第四章：

文本自动分类技术

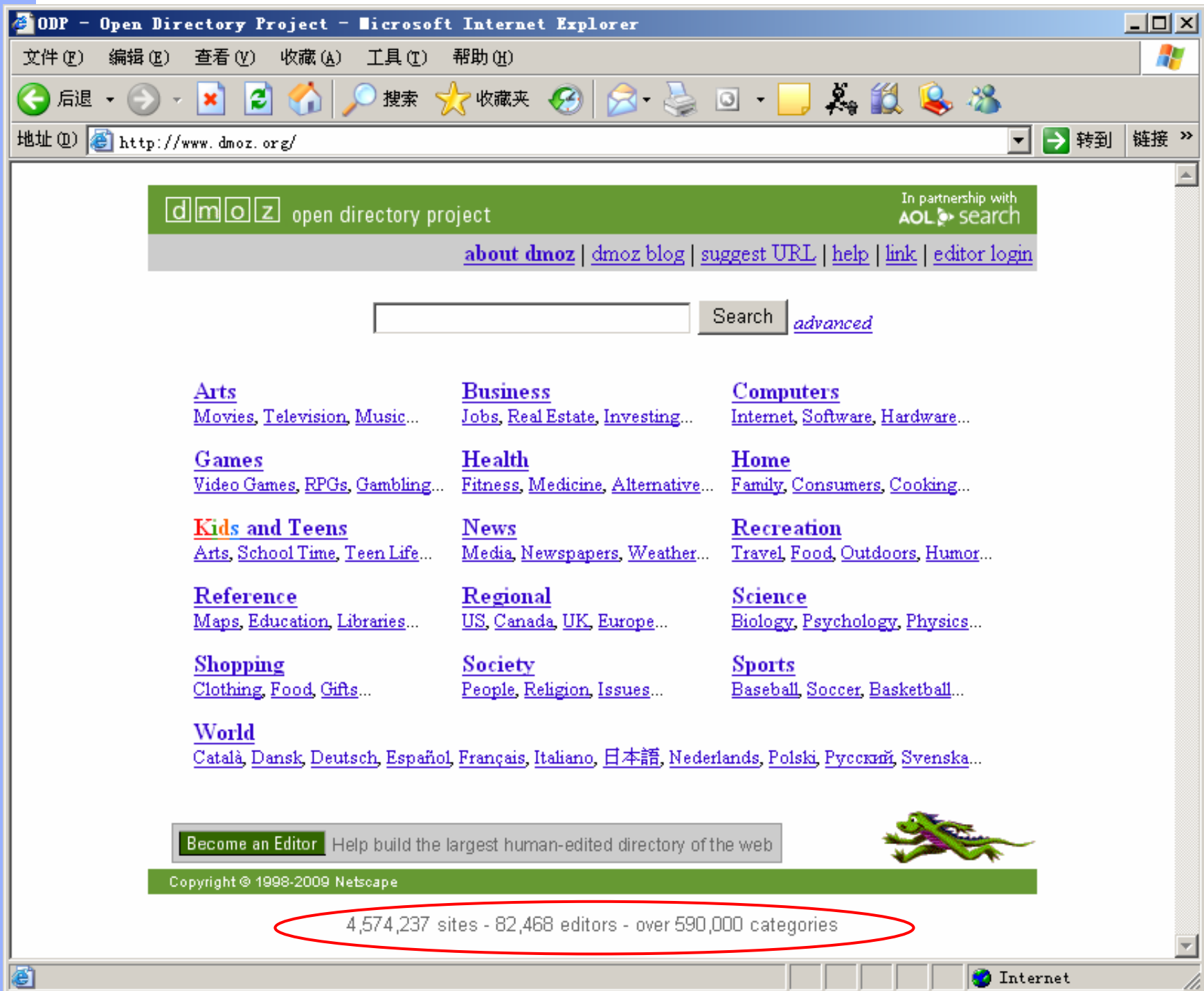
杨建武

北京大学计算机科学技术研究所

Email:yangjianwu@icst.pku.edu.cn

知识的组织

- 知识的结构和知识是孪生的兄弟
  - 结构本身也是知识
- 分类体系
  - 杜威十进制系统（图书分类），
  - 国会图书馆的目录，
  - AMS（美国数学会）的数学知识体系，
  - 美国专利内容的类别体系
- Web catalogs
  - Yahoo以前的主页
  - Open Directory（http://www.dmoz.org/）
    - 志愿者共同维护与建设的最大的全球目录社区

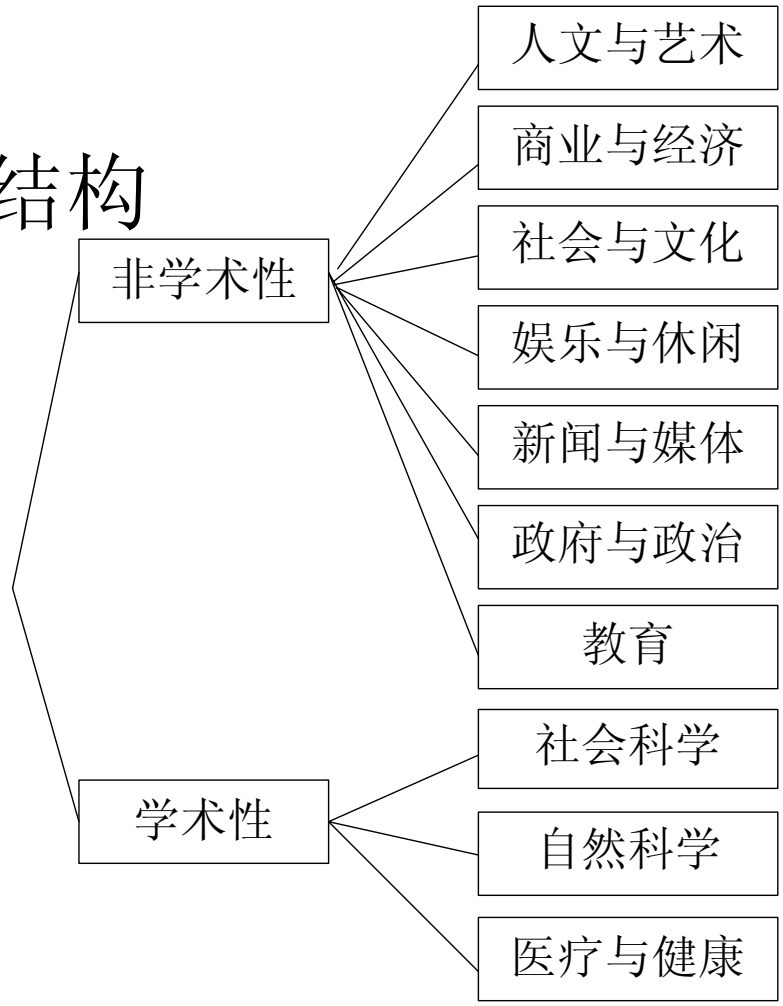

ODP – Open Directory Project – Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)
后退 搜索 收藏夹
地址(D) http://www.dmoz.org/
转到 链接
dmoz open directory project
In partnership with AOL search
about dmoz | dmoz blog | suggest URL | help | link | editor login
Search advanced
Arts
Movies, Television, Music...
Business
Jobs, Real Estate, Investing...
Computers
Internet, Software, Hardware...
Games
Video Games, RPGs, Gambling...
Health
Fitness, Medicine, Alternative...
Home
Family, Consumers, Cooking...
Kids and Teens
Arts, School Time, Teen Life...
News
Media, Newspapers, Weather...
Recreation
Travel, Food, Outdoors, Humor...
Reference
Maps, Education, Libraries...
Regional
US, Canada, UK, Europe...
Science
Biology, Psychology, Physics...
Shopping
Clothing, Food, Gifts...
Society
People, Religion, Issues...
Sports
Baseball, Soccer, Basketball...
World
Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...
Become an Editor
Help build the largest human-edited directory of the web
Copyright © 1998-2009 Netscape
4,574,237 sites - 82,468 editors - over 590,000 categories
Internet

分类的概念

- 分类：对于给定一个对象，从一个事先定好的分类体系中挑出一个（或者多个）最适合该对象的类别。
  - 对象：可以是任何东西
  - 事先定好的分类体系：可能有结构
  - 最适合：判断标准
- 便于今后查找是其最直接、最普遍的应用

分类体系

- 分类体系一般人工构造
  - 政治、体育、军事
- 分类系统可以是层次结构
- 分类模式
  - 2类问题，属于或不属于(binary)
  - 多类问题，多个类别(multi-class)，可拆分成2类问题
  - 一个对象可以属于多类(multi-label)



人工方法和自动方法

- 人工方法
  - 知识工程的方法建立专家系统(80年代末期)
  - 结果容易理解
    - 足球 and 联赛→体育类
  - 费时费力
    - MEDLINE(National Library of Medicine) $2 million/year for manual indexing of journal articles
  - 难以保证一致性和准确性(40%左右的准确率)
  - 专家有时候凭空想象
- 自动的方法(学习)
  - 结果可能不易理解
  - 快速
  - 准确率相对高(准确率可达85%或者更高)
  - 来源于真实文本，可信度高

文本自动分类的定义

- Text Categorization (TC)
- 在给定的分类体系下，根据文本的内容自动地确定文本关联的类别。
- 从数学角度来看，文本分类是一个映射的过程，它将未标明类别的文本映射到已有的类别中，该映射可以是一一映射或一对多的映射。
- 用数学公式表示如下：

