中文文本分类
Chinese Text Classification

苏州大学计算机科学与技术学院

主要内容

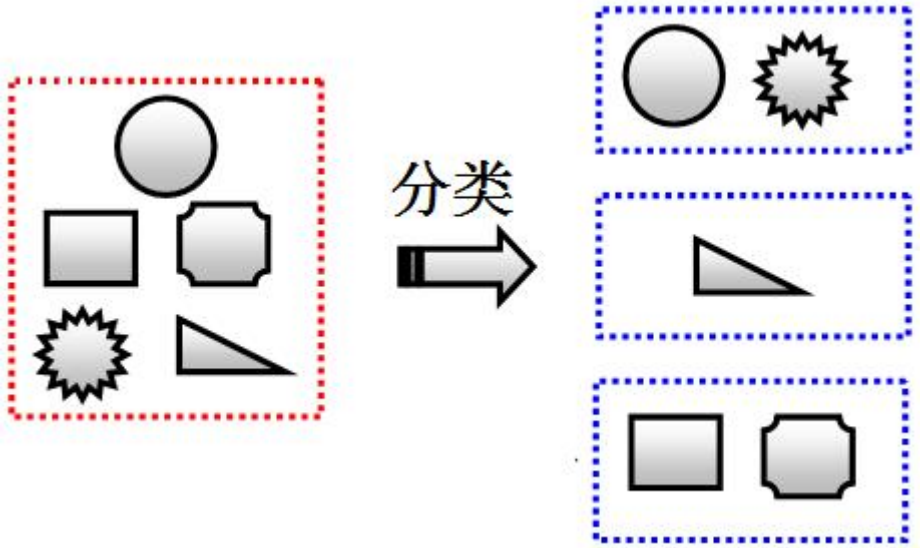
- 什么是文本分类
- 文本分类的应用
- 文本表示
- 分类特征选择
- 文本分类算法
- 文本分类评测

中文信息处理的两个层面

- 字符层
  - 存储：各种字符编码
  - 输入：语音、手写、印刷体识别、键盘编码
  - 输出：显示器、打印机
- 内容层
  - 信息检索：搜索引擎
  - 文本分类：垃圾短信过滤
  - 信息抽取：
  - ……

信息与类

- 类
  - 一组具有某一共同属性的事物对象的集合
- 分类是对信息的一种最基本的认知形式
- 有序化信息有利于信息的
  - 存储
  - 检索
  - 传播
  - 开发利用



何为文本分类？

- 在给定的分类模型下，根据文本的内容让计算机自动判断文本类别的过程
  - 分类体系一般人工构造
  - 因此分类可以认为是有指导的（有监督的）
- 从数学的角度
  - 是一个映射的过程

$f : A \rightarrow B$　　其中，A为待分类的文本集合，B为分类体系中的类别集合

  - 将一个未表明类别的文本映射到已有的类别中
  - 可以一对一，也可以一对多

文本分类应用

- 垃圾邮件的判定
  - 类别 {spam, not-spam}
- 新闻出版按照栏目分类
  - 类别 {政治,体育,军事,…}
- 词性标注
  - 类别 {名词,动词,形容词,…}
- 计算机论文的领域
  - 类别 ACM system
    - H: information systems
    - H.3: information retrieval and storage

- 中国图书馆图书分类法
  - 1975年第一版发行，22类（5个大类）

TP3-0计算机理论与方法
TP30一般性问题
TP31计算机软件
TP32一般计算器和计算机
TP33电子数字计算机（不连续作用电子计算机）
TP34电子模拟计算机（连续作用电子计算机）
TP35混合电子计算机
TP36微型计算机
TP37多媒体技术与多媒体计算机
TP38其他计算机
TP39计算机的应用

百度新闻搜索——全球最大的中文新闻平台 - Windows Internet Explorer

http://news.baidu.com/

Google

文件(F) 编辑(E) 查看(V) 收藏夹(A) 工具(T) 帮助(H)

百度新闻搜索——全球最大的中文新闻平台

页面(P) 工具(O)

