本节提纲

- 情感分析简介 *
- 主要任务：情感分类
- 基于监督学习的情感分类
- 基于半监督学习的情感分类
  - ✓ **基于个人与非个人视图的情感分类**
  - ✓ **基于不平衡数据的半监督情感分类**
  - ✓ **基于集成学习的半监督情感分类**

情感分析简介

- 情感分析和意见挖掘
  - ✓ **对于产品评论和新闻等文本中表达的意见，情感，情绪，主客观性，评价对象等方面的研究**
  - ✓ **情感分析在工业界和学术界都已经有着广泛的应用和研究**

产品评论示例

iPhone 6

iPhone 6 之大，不只是简简单单地放大，而是方方面面都大有提升。它尺寸更大，却变加纤薄；性能更强，却效能非凡。堪称 iPhone 新一代至为出众的大作。

商品评价

96%
好评度

好评(96%)
中评(2%)
差评(2%)

买家印象：
系统流畅(5837) 外观漂亮(5780) 反应快(4744)
功能齐全(4381) 照相不错(4219) 分辨率高(4098)
通话质量好(3494) 音质好(3215) 屏幕大(2854)

苹果手机还是很棒的，发货也很及时 2015-08-05 09:36
反应快 分辨率高
回复（0） 赞（0）

还不错，苹果的嘛，价格高，就是系统还不错，习惯了ios系统，封闭的好处是相对还是安全一点，赞一个～～～ 2015-08-05 09:33
功能齐全
回复（0） 赞（0）

好好好好好好好好好好好好好 2015-08-05 09:32
通话质量好 待机时间长 支持国产机 分辨率高
回复（0） 赞（0）

意见的定义

- 意见(Opinion)
  - ✓ **意见是对于一个实体或者实体的属性的正面或者负面的评价和观点**
- 意见的情感倾向
  - ✓ **正面，负面，或中性（没有情感）**

意见是由哪些成分构成的？

- ✓ 很多天前，我购买了一个**iPhone**手机，这是一个非常好用的手机。屏幕非常的酷。但是就是有些贵了。”
- ✓ 从中我们能发现
  - 评价对象: 实体或者实体的属性
  - 情感: 正面或者负面
  - 意见持有者: 发表意见的人
  - 时间: 意见发表的时间