$f: A \rightarrow B$　其中，A为待分类的文本集合，

B为分类体系中的类别集合

应用领域

- 门户网站（网页）
- 图书馆（电子资料）
- 情报/信息部门（情报处理）
- 政府、企业等（电子邮件）

自动分类的优点

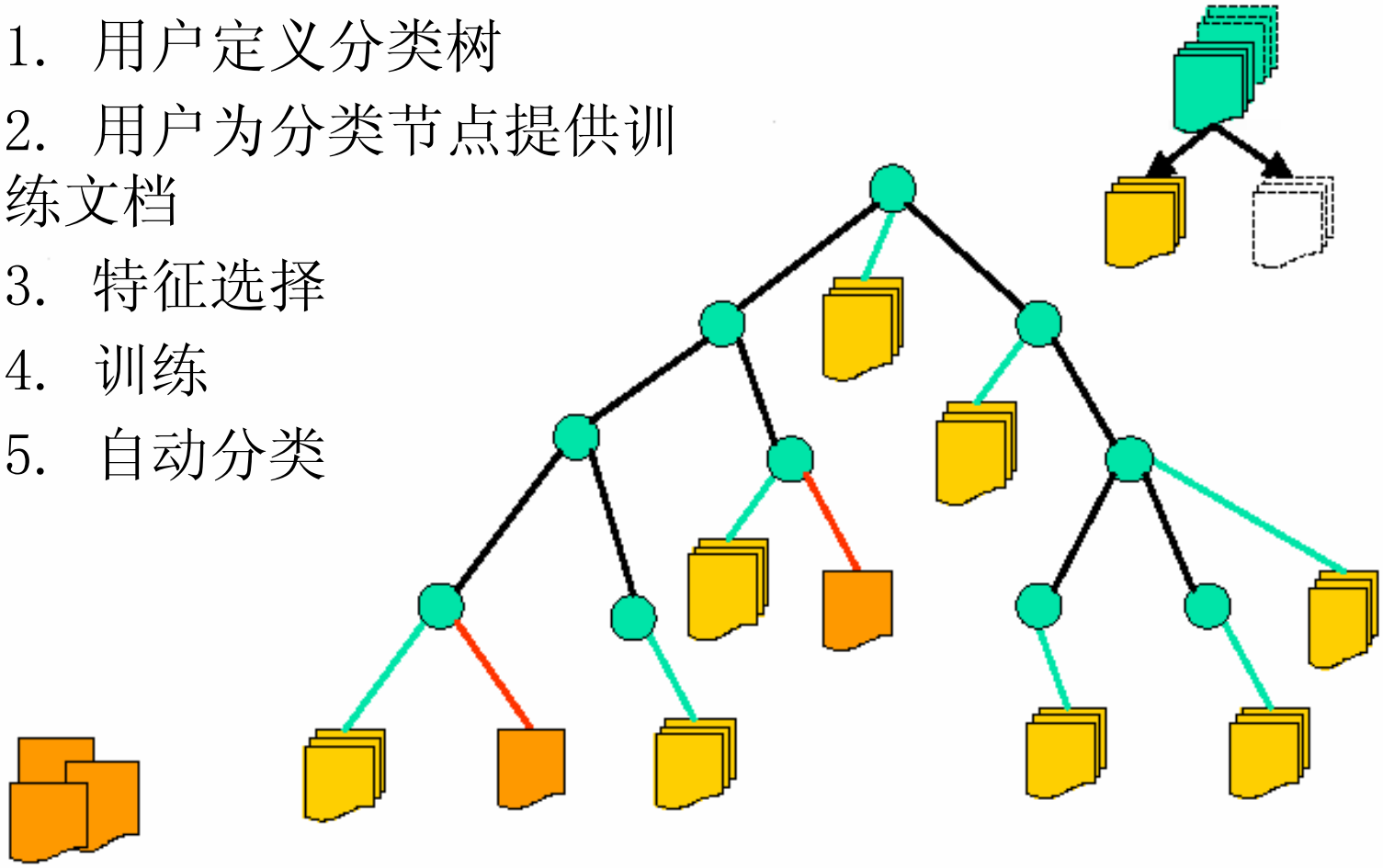
- 减小人工分类的繁杂工作
- 提高信息处理的效率
- 减小人工分类的主观性

文本自动分类

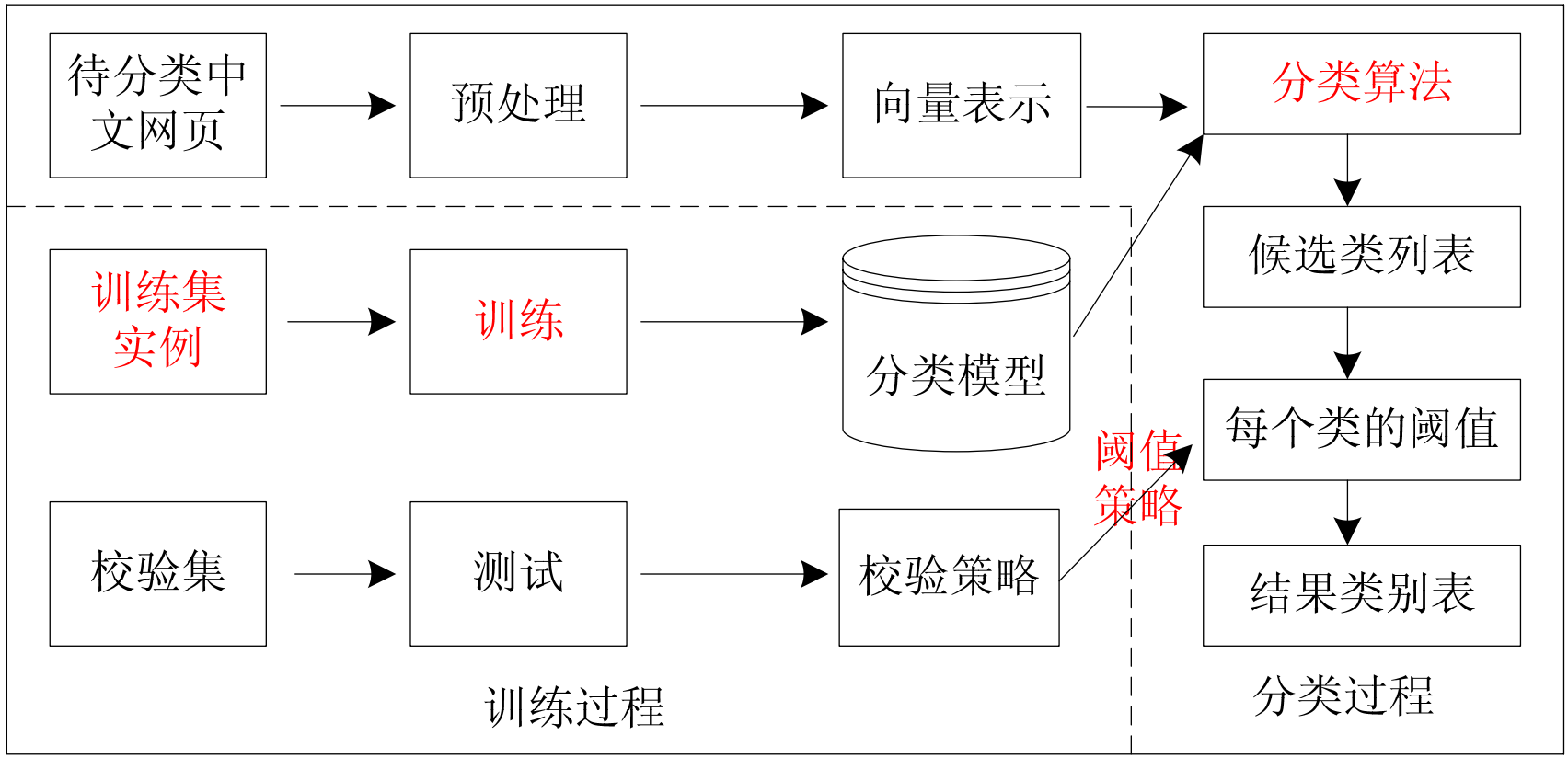
- 基本步骤
  - 定义分类体系
  - 将预先分类过的文档作为训练集
  - 从训练集中得出分类模型（需要测试过程，不断细化）
  - 用训练获得出的分类模型对其它文档加以分类

文本分类基本步骤

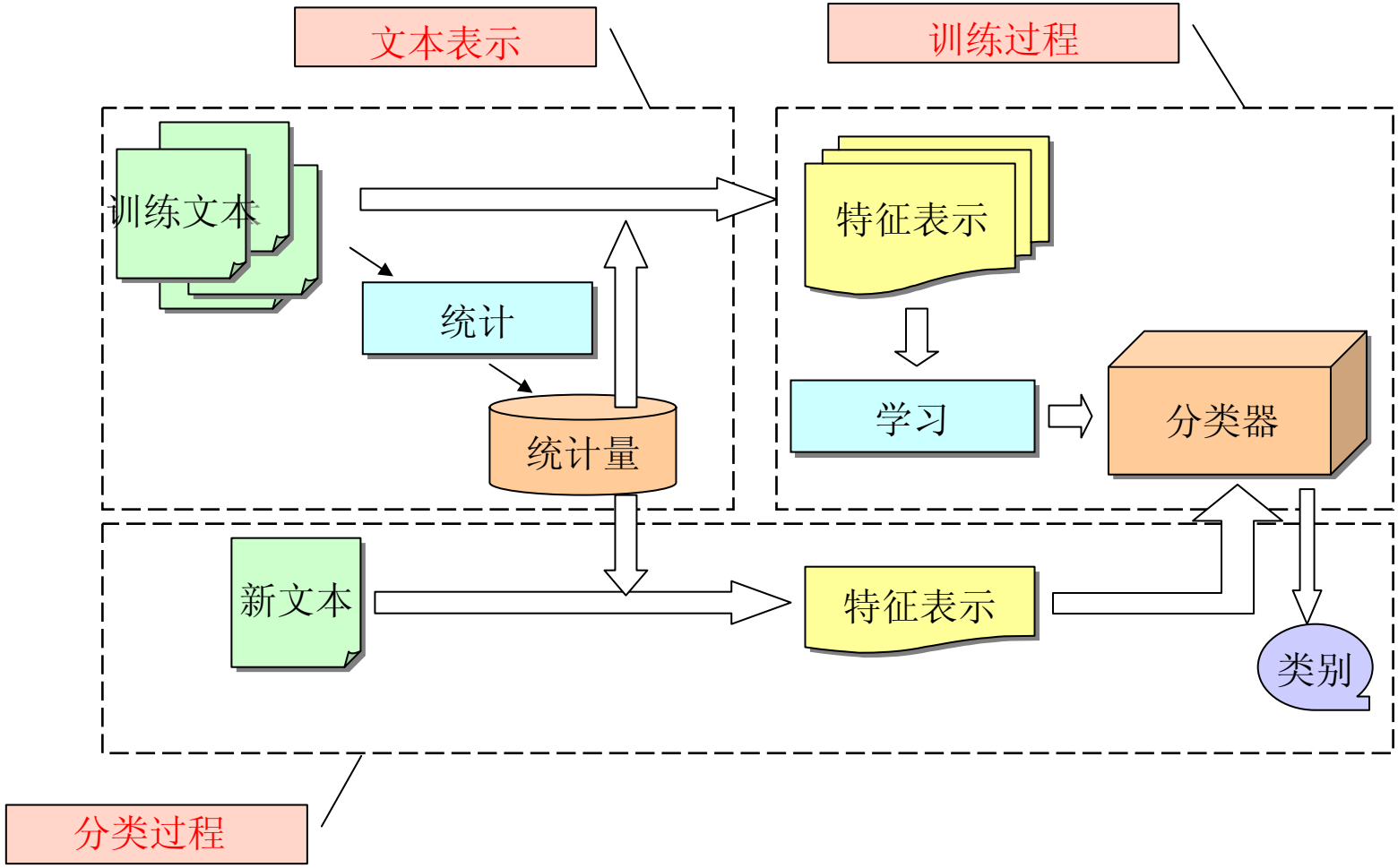
- 1. 用户定义分类树
- 2. 用户为分类节点提供训练文档
- 3. 特征选择
- 4. 训练
- 5. 自动分类



文本分类基本步骤

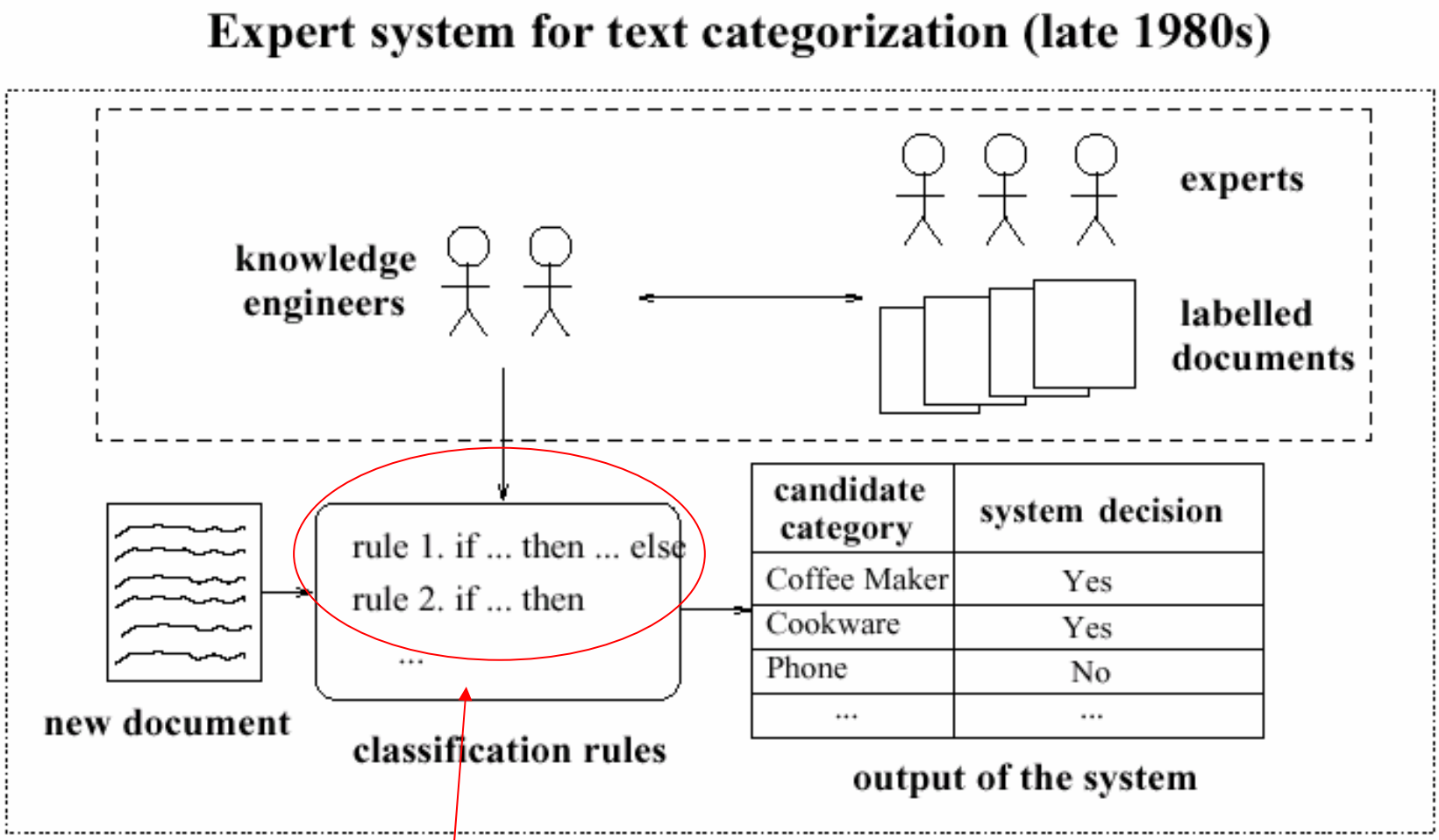

待分类中文网页
预处理
向量表示
分类算法
候选类列表
每个类的阈值
结果类别表
训练集实例
训练
分类模型
阈值策略
校验集
测试
校验策略
训练过程
分类过程

文本分类的过程



自动分类技术的发展

专家系统(late 1980s)