百度首页 | 登录

Baidu新闻

新闻 网页 贴吧 知道 MP3 图片 视频

百度一下 帮助 | 高级搜索 | 偏好设置

新闻全文 新闻标题

以下新闻由机器每5分钟自动选取更新

新闻首页 | 国内 | 国际 | 军事 | 财经 | 互联网 | 房产 | 汽车 | 体育 | 娱乐 | 游戏 | 教育 | 女人 | 科技 | 社会 | 视频 | 图 | 世界杯 | 新 | 定制 | 更多

焦点新闻

汇丰 在线参观中国新总部

韩国罗老号火箭升空至距地面70公里后坠毁

[罗老号发射失败爆炸 韩称将一直发射到成功为止] [脱离轨道并坠] [火箭升空后失去通信联系] [发射遭遇重大险情] [官方曾称是最佳

- 中国5月报告梅毒发病32190例 已成严重公卫问题 20:57
- 南水北调中线工程移民开始大规模搬迁 21:09
- 外交部：望南非妥善处理中国记者遭抢劫事件 21:05
- 新疆克孜勒苏柯尔克孜自治州乌恰县发生5.1级地震 15:13
- 团伙组织名校大学生跨省替考 枪手最高可拿7万 18:08
- 中石油打响央企退地头炮 福建民企出1.26亿接盘 21:04
- 南非世界杯开幕式11日21:30始 中国人首次参与 16:08
- 财政部发出通知 称高校学费未来两年不得上涨 18:51
- 杭州西湖景区回应“名胜故居变私人会所”报道 20:57
- 不符合申诉程序 最高法暂未受理周正龙申诉状 17:51

南非世界杯：民众热情提前引爆

南非世界杯官方主题曲发布

电视台曝光深圳某医院卖号成风

新闻热搜词

更多>>

罗老号坠毁 成都 爆炸
广电总局 相亲节目 新疆地震
最低工资上调 经济数据泄密
世界杯 记者被抢 安理会制裁伊朗
广州最残忍黑帮 鄢颇 李小冉前男友

北京 地铁 票价 英国石油 漏油
绿地 降价 周正龙妻子 申诉 范冰冰 恋情
大连化物所爆炸 王思懿热吻朱时茂
徐怀钰 失踪 兽兽耍大牌 央企 退地

博客 - 论坛

更多>>

- 并非杞人忧天：为开出“一毛钱处方”的医生担忧
- 鄢颇被砍何以变成“摇摆的娱乐”？！
- 把考生考哭是命题者的无能
- 公众的美丽西湖绝不能沦为富人的“洗脚盆”

Internet 100%

文本分类的规则

- 文本分类的映射规则是：
  - 系统根据已经掌握的每类若干样本的数据信息，总结出分类的规律性而建立的判别公式和判别规则；
  - 然后在遇到新文本时，根据总结出的判别规则，确定文本相关的类别。

文本分类的模式

- 根据需要的不同
  - 单类别分类
    - 每个文档必须归属一个类别
    - 二元：属于不属于
    - 多元
      - 可以拆分成多个二元
  - 多类别分类
    - 一篇文档可以属于多个类别
    - 也可以不属于任何类

