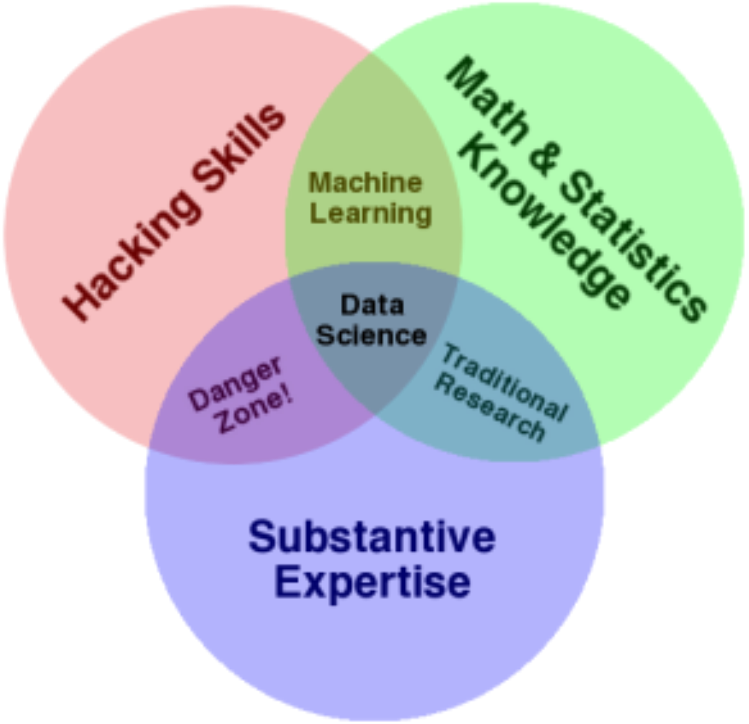


Millions of labeled examples are used to build real-world applications, such as pedestrian detection

机器学习定义-溯源

Arthur Samuel定义的机器学习(1959)：在不直接针对问题进行编程的情况下，赋予计算机学习能力的一个研究领域。

Alpaydin（2004）同时提出自己对机器学习的定义，“机器学习是用数据或以往的经验，以此优化计算机程序的性能标准。”（Machine learning is programming computers to optimize a performance criterion using example data or past experience.）



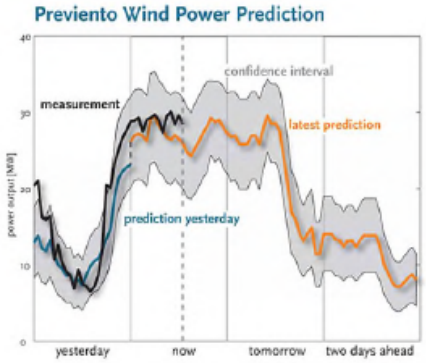
Drew Conway 2004

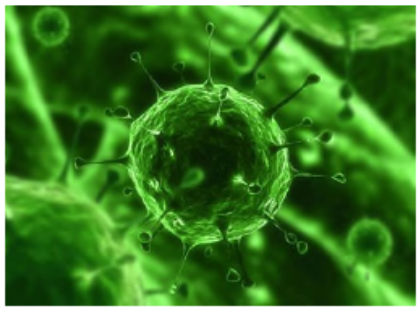
机器学习定义-Oxford

Machine learning

Machine learning deals with the problem of extracting ***features*** from data so as to solve many different ***predictive*** tasks:

- ❑Forecasting (*e.g. Energy demand prediction, finance*)
- ❑Imputing missing data (*e.g. Netflix recommendations*)
- ❑Detecting anomalies (*e.g. Security, fraud, virus mutations*)
- ❑Classifying (*e.g. Credit risk assessment, cancer diagnosis*)
- ❑Ranking (*e.g. Google search, personalization*)
- ❑Summarizing (*e.g. News zeitgeist, social media sentiment*)
- ❑Decision making (*e.g. AI, robotics, compiler tuning, trading*)

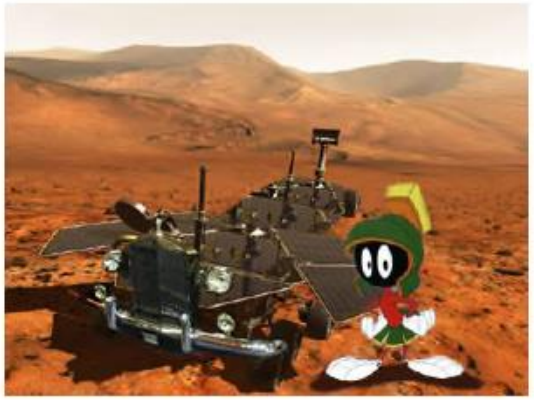


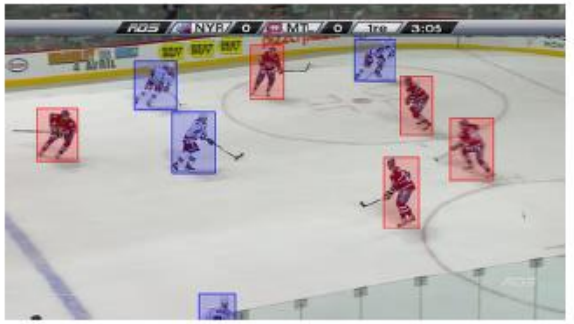


机器学习应用场景-Oxford

When to apply machine learning

- Human expertise is absent (*e.g. Navigating on Mars*)
- Humans are unable to explain their expertise (*e.g. Speech recognition, vision, language*)
- Solution changes with time (*e.g. Tracking, temperature control, preferences*)
- Solution needs to be adapted to particular cases (*e.g. Biometrics, personalization*)
- The problem size is to vast for our limited reasoning capabilities (*e.g. Calculating webpage ranks, matching ads to facebook pages*)





机器学习的典型过程

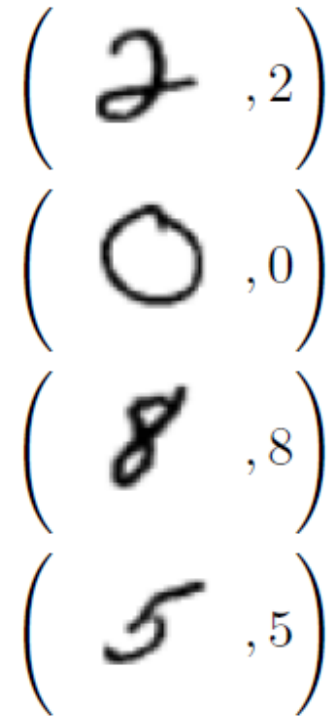
Introduction: digit classification

The task: write a program that, given a 28x28 grayscale image of a digit, outputs the string representation

来源：Zico Kolter，MLSS 2014

机器学习的典型过程-监督方法

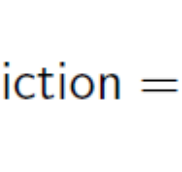
Training Data

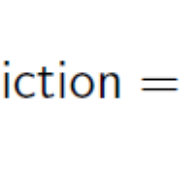


Machine Learning

$\longrightarrow$ Hypothesis function h_θ

Deployment

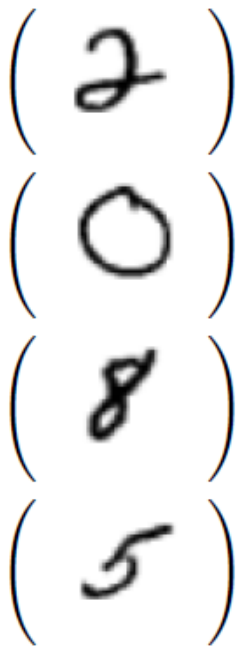
Prediction $= h_\theta$ 

Prediction $= h_\theta$ 

$\vdots$

机器学习的典型过程-无监督

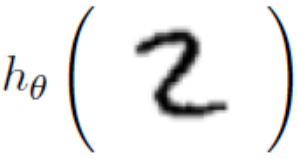
Training Data



Machine Learning

$\longrightarrow$ Hypothesis function h_θ

Deployment

Prediction $= h_\theta$ 

Prediction $= h_\theta$ 

$\vdots$

机器学习的典型过程-小结

- 特征表示
- 特征抽取
- 监督学习
- 无监督学习
- 分类、聚类、预测

交叉验证与泛化

交叉验证

- 10折交叉验证(10-fold cross validation)，将数据集分成十份，轮流将其中9份做训练1份做测试，10次的结果的均值作为对算法精度的估计，一般还需要进行多次10折交叉验证求均值，例如：10次10折交叉验证，以求更精确一点。

泛化能力

- 指机器学习算法对新鲜样本的适应能力。学习的目的是学到隐含在数据对背后的规律，对具有同一规律的学习集以外的数据，经过训练的算法也能给出合适的输出，该能力称为泛化能力。

相关概念辨析

- 机器学习
  - *机器学习强调的是方法*
- 深度机器学习
  - *机器学习的一种，指在云平台上，采用深度神经网络，对大数据进行学习的技术手段*
- 数据挖掘
  - *数据挖掘指的是目标与结果，是机器学习在结构化数据库上的分析结果；是机器学习与数据管理的综合*
- 大数据挖掘
  - *大数据挖掘是数据挖掘的一种，更多的强调的是对超大规模数据集合，采用云计算等大数据架构进行数据挖掘的技术方法*
- 非结构化大数据挖掘
  - *偏重于非结构化的文本、图片、音频、视频等的挖掘*

机器学习与数据挖掘

客户订单数据
会员卡用户数据
客户服务数据
信用卡交易数据
市场调研数据

- 顾客细分（分类、聚类）
  - 根据用户特征（身份、兴趣、收入水平）和消费行为进行分类或聚类。
- 潜在客户发掘和流失预警（分类）
  - 对流失客户和新客户的特征进行监督学习，得到发掘或预警模型。
- 识别顾客需求（分类、推荐）
  - 根据用户特征和消费行为预测用户喜好。
- 交叉销售分析（关联规则挖掘）
  - 根据大量订单数据发掘产品之间的促进或抑制关系。

Target Marketing

客户开发
客户挽留

个性化产品推荐

关联营销策略

数据源

数据挖掘

决策支持

主要方法

- 概念描述: 特征和区分
  - 概化, 汇总, 和比较数据特征, 例如, 干燥和潮湿的地区
- 关联规则挖掘 (相关和因果关系)
  - *age*(*X*, "20..29") ^ *income*(*X*, "20..29K") ⇒*buys*(*X*, "PC")

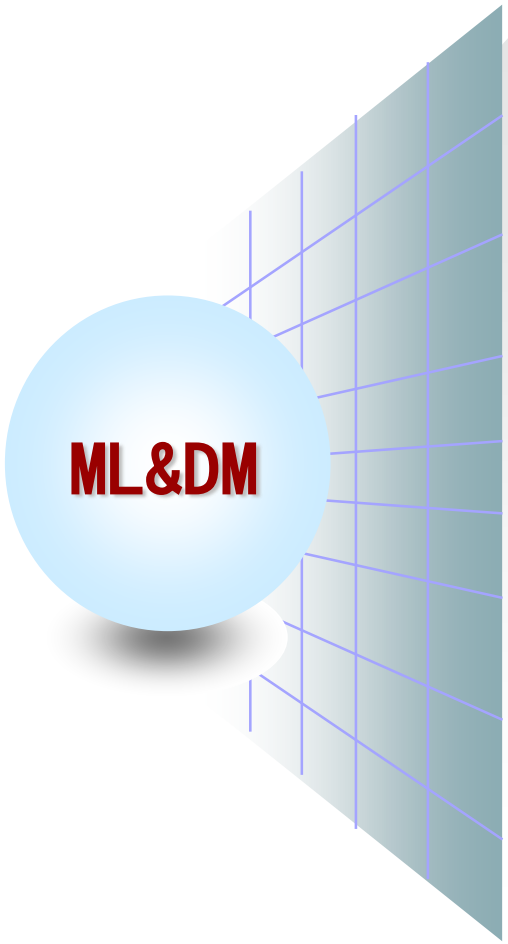
    [*support* = 2%, *confidence* = 60%]
  - *contains*(*T*, "computer") ⇒ *contains*(*T*, "software")

    [*support* = 1%, *confidence* = 75%]
- 分类和预测
  - 找出描述和识别类或概念的模型(函数), 用于将来的预测
  - 例如根据消费行为特征对客户分类, 或根据单位里程的耗油量对汽车分类
  - 表示: 决策树(decision-tree), 分类规则, 神经网络
  - 预测: 预测某些未知或遗漏的属性值

数据挖掘主要方法

- 聚类分析
  - 无监督学习方法
  - 类标号(Class label) 未知: 对数据分组, 形成新的类。例如, 新闻自动聚类、客户细分。
  - 聚类原则: 最大化类内的相似性, 最小化类间的相似性
- 孤立点(Outlier)分析
  - 孤立点: 一个数据对象, 它 与数据的一般行为不一致
  - 孤立点可以被视为例外, 但对于欺骗检测和罕见事件分析, 它是相当有用的
- 趋势和演变分析
  - 趋势和偏离: 回归分析
  - 序列模式挖掘, 周期性分析
  - 基于相似的分析
- 其它基于模式或统计的分析

纲要



I 机器学习与数据挖掘概览

II 关联规则挖掘概念与技术

III 数据分类概念与技术

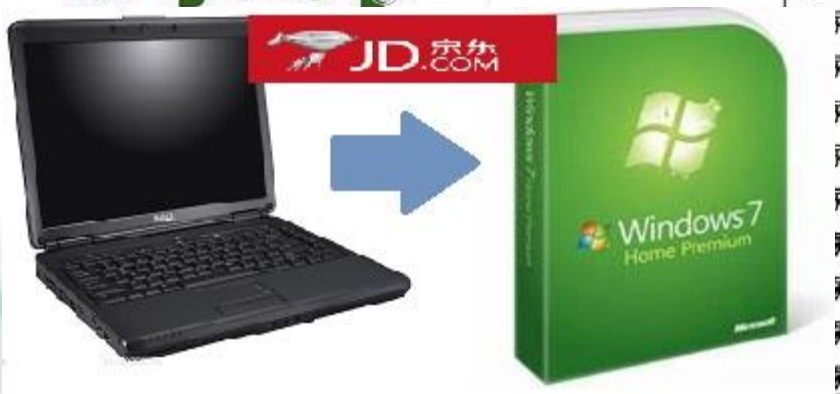
IV 数据聚类概念与技术

关联规则挖掘

- 关联规则挖掘
  - 关联分析就是发现关联规则，这些规则展示属性-值频繁地在给定数据集中一起出现的条件。关联分析广泛用于购物篮或事务数据分析。
- 动机：发现数据中蕴含的内在规律

区域	相关基因	疾病风险基因变异	变异属性	统计显著性
2q13.12	LARP4,DIP2	rs11169552-C	intergenic	2 x 10-10
20q13.33	LAMA5	rs4925386-C	intron	2 x 10-10
1q41	DUSP10	rs6691170-T	intergenic	1 x 10-9
1q41	DUSP10	rs6687758-G	intergenic	2 x 10-9
3q26.2	MYNN	rs10936599-C	cds-synon	3 x 10-8

Antecedent	Consequent	Support %	Confident %
糖尿病诊断--2 型糖尿病	高血压	20.525	69.408
糖尿病诊断--2 型糖尿病	冠心病--心绞痛	16.891	57.895
糖尿病诊断--2 型糖尿病	高脂血症--高甘油三酯	8.063	30.263
糖尿病诊断--IGT	高血压	18.969	57.522
糖尿病诊断-- IGT	冠心病--心绞痛	16.411	50.442
尿病诊断--无	冠心病--无	22.265	45.69
尿病诊断--无	高血压--无	20.914	42.323
尿病诊断--无	冠心病--心绞痛	16.891	37.795
尿病诊断-- IGT	冠心病--无	11.037	33.923
尿病诊断--IGT	高血压--无	10.214	30.973
尿病诊断--NOD	*高血压*	*2.140*	*62.857*
尿病诊断--NOD	*冠心病--心绞痛*	*1.248*	*37.143*
尿病诊断--NOD	*高血压--无*	*0.681*	*20.000*
尿病诊断--NOD	*冠心病--无*	*1.248*	*37.143*



关联规则挖掘

- 关联规则挖掘
  - 关联分析就是发现关联规则，这些规则展示属性-值频繁地在给定数据集中一起出现的条件。关联分析广泛用于购物篮或事务数据分析。
- 动机：发现数据中蕴含的内在规律
  - 那些产品经常被一起购买？
  - 买了PC之后接着都会买些什么？
  - 不同症状之间的并发关系
  - DNA序列的内部联系
- 应用
  - 购物篮分析、WEB日志（点击流）分析、捆绑销售、DNA序列分析等
- 关联挖掘类型
  - 根据规则处理的值的类型，分为布尔的和量化的。
  - 根据规则中数据的维，分为单维和多维的。

关联规则挖掘——基本概念

- 置信度
  - 置信度confidence(.)：是指购物篮分析中有了左边商品，同时又有右边商品的交易次数百分比，也就是说在所有的购买了左边商品的交易中，同时又购买了右边商品的交易概率。

$$Confidence(A \rightarrow B) = P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