人工定义规则

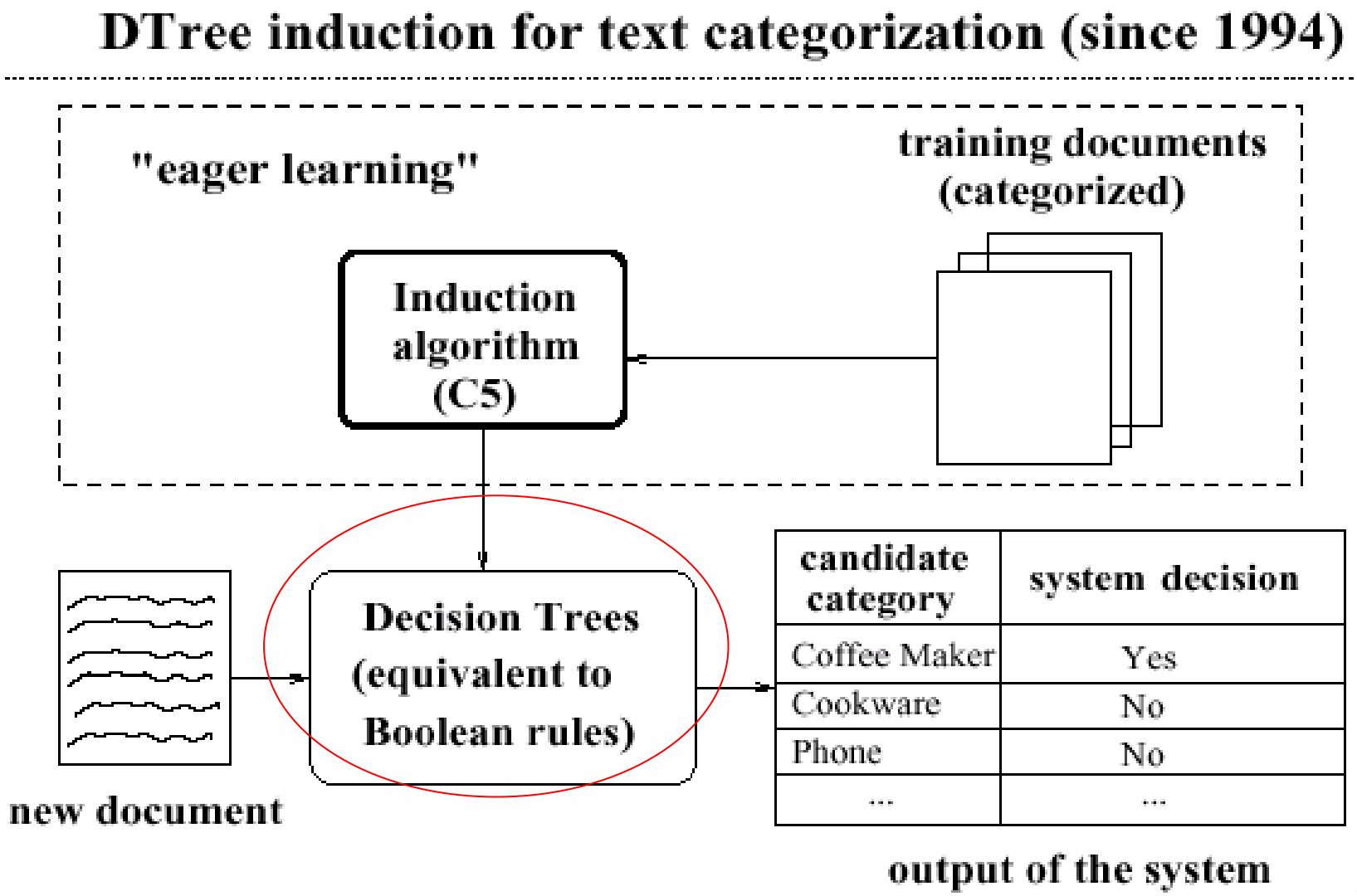
专家系统

- 专家系统（人工定义规则）
  - 太花时间
  - 太难（最初看起来容易）
  - 一致性问题（规则集很大）

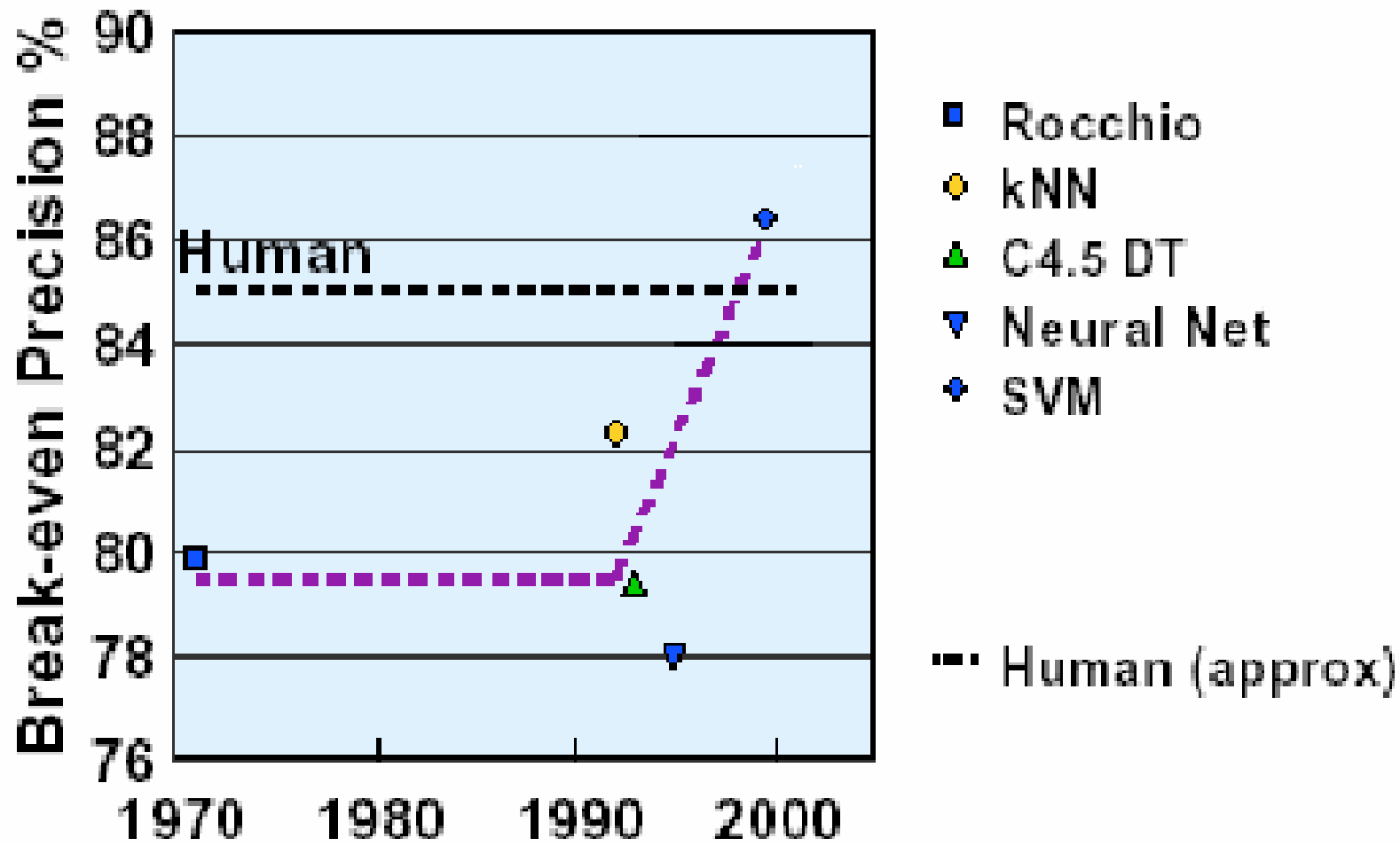
专家系统

- 美国人口调查局（1990）
  - 十年人口统计资料的分析（2200万项资料）
  - 232个产业类别和504个行业类别
  - $15 million if fully done by hand
- 人工定义规则
  - Expert System AIOCS
  - Development time: 192 person-months (2 people, 8 years)
  - Accuracy = 47%
- 基于机器学习的方法
  - 最近邻分类方法 (Creecy '92: 1-NN)
  - Development time: 4 person-months
  - Accuracy = 60%

统计学习取代知识工程


DTree induction for text categorization (since 1994)
"eager learning"
training documents
(categorized)
Induction
algorithm
(C5)
Decision Trees
(equivalent to
Boolean rules)
new document
candidate category
system decision
Coffee Maker
Yes
Cookware
No
Phone
No
...
...
output of the system

自动分类技术发展


Break-even Precision %
90
88
86
84
82
80
78
76
1970
1980
1990
2000
Human
Rocchio
kNN
C4.5 DT
Neural Net
SVM
Human (approx)

文本分类实例

新闻自动分类

- Given: Collection of example news stories already labeled with a category (topic).
- Task: Predict category for news stories not yet labeled.
- For our example, we'll only get to see the headline（标题） of the news story.
- We'll represent categories using colors. (All examples with the same color belong to the same category.)

人工标注的样例

政府事务 | 企业个人事务

Amatil Proposes Two-for-Five Bonus Share Issue	Citibank Norway Unit Loses Six Mln Crowns in 1986	Japan Ministry Says Open Farm Trade Would Hit U.S.	Vieille Montagne Says 1986 Conditions Unfavourable	Jardine Matheson Said It Sets Two-for-Five Bonus Issue Replacing “B” Shares
Anheuser-Busch Joins Bid for San Miguel	Italy’s La Fondiaria to Report Higher 1986 Profits	Isuzu Plans No Interim Dividend	Senator Defends U.S. Mandatory Farm Control Bill	Bowater Industries Profit Exceed Expectations

什么没看到之前

能给一个新闻赋予什么颜色？

?

分类预测：
取多数？

Amatil Proposes Two-for-Five Bonus Share Issue	Citibank Norway Unit Loses Six Mln Crowns in 1986	Japan Ministry Says Open Farm Trade Would Hit U.S.	Vieille Montagne Says 1986 Conditions Unfavourable	Jardine Matheson Said It Sets Two-for-Five Bonus Issue Replacing “B” Shares
Anheuser-Busch Joins Bid for San Miguel	Italy’s La Fondiaria to Report Higher 1986 Profits	Isuzu Plans No Interim Dividend	Senator Defends U.S. Mandatory Farm Control Bill	Bowater Industries Profit Exceed Expectations

看见标题

Senate Panel Studies Loan Rate, Set Aside Plans

Amatil Proposes Two-for-Five Bonus Share Issue	Citibank Norway Unit Loses Six Mln Crowns in 1986	Japan Ministry Says Open Farm Trade Would Hit U.S.	Vieille Montagne Says 1986 Conditions Unfavourable	Jardine Matheson Said It Sets Two-for-Five Bonus Issue Replacing “B” Shares
Anheuser-Busch Joins Bid for San Miguel	Italy’s La Fondiaria to Report Higher 1986 Profits	Isuzu Plans No Interim Dividend	Senator Defends U.S. Mandatory Farm Control Bill	Bowater Industries Profit Exceed Expectations

得到分类：政府事务

Senate Panel Studies Loan Rate, Set Aside Plans

Amatil Proposes Two-for-Five Bonus Share Issue	Citibank Norway Unit Loses Six Mln Crowns in 1986	Japan Ministry Says Open Farm Trade Would Hit U.S.	Vieille Montagne Says 1986 Conditions Unfavourable	Jardine Matheson Said It Sets Two-for-Five Bonus Issue Replacing “B” Shares
Anheuser-Busch Joins Bid for San Miguel	Italy’s La Fondiaria to Report Higher 1986 Profits	Isuzu Plans No Interim Dividend	Senator Defends U.S. Mandatory Farm Control Bill	Bowater Industries Profit Exceed Expectations

评价指标

评价指标

- 「准确率」（P，precision）
- 「召回率」（R，recall）
- F—Measure

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

$$F_1 = \frac{2PR}{P+R}$$

		Human	
		True	False
Classifier	Yes	a	b
	No	c	d

准确率 P = a/(a+b)

召回率 R = a/(a+c)

评价指标

- 每个类
  - Precision=a/(a+b)
  - Recall=a/(a+c),
  - miss rate=1-recall
  - accuracy=(a+d)/(a+b+c+d), error=(b+c)/(a+b+c+d)=1-accuracy
  - fallout=b/(b+d)=false alarm rate,
  - F=(β^2+1)p.r/(β^2p+r)
  - Break Even Point, BEP, p=r的点
  - interpolated 11 point average precision (p-r曲线)

		Human	
		True	False
Classifier	Yes	a	b
	No	c	d

准确率 P = a/(a+b)
召回率 R = a/(a+c)

评价指标

- 所有类的总体评价

$$F_i = \frac{1}{\alpha \frac{1}{P_i} + (1-\alpha)\frac{1}{R_i}}$$

- 宏平均 Macro

$$F_{1i} = \frac{2P_iR_i}{P_i + R_i}$$

$$Macro - F = \frac{1}{m}\sum_{i=1}^{m} F_i$$

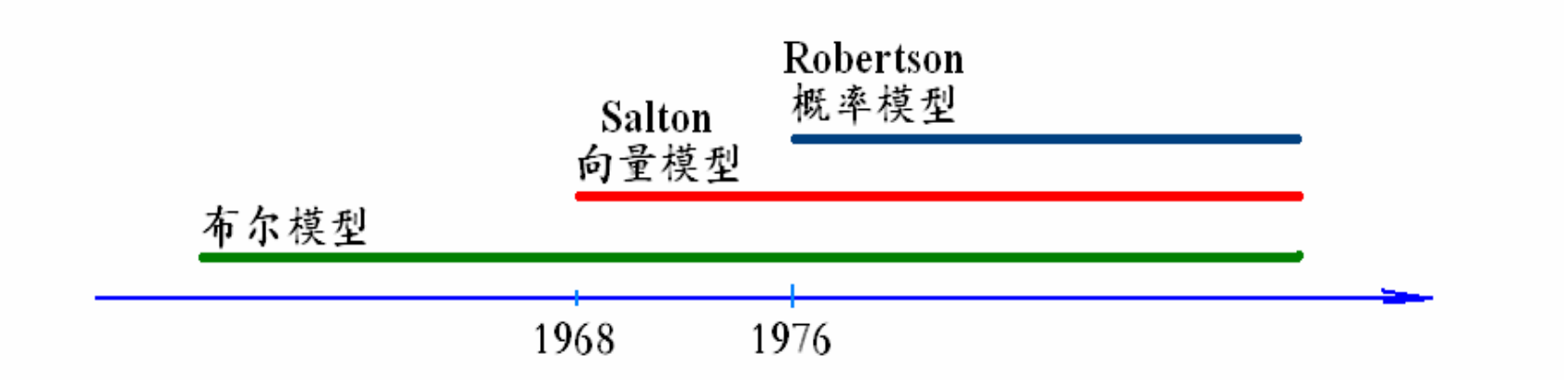
- 微平均 Micro

$$Micro - F = \sum_{i=1}^{m}(n_i \cdot F_i) \Big/ \sum_{i=1}^{m} n_i$$

特征抽取

文档模型

- 布尔模型
- 向量空间模型
- 概率模型

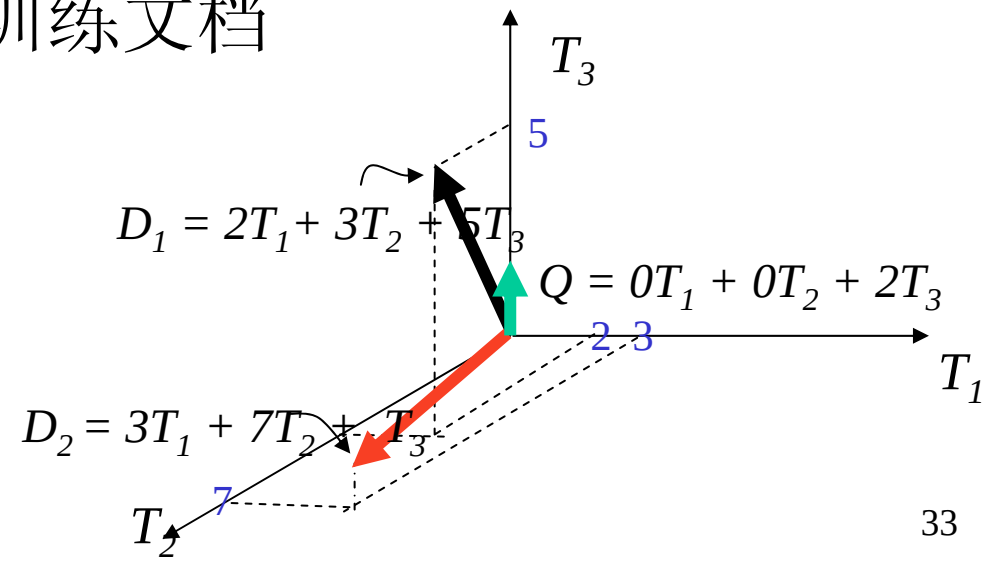


特征抽取(feature extraction)