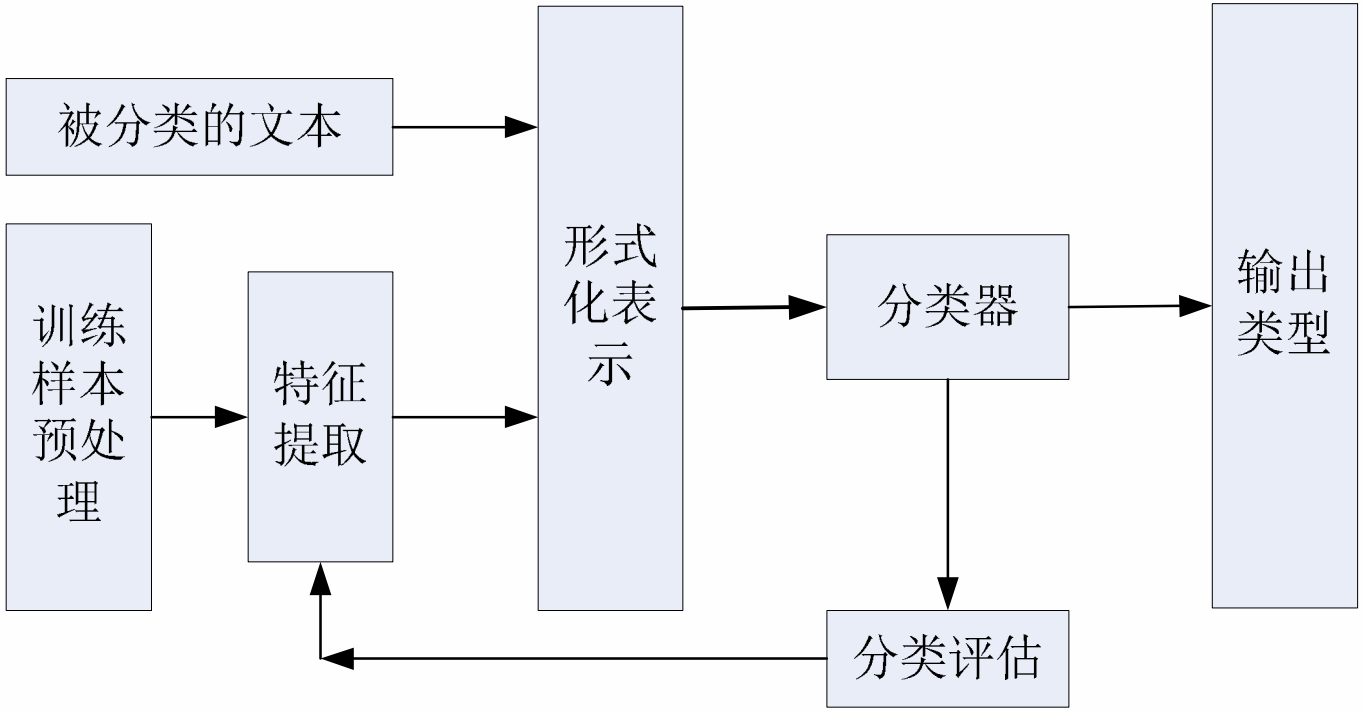
文本分类的方式

- 以文档为中心的分类
  - Document-Pivoted Text Categorization
  - 给定一篇文档，遍历所有类别，判断它属于的类
  - 文档陆陆续续过来：例如邮件过滤
- 以类别为中心的分类
  - Category-Pivoted Text Categorization
  - 假定某个类别，在给定的文档集中找出属于该类的文档子集

中文文本分类系统

- 自动分词
- 特征选择与抽取
- 文本计算模型
  - 将特征表示为可处理的数学模型
- 文本识别算法
  - 根据文本计算模型，计算出类别

中文文本分类系统结构图



文本分类的发展

- 可行性研究阶段
  - 1958-1964
- 试验研究阶段
  - 1965-1974
- 实用化阶段
  - 1975至今
  - 邮件分类、信息过滤等
- 中文始于20世纪80年代：
  - 目前正确率60%-90%

文本特征的选择

- 预处理
  - 去除格式标记：<h1>这样的标记
  - 去除停用词：
    - 高频不带区分度的词：的 是
    - 低频但几乎不带分析信息生僻词
  - 自动分词：
  - 词性标注：tagging
  - 词频统计：
  - 句法分析：Parser

文本表示

- 词袋(Bag of words)模型
  - 不考虑词在文档中出现的顺序
    - 我 爱 打 篮球
    - 篮球 爱 打 我
  - 在某种意思上说，这种表示方法是一种“倒退”，因为丢失了位置信息
  - 但是问题得以大大简化

文本的表示（续1）

- 向量空间模型(Vector Space Model)
  - 把长度不相等的文本转换为长度相等的向量
- 例如：
  - 1.我 爱 打 篮球
  - 2.我 爱 游泳
  - 3. 她 喜欢 跳舞