- 支持度
  - 支持度sup(.)：表示在购物篮分析中同时包含关联规则左右两边物品的交易次数百分比，即支持这个规则的交易的次数百分比 。

$$Support(A \rightarrow B) = P(B \cap A)$$

Transaction ID	Items Bought
1000	A,B,C
2000	A,C
3000	A,D
4000	B,E,F

- 对于规则A⇒ C
  - support = support({A, C}) = 50%
  - confidence = support({A, C})/ support({A}) = 66.6%

关联规则挖掘——基本思想

关联规则挖掘基本思想

- 关联规则的挖掘问题，即发现所有的强关联规则，即发现所有同时满足最小支持度阈值的最小置信度值的规则。此过程分为两步：
- 第一步：识别所有的频繁K-项集，并统计其频率;
- 第二步：由频繁K-项集产生强关联规则。依据搜索到的频繁K-项集，导出满足给定条件的关联规则。

Apriori算法描述

- 使用一种称作逐层搜索的迭代方法，K-项集用于探索（K＋1）-项集。首先，找出频繁1-项集的集合，记为l1。l1用于找频繁2-项集的集合l2，而l2用于找l3，如此下去，直到不能找到频繁K-项集LK。找每个LK需要一次数据库扫描。最后由频繁K－项集可直接产生强关联规则。

Apriori的性质：

- 任何频繁项集的所有非空子集都必须也是频繁的
- 例：如果｛啤酒，尿布，坚果｝是一个频繁的，
- 则其子集｛啤酒，尿布｝｛啤酒，坚果｝｛尿布，坚果｝都是频繁的。

关联规则挖掘——原理举例

例：设有一个Electronics的事务数据库(如图1示)。数据库中有9个事务，即｜D｜＝9。Apriori假定事务中的项按字典次序存放。我们使用图2解释Apriori算法寻找D中的频繁项集。

TID	项ID的列表
T100	L1,L2,L5
T200	L2,L4
T300	L2,L3
T400	L1,L2,L4
T500	L1,L3
T600	L2,L3
T700	L1,L3
T800	L1,L2, L3,L5
T900	L1,L2,L3

（图1）

关联规则挖掘——原理举例

扫描D,C_1 对每个候选计数

项集	支持度计数
{L1}	6
{L2}	7
{L3}	6
{L4}	2
{L5}	2

L_1 比较候选支持度计数与最小支持度计数，设最小支持度为2

项集	支持度计数
{L1}	6
{L2}	7
{L3}	6
{L4}	2
{L5}	2

C_2 由L1产生候选C_2

项集
{L1, L2}
{L1, L3}
{L1, L4}
{L1, L5}
{L2, L3}
{L2, L4}
{L2, L5}
{L3, L4}
{L3, L5}
{L4, L5}

C_2 扫描D，对每个候选计数

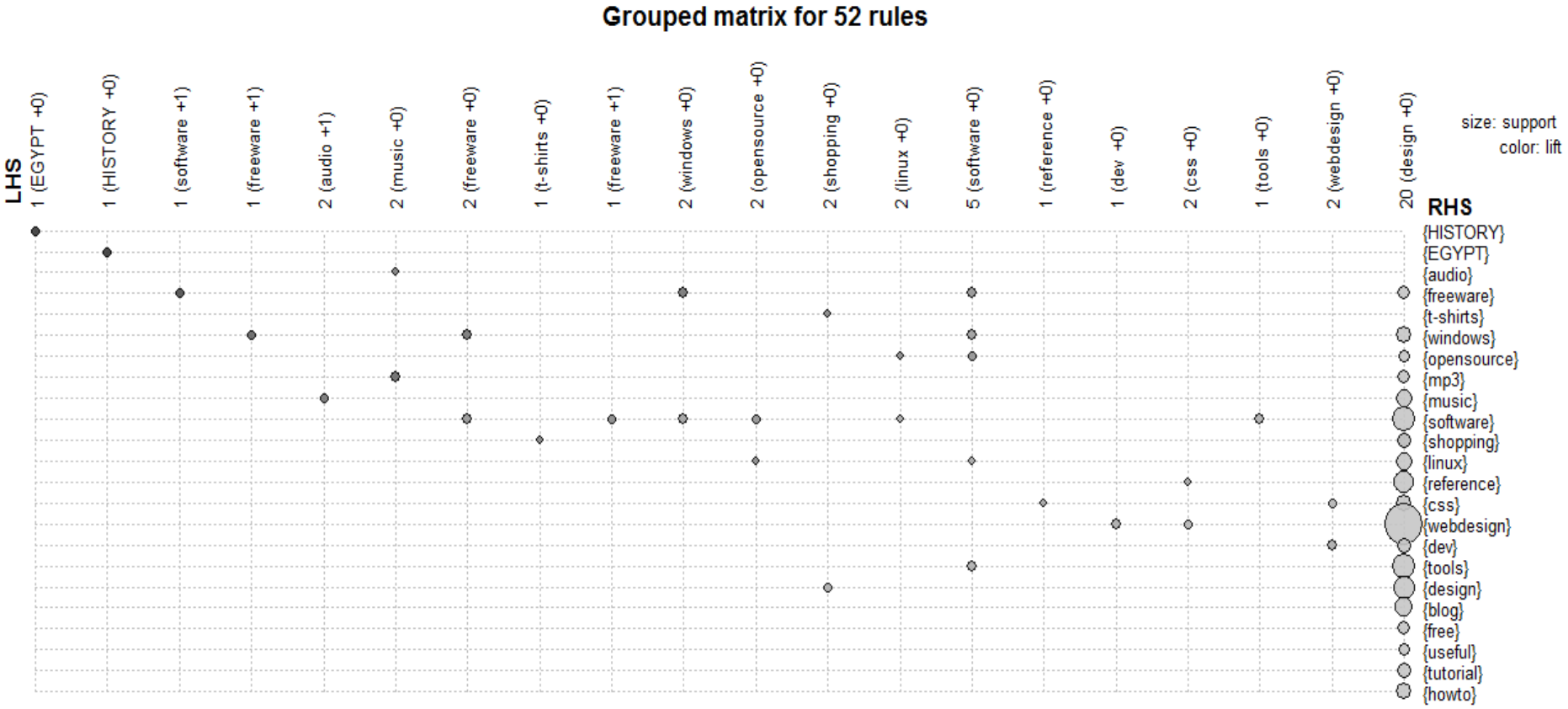
项集	支持度计数
{L1, L2}	4
{L1, L3}	4
{L1, L4}	1
{L1, L5}	2
{L2, L3}	4
{L2, L4}	2
{L2, L5}	2
{L3, L4}	0
{L3, L5}	1
{L4, L5}	0

比较候选支持度计数与最小支持度计数

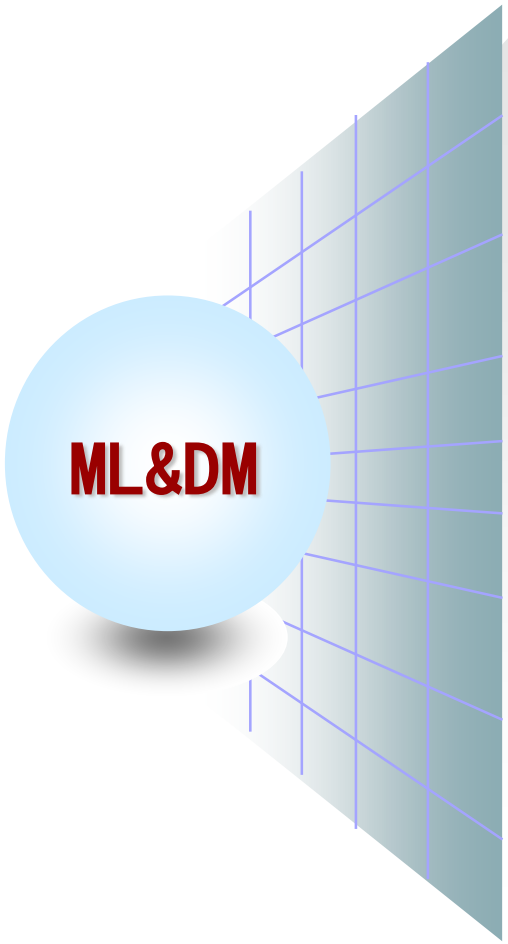
L_2

项集	支持度计数
{L1, L2}	4
{L1, L3}	4
{L1, L5}	2
{L2, L3}	4
{L2, L4}	2
{L2, L5}	2

关联挖掘案例——购物篮分析



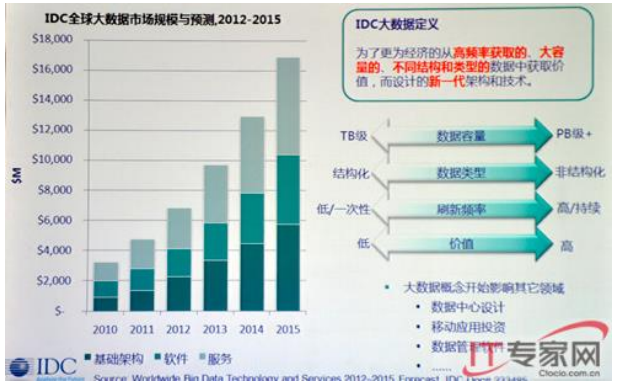
纲要



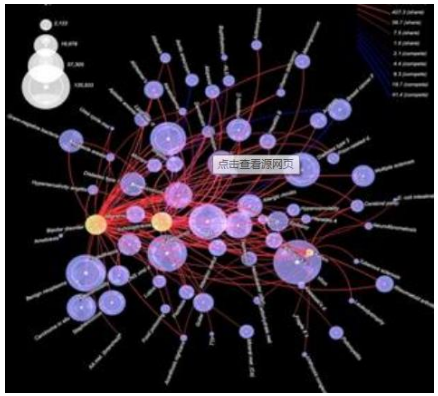
分类和预测

- 分类和预测是两种数据分析形式，用于提取描述重要数据类或预测未来的数据趋势 的模型
- 应用：欺诈检测、市场定位、性能预测、医疗诊断

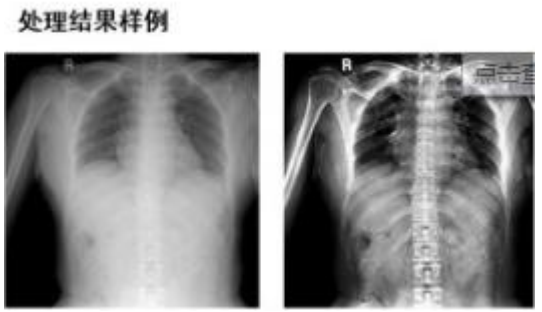
市场预测



X线影像数据挖掘



CT影像数据挖掘



运用数据挖掘技术检测金融欺诈行为

Kohonen神经网络算法在电信欺诈预测中的研究

罗格斯大学商学院　肖可砾　熊辉

分类和预测

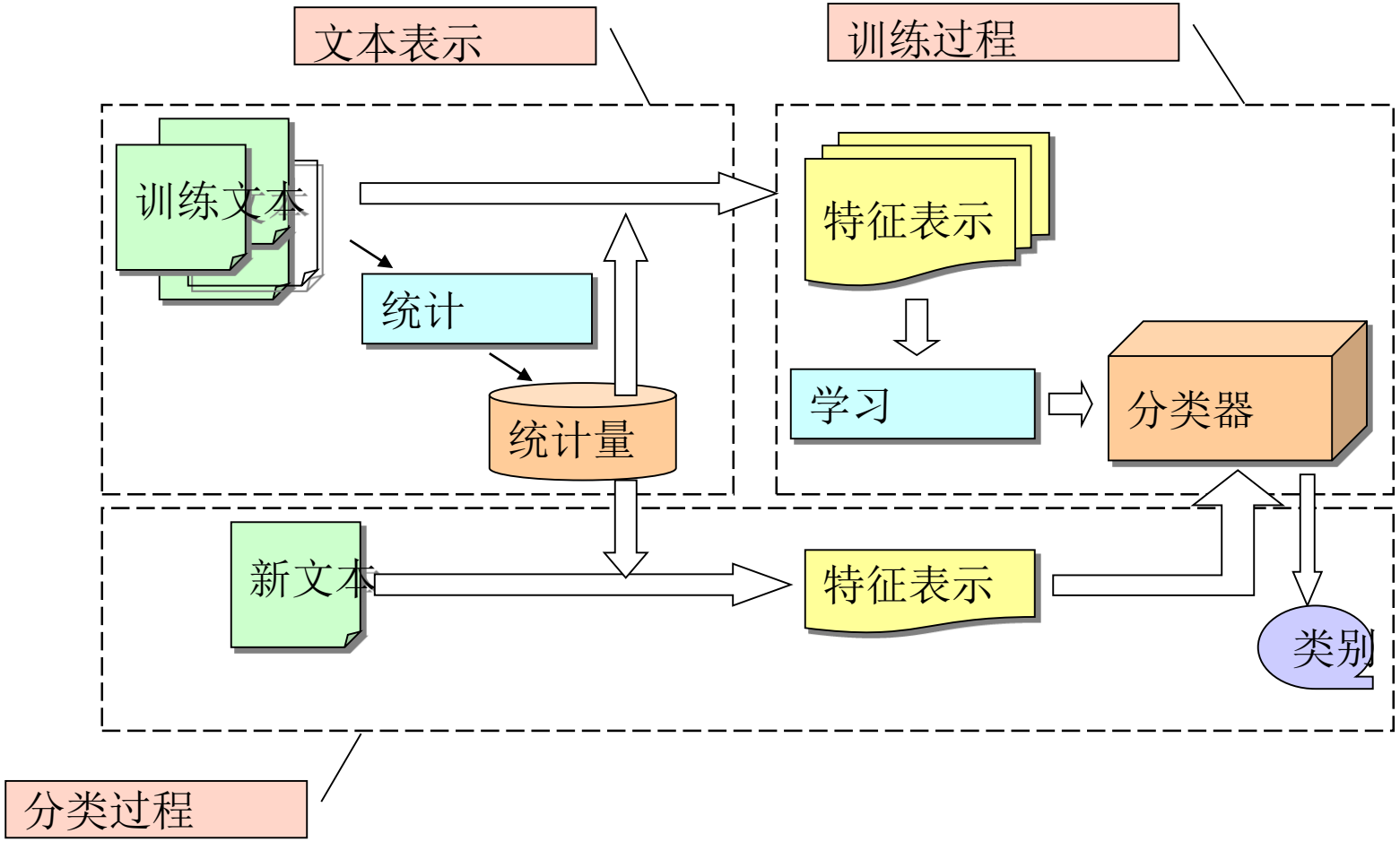
- 分类和预测是两种数据分析形式，用于提取描述重要数据类或预测未来的数据趋势 的模型
  - 分类：
    - 预测类对象的分类标号（或离散值）
    - 根据训练数据集和类标号属性，构建模型来分类现有数据，并用来分类新数据
  - 预测：
    - 建立连续函数值模型
    - 比如预测空缺值，或者预测顾客在计算机设备上的花费
- 典型应用
  - 欺诈检测、市场定位、性能预测、医疗诊断
- 常用方法
  - 决策树、贝叶斯、K近邻、SVM、神经网络、回归模型

分类和预测步骤

- 第一步，也成为**学习步**，目标是建立描述预先定义的数据类或概念集的分类器
  - 分类算法通过分析或从训练集“学习”来构造分类器。
  - 训练集由数据库元组（用n维属性向量表示）和他们相对应的类编号组成；假定每个元组属于一个预定义的类。
    - **训练元组**：训练数据集中的单个元组
  - 学习模型可以用分类规则、决策树或数学公式的形式提供。
- 第二步，使用模型，对将来的或未知的对象进行分类
  - 首先评估模型的预测准确率。
    - 对每个测试样本，将已知的类标号和该样本的学习模型预测比较
    - 模型在给定测试集上的准确率是正确被模型分类的测试样本的百分比
    - 测试集要独立于训练样本集，否则会出现“过分拟合”的情况

文本分类步骤

情感分类、新闻分类



文本分类器设计

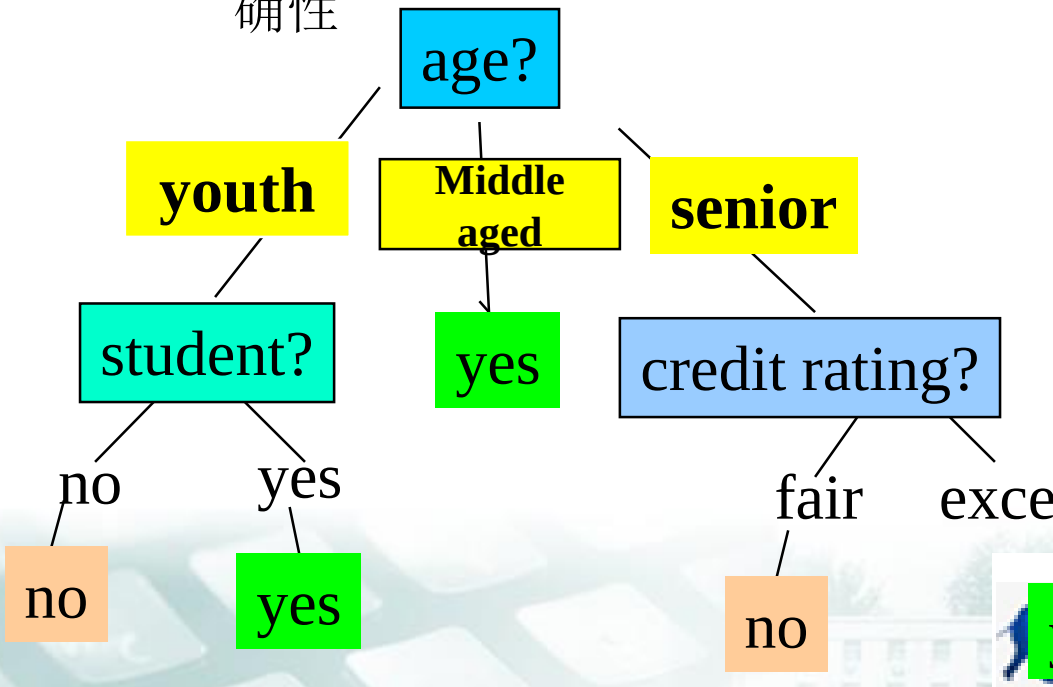
- 文本分类的方法大部分来自于模式分类，基本上可以分为三大类：
  - 一种是基于统计的方法，如Naïve Bayes，KNN、类中心向量、回归模型、支持向量机、最大熵模型等方法
  - 另一种是基于连接的方法，即人工神经网络
  - 还有一种是基于规则的方法，如决策树、关联规则等，这些方法的主要区别在于规则获取方法