谷歌的意见挖掘系统

Google product search

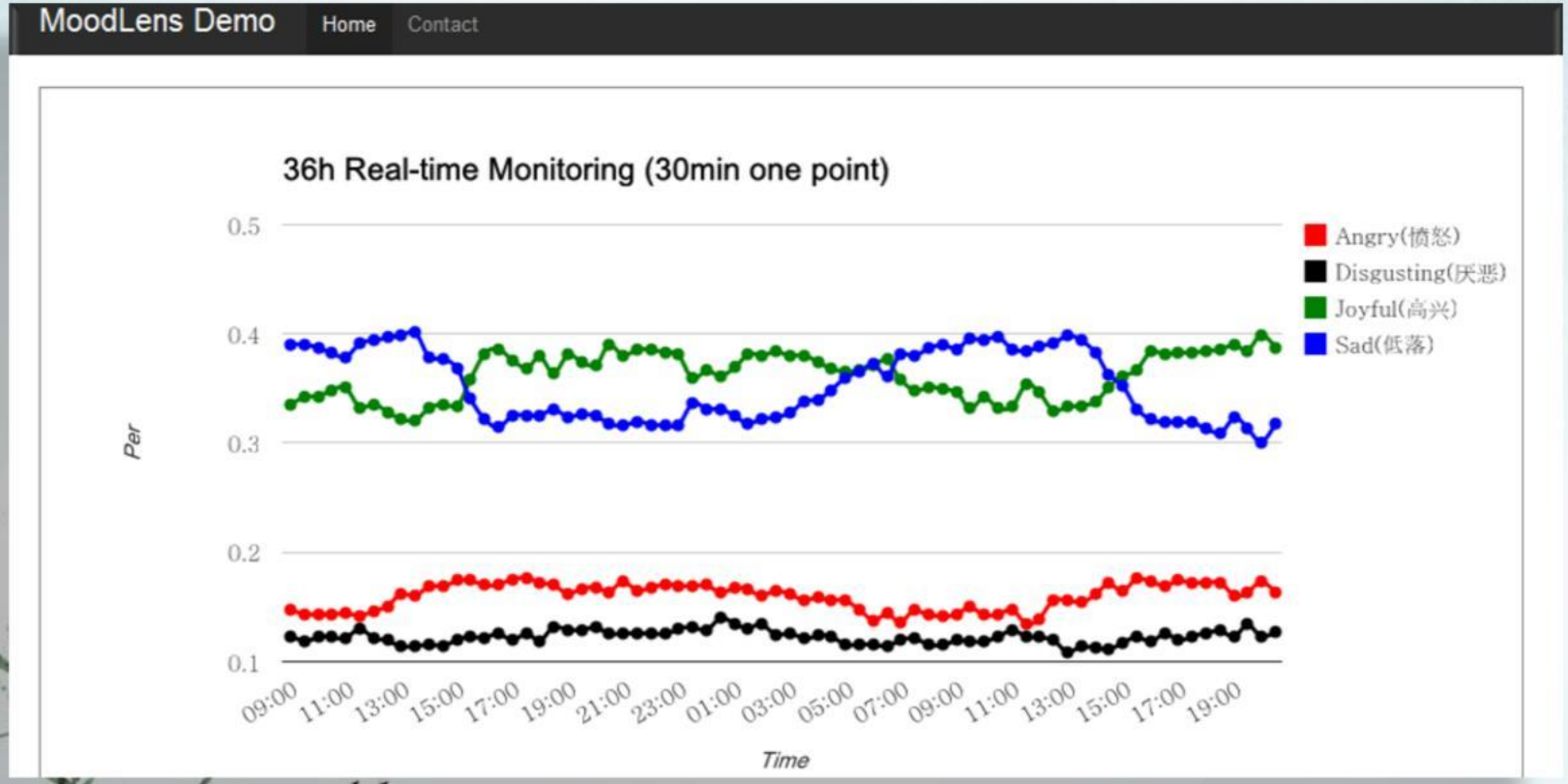
天猫的意见挖掘系统


与描述相符
4.4
1 2 3 4 5
非常不满 不满意 一般 满意 非常满意
大家认为
偏小 适合
尺码: 78%
适合:78%
有色差 无色差
色差: 91%
查看追加 (4)
有内容评价
按时间
按信用
Beta
按推荐
物流速度太慢啦。。。但是应该也跟双十一有关，不过衣服很喜欢，质量很好 就是感觉有一点点薄，不知道穿着会不会冷
今天
颜色：黑色
尺码：S(165)
刘***1（匿名）
还行
今天
颜色：黑色
尺码：XL(180)
h***a（匿名）
很好，比图片还要好看
今天
颜色：黑色
尺码：XL(180)
体重：75kg
身高：178cm
alpswei

天猫的意见挖掘系统


商品详情
包装和参数
累计评价 6971
月成交记录1396件
电器城服务详情
与描述相符
4.8
性价比很高 (197)
系统流畅(185)
屏幕不错(142)
包装不错哦(109)
电池耐用(88)
外观靓丽(80)
电池一般(167)
系统不流畅(46)
查看追加 (281)
有内容评价
按时间
按信用
Beta
按推荐
商品：谁说的光棍节哪天买便宜些，上当了，一订购，没过几天就降价了，伤心
服务：客服人员还可以了，就是价格也太让人伤心了，都说光棍节购东西便宜，结果那天要贵些。
11.20
机身颜色：皓月白
手机套餐：官方标配
筱***婷（匿名）
手机手感很好，功能强大，价格实惠，这次网购很开心，希望店子兴隆。
11.19
机身颜色：皓月白
手机套餐：官方标配
唐***8（匿名）
性价比很高 系统流畅
11.17
机身颜色：皓月白
手机套餐：官方标配
c***8（匿名）
凭什么双十一之后还降价了，不是说最便宜吗？别的网站怎么更便宜，