	我	她	爱	打	喜欢	篮球	游泳	跳舞
句1	1	0	1	1	0	1	0	0
句2	1	0	1	0	0	0	1	0
句3	0	1	0	0	1	0	0	1

文本的表示（续2）

- 变成了等长向量
  - 句1 (1,0,1,1,0,1,0,0)
  - 句2 (1,0,1,0,0,0,1,0)
  - 句3 (0,1,0,0,1,0,0,1)
- 实际向量维度很高
  - 中文常用词在8万以上
  - 很多输入法包含40万词

特征的权重

- 布尔权重
  - 出现为1，不出现为0
- 词项频率Term Frequency
  - 一个Term在文档中出现的次数
- TFIDF型权重
  - DF（文档频率，Document Frequency）
  - IDF（逆文档频度，$\log(N/DF_i)$） N是总文档数
  - TF*IDF 降低了TF的作用

特征选择

- 文本特征的选择:降维
  - 目的提高分类效率、减少计算复杂度
  - 去除不带分类信息和信息量较少的词
  - 一个特征词条在一个文档中出现的次数越多，它与该文档对应的主题越相关
  - 一个特征词在越多的文档中出现，它对类别区分度的作用越小
  - 用权值来表示一个词的作用

类别的代表term

- 为每个类别抽取n个最有区别能力的term
- 例如：
  - 计算机领域：
    - CPU、芯片、操作系统、编译、 …
  - 汽车领域:
    - 轮胎、方向盘、底盘、气缸、发动机…

特征选择之DF

- 文档频率
  - DF表示在文档集中包含某个特征项t的文档数
- 选择方法
  - Term的DF小于某个阈值去掉
    - 太少，没有代表性
  - Term的DF大于某个阈值也去掉
    - 太多，没有区分度

DF（续）

- 这种策略不符合被广泛接受的信息检索理论：
  - 高频词没有低频词对文档特征贡献大
- DF是最简单的特征项选取方法，而且该方法的计算复杂度低，能够胜任大规模的分类任务

特征选择之IG

- 信息增益(Information Gain)
  - 根据某个特征项在文档中出现与否来计算它为文档类别预测所贡献的信息量
  - 不考虑特征的熵和考虑该特征后的熵的差值

$$G(t) = -\sum_{i=1}^{m} P_r(c_i)\log P_r(c_i) + \mathrm{p_r(t)}\sum_{i=1}^{m} P_r(c_i \mid t)\log P_r(c_i \mid t) + P_r(\bar{t})\sum_{i=1}^{m} P_r(c_i \mid \bar{t})\log P_r(c_i \mid \bar{t})$$

- 互信息
  - Mutual Information
  - 计算特征词条t和类别c之间的相关性
  - 如果有m个类，则对于每个t会有m个值

- X^2统计量
  - 如果特征项和类别反相关，就说明含有特征项的文档不属于的概率要大一些，这对于判断一篇文档是否不属于类别也是很有指导意义的

	c	~c
t	A	B
~t	C	D

$$c^2(t,c)=\frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

文本分类方法（一）

- Rocchio方法
  - 相似度方法
  - 为每类文本集生成中心向量
  - 确定文本向量
  - 计算文本向量与每类文本集中心向量的相似度
  - 取相似度的最大值

Rocchio方法优点

- 每一类确定一个中心点（代表元），计算待分类的文档与各类代表元间的距离，并作为判定是否属于该类的判据。
- Rocchio算法的突出优点
  - 容易实现
  - 计算（训练和分类）特别简单
  - 它通常用来实现衡量分类系统性能的基准系统
  - 实用的分类系统很少采用这种算法解决具体的分类问题