分类属性选择度量

- 属性选择度量是一种选择分裂准则，将给定类标号的训练元组最好的进行划分的方法
  - 理想情况，每个划分都是“纯”的，即落在给定划分内的元组都属于相同的类
  - 属性选择度量又称为分裂准则
- 常用的属性选择度量
  - 信息增益
  - 增益率
  - Gini指标

决策树归纳分类

- 第一步：利用训练集生成决策树
  - 决策树构建
    - 使用属性选择度量来选择将元组最好的划分为不同的类的属性
    - 递归的通过选定的属性，来划分样本 。
  - 树剪枝
    - 决策树建立时，许多分枝反映的是训练数据中的噪声和离群点点，树剪枝试图识别并剪去这种分枝，以提高对未知数据分类的准确性

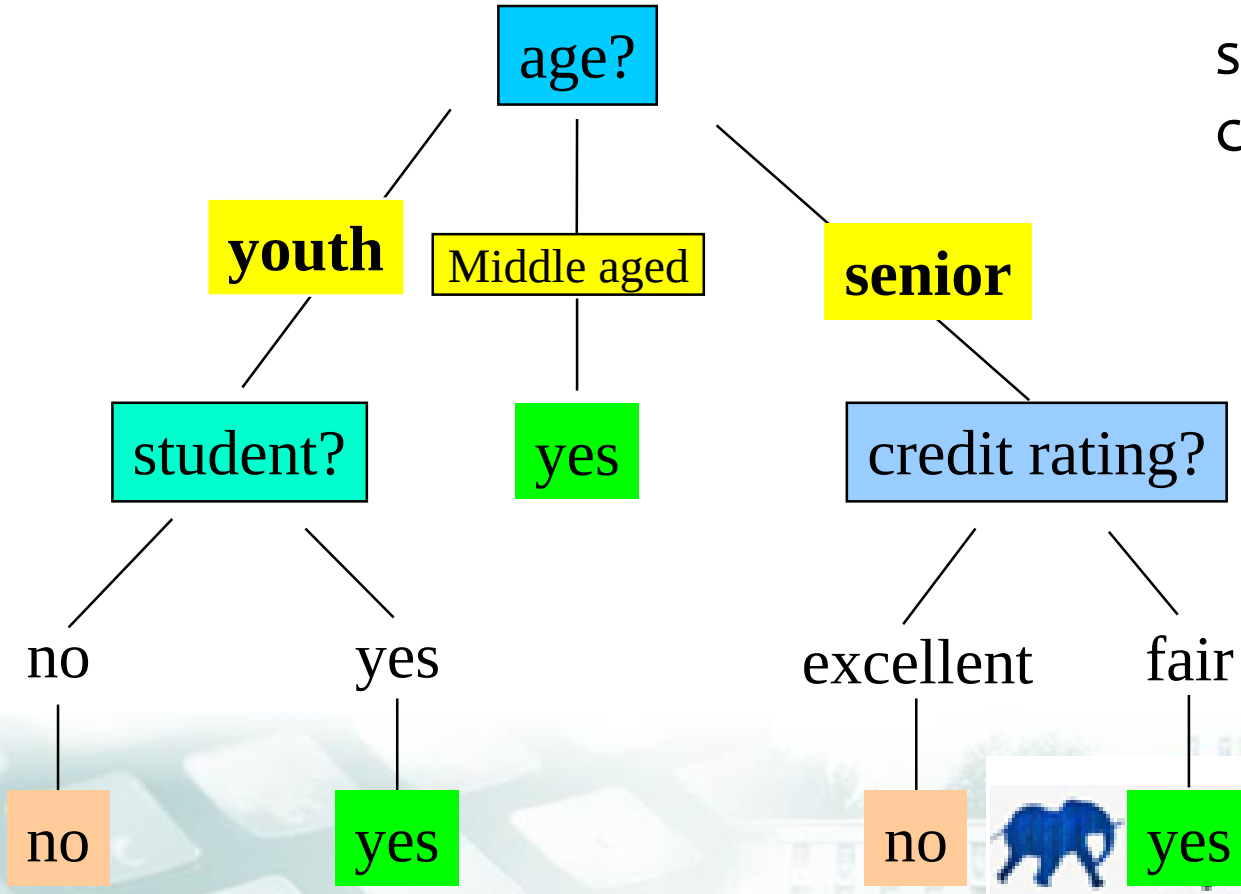


决策树：
Buys_computer

决策树归纳分类

- 第二步：对类别未知元祖进行分类
  - 给定一个类标号未知的元组X，在决策树上测试元组的属性值，跟踪一条由根到叶节点的路径，叶节点存放该元组的类预测。
  - 决策树容易转换为分类规则

^o^ = < age=youth,
student=yes,
credit rating=none >



由决策树提取分类规则

- 可以提取决策树表示的知识，并以IF-THEN形式的分类规则表示
- 对从根到树叶的每条路径创建一个规则
- 沿着给定路径上的每个属性-值对形成规则前件（"IF"部分）的一个合取项
- 叶节点包含类预测，形成规则后件（"THEN"部分）
- IF-THEN规则易于理解，尤其树很大时
- 示例：
  - IF age = “youth” AND student = “no” THEN buys_computer = “no”
  - IF age = “youth” AND student = “yes” THEN buys_computer = “yes”
  - IF age = “middle_aged” THEN buys_computer = “yes”
  - IF age = “senior” AND credit_rating = “excellent” THEN buys_computer = “yes”
  - IF age = “senior” AND credit_rating = “fair” THEN buys_computer = “no”

ID3决策树算法

特征选择指标：信息增益

信息增益可以定义为样本按照某属性划分时造成熵减少的期望，可以区分训练样本中正负样本的能力，其计算公式是

$$Gain(S,A) = Entropy(S) - \sum_{v \in V(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

- V(A)是属性A的值域
- S是样本集合
- S_v 是S中在属性A上值等于v的样本集合

ID3算法实例

小明问题：
今天阳光不错，气温凉爽，PM2.5正常，但是风比较大，小明会不会出去打篮球呢?

	A	B	C	D	E
1	Outlook	Temperature	PM2.5	Windy	Class
2	sunny	hot	high	weak	N
3	sunny	hot	high	strong	N
4	overcast	hot	high	weak	P
5	rain	mild	high	strong	P
6	rain	cool	normal	weak	P
7	rain	cool	normal	strong	N
8	overcast	cool	normal	strong	P
9	sunny	mild	high	weak	N
10	sunny	cool	normal	weak	P
11	rain	mild	normal	weak	P
12	sunny	mild	normal	strong	P
13	overcast	mild	high	strong	P
14	overcast	hot	normal	weak	P
15	rain	mild	high	true	N

计算信息增益

$$Values(Wind) = Weak, Strong$$

$$S = [9+, 5-]$$

$$S_{Weak} \leftarrow [6+, 2-]$$

$$S_{Strong} \leftarrow [3+, 3-]$$

$$Gain(S, Wind) = Entroy(S) - \sum_{v \in \{Weak, Strong\}} \frac{|S_v|}{|S|} Entropy$$

$$= Entropy(S) - (8/14)Entropy(S_{Weak}) - (6/14)Entropy(S_{Strong})$$

$$= 0.949 - (8/14)0.811 - (6/14)1.00$$

$$= 0.048$$

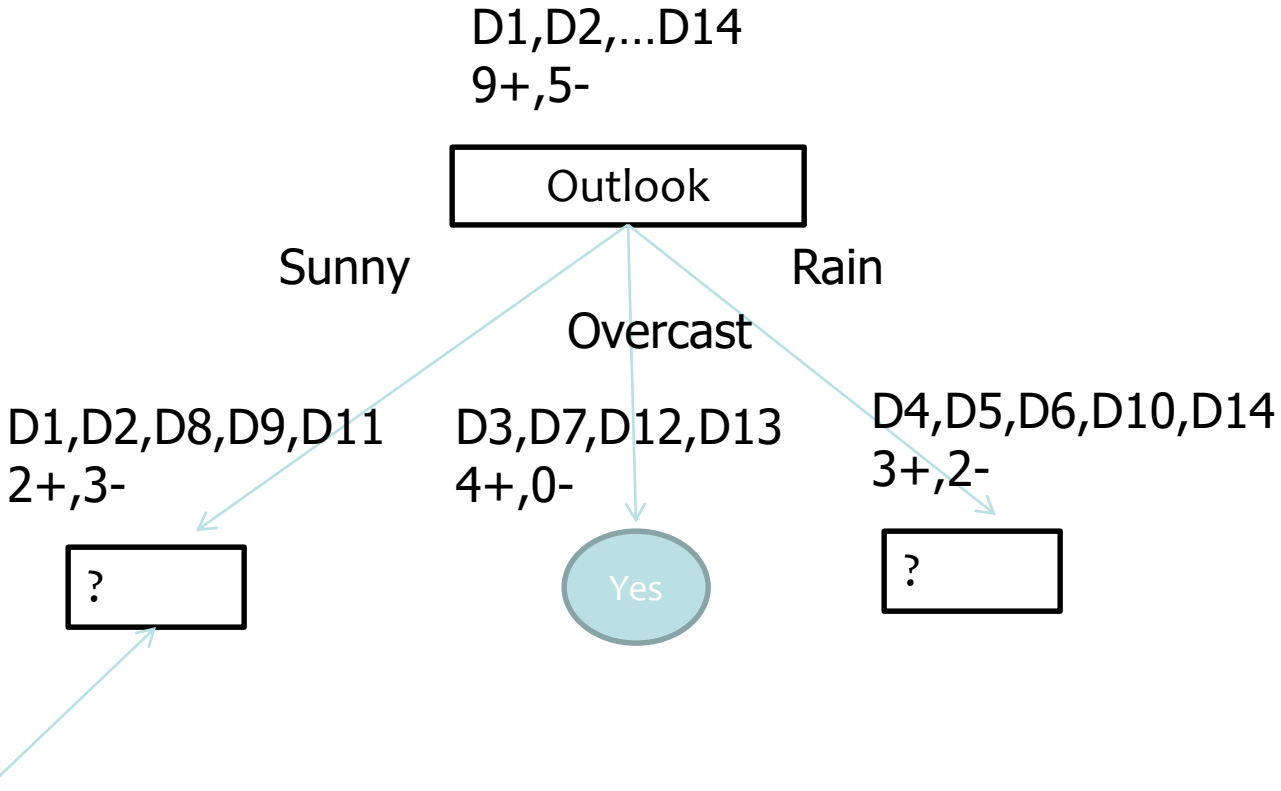
PS: Entropy(X)=-∑p(xi)log(2,p(xi)) (i=1,2,..n)

ID3算法实例

不同属性的信息增益

- 计算各属性的熵值
  - Gain(S,Outlook)=0.246
  - Gain(S,PM2.5)=0.151
  - Gain(S,Wind)=0.048
  - Gain(S,Temperature)=0.029
- 可以看到，Outlook得信息增益最大

ID3算法实例

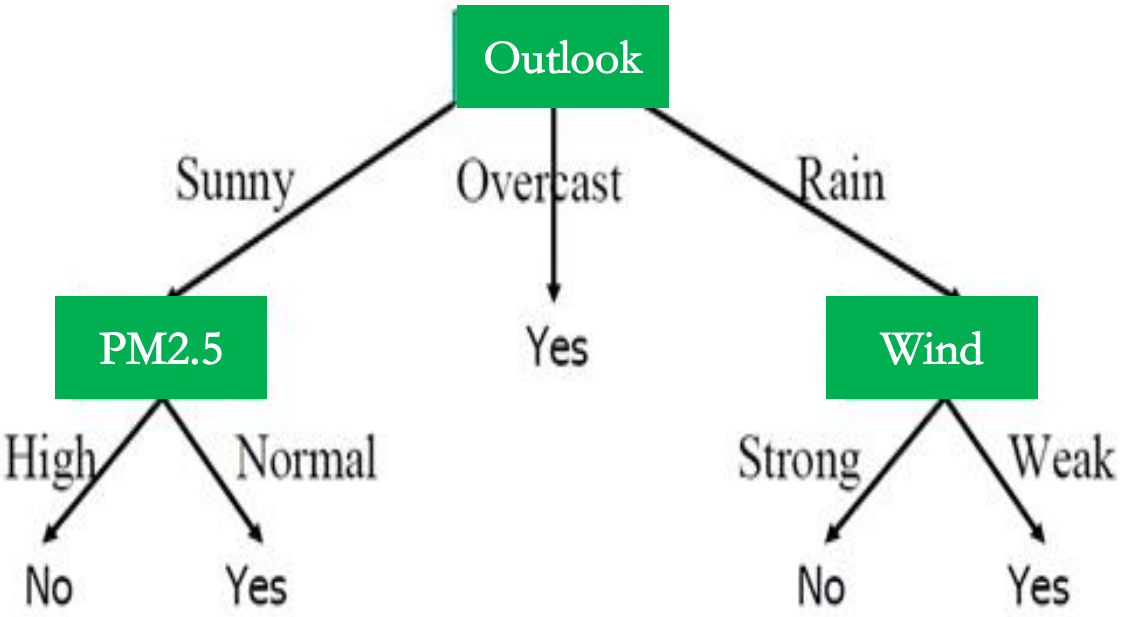


哪一个属性在这里被测试？

S_{sunny}={D1,D2,D8,D9,D11}
Gain（S_{sunny},PM2.5）=0.970-(3/5)0.0-(2/5)0.0=0.970
Gain(S_{sunny}, Temperature)=0.970-(2/5)0.0-(2/5)1.0-(1/5)0.0=0.570
Gain(S_{sunny}, Wind)=0.970-(2/5)1.0-(3/5)0.918=0.019

ID3算法实例

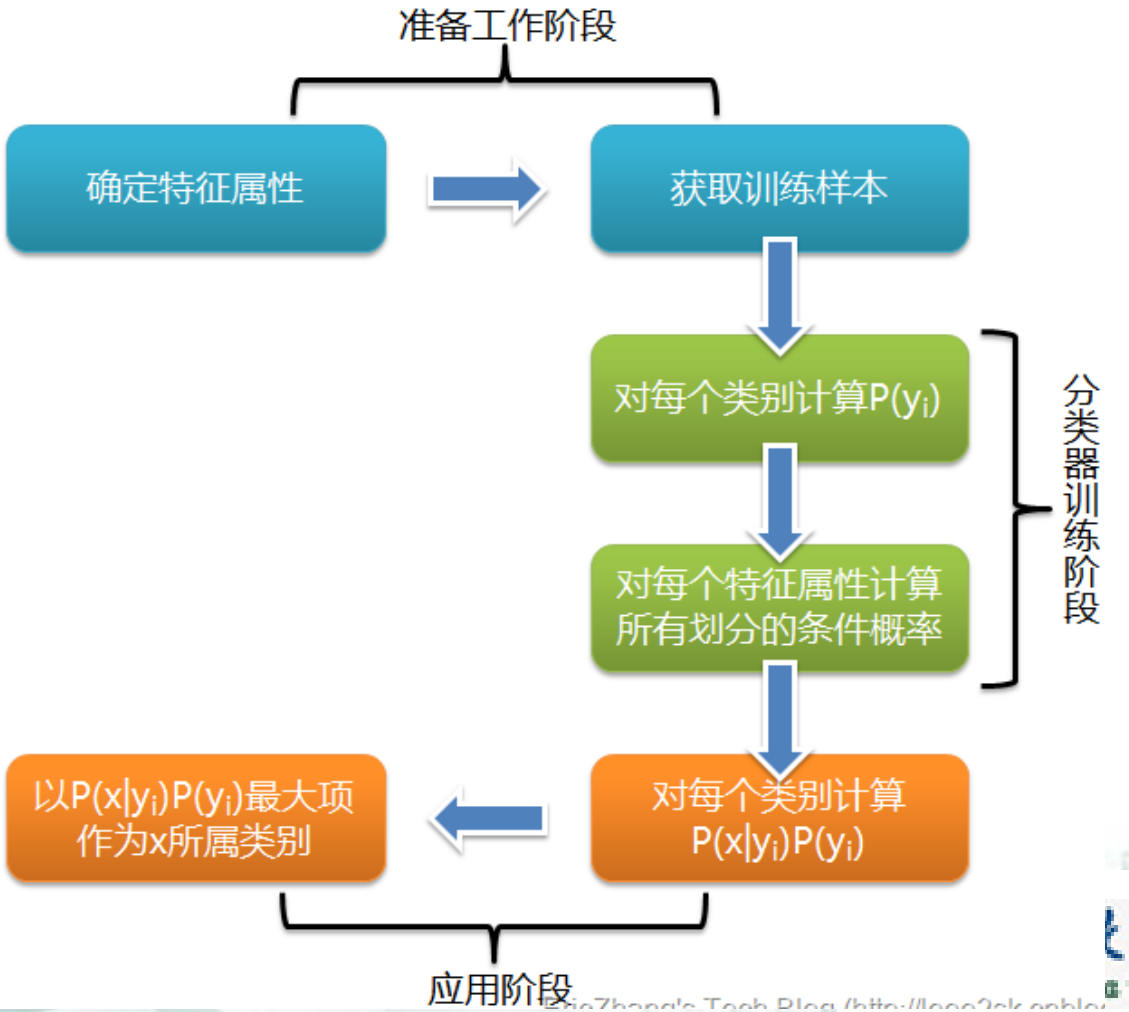
最终得到的决策树



有了决策树后，就可以根据气候条件做预测了:
例如如果气候数据是{Sunny,Cool,Normal,Strong} ,根据决策树到左侧的yes叶节点，可以判定属于P。