北航的舆情监控系统


MoodLens Demo
Home
Contact
36h Real-time Monitoring (30min one point)
Per
0.5
0.4
0.3
0.2
0.1
09:00
11:00
13:00
15:00
17:00
19:00
21:00
23:00
01:00
03:00
05:00
07:00
09:00
11:00
13:00
15:00
17:00
19:00
Time
Angry(愤怒)
Disgusting(厌恶)
Joyful(高兴)
Sad(低落)

情感分析的任务

- 情感分类 (Sentiment Classification)
- 评价对象抽取(Opinion Target Extraction)
- 主客观分析(Subjective Analysis)
- 情绪分析(Emotion Analysis)
- 垃圾文本过滤(Opinion spam detection)
- …

本节提纲

- 情感分析简介
- 主要任务：情感分类 *
- 基于监督学习的情感分类
- 基于半监督学习的情感分类
  - ✓ **基于个人与非个人视图的情感分类**
  - ✓ **基于不平衡数据的半监督情感分类**
  - ✓ **基于集成学习的半监督情感分类**

情感分类

- 基于整体的情感倾向，将整个文档（比如产品评论），区分为正面或负面情感的。
  - ✓ 获得最多研究的情感分析问题

一个文本分类任务

- 这是一个基本的文本分类任务
- 但是和基于主题的文本分类有很大的不同
  - ✓ 在基于主题（科技，人文，运动）的文本分类中，主题词是很重要的。
  - ✓ 但是在情感分类中，情感词是很重要的。比如‘高兴’，‘出色’，‘恐怖’，‘差劲’等。

一些产品评论示例

IT168网友 2013年04月03日 14:45 投稿数：5596

外观	★★★★☆ 4
操控	★★★★★ 5
成像	★★★★★ 5
手感	★★★★★ 5
色彩	★★★★☆ 4
性能	★★★★★ 5
满意度	☆☆☆☆☆ 5
您的水平	·纯粹小白
主要用途	·风景 ·夜景

很好。

好是好，高感抑噪连拍速度都可以，对焦精度和速度也仅次于7D，画质优秀。可就是买的时候一定要注意，我春节过后在沈阳三好街赛博数码广场裕宁也买了一台，当时没太看仔细，回家后发现连屈光度都被调整了。电池也根本就不是原厂的，而且拍不到100张就剩40%，再次充电只能充90%，换了四次才换回来一个南韩的电池。而且充电器的连线都是废弃的。幸好我有个7D机子，否则充电都充不了。我还是老客户，买了多架相机的我，这次是最不满意的一次，有被耍的感觉，大家一定要注意，千万别上当。

对您有帮助吗？ 有用 8 没用 2

kissphoto 2012年08月06日 22:05 投稿数：13

外观	★★★★★ 5
操控	★★★★☆ 4
成像	★★★★★ 5
手感	★★★★★ 5
色彩	★★★★★ 5
性能	★★★★★ 5
满意度	☆☆☆☆☆ 5
您的水平	·入门用户
主要用途	·风景 ·人像 ·新闻摄影 ·运动