文本分类方法（二）

- 贝叶斯方法
  - 一种简单有效的分类方法
  - 计算文本属于某个类别的概率
  - 具体步骤：
    - 计算特征词属于每个类别的概率向量
    - 新文本到达，根据切分出的特征词，计算该文本属于不同类的概率
    - 比较计算出的多个概率，并决定类型

$$P(c_i \mid d_j) = \frac{P(d_j \mid c_i)P(c_i)}{P(d_j)}$$

- 利用词条独立假设，这样做不严格

贝叶斯参数计算

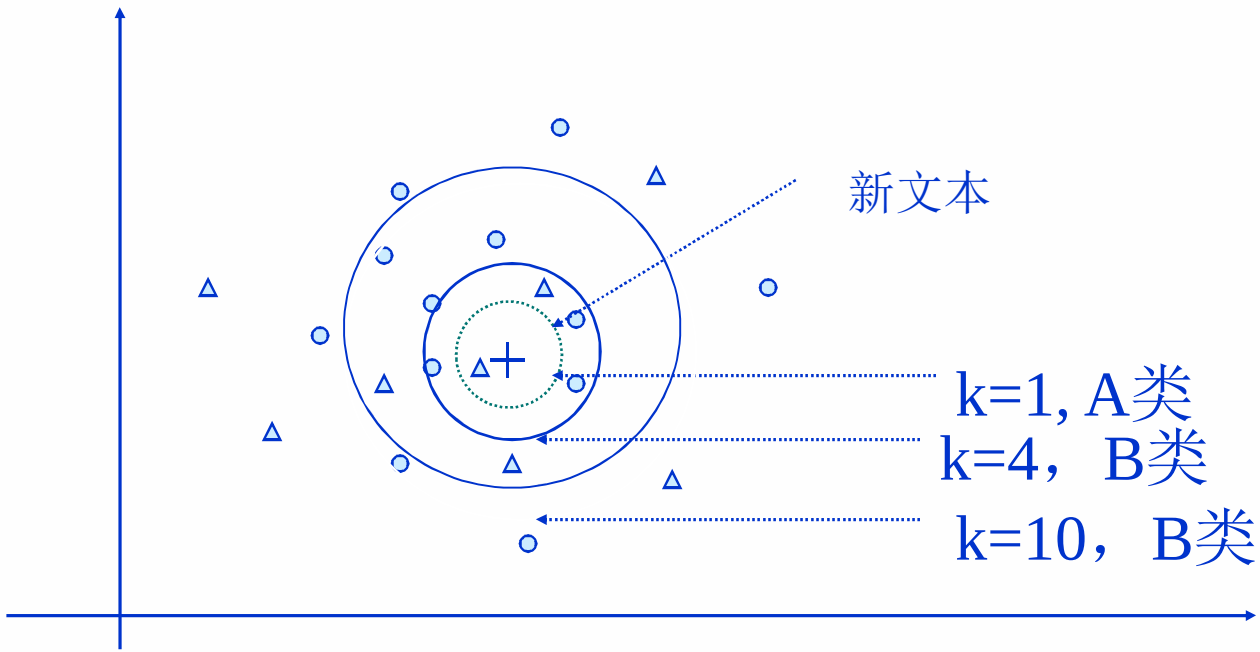
$$P(d_i \mid c_j) = \prod_{k=1}^{r} P(w_{ik} \mid c_j), \quad \text{独立性假设}$$

$$P(c_j) = \frac{c_j\text{的文档个数}}{\text{总文档个数}}$$

$$P(w_i \mid c_j) = \frac{w_i\text{在}c_j\text{类别文档中出现的次数}}{\text{在}c_j\text{类所有文档中出现的词的次数}}$$

文本分类方法（三）

- KNN方法
  - K近邻方法（ k-nearest neighbor ）
  - 将每个文本看称平面上的一个点
  - 小于指定的K值，则为它们的邻居，并观察这些样本点所属类别，少数服从多数
    - 加入权值信息