贝叶斯分类器

- 数据挖掘中以贝叶斯定理为基础，用于分类的技术有朴素贝叶斯分类和贝叶斯信念网络两种。



贝叶斯分类器

贝叶斯定理：

$$P(C_i|X)=\frac{P(C_i\cap X)}{P(X)}=\frac{P(C_i)\times P(X|C_i)}{P(X)}$$

X = <头发=长， 身高=170-175， 体重=50-60，专业=法律>

C_1 = 男　　C_2 = 女

$P(C_i)$表示先验概率(Prior probability)。

$P(C_i|X)$表示后验概率(Posteriori probability)，

先验概率是由以往的数据分析得到的。根据样本数据得到更多的信息后，对其重新修正，即是后验概率。

贝叶斯分类器

贝叶斯分类过程：

1、每个数据样本用一个n维特征向量$X=\{x_1, x_2, \ldots, x_n\}$表示，分别描述对$n$个属性$A_1, A_2, \ldots, A_n$样本的$n$个度量。

X = <头发=长， 身高=170-175， 体重=50-60，专业=法律>

C_1 = 男　　C_2 = 女

贝叶斯分类器

贝叶斯分类过程：

2、假定有m个类$C_1, C_2, \ldots, C_m$。给定一个未知的数据样本X，分类法将预测X属于具有最高后验概率（条件X下）的类。即是说，朴素贝叶斯分类将未知的样本分配给类C_i，当且仅当

X = <头发=长， 身高=170-175， 体重=50-60， 专业=法律>

C_1 = 男　　C_2 = 女

$$P(C_i|X) > P(C_j|X), 1 \le j \le m, j \ne i$$

根据贝叶斯定理

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

因此，由于$P(X)$对于所有类为常数，只需要$P(X|C_i)P(C_i)$最大即可。

贝叶斯分类器

贝叶斯分类过程：

X = <头发=长， 身高=170-175， 体重=50-60，专业=法律>

C_1 = 男　　C_2 = 女

3、假定属性值相互条件独立，即在属性间不存在依赖关系，这样，

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

概率$P(x_k|C_i)$可以由训练样本估值，其中

（1）如果A_k是离散型属性，则$P(x_k|C_i) = s_{ik}/s_i$，

其中s_{ik}是在属性A_k上具有x_k的类C_i的训练样本数，而s_i是C_i中的训练样本数。

贝叶斯分类器

贝叶斯分类过程：

X = <头发=长， 身高=170-175， 体重=50-60，专业=法律>

C_1 = 男　　C_2 = 女

（2）如果A_k是连续型属性，则通常假定该属性服从高斯分布。因而，

$$P(x_k|C_i)=g(x_k,\mu_{C_i},\sigma_{C_i})=\frac{1}{\sqrt{2\pi\sigma_{C_i}}}e\frac{(x_k-\mu_{C_i})^2}{2\sigma_{C_i}^2}$$

其中，给定类C_i的训练样本属性A_k的值，$g(x_k,\mu_{C_i},\sigma_{C_i})$是属性$A_k$的高斯密度函数。

贝叶斯分类器

贝叶斯分类过程：

$\mathbf{X}$ = <头发=长， 身高=170-175， 体重=50-60，专业=法律>

C_1 = 男　　C_2 = 女

4、为对未知样本X分类，对每个类C_i，计算$P(X|C_i)P(C_i)$。样本被指派到类C_i，当且仅当

$$P(X|C_i)P(C_i) > P(X|C_j)P(C_j)$$

即是说，X被指派到$P(X|C_i)P(C_i)$最大的类。

贝叶斯分类器

优点：

- 计算速度最快的演算法；
- 规则清楚易懂；
- 独立事件的假设，大多数问题上不至于发生太大偏误；

缺点：

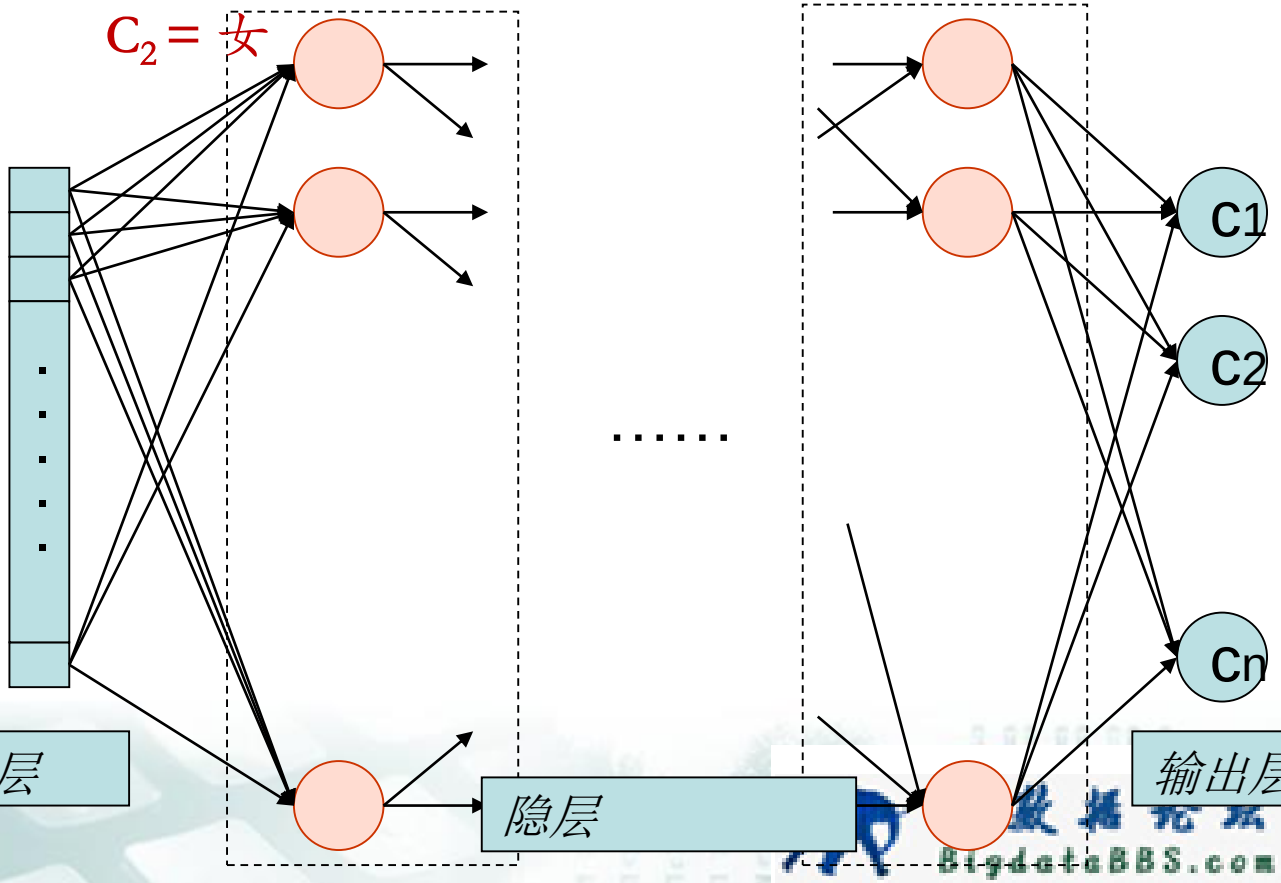
- 仅适用于类别变量；
- 仅能应用于分类问题；
- 假设变量间为独立互不影响，因此使用时需要谨慎分析变量间的相关性。

神经网络分类

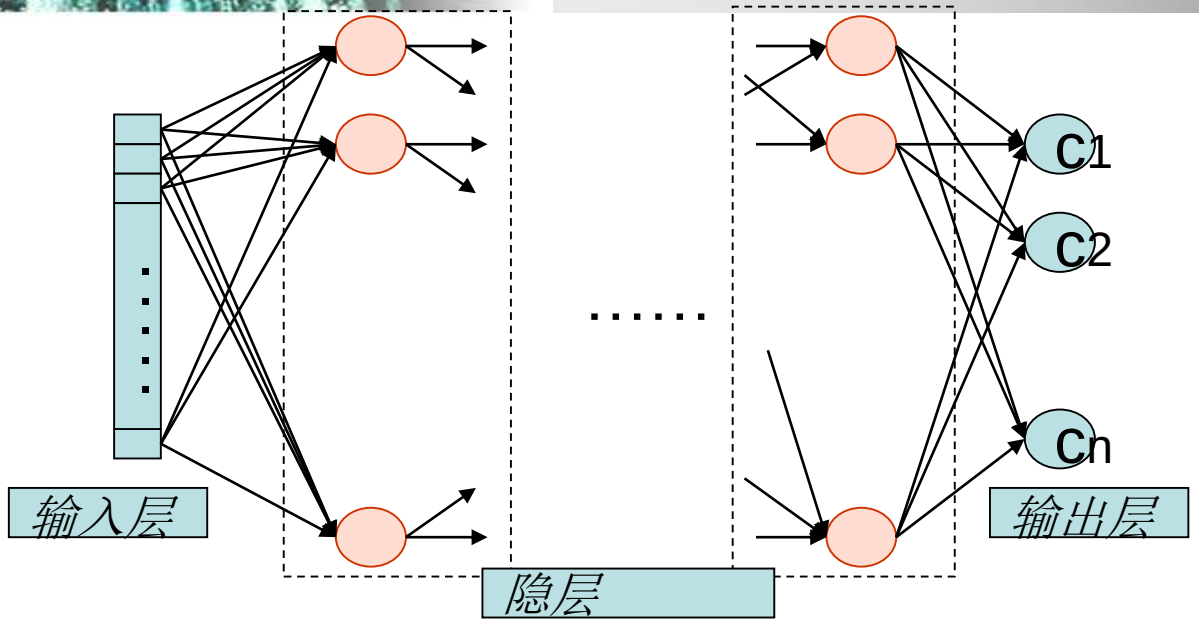
- 利用后向传播的神经网络学习算法
  - 神经网络是一组连接的输入/输出单元，每个连接都与一个权相连。在学习阶段，通过调整神经网络的权，使得能够预测输入样本的正确标号来学习。

X = <头发=长， 身高=170-175， 体重=50-60， 专业=法律>

C_1 = 男　　C_2 = 女



神经网络分类



X = <头发=长，
身高=170-175，
体重=50-60，
专业=法律 >

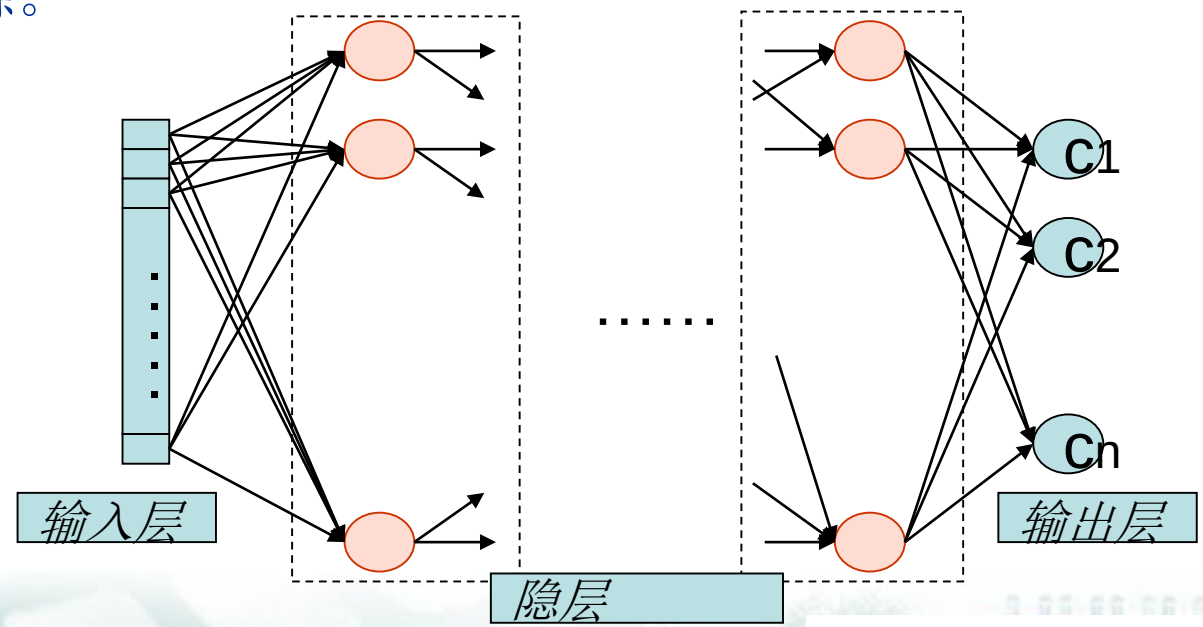
C_1 = 男　　C_2 = 女

BP神经网络模型拓扑结构包括输入层、隐层和输出层。输入层神经元的个数由样本属性的维度决定，输出层神经元个数由样本分类数决定。隐藏层的层数和每层的神经元个数由用户指定。网络中的弧线表示前一层神经元和后一层神经元之间的权值。每个神经元都有输入和输出。输入层的输入和输出都是训练样本的属性值。

神经网络分类

算法基本流程：

1. 初始化网络权值和神经元的阈值
2. 前向传播：按照公式一层一层的计算隐层神经元和输出层神经元的输入和输出。
3. 后向传播：根据公式修正权值和阈值
4. 若不满足终止条件，则回到第二步；满足终止条件，则完成训练。



神经网络分类

X = <头发=长，
身高=170-175，
体重=50-60，
专业=法律 >

C_1 = 男　　C_2 = 女

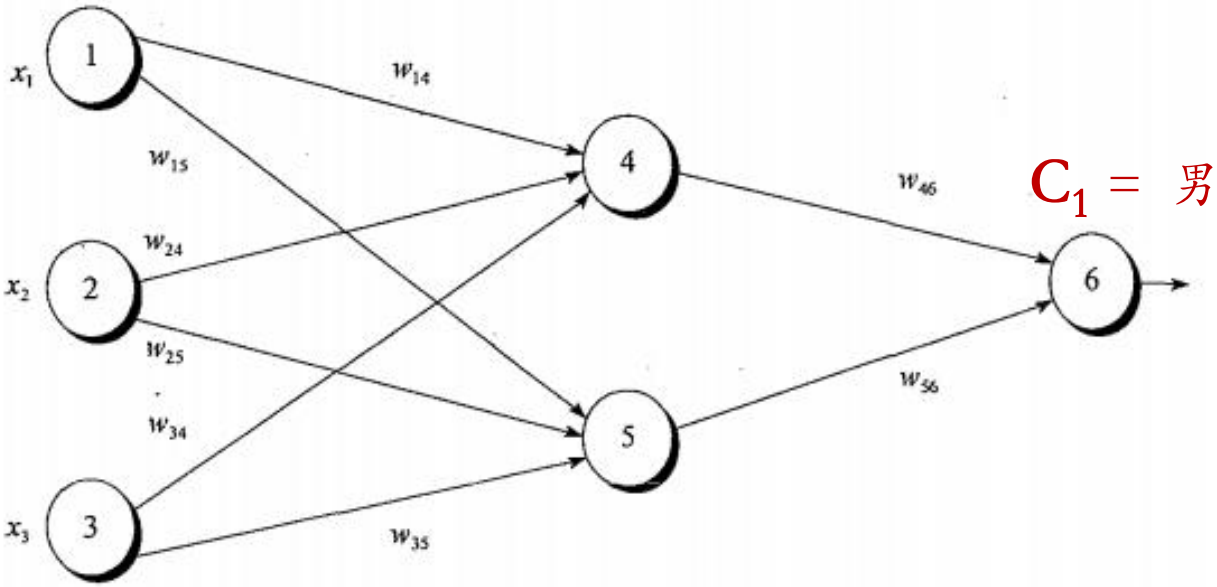


图 9.5　多层前馈神经网络的一个例子

表 9.1　初始输入、权重和偏倚值

x_1	x_2	x_3	w_{14}	w_{15}	w_{24}	w_{25}	w_{34}	w_{35}	w_{46}	w_{56}	θ_4	θ_5	θ_6
1	0	1	0.2	−0.3	0.4	0.1	−0.5	0.2	−0.3	−0.2	−0.4	0.2	0.1