各个方面性能比上一代的无敌兔有所提升

【外观】采用镁合金机身，具有防水窗防尘性能，采用显示效果出色的3.2英寸104万像素的液晶屏，依然没有内置闪光灯

【操控】佳能的主菜单非常好用，相比尼康单反主菜单的密集，佳能的主菜单更为简洁，使用起来更方便些。新增多重曝光和HDR功能，多重曝光最多可曝光9次

【成像】是目前最高像素的佳能单反相机——2230万有效像素，对于摄影师和发烧友来说，5D Mark III能在满足需要的前提下提供更加出色的控噪水平。

【手感】与上代无敌兔一样，握感不错

【色彩】颜色能够真是的还原，色彩锐利

【性能】61点AF系统，5D Mark III有着目前理论上最顶级的对焦速度和精度，可用感光度范围达到ISO 50-3200

【综合点评】继无敌兔后佳能全幅系列中又一力作，各方面性能突出，价格较之无敌兔偏高

对您有帮助吗？ 有用 51 没用 24

情感分类的研究领域

- 按机器学习方法分类
  - ✓ **基于监督学习的情感分类**
  - ✓ **基于半监督学习的情感分类**
  - ✓ **基于无监督学习的情感分类**
- 按研究问题分类
  - ✓ **基于不平衡数据的情感分类**
  - ✓ **跨领域情感分类**
  - ✓ **跨语言情感分类**

本节提纲

- 情感分析简介
- 主要任务：情感分类
- 基于监督学习的情感分类 *
- 基于半监督学习的情感分类
  - ✓ **基于个人与非个人视图的情感分类**
  - ✓ **基于不平衡数据的半监督情感分类**
  - ✓ **基于集成学习的半监督情感分类**

基于监督学习的情感分类

- 监督学习
  - ✓ 训练和测试数据
    - 基于打星的电影评论
      - ✓ ☆☆☆☆-☆☆☆☆☆为**正面**
      - ✓ ☆-☆☆为**负面**
  - ✓ 支持向量机分类模型（SVM）能获得了最好的分类性能
    - 准确率：83%
    - 特征：每个单词作为特征

(Pang et al, 2002)

基于监督学习的情感分类

- 向量空间模型（Vector Space Model，VSM）
  - ✓ 自然语言处理中常用的模型，可以用来衡量两个向量之间的相关程度
  - ✓ 涉及到的一些基本概念：
    - 文档（Document）
    - 项/特征项（Term/Feature）
    - 项的权重（Term Weight）