- 预处理
  - 去掉html一些tag标记
  - 停用词(stop words)去除、词根还原(stemming)
  - (中文)分词、词性标注、短语识别、…
  - 词频统计 (TF DF)
  - 数据清洗：去掉噪声文档或文档内垃圾数据
- 文本表示
  - 向量空间模型
- 降维技术
  - 特征选择(Feature Selection)
  - 特征重构(Re-parameterisation，如LSI)

向量空间模型

- 向量空间模型(Vector Space Model)
  - M个无序标引项t_i（特征），词根/词/短语/其他
  - 每个文档d_j可以用标引项向量来表示
    - $(a_{1j}, a_{2j}, \cdots, a_{Mj})$
  - 权重计算，N个训练文档
    - $A_{M*N}= (a_{ij})$
  - 相似度比较
    - Cosine计算
    - 内积计算



Term的粒度

- Character，字：中
- Word，词：中国
- Phrase，短语：中国人民银行
- Concept，概念
  - 同义词：开心 高兴 兴奋
  - 相关词cluster，word cluster：葛非/顾俊
- N-gram，N元组：中国 国人 人民 民银 银行
- 某种规律性模式：比如某个window中出现的固定模式
- David Lewis等一致地认为：(英文分类中)使用优化合并后的Words比较合适

权重计算方法

- 布尔权重(boolean weighting)
  - a_{ij}=1(TF_{ij}>0) or 0 (TF_{ij}=0)
- TFIDF型权重
  - TF: $a_{ij}=TF_{ij}$
  - TF*IDF: $a_{ij}=TF_{ij}*\log(N/DF_i)$
  - TFC: 对上面进行归一化 $a_{ij}=\frac{TF_{ij}*\log(N/DF_i)}{\sqrt{\sum_k [TF_{kj}*\log(N/DF_k)]^2}}$
  - LTC: 降低TF的作用

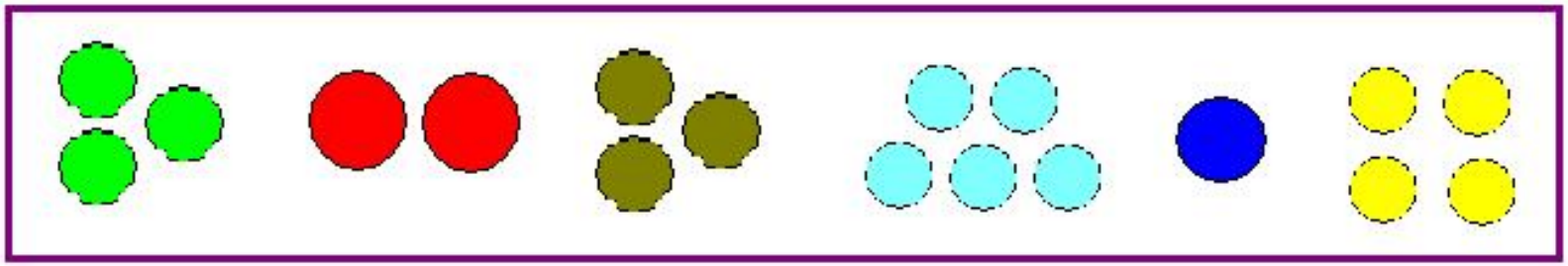
$$a_{ij}=\frac{\log(TF_{ij}+1.0)*\log(N/DF_i)}{\sqrt{\sum_k [\log(TF_{kj}+1.0)*\log(N/DF_k)]^2}}$$

特征选择

总体属性（词项）集合

Σ: t_1 t_2 t_3 t_4 ... t_k ... t_N

我们需要确定其中哪一些具有较强的"类别区分"能力



已分类的文档训练集

特征选择

- 目的
  - 避免过拟合（over fitting），提高分类准确度
    - 如果经过某种学习之后的分类模型，使得训练文档适应得很好（导致很高的自动分类精度），但对训练集之外的文档显得差许多，则称此时的学习模型有

      Over-fitting problem
    - 希望模型的表现对训练集和未知文档基本一致。
  - 通过降维，大大节省计算时间和空间
    - 样例空间涉及的总词项数很大（N在10万量级），但每篇文档只涉及其中的一小部分（例如一篇网页通常只有几百个词）（到1/10 – 1/100，甚至更多）
- 基本信念：除那些stop words外，还有许多词实际上对分类没什么贡献
- 但如何确定这些词？

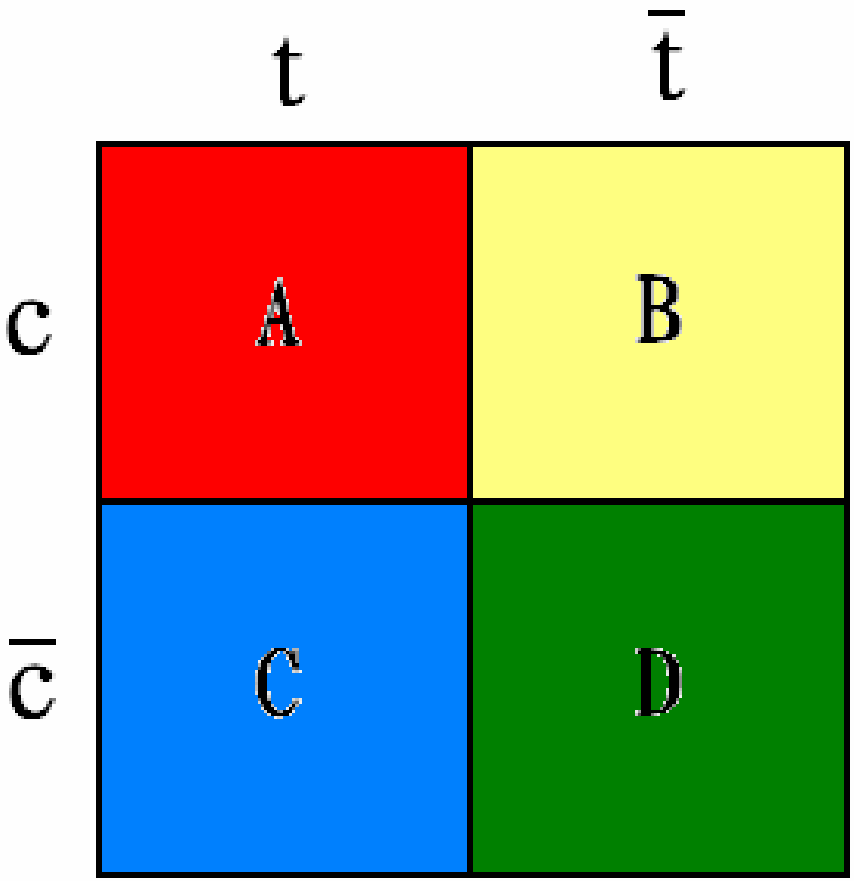
特征选取的方法

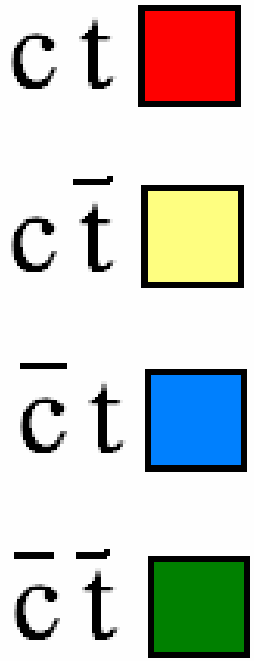
- 文档频率法（DF, document frequency）
- 信息增益法（information gain）
- 互信息法（mutual information）
- The χ^2 test（chi-square，开方拟合检验）

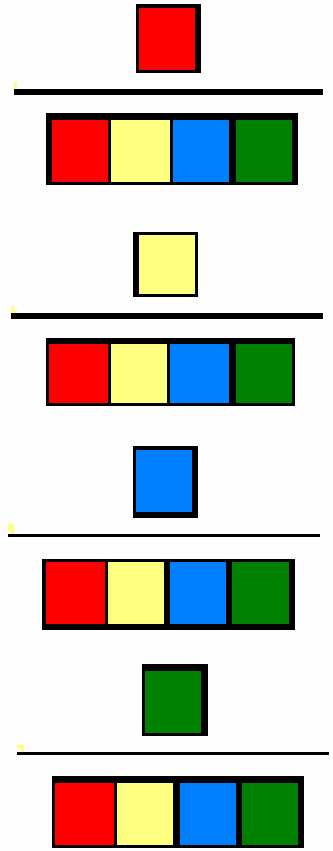
特征选择--DF

- 基于DF的启发式要点
  - 太频繁的词项没有区分度
    - Term的DF大于某个阈值去掉
  - 太稀有的词项独立表达的类别信息不强
    - 稀有词项的全局影响力不大
    - 在训练集中，某些文档如果有某个稀有词项，它们通常也会有一些常见词项（对那一类）
    - 和通常信息获取观念有些抵触：稀有的更有代表性（这是一种*ad hoc*方法，不依据什么理论）
- 最容易实现，可扩展性好

相关表
Contingency table (matrix)







（类，特征）相关表

特征选择--熵

- 设信息出现（例如“硬币出现某一面”，“一篇文档属于某一类”）的概率空间

$$p = \{p_1, p_2, \ldots, p_m\}$$

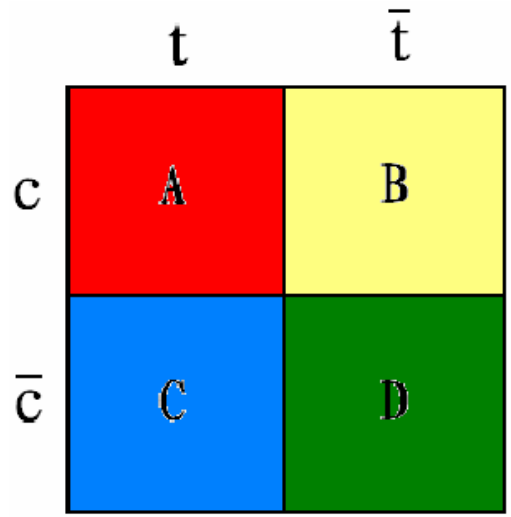
- 在引入某个词项t之前，系统的熵（即一个随机文档落入某个类的概率空间的熵）

$$Entropy(t) = -\sum_i P_i \log P_i$$

$$P_i = c_i / c$$

$$P_i = (A + B) / n$$

$$n = A + B + C + D$$



特征选择--熵

- 在观察到t以后，文档落入某个类c_i的概率就应该是条件概率$P(c_i \mid t)$
  - 对应于相关表中的A/A+C
  - 注意，对不同的c_i，每个分量不一定相同

$$Entropy(t) = -\sum_i P(c_i \mid t) \log P(c_i \mid t)$$

- term类别分布的熵：
  - 该值越大，说明分布越均匀，越有可能出现在较多的类别中；
  - 该值越小，说明分布越倾斜，词可能出现在较少的类别中

特征选择--信息增益IG

- 信息增益(Information Gain, IG)：
  - 该term为整个分类所能提供的信息量
  - t出现与否导致的熵的变化
  - 不考虑任何特征的熵和考虑该特征后的熵的差值

$$\text{Gain(t)} = \text{Entropy}(S) - \text{Expected Entropy}(S_t)$$

$$\begin{aligned} &= \{-\sum_{i=1}^{M} P(c_i)\log P(c_i)\} \\ &-[P(t)\{-\sum_{i=1}^{M} P(c_i \mid t)\log P(c_i \mid t)\} \\ &+ P(\bar{t})\{-\sum_{i=1}^{M} P(c_i \mid \bar{t})\log P(c_i \mid \bar{t})\}] \end{aligned}$$