神经网络分类

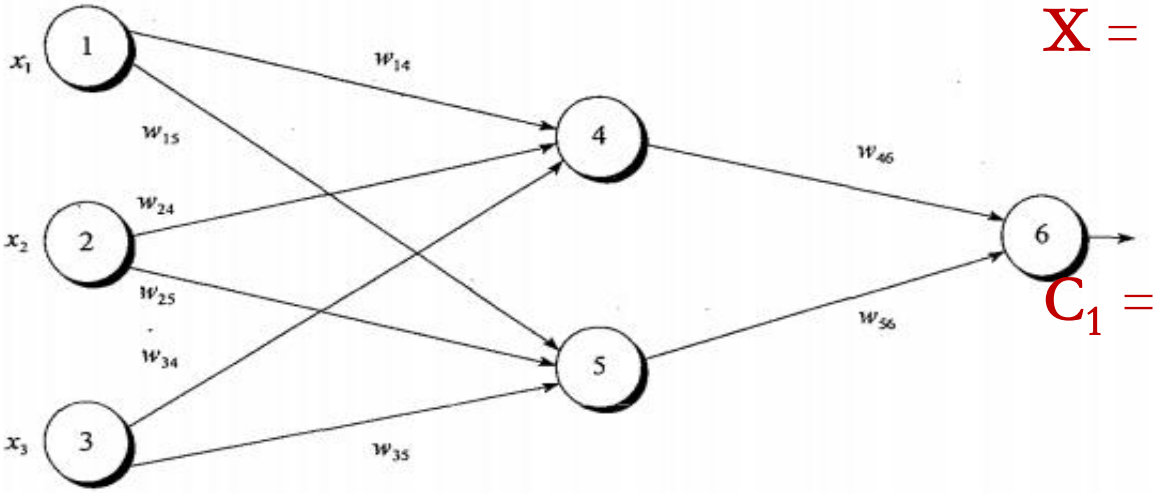


图 9.5　多层前馈神经网络的一个例子

X = <头发=长，
身高=170-175，
体重=50-60，
专业=法律 >

C_1 = 男　　C_2 = 女

表 9.1　初始输入、权重和偏倚值

x_1	x_2	x_3	w_{14}	w_{15}	w_{24}	w_{25}	w_{34}	w_{35}	w_{46}	w_{56}	θ_4	θ_5	θ_6
1	0	1	0.2	−0.3	0.4	0.1	−0.5	0.2	−0.3	−0.2	−0.4	0.2	0.1

表 9.2　净输入和输出的计算

单元 j	净输入 I_j	输出 O_j
4	$0.2+0-0.5-0.4=-0.7$	$1+(1+e^{0.7})=0.33$
5	$-0.3+0+0.2+0.2=0.1$	$1+(1+e^{-0.1})=0.525$
6	$-(0.3)(0.332)-(0.2)(0.525)+0.1=-0.105$	$1+(1+e^{0.105})=0.474$

神经网络分类

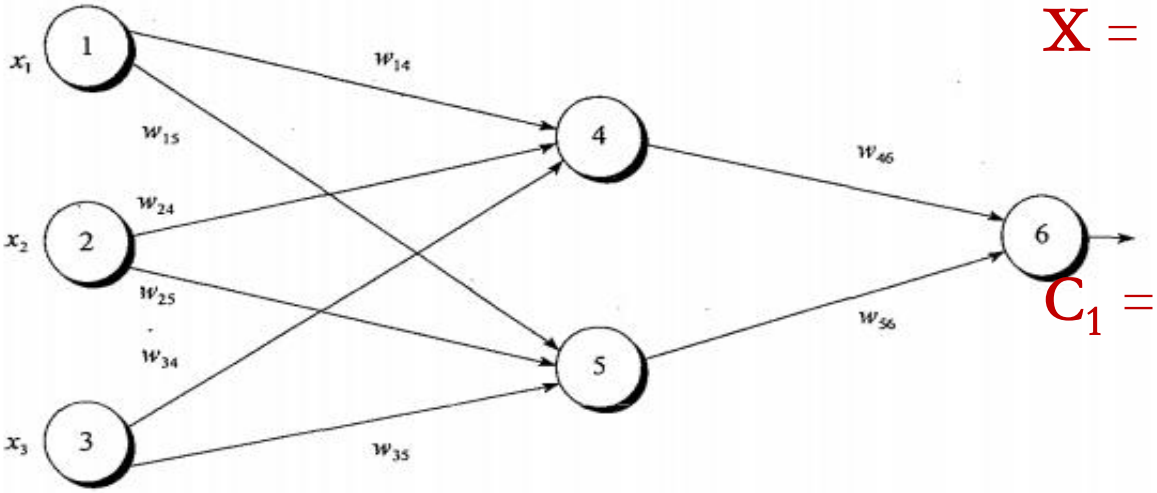


图 9.5　多层前馈神经网络的一个例子

X = <头发=长，
身高=170-175，
体重=50-60，
专业=法律 >

C_1 = 男　　C_2 = 女

表 9.1　初始输入、权重和偏倚值

x_1	x_2	x_3	w_{14}	w_{15}	w_{24}	w_{25}	w_{34}	w_{35}	w_{46}	w_{56}	θ_4	θ_5	θ_6
1	0	1	0.2	−0.3	0.4	0.1	−0.5	0.2	−0.3	−0.2	−0.4	0.2	0.1

表 9.2　净输入和输出的计算

单元 j	净输入 I_j	输出 O_j
4	$0.2+0-0.5-0.4=-0.7$	$1+(1+e^{0.7})=0.33$
5	$-0.3+0+0.2+0.2=0.1$	$1+(1+e^{-0.1})=0.525$
6	$-(0.3)(0.332)-(0.2)(0.525)+0.1=-0.105$	$1+(1+e^{0.105})=0.474$

神经网络分类

表 9.2　净输入和输出的计算

单元 j	净输入 I_j	输出 O_j
4	$0.2+0-0.5-0.4=-0.7$	$1+(1+e^{0.7})=0.33$
5	$-0.3+0+0.2+0.2=0.1$	$1+(1+e^{-0.1})=0.525$
6	$-(0.3)(0.332)-(0.2)(0.525)+0.1=-0.105$	$1+(1+e^{0.105})=0.474$

表 9.3　每个节点误差的计算

$$Err_j = O_j(1-O_j)(T_j-O_j)$$

单元 j	Err_j
6	$(0.474)(1-0.474)(1-0.474)=0.1311$
5	$(0.525)(1-0.525)(0.1311)(-0.2)=-0.0065$
4	$(0.332)(1-0.332)(0.1311)(-0.3)=-0.02087$

表 9.4　权重和偏倚更新的计算

$$Err_j = O_j(1-O_j)\sum_k Err_k w_{jk}$$

$$\Delta w_{ij} = (l)Err_j O_i$$

$$w_{ij} = w_{ij} + \Delta w_{ij}$$

权重或偏差	新值
w_{46}	$-0.3+(0.9)(0.1311)(0.332)=-0.261$
w_{56}	$-0.2+(0.9)(0.1311)(0.525)=-0.138$
w_{14}	$0.2+(0.9)(-0.0087)(1)=0.192$
w_{15}	$-0.3+(0.9)(-0.0065)(1)=-0.306$
w_{24}	$0.4+(0.9)(-0.0087)(0)=0.4$
w_{25}	$0.1+(0.9)(-0.0065)(0)=0.1$
w_{34}	$-0.5+(0.9)(-0.0087)(1)=-0.508$
w_{35}	$0.2+(0.9)(-0.0065)(1)=0.194$
θ_6	$0.1+(0.9)(0.1311)=0.218$
θ_5	$0.2+(0.9)(-0.0065)=0.194$
θ_4	$-0.4+(0.9)(-0.0087)=-0.408$

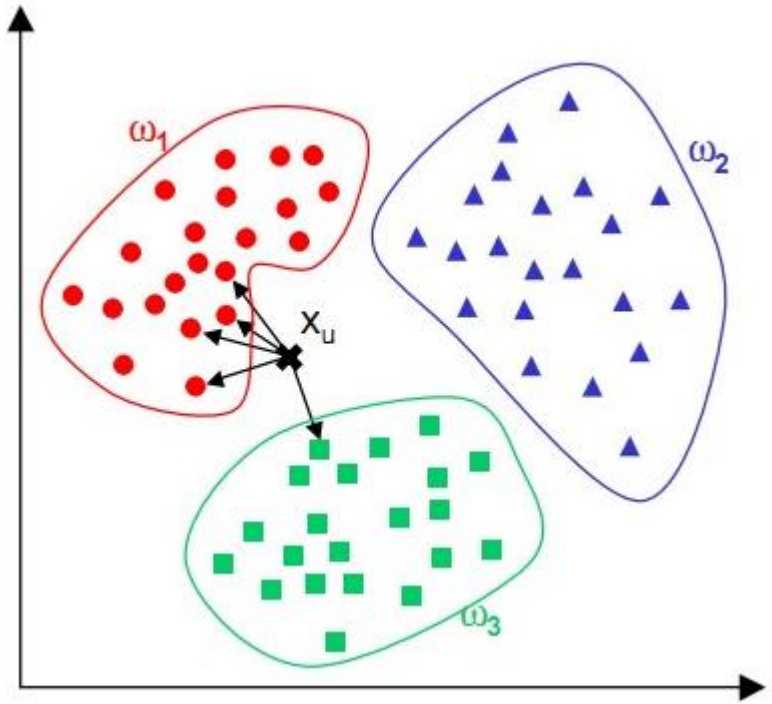
神经网络分类

- 利用后向传播的神经网络学习算法
  - 神经网络是一组连接的输入/输出单元，每个连接都与一个权相连。在学习阶段，通过调整神经网络的权，使得能够预测输入样本的正确标号来学习。
- 优点
  - 预测精度总的来说较高
  - 健壮性好，训练样本中包含错误时也可正常工作
  - 输出可能是离散值、连续值或者是离散或量化属性的向量值
  - 对目标进行分类较快
- 缺点
  - 训练（学习）时间长
  - 蕴涵在学习的权中的符号含义很难理解
  - 很难根专业领域知识相整合

KNN分类方法

k-最临近分类

- 给定一个未知样本，k-最临近分类法搜索模式空间，找出最接近未知样本的k个训练样本；然后使用k个最临近者中最公共的类来预测当前样本的类标号

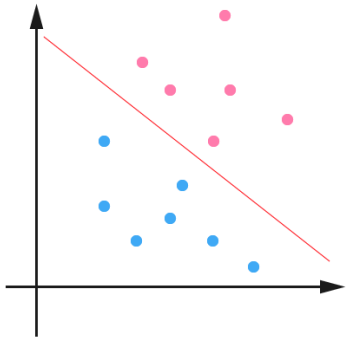


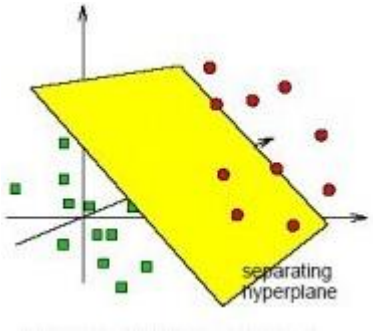
算法步骤：

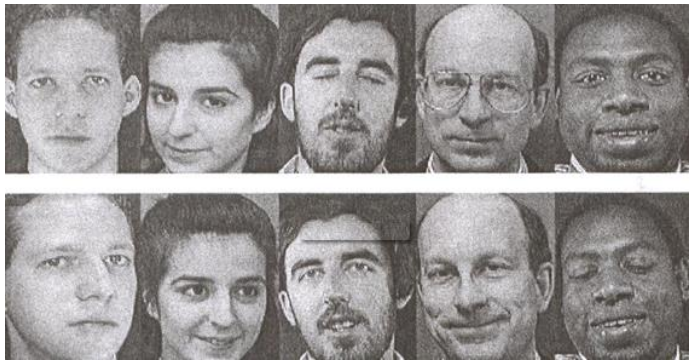
1. 初始化距离为最大值
2. 计算未知样本和每个训练样本的距离dist
3. 得到目前K个最临近样本中的最大距离maxdist
4. 如果dist小于maxdist，则将该训练样本作为K-最近邻样本
5. 重复步骤2、3、4，直到未知样本和所有训练样本的距离都算完
6. 统计K-最近邻样本中每个类标号出现的次数
7. 选择出现频率最大的类标号作为未知样本的类标号

支持向量积SVM分类

- 支持向量机（SVM）是90年代中期发展起来的基于统计学习理论的一种机器学习方法，通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。



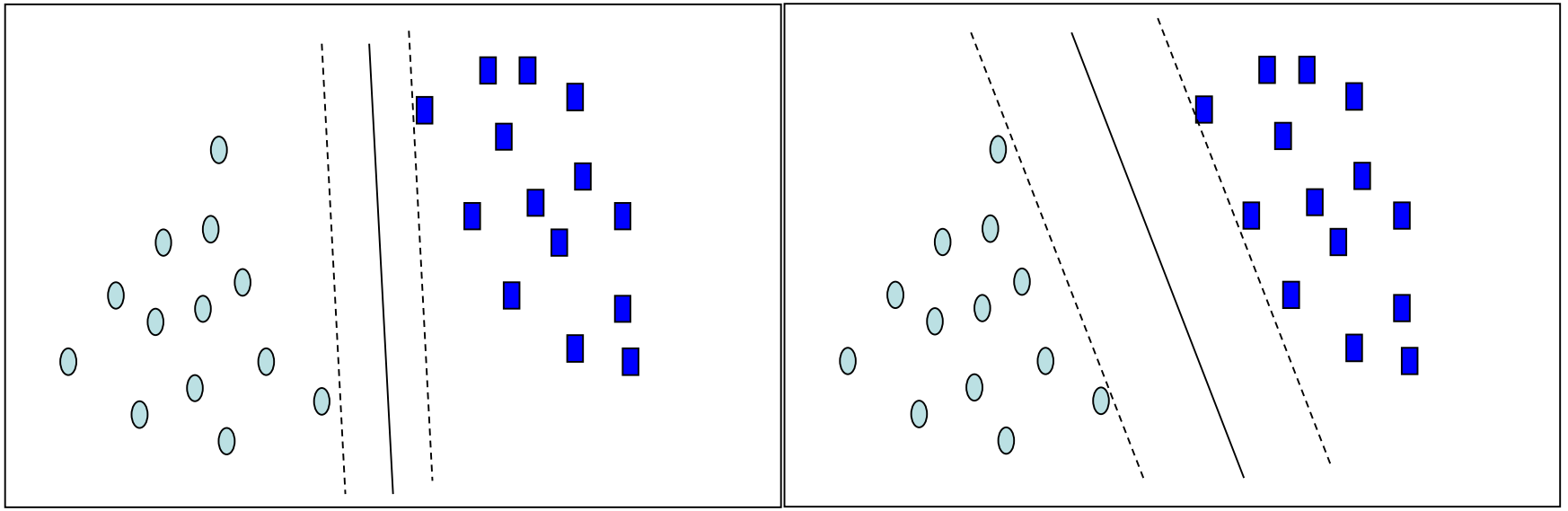




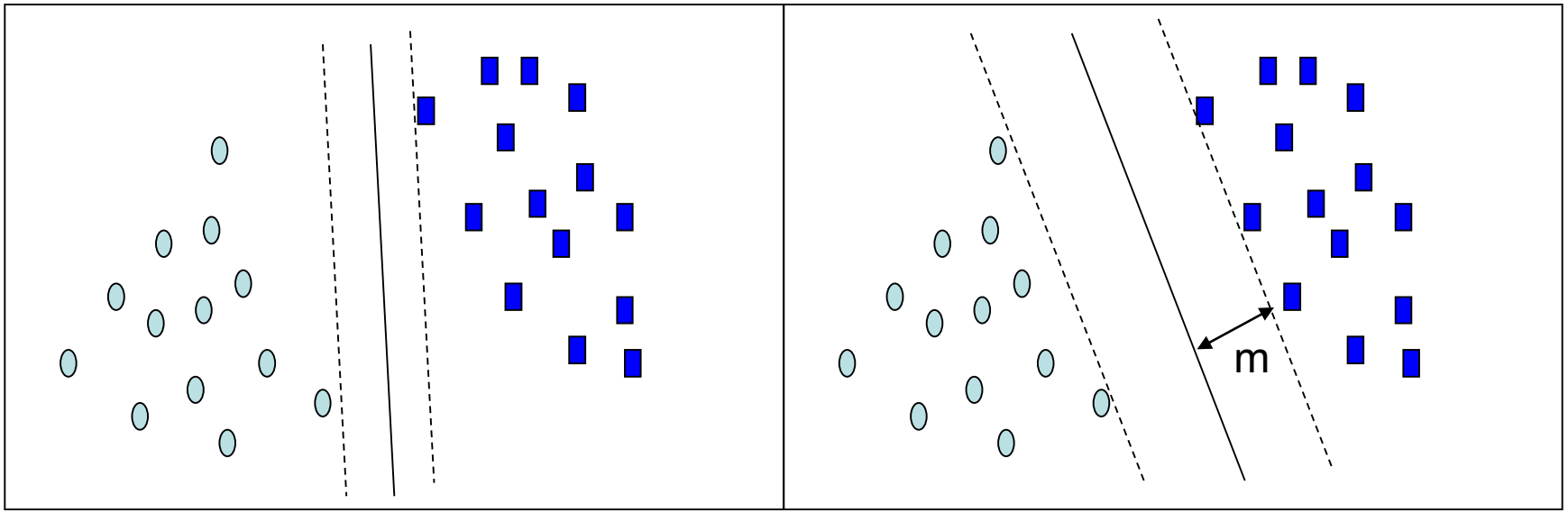
- 特点: 训练时间非常长，但对复杂的非线性决策边界的建模能力是高度准确的（使用最大边缘）
- 应用:
  - 手写数字识别，对象识别，语音识别, 以及基准时间序列预测检验。