基于监督学习的情感分类

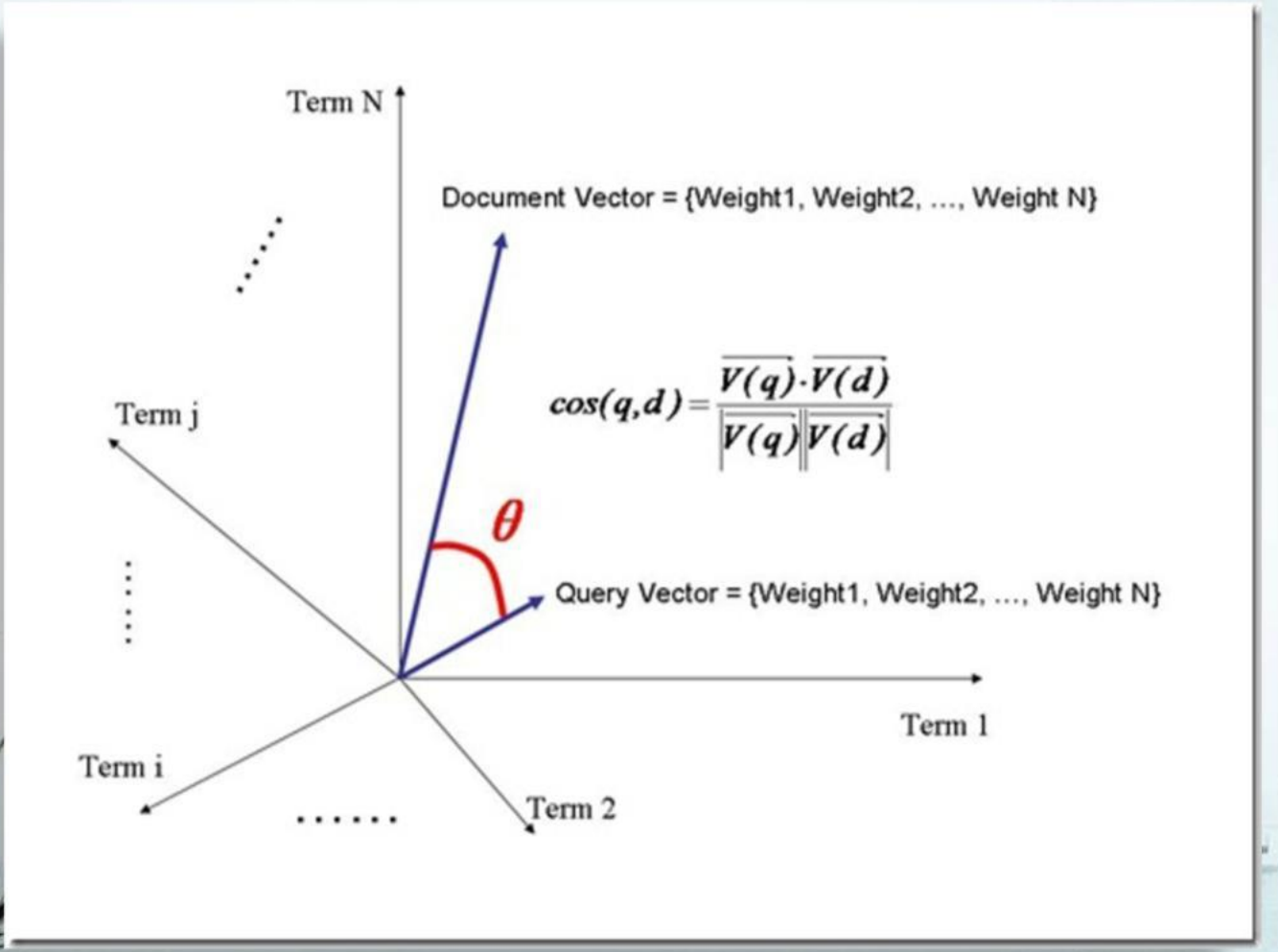
- 向量的相似度度量（similarity）
  - ✓ **文档** $D_1 = D_1(w_{11}, w_{12}, \ldots\ldots w_{1n})$

    $D_2 = D_2(w_{21}, w_{22}, \ldots\ldots w_{2n})$
  - ✓ **两个文档** D_1 **和** D_2 **内容的相似程度** $Sim(D_1, D_2)$ **如下：**

$$Sim(D_1, D_2) = \cos\theta = \frac{\sum_{k=1}^{n} w_{1k} \times w_{2k}}{\sqrt{(\sum_{k=1}^{n} w_{1k}^{\ 2})(\sum_{k=1}^{n} w_{2k}^{\ 2})}}$$