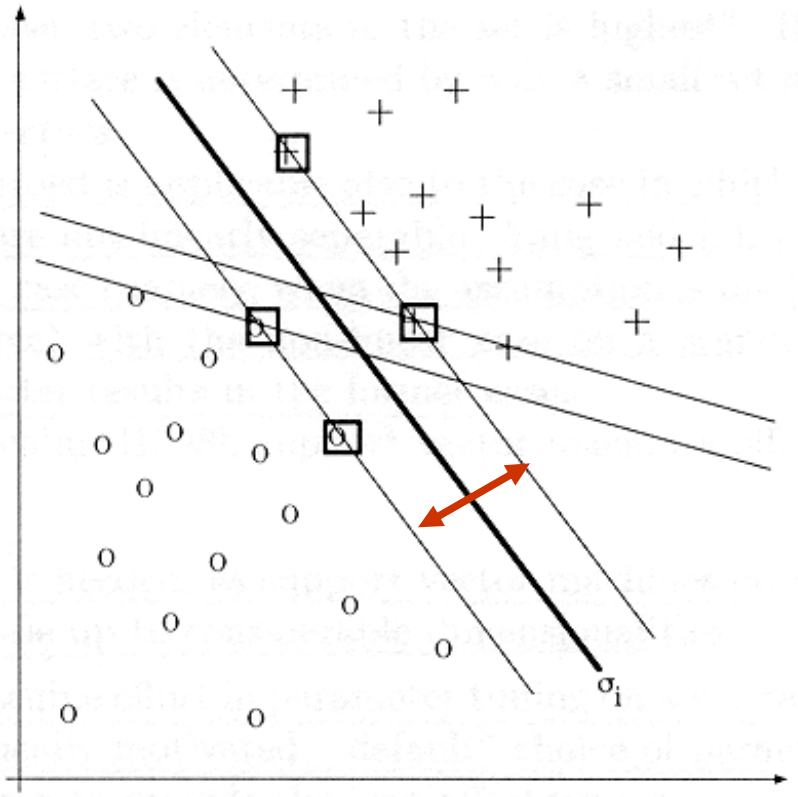
带权重计算，计算权重和最大的类。k常取3或者5。

KNN方法（续）

- KNN算法本身简单有效
- 它是一种lazy-learning算法
- 分类器不需要使用训练集进行训练
- 训练时间复杂度为0
- KNN分类的计算复杂度和训练集中的文档数目成正比
- 如果训练集中文档总数为n，那么KNN的分类时间复杂度为O(n)

文本分类方法（四）

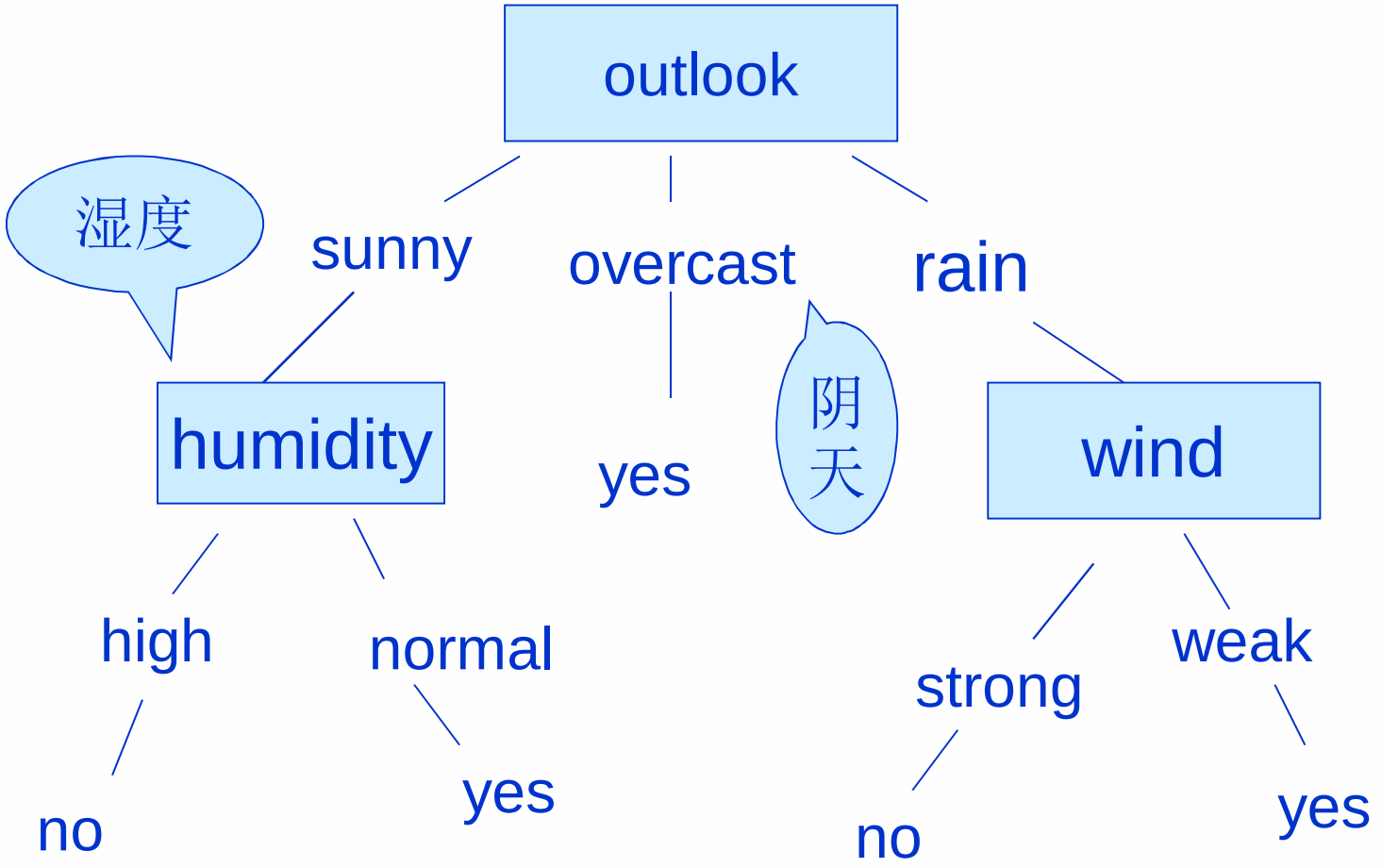
- SVM
  - Support Vector Machine
  - 支持向量机
    - 寻找超平面