特征选择--互信息MI

- 互信息MI是统计建模的一种方法，是评估两个随机变量X、Y相关程度的一种度量

$$MI(X,Y)=\sum_{x}\sum_{y}P(x,y)\log\frac{P(x,y)}{P(x)P(y)}$$

其中P (x,y)是变量取值（x,y）的概率

特征选择--互信息MI

- X，Y分别对应词项t的出现情况和类别的出现情况
  - 看成随机事件，t可能出现0,1,2,...次（取决于要用的文档模型）；
  - 类别有m种可能$c=\{c_1, c_2, ..., c_m\}$
  - 不混淆情况下，用t，c表示这两个随机变量
- 关心的P(t)，P(c)，P(t,c)都可以通过统计训练集中的数据情况得到
  - 如果用二元模型，即前面的“相关表”就足够，
  - 如果是多元模型，也容易推广

特征选择--互信息MI

- 互信息(Mutual Information)：MI越大t和c共现程度越大
- (N=A+B+C+D)

	c	~c
t	A	B
~t	C	D

$$I(t,c)=\log\frac{P(t\wedge c)}{P(t)P(c)}$$

$$=\log\frac{A/N}{((A+B)/N)*((A+C)/N)}=\log\frac{A\times N}{(A+B)(A+C)}$$

$$I_{AVG}(t)=\sum_{i=1}^{m}P(c_i)I(t,c_i)$$

$$I_{MAX}(t)=\max_{i=1}^{m}P(c_i)I(t,c_i)$$

特征选择--互信息MI

- 特点
  - MI(t,C)的值越大，t对于C的区分能力越强
  - 对同一个类，不同的词项，在同样P(t|c)情况下，相对稀有的 t 会得到较大的值
  - MI还可以解释为给定一个随机变量后另外一个随机变量上的减少

特征选择 -- χ^2（卡方）

- 源于统计学的卡方分布(chi-square)
- 从（类，词项）相关表出发
  - 考虑每一个类和每一个词项的相关情况，chi-square(t,c)

特征选择--χ^2(卡方)

- χ^2 统计量:
  - 度量两者(term和类别)独立性的缺乏程度
  - χ^2 越大，独立性越小，相关性越大
  - 若AD<BC,则类和词独立， N=A+B+C+D

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

$$\chi^2_{AVG}(t)=\sum_{i=1}^{m}P(c_i)\chi^2(t,c_i)$$

	c	~c
t	A	B
~t	C	D

$$\chi^2_{MAX}(t)=\max_{i=1}^{m}\{\chi^2(t,c_i)\}$$

特征选择—术语强度（TS）

- term strength
- 一种新颖的角度，考察一个词项在“相似文档”中出现的可能性
- 假定我们已经有了一个相似文档的集合S，设x，y为其中任意两个文档，那么词项t对这个集合的术语强度为

$$s(t) = Pr(t \varepsilon y \mid t \varepsilon x)$$

- S中的文档要满足一定的“相似度”指标，S不一定就是训练集中的那些类。通常由聚类来确定S。

特征选择

- DF: Document Frequency
- IG: Information Gain

$$G(t) = -\sum_{i=1}^{m} P_r(c_i)\log P_r(c_i)$$

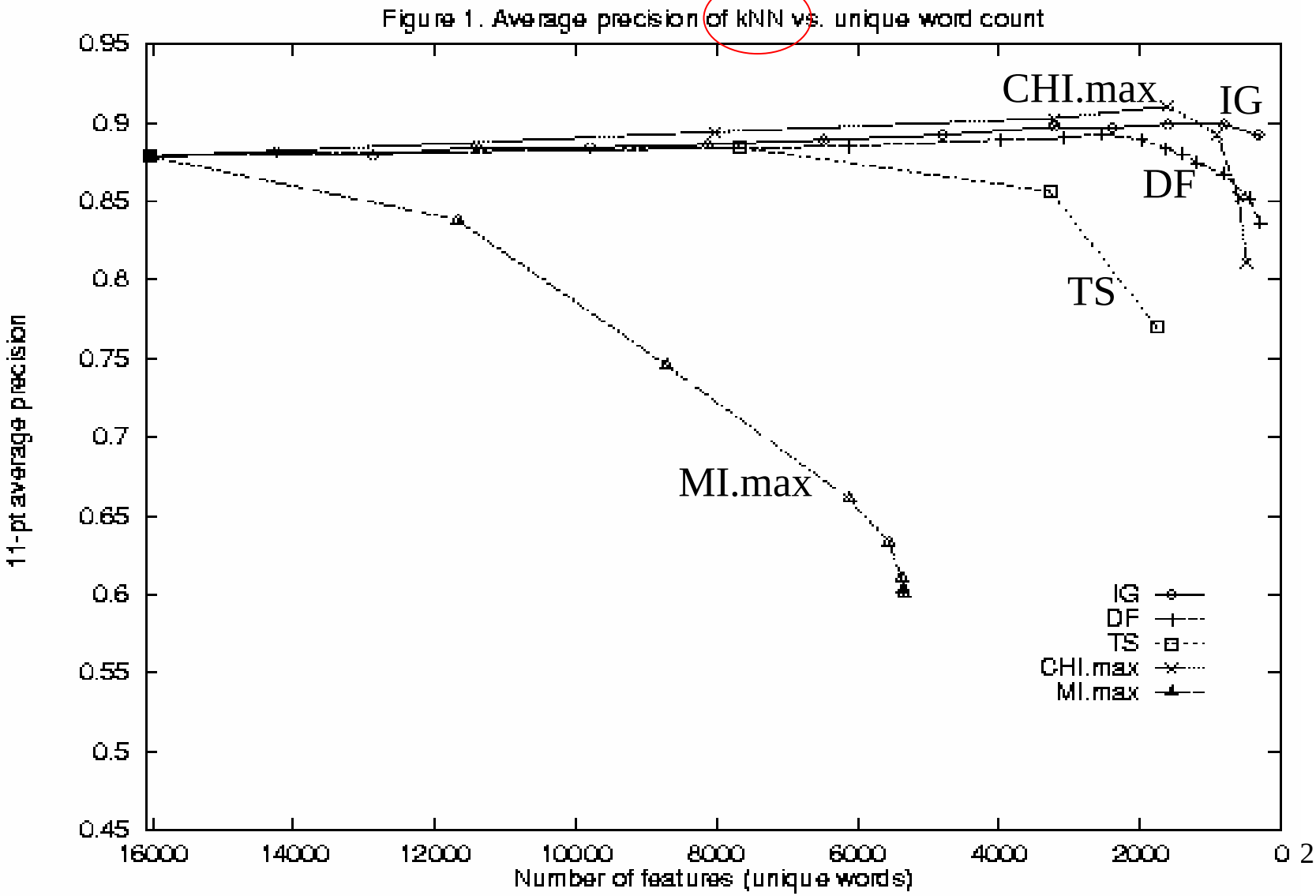
$$+p_r(t)\sum_{i=1}^{m} P_r(c_i \mid t)\log P_r(c_i \mid t) + P_r(\bar{t})\sum_{i=1}^{m} P_r(c_i \mid \bar{t})\log P_r(c_i \mid \bar{t})$$

- MI: Mutual Information

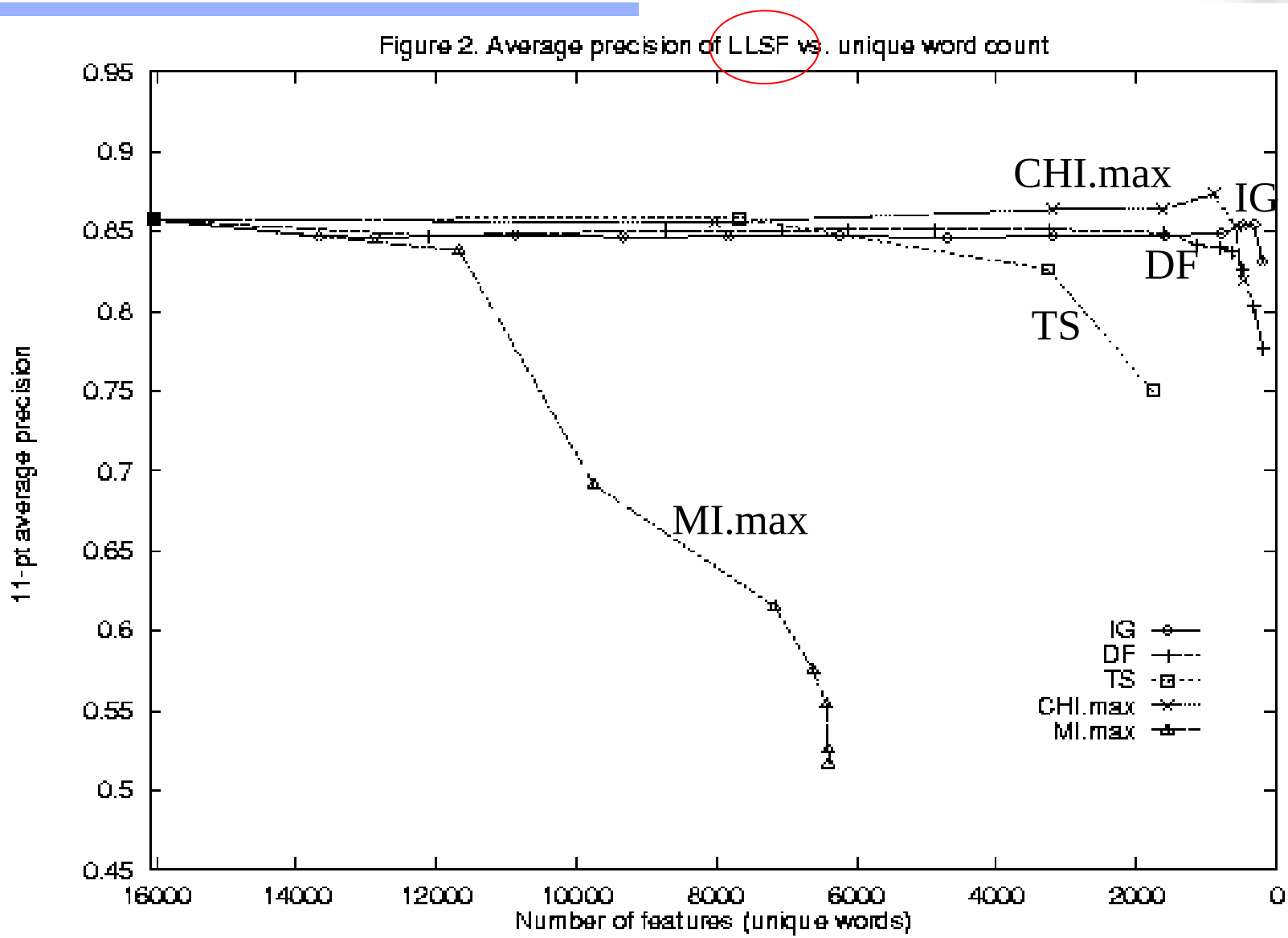
$$I(t,c) \approx \log \frac{A \times N}{(A+C)\times(A+B)}$$

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

特征选择方法的性能比较(kNN)

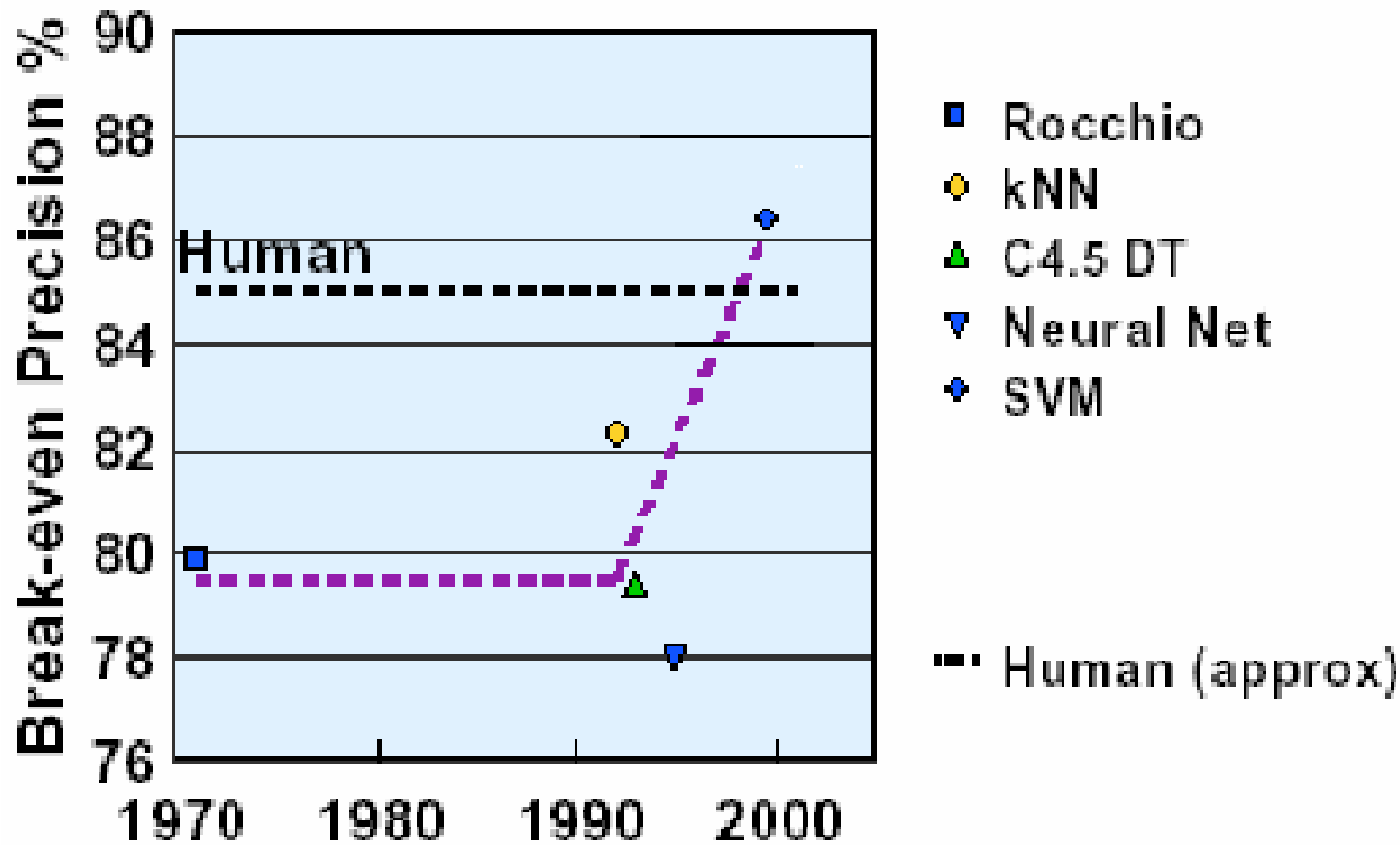


特征选择方法的性能比较 (LLSF)