支持向量积SVM分类

- 使用一种非线性的映射，将原训练数据映射到较高的维
- 一个数据被认为是p维向量，数据在这个p维向量空间中被分为两类；SVM的目的是找到一个p-1维的超平面，来划分p维向量空间的数据
  - 在新的维上，它搜索线性最佳分离超平面 (即将一类的元组与其他类分离的“决策边界”)。



支持向量积SVM分类



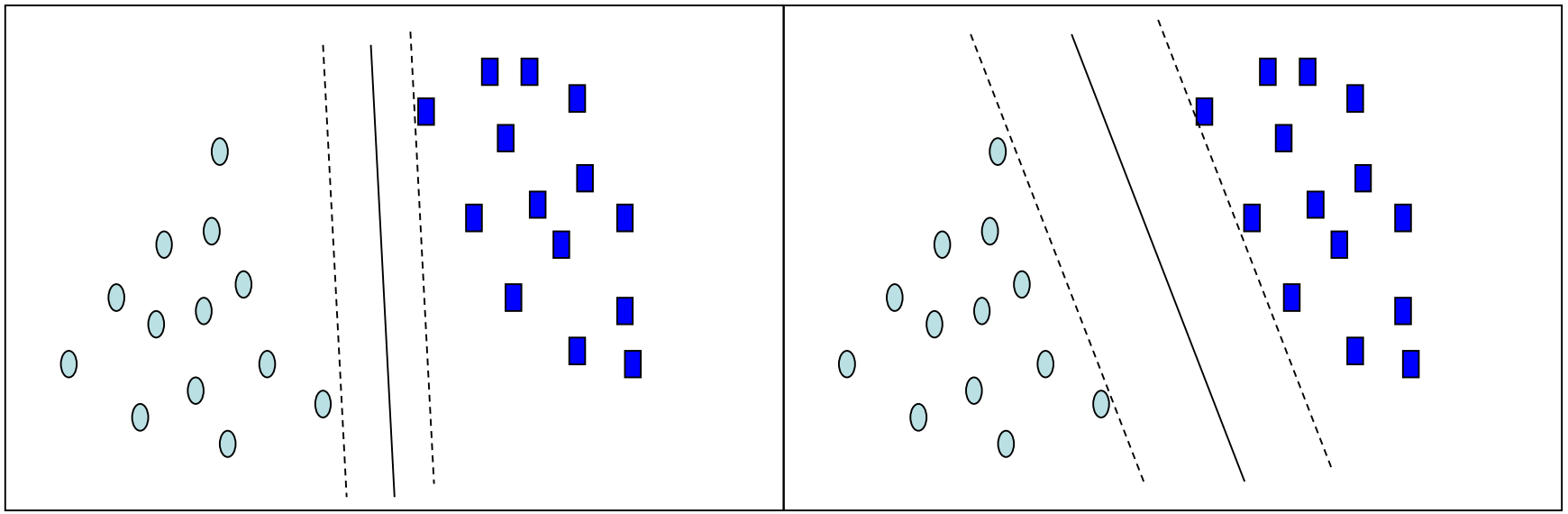
设给定的数据集 D 为 $(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_{|D|}, y_{|D|})$, 其中$\mathbf{X}_i$是训练元组，具有相关联的类标号yi。

可以画出无限多条分离直线（或超平面）将类+1的元组与类-1的元组分开，我们想找出“最好的”那一条 (对先前未见到的元组具有最小分类误差的那一条）。

SVM 要搜索具有最大边缘的超平面，即**最大边缘超平面** (MMH)

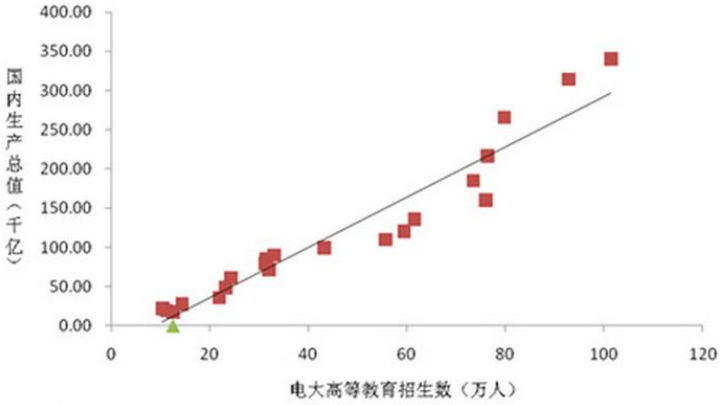
支持向量积SVM分类

- 对于非线性可分的数据，使用一个适当的对足够高维的非线性映射，两类的数据总可以被超平面分开。
- SVM 使用支持向量（“基本”训练元组）和边缘（由支持向量定义）发现该超平面

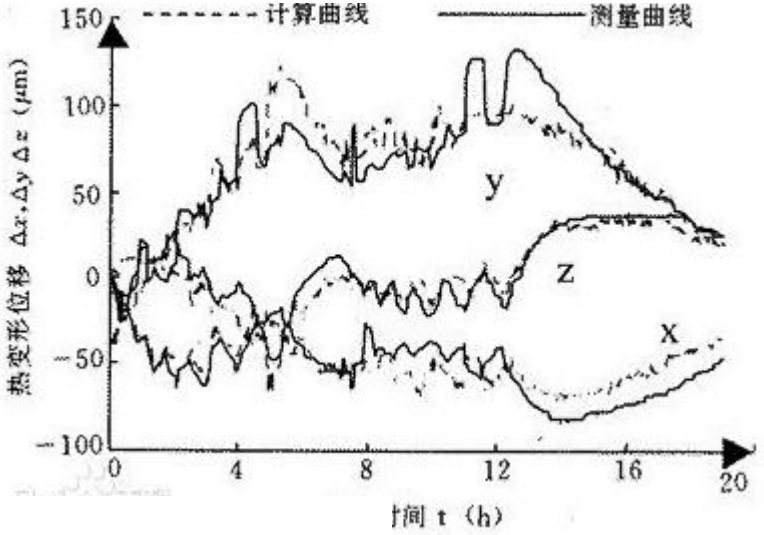


线性回归、多元回归和非线性回归

- 线性回归： $Y = \alpha + \beta X$



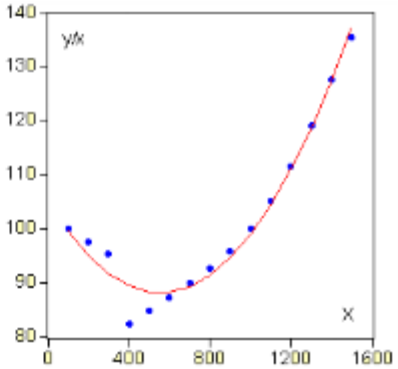
- 多元回归： $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$



- 非线性回归： $Y = \alpha + \beta_1 X_1 + \beta_2 X_2^2 + \beta_3 X_3^3$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	105.1552	2.475876	42.47191	0.0000
X	-0.061751	0.007121	-8.672084	0.0000
X^2	5.55E-05	4.33E-06	12.82378	0.0000

R-squared	0.972820	Mean dependent var	101.6315
Adjusted R-squared	0.968290	S.D. dependent var	15.60920
S.E. of regression	2.779592	Akaike info criterion	5.059342
Sum squared resid	92.71358	Schwarz criterion	5.200952
Log likelihood	-34.94506	F-statistic	214.7481
Durbin-Watson stat	1.862198	Prob(F-statistic)	0.000000



线性回归、多元回归和非线性回归

- 线性回归：$Y = \alpha + \beta X$
  - 其中α和β是回归系数，可以根据给定的数据点，通过最小二乘法来求得

$$\alpha = \bar{y} - \beta\bar{x} \qquad \beta = \frac{\sum_{i=1}^{S}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{S}(x_i - \bar{x})^2}$$

- 多元回归：$Y = \alpha + \beta_1 X_1 + \beta_2 X_2$
  - 线性回归的扩展，设计多个预测变量，可以用最小二乘法求得上式中的α，β_1和β_2
- 非线性回归：$Y = \alpha + \beta_1 X_1 + \beta_2 X_2^2 + \beta_3 X_3^3$
  - 对不呈线性依赖的数据建模
  - 使用多项式回归建模方法，然后进行变量变换，将非线性模型转换为线性模型，然后用最小二乘法求解

分类预测案例：客户流失预警模型

- 采用LOGIT回归模型对新入网用户在4个月内流失进行预警

建模变量

基本情况					语音通话情况				数据业务使用情况			
品牌	ARPU	余额	套餐名称	…	本地通话次数	长途通话次数	漫游通话次数	…	数据业务费用	数据业务消费占比	数据业务使用种类数	…

入网月份　　取数月份

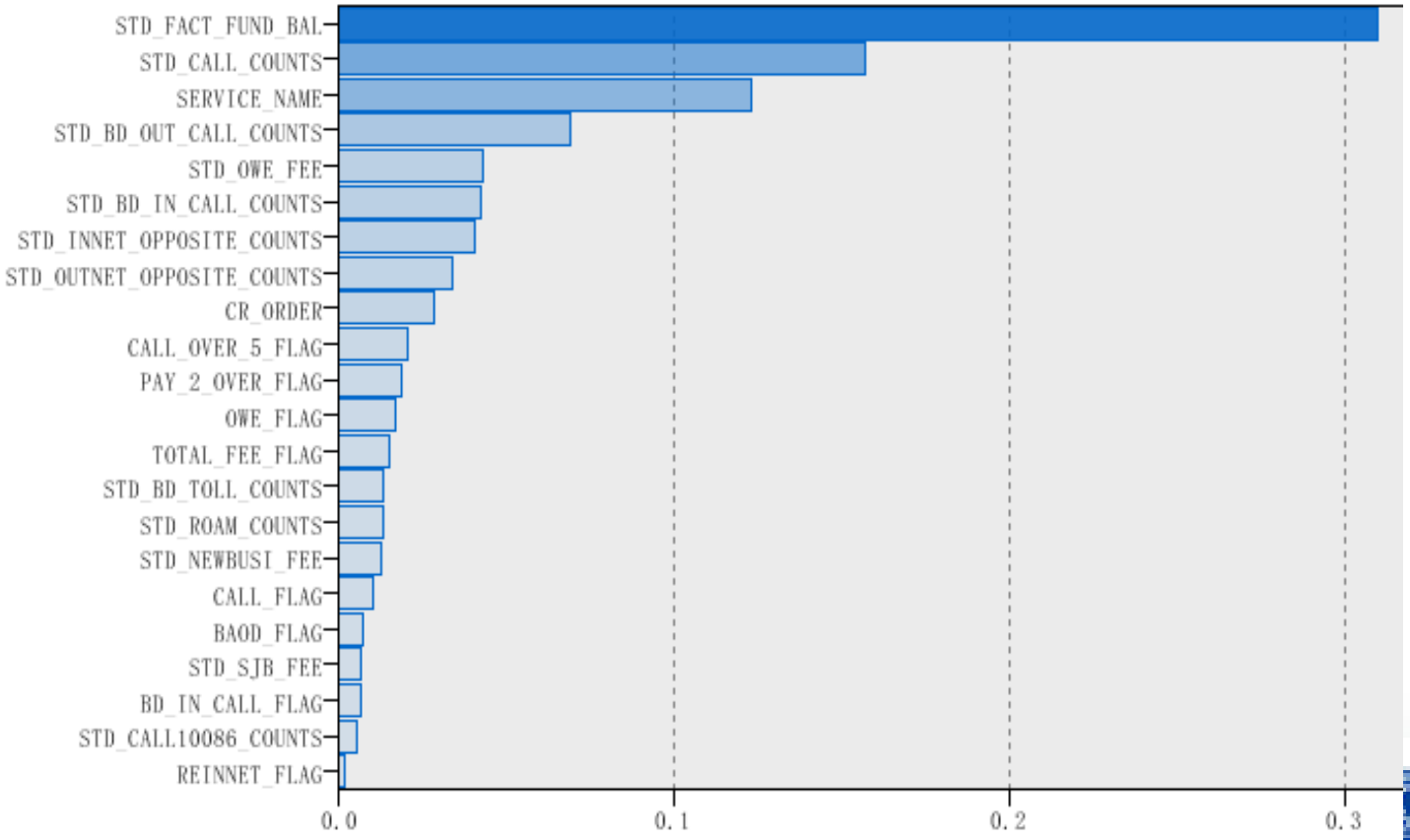
流失月份　　流失观察月份

	3月	4月	5月	6月	7月	8月	9月	10月	11月
建模时间窗口	3月	4月	5月	6月	7月	8月	9月	10月	11月
	3月	4月	5月	6月	7月	8月	9月	10月	11月
测试时间窗口	3月	4月	5月	6月	7月	8月	9月	10月	11月

客户流失预警模型

- 所有的数据经过对数变换和标准化变换，消除不同变量和不同月份对模型的影响
- 模型的结果为对数流失风险比的线性拟合表达式，应用模型时输入为用户当前在建模变量上的值，输出为用户的流失概率
- 下图展示了建模变量的重要性（调整后的实际建模变量）

余额
通话次数
套餐名称
本地被叫次数
欠费金额
本地主叫次数
网内通话对端数
网外通话对端数
彩铃订购标志
通话时长大于5分钟标志
多次充值标志
欠费标志
ARPU大于0标志
长途通话次数
漫游通话次数
新业务费
通话标志
包打套餐标志
手机报费用
本地主叫通话标志
拨打10086次数
重入网标志



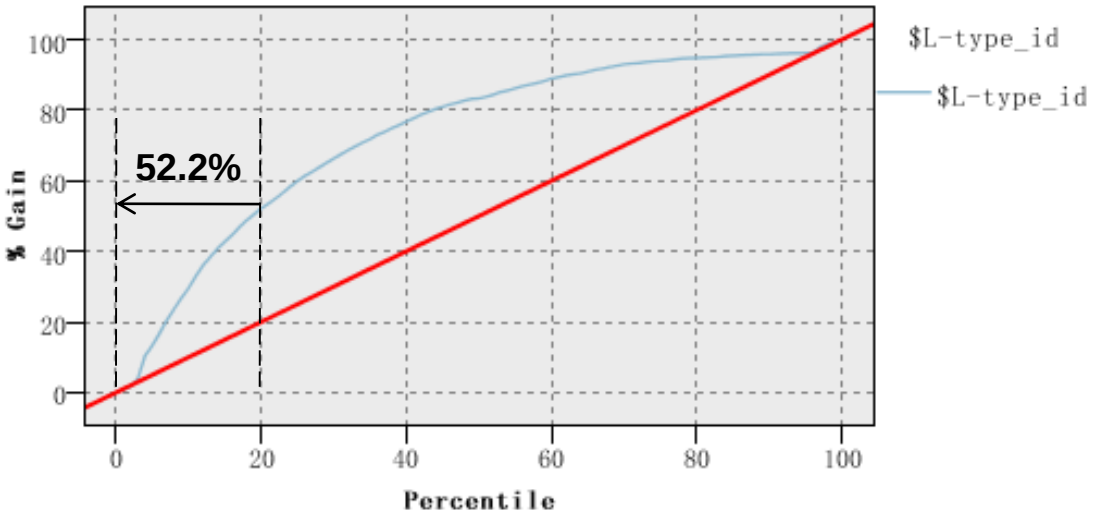
客户流失预警模型

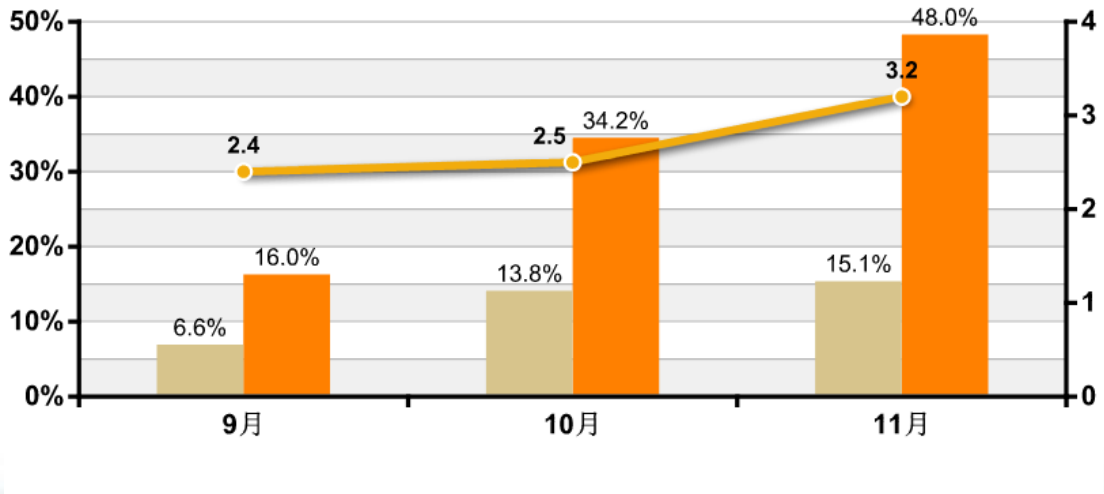
用户集：09年07月入网、08月仍然正常的用户；

模型筛选用户：通过模型筛选出的用户集中20%的用户，其中，

✓ **包含了用户集中52.2%的9月流失用户；**

✓**模型筛选用户9月、10月和11月的流失率（模型准确率）分别为16%、34%和48%，比用户集中用户的流失率（不使用模型筛选的准确率）提升度分别为2.4、2.5和3.2倍。**