文本分类方法（五）

- Decision Tree
  - 决策树方法
  - 非常有效的机器学习算法

决策树表示举例


outlook
湿度
sunny
overcast
rain
humidity
yes
阴天
wind
high
normal
strong
weak
no
yes
no
yes

多分类器融合

- 基于投票的方法
  - Bagging方法
    - 训练R个分类器f_i，分类器之间其他相同就是参数不同。其中f_i是通过从训练集合中(N篇文档)随机取(取后放回)N次文档构成的训练集合训练得到的。
    - 对于新文档d，用这R个分类器去分类，得到的最多的那个类别作为d的最终类别
  - Boosting方法
    - 类似Bagging方法，但是训练是串行进行的，第k个分类器训练时关注对前k-1分类器中错分的文档，即不是随机取，而是加大取这些文档的概率

选择阈值

- 文本分类的方法化分类为数学计算
  - 通常是计算出一个值
  - 该值小于指定的阈值则为该类
  - 如何确定阈值？
- 平均法
- CSV阈值法
- 均衡阈值法
- 固定阈值法
- 各种方法都有优缺点、根据不同需要选择使用

训练集测试集

- 训练集用于建立模型
- 测试集评估模型的预测等能力
- N折交叉测试
  - 避免偶然现象
  - N经常取5、10

评估方法

- 查全率
  - 分类的正确文本数/应有的文本数
- 准确率
  - 分类的正确文本数/实际分类的文本数
- F1测试值
  - 查全率×查准率×2/(查全率+查准率)

分类的评测

- 偶然事件表（Contingency Table）

	属于此类	不属于此类
判定属于此类	A	B
判定不属于此类	C	D

- 对一个分类器的度量
  - 准确率(precision) = A / (A + B)
  - 召回率(recall) = A / (A + C)

文本情感分类

- 81% 网络用户会在线查找产品的信息
- 已有的用户评价会显式影响潜在客户
  - 评价作弊

- 公司很想知道用户对自己产品的评价

- 情感分析
- 情感分类
- 褒贬分类
- 观点分类

研究意义

- 具有很大的研究意义与实用价值
- 提供决策支持
  - 美国总统候选人
    - 希拉里·克林顿
    - 唐纳德·特朗普
    - 大量Tweets
      - 谁的支持高？评价所针对的主要优缺点？