分类算法

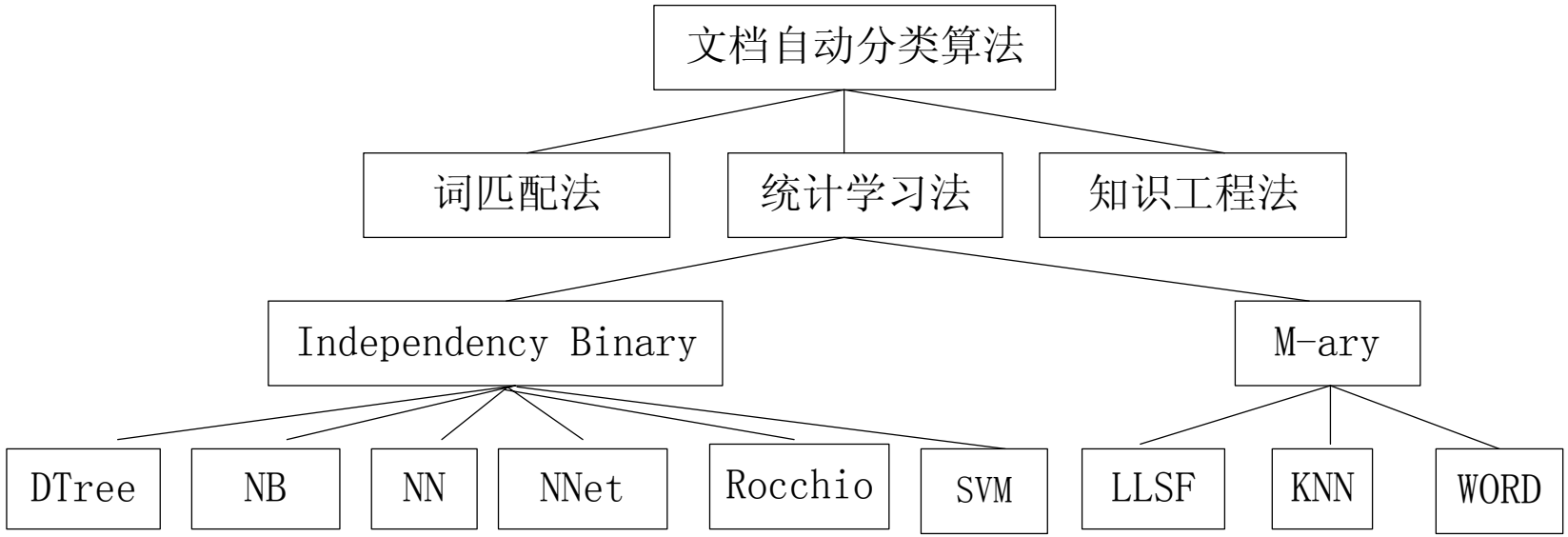
分类技术发展


Break-even Precision %
90
88
86
84
82
80
78
76
1970
1980
1990
2000
Human
Rocchio
kNN
C4.5 DT
Neural Net
SVM
Human (approx)

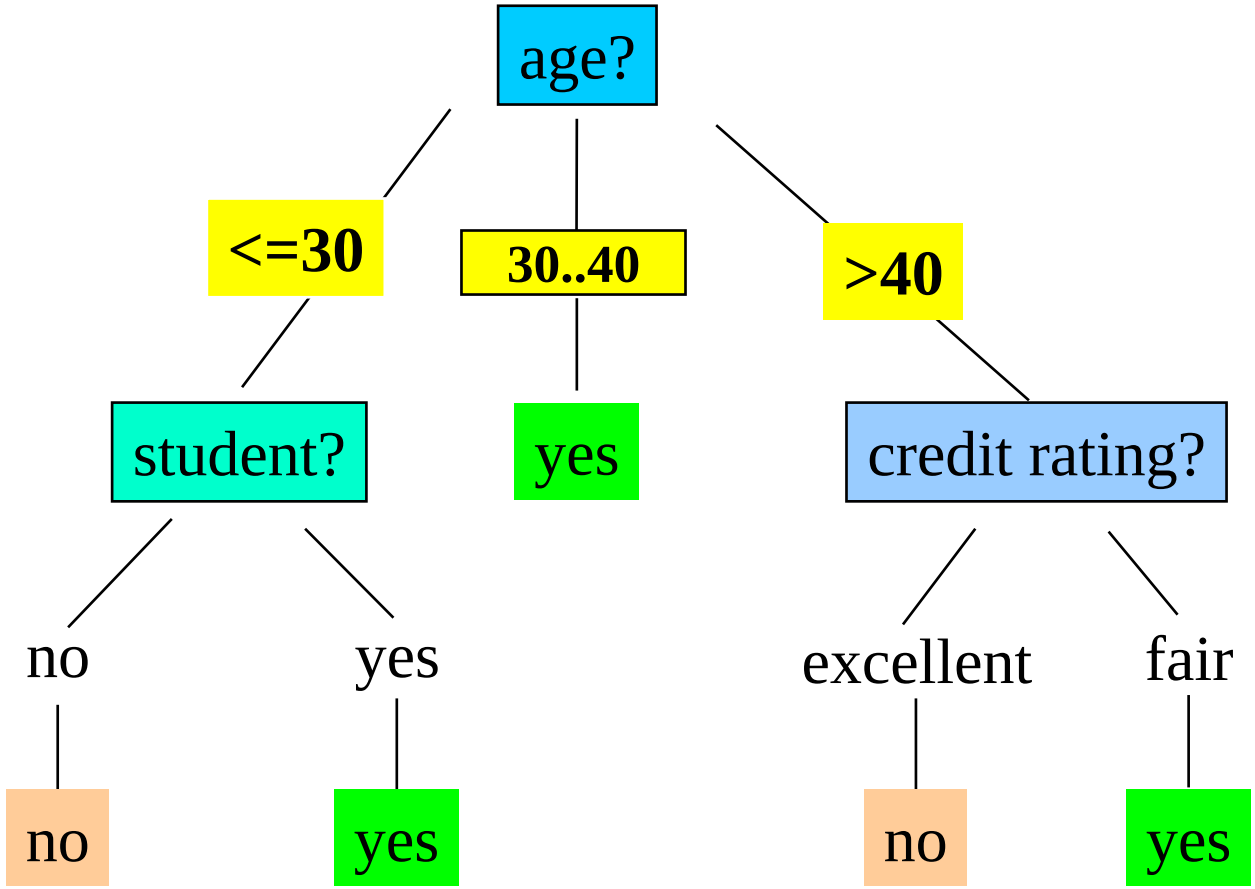
分类算法

- 决策树（Decision Trees）
- KNN算法(K-Nearest Neighbour)
- 贝叶斯网络（Bayes Network）
- 神经网络（Neural Networks）
- Boosting
- 支持向量机（SVM）

自动分类算法分类


文档自动分类算法
词匹配法
统计学习法
知识工程法
Independency Binary
M-ary
DTree
NB
NN
NNet
Rocchio
SVM
LLSF
KNN
WORD

决策树方法


age?
<=30
30..40
>40
student?
yes
credit rating?
no
yes
excellent
fair
no
yes
no
yes

决策树方法

- 构造决策树
  - CART
  - C4.5（由ID3发展而来）
  - CHAID
- 决策树的剪枝(pruning)

Attribute Selection Measure: Information Gain(ID3/C4.5)

- 选择信息增益最大的属性
- S contains si tuples of class Ci for i = {1, …, m}
- information measures info required to classify any arbitrary tuple

$$I(s_1,s_2,\ldots,s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \log_2 \frac{s_i}{s}$$

Attribute Selection Measure: Information Gain(ID3/C4.5)

- entropy of attribute A with values {a1, a2, …, av}

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} I(s_{1j}, \ldots, s_{mj})$$

- information gained by branching on attribute A

$$Gain(A) = I(s_1, s_2, \ldots, s_m) - E(A)$$

选择信息增益最大的属性作为判定的分支节点

Rocchio方法

- Rocchio公式

$$w'_{jc} = \alpha w_{jc} + \beta \frac{\sum_{i \in C} x_{ij}}{n_C} - \gamma \frac{\sum_{i \notin C} x_{ij}}{n - n_C}$$

类C中心向量的权重　　训练样本中正例个数　　文档向量的权重

- 分类　$CSV_c(d_i) = \mathrm{w}_c \cdot \mathrm{x}_i = \dfrac{\sum w_{cj} \cdot x_{ij}}{\sqrt{\sum w_{cj}^2}\sqrt{\sum x_{ij}^2}}$

- 可以认为类中心向量法是它的特例

K-NN分类方法

1-Nearest Neighbor

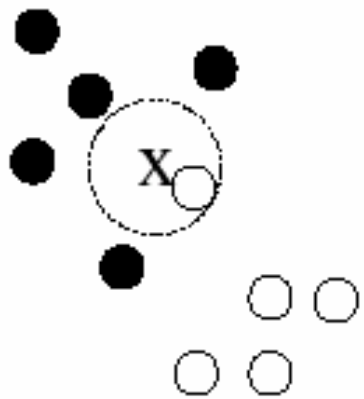
- 回顾前面的例子
  - Did anyone try to find the most similar labeled item and then just guess the same color?
  - This is 1-Nearest Neighbor

Senate Panel Studies Loan Rate, Set Aside Plans

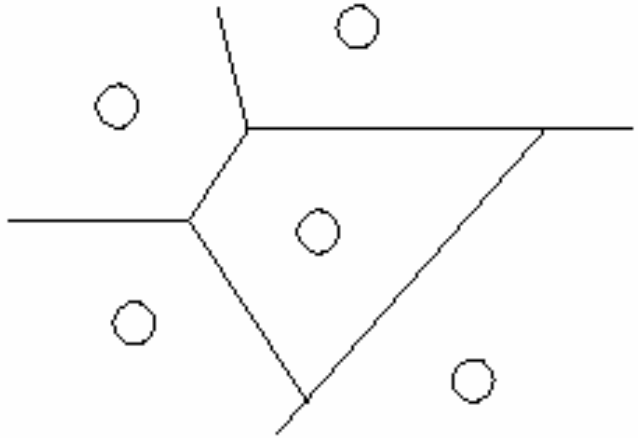
Amatil Proposes Two-for-Five Bonus Share Issue	Citibank Norway Unit Loses Six Mln Crowns in 1986	Japan Ministry Says Open Farm Trade Would Hit U.S.	Vieille Montagne Says 1986 Conditions Unfavourable	Jardine Matheson Said It Sets Two-for-Five Bonus Issue Replacing "B" Shares
Anheuser-Busch Joins Bid for San Miguel	Italy's La Fondiaria to Report Higher 1986 Profits	Isuzu Plans No Interim Dividend	Senator Defends U.S. Mandatory Farm Control Bill	Bowater Industries Profit Exceed Expectations

1-Nearest Neighbor (graphically)

1-NN: assign "x" (new point) to the class of it nearest neighbor



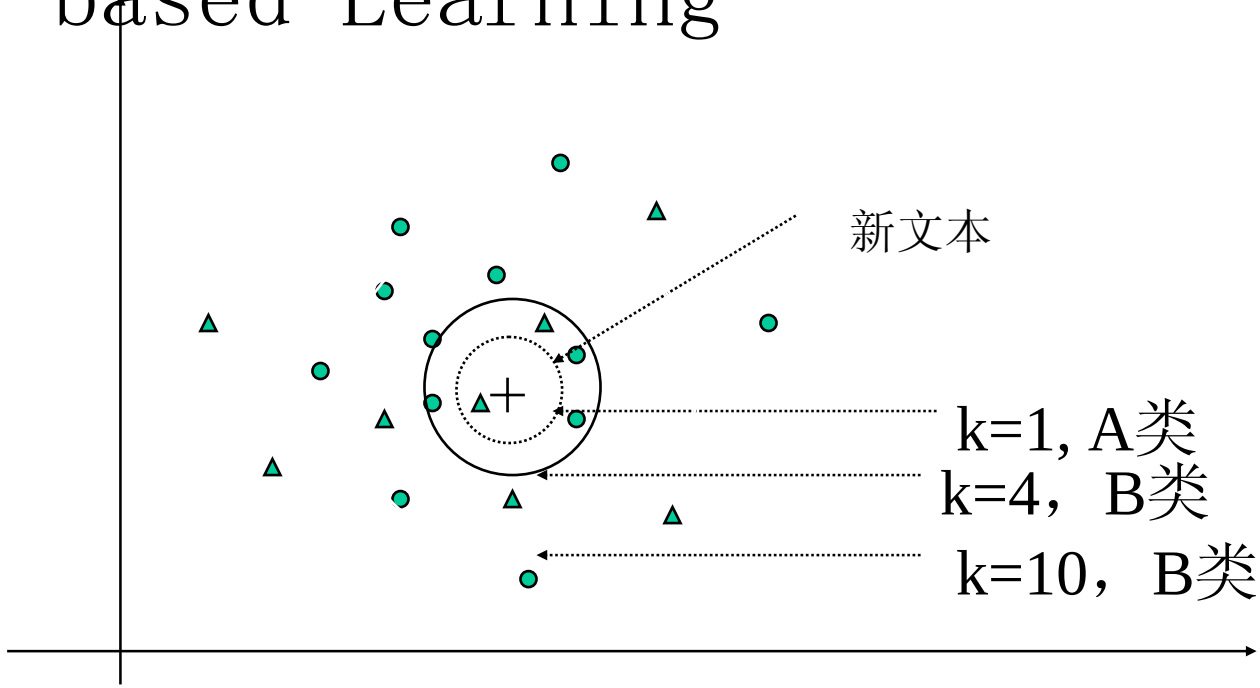
assign "x" to "white"



decision surface divided by points ("Voronoi diagram")

kNN方法

- 一种Lazy Learning, Example-based Learning



带权重计算，计算权重和最大的类。k常取3或者5。

kNN方法的发展

- *Instance-Based Learning, Lazy Learning*
- well-known approach to pattern recognition
- initially by Fix and Hodges (1951)
- theoretical error bound analysis by Duda & Hart (1957)
- applied to text categorization in early 90's (YYM)
- among top-performing methods in TC evaluations
- scalable to large TC applications

kNN for Text Categorization (Yang YM, SIGIR-1994)