* 提升度=模型筛选用户流失率/用户集中用户流失率

纲要

ML&DM

I 机器学习与数据挖掘概览

II 关联规则挖掘概念与技术

III 数据分类概念与技术

IV 数据聚类概念与技术

聚类分析

- 聚类分析
  - 将物理或抽象对象的集合分组成为由类似的对象组成的多个类的过程。聚类是一种无指导的学习：没有预定义的类编号。
- 聚类（簇）：数据对象的集合
  - 在同一个聚类（簇）中的对象彼此相似
  - 不同簇中的对象则相异
- 聚类算法的选择取决于数据类型，主要包括:
  - 划分方法、层次的方法、基于密度的方法、基于网格的方法、基于模型的方法
- 聚类分析的应用
  - 模式识别、空间数据分析（主题地图、空间聚类）
  - 按客户特征、消费行为聚类，实现客户细分。
  - Web日志聚类，发现用户行为模式.
  - Web新闻、博客等文档内容，实现主题挖掘。

聚类分析

$X=\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$

x_1 = <运动=多， 身高=175-180， 体重=60-70， 爱好=旅行>

x_2 = <运动=多， 身高=170-175， 体重=50-60， 爱好=游泳>

x_3 = <运动=多， 身高=175-180， 体重=60-70， 爱好=听音乐>

x_4 = <运动=少， 身高=170-175， 体重=60-70， 爱好=coding>

x_5 = <运动=少， 身高=170-175， 体重=50-60， 爱好=coding>

x_6 = <运动=少， 身高=170-175， 体重=60-70， 爱好=coding>

----------聚类之后----------

类1 = <x1, x2, x3> 总结：经常运动，身材匀称，兴趣爱好广泛

类2 = <x4, x5, x6> 总结：缺乏运动，身形微福，兴趣爱好贫瘠

……

……

聚类分析——划分方法

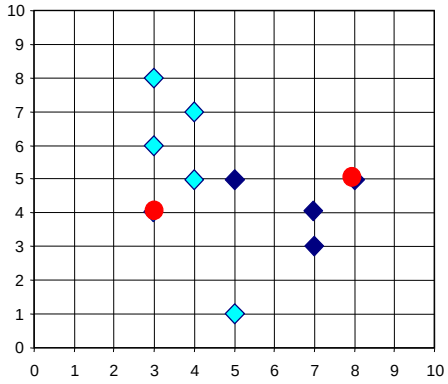
- 给定一个n个对象或元组的数据库，一个划分方法构建数据的k个划分，每个划分表示一个簇，并且k<=n。
  - 每个组至少包含一个对象
  - 每个对象属于且仅属于一个组
- 划分准则：同一个聚类中的对象尽可能的接近或相关，不同聚类中的对象尽可能的原理或不同
- 簇的表示
  - k-平均算法
    - 由簇的平均值来代表整个簇
  - k中心点算法
    - 由处于簇的中心区域的某个值代表整个簇

k均值方法步骤

- 簇的相似度是关于簇中对象的均值度量，可以看作簇的质心（centroid）
- k均值算法流程
  - 随机选择k个对象，每个对象代表一个簇的初始均值或中心
  - 对剩余的每个对象，根据它与簇均值的距离，将他指派到最相似的簇
  - 计算每个簇的新均值
  - 回到步骤2，循环，直到准则函数收敛
    - 常用准则函数：平方误差准则

$$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} \left| p - m_i \right|^2$$ （p是空间中的点，m_i是簇C_i的均值）

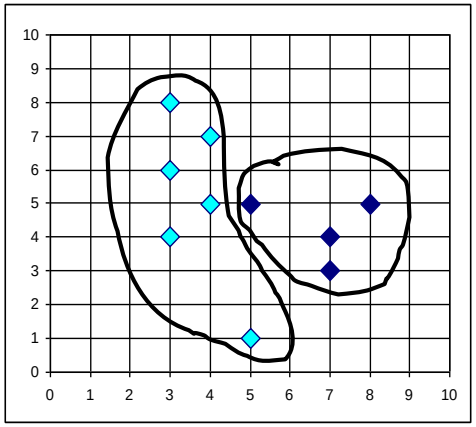
k均值方法---示例



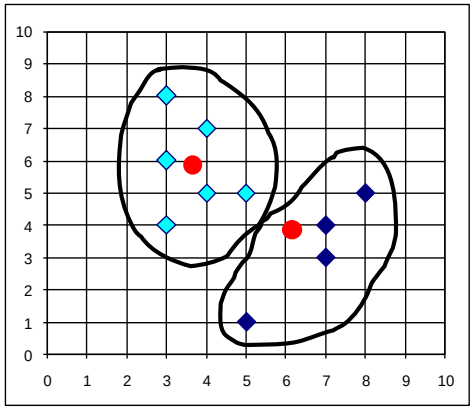
K=2

随机选择2个对象，作为簇的中心

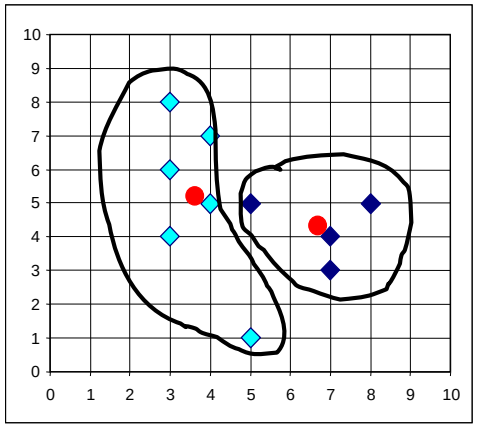
将每个对象指派到最相似的簇



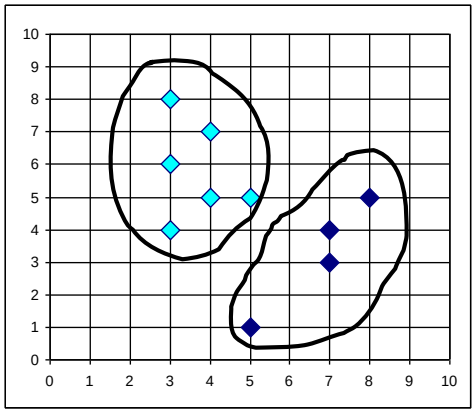
重新分派...



更新每个簇的均值



更新每个簇的均值



k中心点方法步骤

- k中心点方法仍然基于最小化所有对象与其对应的参照点之间的相异度之和原则，使用的是绝对误差标准

$$E = \Sigma_{j=1}^{k} \Sigma_{p \in C_j} \left| p - o_j \right|$$

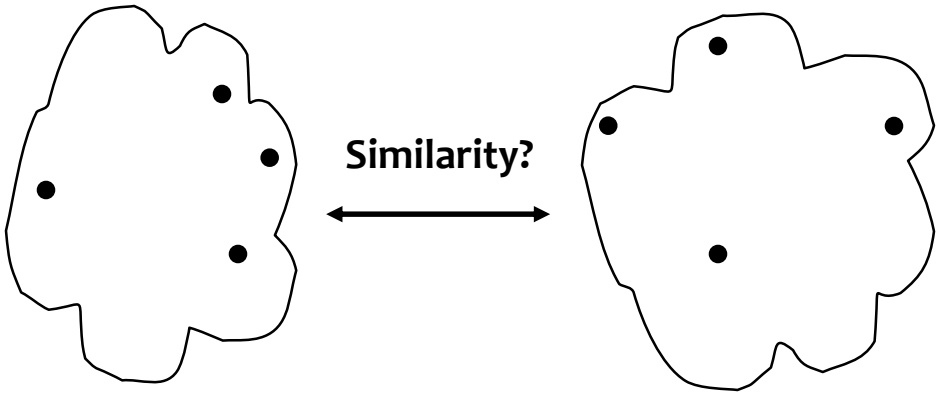
(p是空间中的点，代表簇Cj中一个给定对象；oj是簇Cj中的代表对象)

- 重复迭代，直到每个代表对象都成为它的簇的实际中心点
  - 首先随意选择初始代表对象。
  - 只要能够提高结果聚类质量，迭代过程就使用非代表对象替换代表对象。
    - 聚类结果的质量用代价函数评估，该函数度量对象与其簇的代表对象之间的平均差异度。

聚类分析——层次方法

- 层次聚类方法对给定的数据集进行层次的分解，直到某种条件满足为止。具体又可分为：
  - 凝聚的层次聚类：一种自底向上的策略，首先将每个对象作为一个簇，然后合并这些原子簇为越来越大的簇，直到某个终结条件被满足。
  - 分裂的层次聚类：采用自顶向下的策略，它首先将所有对象置于一个簇中，然后逐渐细分为越来越小的簇，直到达到了某个终结条件。
- 层次凝聚的代表是AGNES算法。层次分裂的代表是DIANA算法。

类间距离的衡量标准

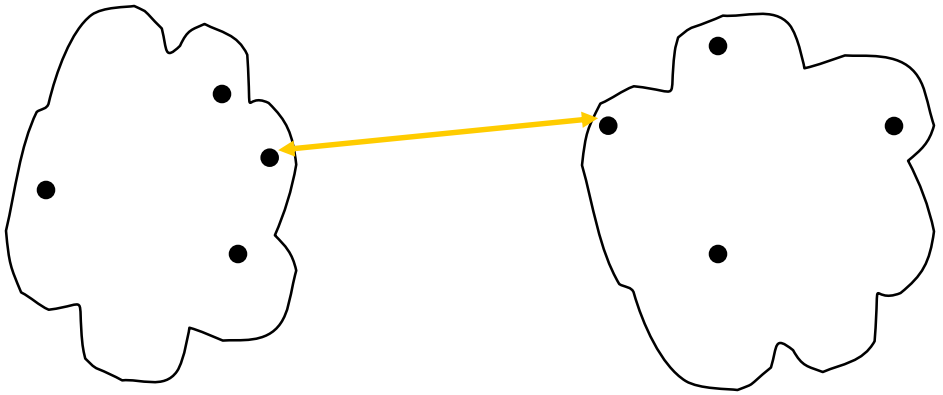


	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

类间距离的衡量标准

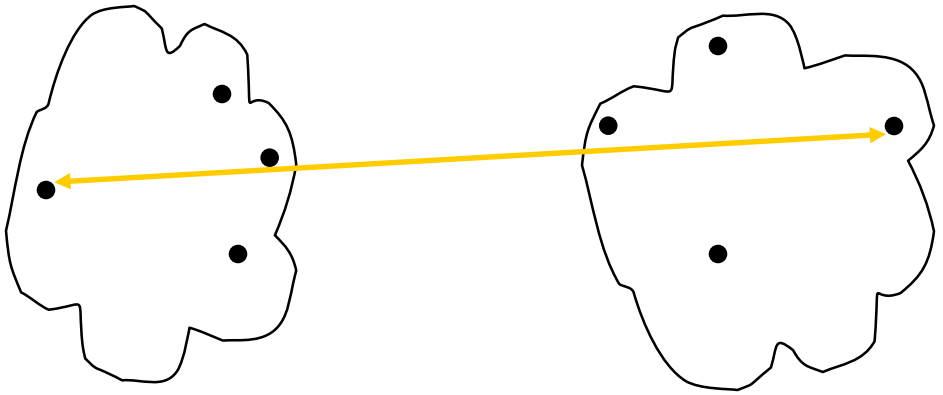


	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

类间距离的衡量标准

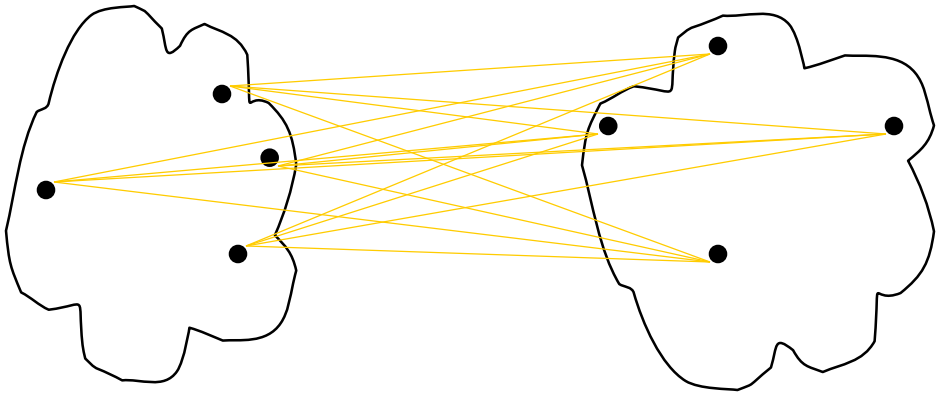


	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX (两集合之间）
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

类间距离的衡量标准

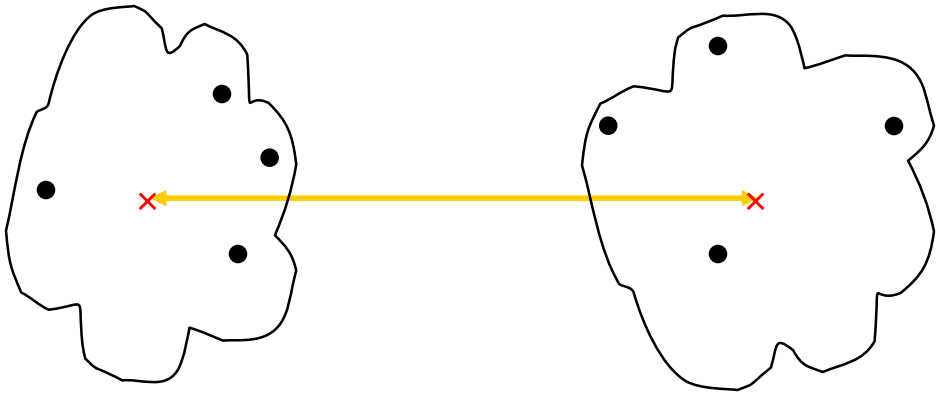


	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

类间距离的衡量标准



	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

AGNES算法

- AGNES(Agglomerative NESting)算法最初将每个对象作为一个簇，然后这些簇根据某些准则被一步步地合并。两个簇间的相似度由这两个不同簇中距离最近的数据点对的相似度来确定。聚类的合并过程反复进行直到所有的对象最终满足簇数目。

自底向上凝聚算法（AGNES):