- 分析思潮变化
  - 国民党将军张灵甫
    - 之前：负面评价文本主流
    - 现在：正面评价文本主流

事实与观点

- 文本信息的主要形式
  - 事实
    - 目前大多数文本信息处理针对事实
    - 事实可以用关键词表达
  - 观点
    - 很难用少量关键词表达

情感计算

- 对象
  - 主观性文本
    - 断言
    - 评论
    - ...
  - 两种粒度
    - 文档
      - 单文档
      - 多文档
    - 句子
- 例子：
  - 我昨天买了一辆车，它不仅非常漂亮，而且性能特别好。

主观性文本识别技术（分类）

- 以观点倾向词为主，辅以各种词汇以及文法信息
- 然后送入标准分类器
- 目前：
  - 文档粒度的主观性识别准确率97%
  - 句子粒度的主观性识别准确率55%

观点的三要素

- 持有者
  - 人
  - 组织机构
- 对象
  - 主题
- 观点倾向性
  - 情感的方向
  - 情感的强度
- 主题与情感词具有领域相关的特点

观点的模型

- O 是一个对象
- p 是一个观点持有者
- $F=\{f_1,f_2,\ldots,f_n\}$
  - 是O的特性集合
- $W=\{w_1,w_2,\ldots w_n\}$
  - 是对于特性的评价集合
- p对O持有观点描述如下
  - 集合$Sp\subseteq F$，对于每一个$f_k\in Sp$ 用W中的一个元素来描述f_k

基于前述模型的三种情况

- F和W都是未知
- F已知，W未知
- F已知，W已知

两种级别的情感分类

- 文档级
  - 粗糙
  - 假设：一个文档对应一个观点持有者对于一个对象的观点
- 句子级
  - 识别主观性句子
    - 分类问题
  - 假设：一个句子包含一个观点
- 大多数应用需要句子级

三个子任务

- 意见持有者识别
  - Holder Identification
- 主题抽取
  - Topic Extraction
- 情感分析
  - Sentimental Analysis

主题的情感

- 情感的方向
  - 褒
  - 贬
  - 中
- 情感的强度
  - 离散量
- 情感的表示主要是情感词

情感词

- 情感词汇
  - 利用词汇与语义
    - 具有情感趋向的词汇
    - 借助同义词、近义词找潜在情感词
    - HowNet 6564词组
      - 人工标注情感词的强度与极性
  - 借助频繁模式挖掘
    - 利用关联规则
    - 例如：产品评价分析，大量评论中频繁出现的词，必然和产品特征相关

词汇的极性

- 原极性
  - 稳定
- 上下文极性
  - 变化
    - 否定前缀
      - 导致方向 相反
      - 但 不漂亮 ！= 丑陋
    - 强调前缀
      - 很 非常
  - 构造 否定词典 和 强调词典

分类方法

- 分类方法
  - 基于机器学习的方法
    - 朴素贝叶斯算法
    - 最大熵算法
    - SVM算法
  - 基于语义分析的方法
    - 根据词的语义倾向来判断文本的类别
    - 计算词语义倾向的方法
      - 一个词与一个观点的关联程度
    - SO-PMI （Semantic Orientation Pointwise Mutual Information）

SO-PMI

- PMI计算公式如下：

$$PMI(word_1, word_2) = \log(\frac{P(word_1 \& word_2)}{P(word_1)P(word_2)})$$

其中，P（$word_1$&$word_2$）表示$word_1$和$word_2$同时出现的概率。

  - SO-PMI计算公式如下

$$SO-PMI(word) = \sum_{pword \in Pset} PMI(word, pword) - \sum_{Nword \in Nset} PMI(word, nword)$$

其中，$Pset$ 和 $Nset$ 分别是褒义和贬义种子情感词的集合。

- 计算两个词同现的频率来衡量二者的关联度

情感词

- 一些现象
  - 相同极性情感描述项经常同时出现
  - 用连词连接的情感词往往要么相同、要么相反
- 可以基于这些现象扩展情感词