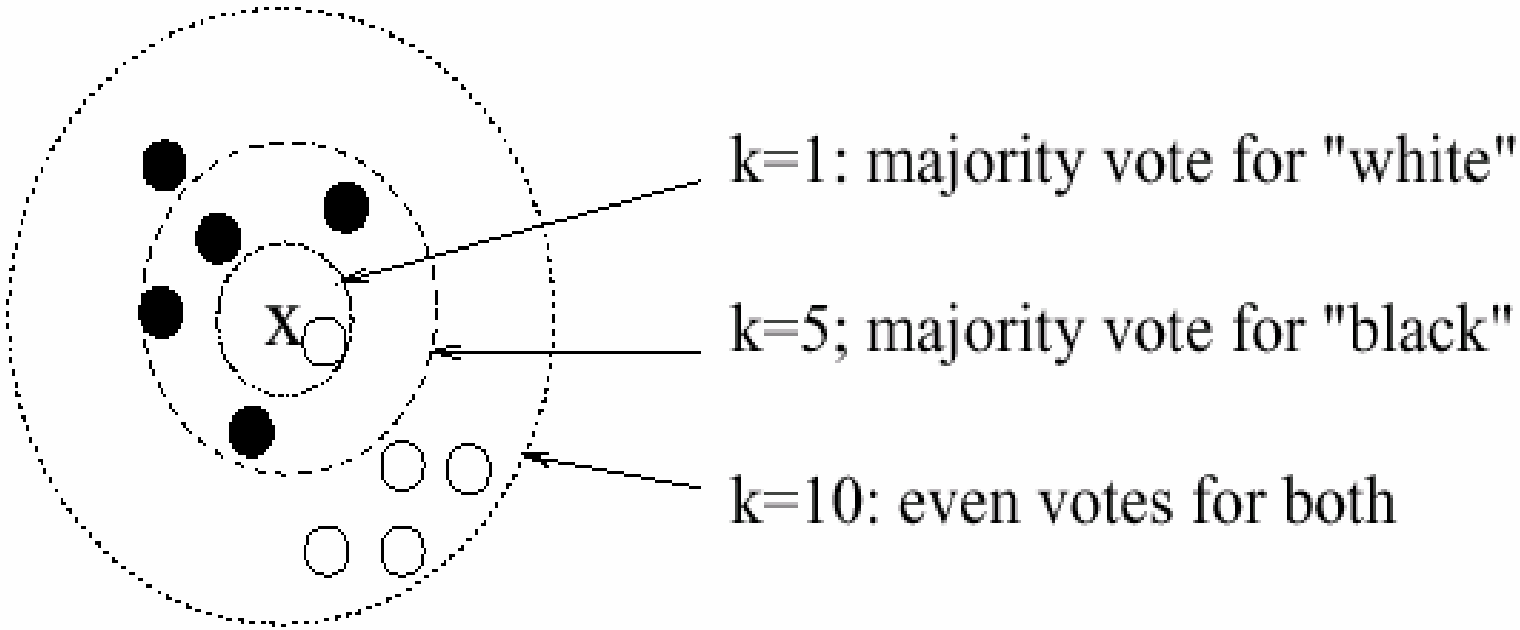
- Represent documents as points (vectors).
- Define a similarity measure for pair wise documents.
- Tune parameter *k* for optimizing classification effectiveness.
- Choose a voting scheme (e.g., weighted sum) for scoring categories
- Threshold on the scores for classification decisions（不是简单排序取最高的，需要有个门槛）.

Nearest Neighbor方法的要点

- “Similar” item: We need a functional（可操作的）definition of “similarity” if we want to apply this automatically.
- 要考虑多少邻居？
- Does each neighbor get the same weight?
- 近邻的所有类别都算？还是只取那些出现次数多的？如何作出最后的决定？

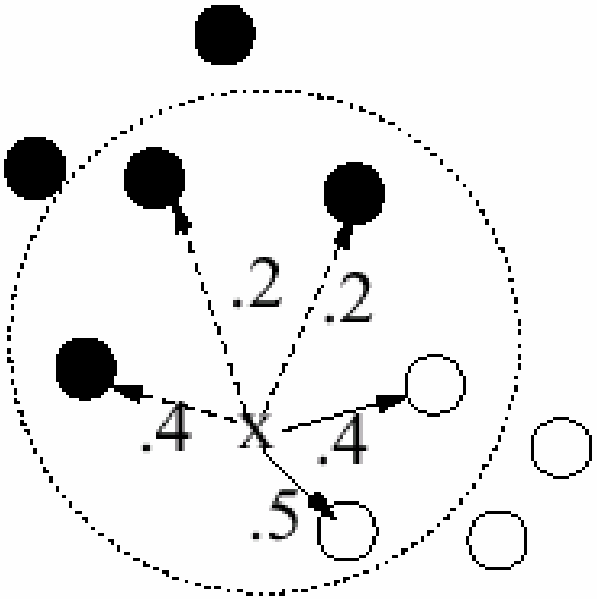
k的作用

K-Nearest Neighbor using a *majority* voting scheme



K-NN using a weighted-sum voting Scheme

k-NN using a weighted-sum voting scheme



kNN (k = 5)

Assign "white" to x because the weighted sum of "whites" is larger then the sum of "blacks".

Each neighbor is given a weight according to its nearness.

Category Scoring for Weighted-Sum

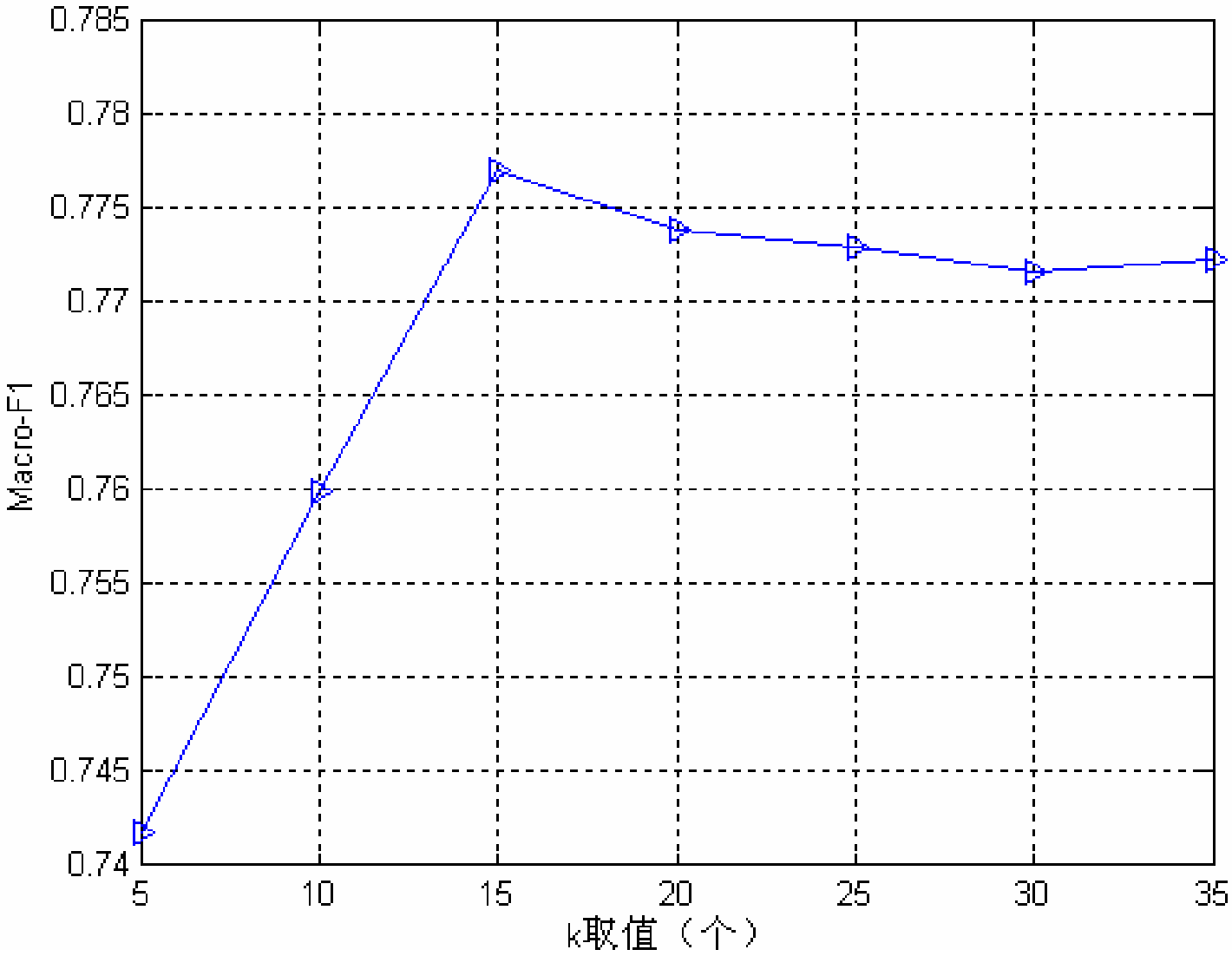
- The score for a category is the sum of the similarity scores between the point to be classified and all of its k-neighbors that belong to the given category.

- To restate: $$score(c \mid x) = b_c + \sum_{d \in kNN\ of\ x} sim(x,d)\, I(d,c)$$

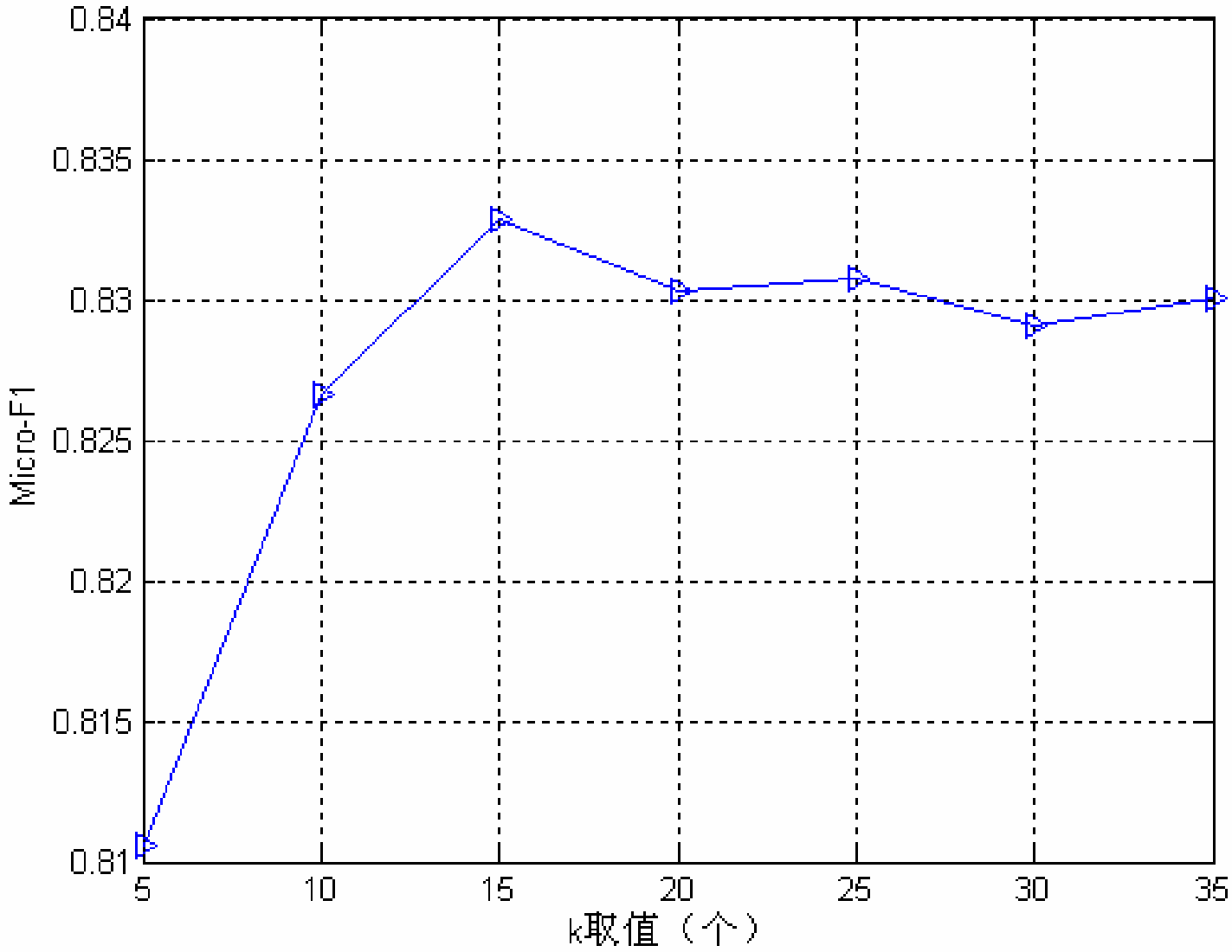
  where
  - x is the new point; c is a class
  - d is a classified point among the k-nearest neighbors of x;
  - $sim(x, d)$ is the similarity between x and d;
  - $I(d, c)$ = 1 iff point d belongs to class c;