输入：包含n个对象的数据库，终止条件簇的数目k。

输出：k个簇，达到终止条件规定簇数目。

（1）将每个对象当成一个初始簇；

（2）REPEAT

（3）根据两个簇中最近的数据点找到最近的两个簇；

（4）合并两个簇，生成新的簇的集合；

（5）UNTIL达到定义的簇的数目；

AGNES算法例题

序号	属性1	属性2
1	1	1
2	1	2
3	2	1
4	2	2
5	3	4
6	3	5
7	4	4
8	4	5

第1步：根据初始簇计算每个簇之间的距离，随机找出距离最小的两个簇，进行合并，最小距离为1，合并后1,2两个点合并为一个簇。

第2步：对上一次合并后的簇计算簇间距离，找出距离最近的两个簇进行合并，合并后3,4点成为一簇。

第3步：重复第2步的工作，5,6点成为一簇。

第4步：重复第2步的工作，7,8点成为一簇。

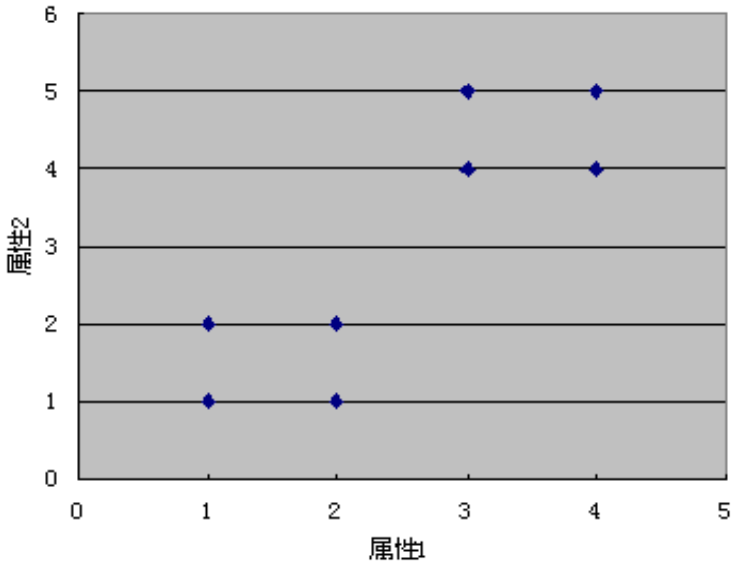
第5步：合并{1,2}，{3,4}成为一个包含四个点的簇。

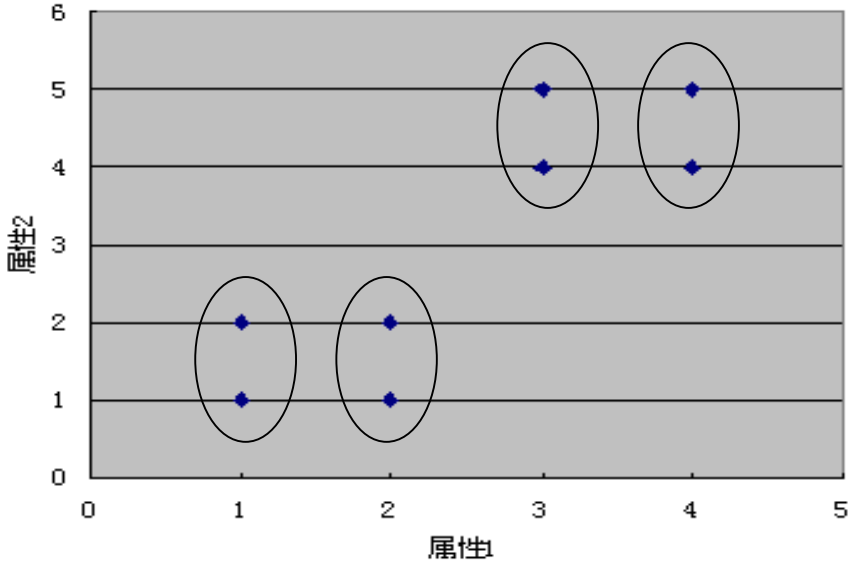
第6步：合并{5,6}，{7,8}，由于合并后的簇的数目已经达到了用户输入的终止条件，程序终止。

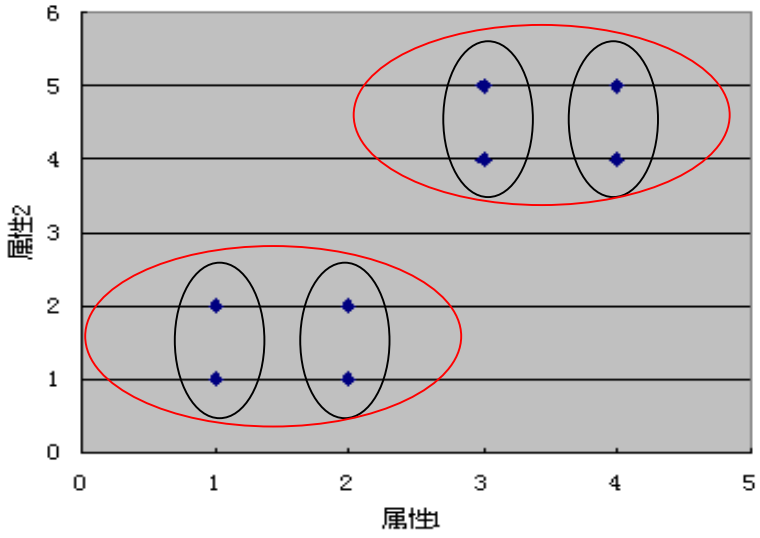
步骤	最近的簇距离	最近的两个簇	合并后的新簇
1	1	{1}，{2}	{1,2}，{3}，{4}，{5}，{6}，{7}，{8}
2	1	{3}，{4}	{1,2}，{3,4}，{5}，{6}，{7}，{8}
3	1	{5}，{6}	{1,2}，{3,4}，{5,6}，{7}，{8}
4	1	{7}，{8}	{1,2}，{3,4}，{5,6}，{7,8}
5	1	{1,2},{3,4}	{1,2,3,4}，{5,6}，{7,8}
6	1	{5,6}，{7,8}	{1,2,3,4}，{5,6,7,8}结束

AGNES算法例题

序号	属性1	属性2
1	1	1
2	1	2
3	2	1
4	2	2
5	3	4
6	3	5
7	4	4
8	4	5







AGNES性能分析

- AGNES算法比较简单，但经常会遇到合并点选择的困难。假如一旦一组对象被合并，下一步的处理将在新生成的簇上进行。已做处理不能撤销，聚类之间也不能交换对象。如果在某一步没有很好的选择合并的决定，可能会导致低质量的聚类结果。
- 这种聚类方法不具有很好的可伸缩性，因为合并的决定需要检查和估算大量的对象或簇。
- 假定在开始的时候有n个簇，在结束的时候有1个簇，因此在主循环中有n此迭代，在第i次迭代中，我们必须在n-i+1个簇中找到最靠近的两个聚类。另外算法必须计算所有对象两两之间的距离，因此这个算法的复杂度为$O(n^2)$，该算法对于n很大的情况是不适用的。

DIANA分裂层次聚类算法

- DIANA（Divisive ANAlysis)算法是典型的分裂聚类方法。
- 在聚类中，用户能定义希望得到的簇数目作为一个结束条件。同时，它使用下面两种测度方法：

  簇的直径：在一个簇中的任意两个数据点的距离中的最大值。

  平均相异度

DIANA算法

- 算法 DIANA（自顶向下分裂算法）
- 输入：包含n个对象的数据库，终止条件簇的数目k。
- 输出：k个簇，达到终止条件规定簇数目。
  - 将所有对象整个当成一个初始簇；
  - FOR （i=1; i≠k; i++) DO BEGIN
    - 在所有簇中挑出具有最大直径的簇C；
    - 找出C中与其它点平均相异度最大的一个点p并把p放入splinter group，剩余的放在old party中；
    - REPEAT上一步骤
    - 在old party里找出到最近的splinter group中的点的距离不大于到old party中最近点的距离的点，并将该点加入splinter group。
    - UNTIL 没有新的old party的点被分配给splinter group；
    - splinter group和old party为被选中的簇分裂成的两个簇，与其它簇一起组成新的簇集合。
    - END.

DIANA算法

- 算法 DIANA（自顶向下分裂算法）
- 输入：包含n个对象的数据库，终止条件簇的数目k。
- 输出：k个簇，达到终止条件规定簇数目。
  - 将所有对象整个当成一个初始簇；
  - FOR （i=1; i≠k; i++) DO BEGIN
    - 在所有簇中挑出具有最大直径的簇C；
    - 找出C中与其它点平均相异度最大的一个点p并把p放入splinter group，剩余的放在old party中；
    - REPEAT上一步骤
    - 在old party里找出到最近的splinter group中的点的距离不大于到old party中最近点的距离的点，并将该点加入splinter group。
    - UNTIL 没有新的old party的点被分配给splinter group；
    - splinter group和old party为被选中的簇分裂成的两个簇，与其它簇一起组成新的簇集合。
    - END.

DIANA算法例题

序号	属性 1	属性 2
1	1	1
2	1	2
3	2	1
4	2	2
5	3	4
6	3	5
7	4	4
8	4	5

第1步，找到具有最大直径的簇，对簇中的每个点计算平均相异度（假定采用是欧式距离）。

1的平均距离：（1+1+1.414+3.6+4.24+4.47+5）/7=2.96

类似地，2的平均距离为2.526；3的平均距离为2.68；4的平均距离为2.18；5的平均距离为2.18；6的平均距离为2.68；7的平均距离为2.526；8的平均距离为2.96。

挑出平均相异度最大的点1放到splinter group中，剩余点在old party中。

第2步，在old party里找出到最近的splinter group中的点的距离不大于到old party中最近的点的距离的点，将该点放入splinter group中，该点是2。

第3步，重复第2步的工作，splinter group中放入点3。

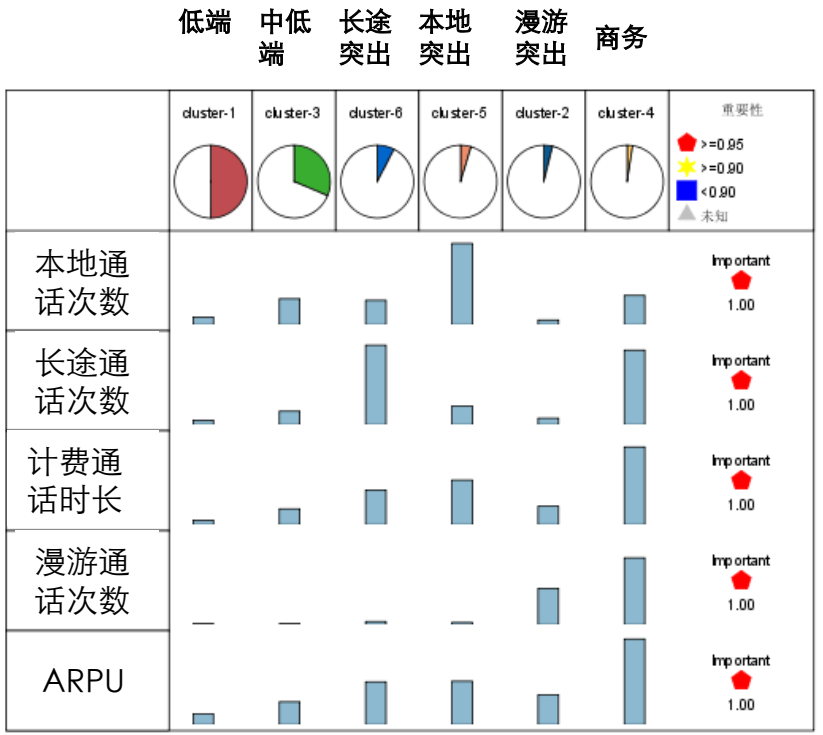
第4步，重复第2步的工作，splinter group中放入点4。

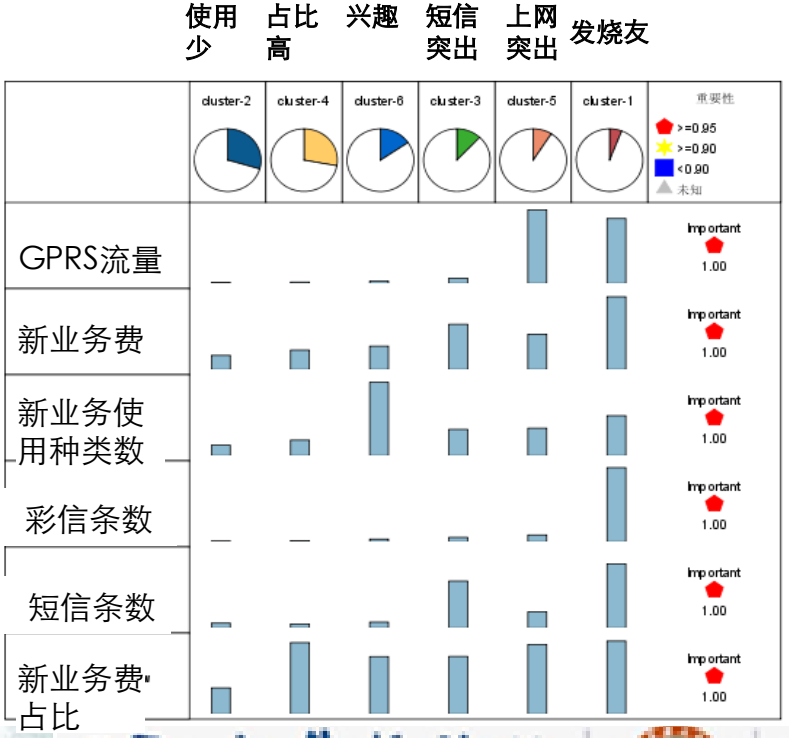
第5步，没有在old party中的点放入了splinter group中且达到终止条件（k=2），程序终止。如果没有到终止条件，因该从分裂好的簇中选一个直径最大的簇继续分裂。

步骤	具有最大直径的簇	splinter group	Old party
1	{1，2，3，4，5，6，7，8}	{1}	{2，3，4，5，6，7，8}
2	{1，2，3，4，5，6，7，8}	{1，2}	{3，4，5，6，7，8}
3	{1，2，3，4，5，6，7，8}	{1，2，3}	{4，5，6，7，8}
4	{1，2，3，4，5，6，7，8}	{1，2，3，4}	{5，6，7，8}
5	{1，2，3，4，5，6，7，8}	{1，2，3，4}	{5，6，7，8} 终止

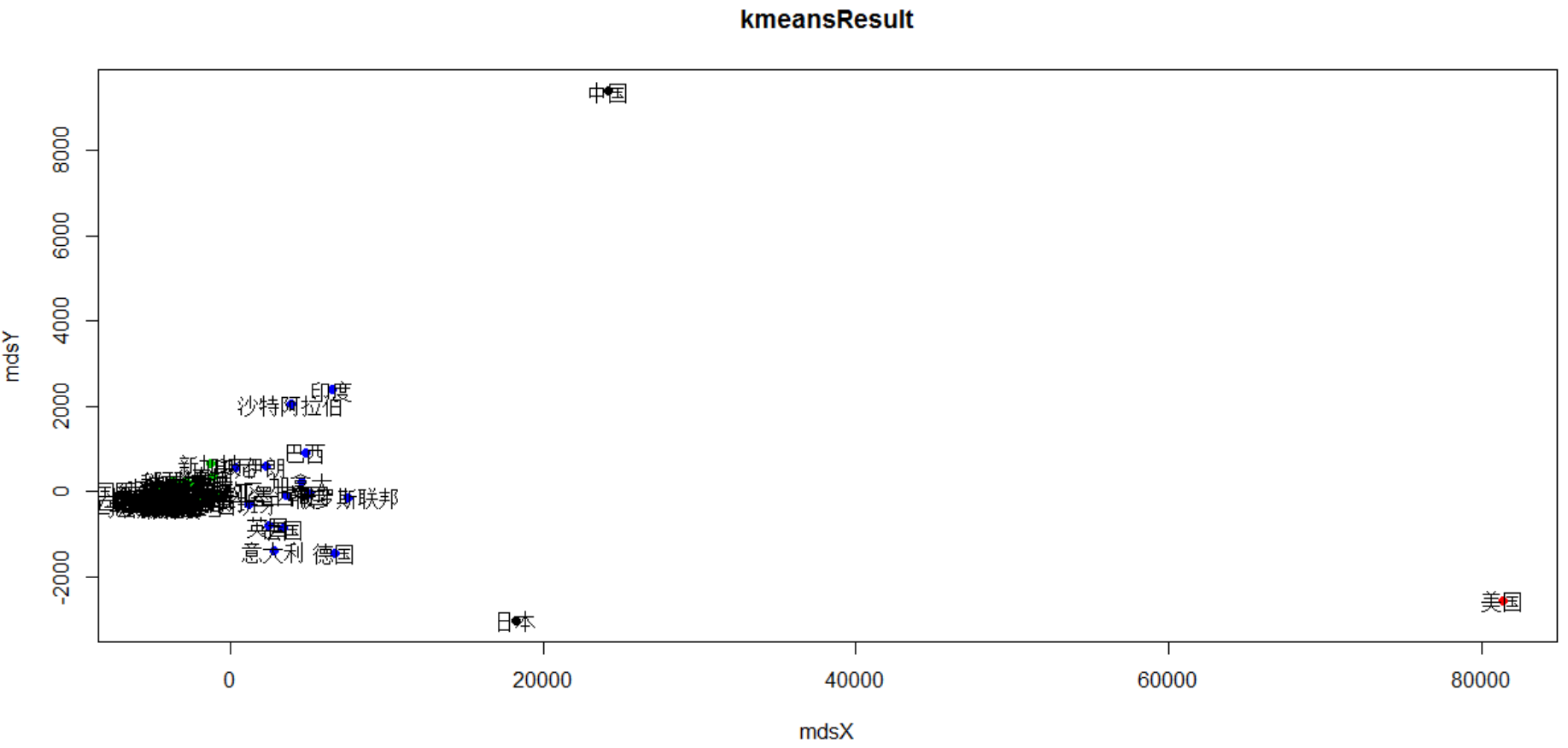
聚类案例：用户细分模型

- 根据用户基础数据和消费行为数据采用Two Step聚类法对用户进行细分
- 聚类数据集为2009年1-9月新增用户入网后第二个月，且第二月状态正常的用户的基础数据和消费行为数据
- 使用细分矩阵，按照语音消费行为和数据业务消费行为两次聚类的方法分别聚类，多维聚类的方法较传统单维聚类方法，聚类后的用户细分特征更明显





利用R进行K均值和K中心聚类



感谢关注聆听！

张华平
Email: kevinzhang@bit.edu.cn
微博：@ICTCLAS张华平博士
实验室官网：
http://www.nlpir.org

大数据千人会