    $I(d, c)$ = 0 otherwise.

kNN算法中k的取值


0.785
0.78
0.775
0.77
0.765
0.76
0.755
0.75
0.745
0.74
Macro-F1
5
10
15
20
25
30
35
k取值（个）

kNN算法中k的取值



kNN的优点

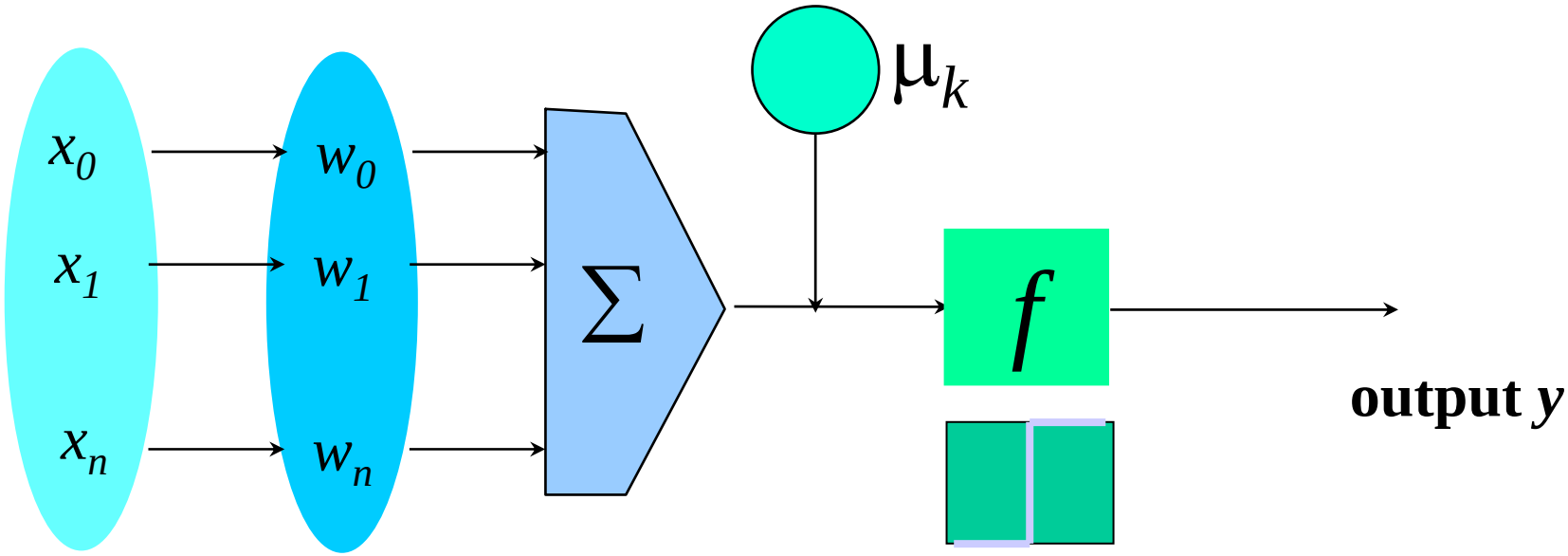
- 简单、有效（among top-5 in benchmark evaluations）
- 重新训练的代价较低（包括类别体系的变化和训练集的变化，在Web环境和电子商务应用中是很常见的）
- 计算时间和空间线性于训练集的规模（在一些场合不算太大）

kNN的不足

- 不很适合在线分类，响应速度较慢
  - kNN是懒散学习方法（*lazy learning*，基本上不学习），一些积极学习（*eager learning*）的算法要快很多
  - 其实是一个时间分配问题：学习时间和工作时间的权衡，“磨刀不误砍柴功”？
- 类别评分不是规格化的（不像概率评分）
  - 和其他评分方法的比较，以及和其他评分方法的结合是一个 open problem.
- 输出的可解释性不强
  - 例如决策树方法的可解释性较强

神经网络分类

- The *n*-dimensional input vector *x* is mapped into variable *y* by means of the scalar product and a nonlinear function mapping



贝叶斯分类法

Bayesian公式

$$P(c_j \mid d_i) = \frac{P(d_i \mid c_j)P(c_j)}{P(d_i)} \propto P(d_i \mid c_j)P(c_j)$$

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

参数计算

$$P(c_j) = \frac{c_j\text{的文档个数}}{\text{总文档个数}} = \frac{N(c_j)}{\sum_k N(c_k)} \approx \frac{1 + N(c_j)}{|c| + \sum_{k=1} N(c_k)}$$

$$P(w_i \mid c_j) = \frac{w_i\text{在}c_j\text{类别文档中出现的次数}}{\text{在}c_j\text{类所有文档中出现的词 的次数}} \approx \frac{1 + N_{ij}}{\text{不同词个数} + \sum_k N_{kj}}$$

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
  - AdaBoost

基于SVM的文本分类

SVM 基本原理

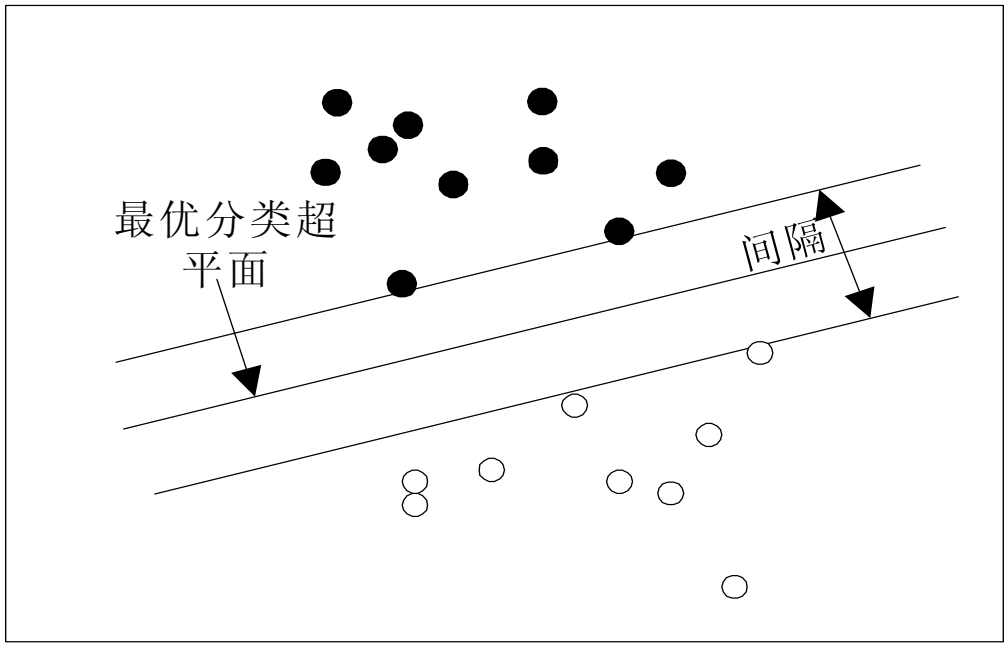
对于一组训练样本

$$(x_1, y_1),\ldots(x_l, y_l), \quad x \in R^n, \quad y \in \{+1,-1\}$$

在线性可分的情况下会有一个超平面

$$(w \cdot x) + b = 0$$

将这两类样本完全分开；并且离超平面最近的向量与超平面之间的距离最大。



最优分类超平面

为描述分类超平面，可采用下面的形式：

$$\begin{cases}(w\cdot x_i)+b\geq+1, when: y_i=+1\\(w\cdot x_i)+b\leq-1, when: y_i=-1\end{cases}$$

即：
$$y_i[(w\cdot x_i)+b]\geq 1,\quad i=1,...,l \qquad (1)$$

类别之间的分类间隔是：
$$2/\|w\|$$

为求解最优超平面，可以看成解二次型规划问题：

约束条件：
$$y_i[(x_i\bullet w)-b]\geq 1, i=1,2,...,l$$

最小化：
$$\Phi(w)=\frac{1}{2}(w\cdot w) \qquad (2)$$

鞍点理论

最小化： $\Phi(w)=\frac{1}{2}(w\cdot w)$

约束条件： $y_i[(x_i \bullet w)-b]\geq 1, i=1,2,...,l$

转化为如下的拉格朗日的极值问题（鞍点理论）：

$$L(w,b,\alpha)=\frac{1}{2}(w\cdot w)-\sum_{i=1}^{l}\alpha_i\{y_i[(x_i\cdot w)+b]-1\} \quad (3)$$

根据鞍点理论(导数求极值)：

$$\begin{cases}\dfrac{\partial L(w_0,b_0,\alpha^0)}{\partial b}=0\Rightarrow\sum_{i=1}^{l}\alpha_i^0=0, \quad i=1,\cdots,l & (4)\\ \dfrac{\partial L(w_0,b_0,\alpha^0)}{\partial w_0}=0\Rightarrow w_0=\sum_{i=1}^{l}y_i\alpha_i^0x_i, \quad i=1,\cdots,l & (5)\end{cases}$$

将式(5)代入式(3)，

对偶理论

根据对偶理论与K-T条件，(参见非线性规划理论)原问题转化为其对偶问题，求解如下泛函的最大化:

$$W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{i=1}^{l}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j) \qquad (6)$$

满足约束：

$$\sum_{i=1}^{l}y_i\alpha_i=0, \qquad (7)$$

$$\alpha_i\geq 0, \quad i=1,...,l \qquad (8)$$

分类函数

该类约束优化问题的解必须满足

$$\alpha_i\{y_i(w\cdot x_i+b)-1\}=0, \qquad (9)$$

其中，$\alpha_i=0$ 的样本对分类不起什么作用，有用的是 $\alpha_i>0$ 样本， 这些样本称为支持向量。

最后的获得的分类函数为：

$$f(x)=\mathrm{sgn}\left(\sum_{\text{支持向量}} y_i\alpha_i^*(x_i\cdot x)+b^*\right) \qquad (10)$$

松弛因子

而对于线性不可分的情况，通过引入松弛变量 $\xi_i \geq 0$

相应的目标函数变为：

最小化泛函：
$$\Phi(w,\xi) = \frac{1}{2}(w \cdot w) + C\left(\sum_{i=1}^{l} \xi_i\right) \qquad (11)$$

满足约束
$$y_i[(w \cdot x_i) + b] \geq 1 - \xi_i, \quad i = 1,...,l \qquad (12)$$

优化过程与线性可分情况基本一致，(6)(7)不变，约束条件(8)变为：
$$0 \leq \alpha_i \leq C \qquad (13)$$

其中，C为预定义的常数（惩罚因子）

非线性SVM

Classification using SVM (*w*,*b*)

$$x_i \cdot w + b \overset{?}{>} 0$$

In non linear case we can see this as

$$K(x_i, w) + b \overset{?}{>} 0$$

Kernel – Can be thought of as doing dot product in some high dimensional space

内积回旋： $(Z_i \bullet Z) = K(x, x_i)$ (核函数)

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{i=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

训练算法

- 由于支持向量机在训练时需要进行样本数L*L的二次方的矩阵的求解，因此，训练速度慢，内存占用大。为此，需用一些改进的算法来加快训练的进程，常用的主要有：
  - 块算法： 删除算法矩阵中对应Lagrange乘数为零的行和列将不会影响最终的结果，逐步排除非支持向量。
  - 分解法：将大的矩阵计算分解成小
  - SMO方法：每次只进行两个样本的优化，直接代数解

SVM 优缺点

- 现有样本有限信息情况下寻最优
- 转化为二次型寻优问题，全局最优（避免局部极值）
- 在高维空间构造线性判别函数，实现原空间的非线性判别函数

- 结构风险最小原理并不能严格证明好的推广能力
- 学习机的VC维的分析尚没有通用方法

其他分类方法

- **Regression based on Least Squares Fit (1991)**
- **Nearest Neighbor Classification (1992) ***
- **Bayesian Probabilistic Models (1992) ***
- **Symbolic Rule Induction (1994)**
- **Decision Tree (1994) ***
- **Neural Networks (1995)**
- **Rocchio approach (traditional IR, 1996) ***
- **Support Vector Machines (1997)**
- **Boosting or Bagging (1997)***
- **Hierarchical Language Modeling (1998)**
- **First-Order-Logic Rule Induction (1999)**
- **Maximum Entropy (1999)**
- **Hidden Markov Models (1999)**
- **Error-Correcting Output Coding (1999)**
- …

小结

- 自动分类的概念
- 分类效果的评价
- 特征选择
  - 文档频率法（DF, document frequency）
  - 信息增益法（information gain）
  - 互信息法（mutual information）
  - The x^2 test（chi-square）
- 分类算法
  - KNN
  - SVM

Any Question?