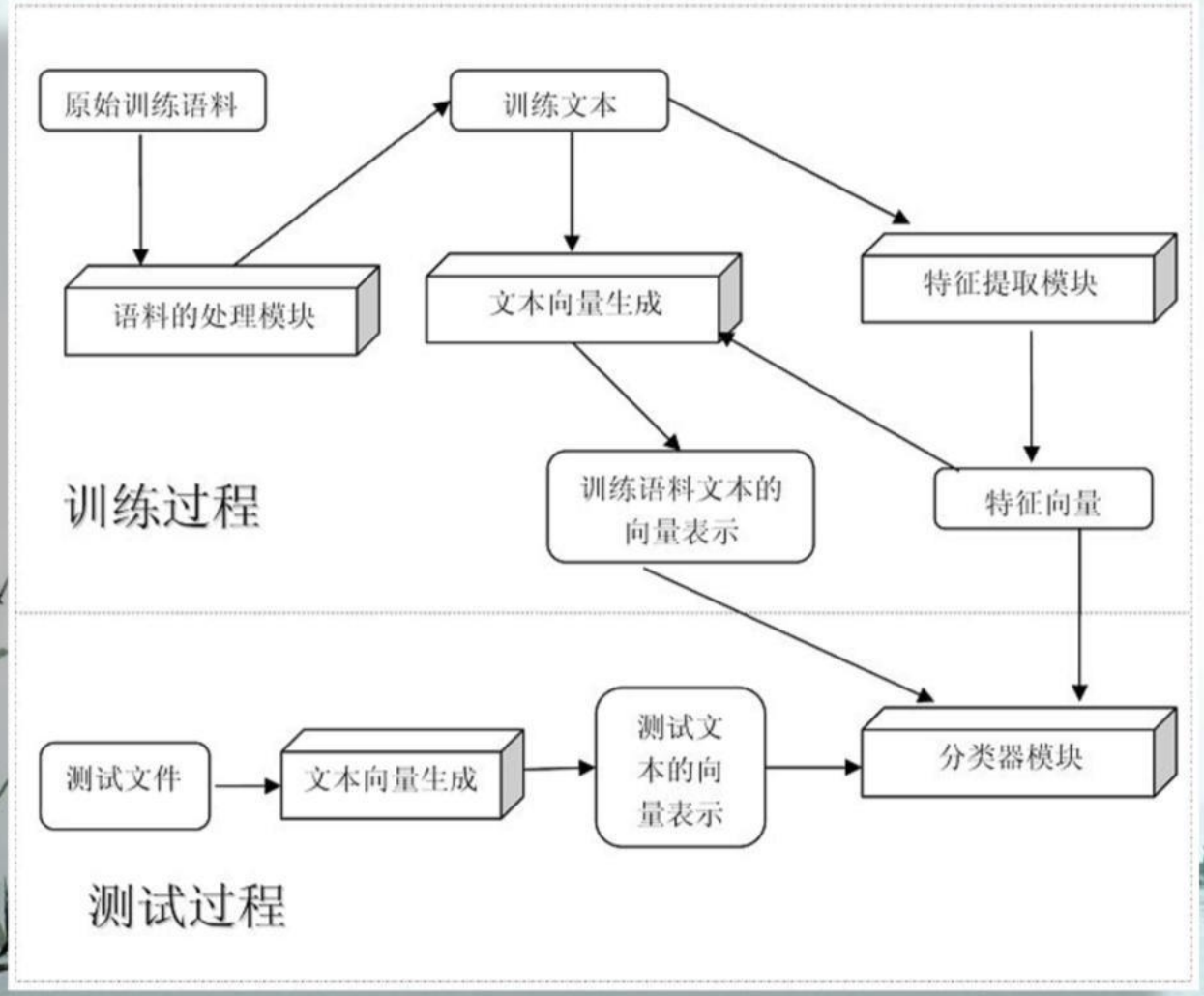
基于监督学习的情感分类

- 文档的向量空间模型示意图



基于监督学习的情感分类

- 基于的向量空间模型（VSM）的分类系统框架



基于监督学习的情感分类

- 监督学习的特征选择
  - ✓ **基于特征工程的方法，很多种类的特征被证明是有效的，例如：**
    - **单词**
    - **词性标记**
    - **情感词和短语**
    - **否定词**
    - **句法和依存关系**

基于监督学习的情感分类

- 监督学习的分类算法
  - ✓ k-最近邻法（k-Nearest Neighbor，kNN）
  - ✓ 朴素贝叶斯法（Naïve Bayesian，NB）
  - ✓ 支持向量机法（Support Vector Machines，SVM）
  - ✓ …

基于监督学习的情感分类

- k-最近邻法（k-Nearest Neighbor，kNN）
  - 基本思想是：给定一个测试文档，系统在训练集中查找离他最近的k个邻居，并根据这些邻居的分类来给该文档的候选分类评分。决策规则如下：

$$y(\vec{x}, c_i) = \sum_{\vec{d}_j \in kNN} sim(\vec{x}, \vec{d}_j) y(\vec{d}_j, c_i) - b_i$$

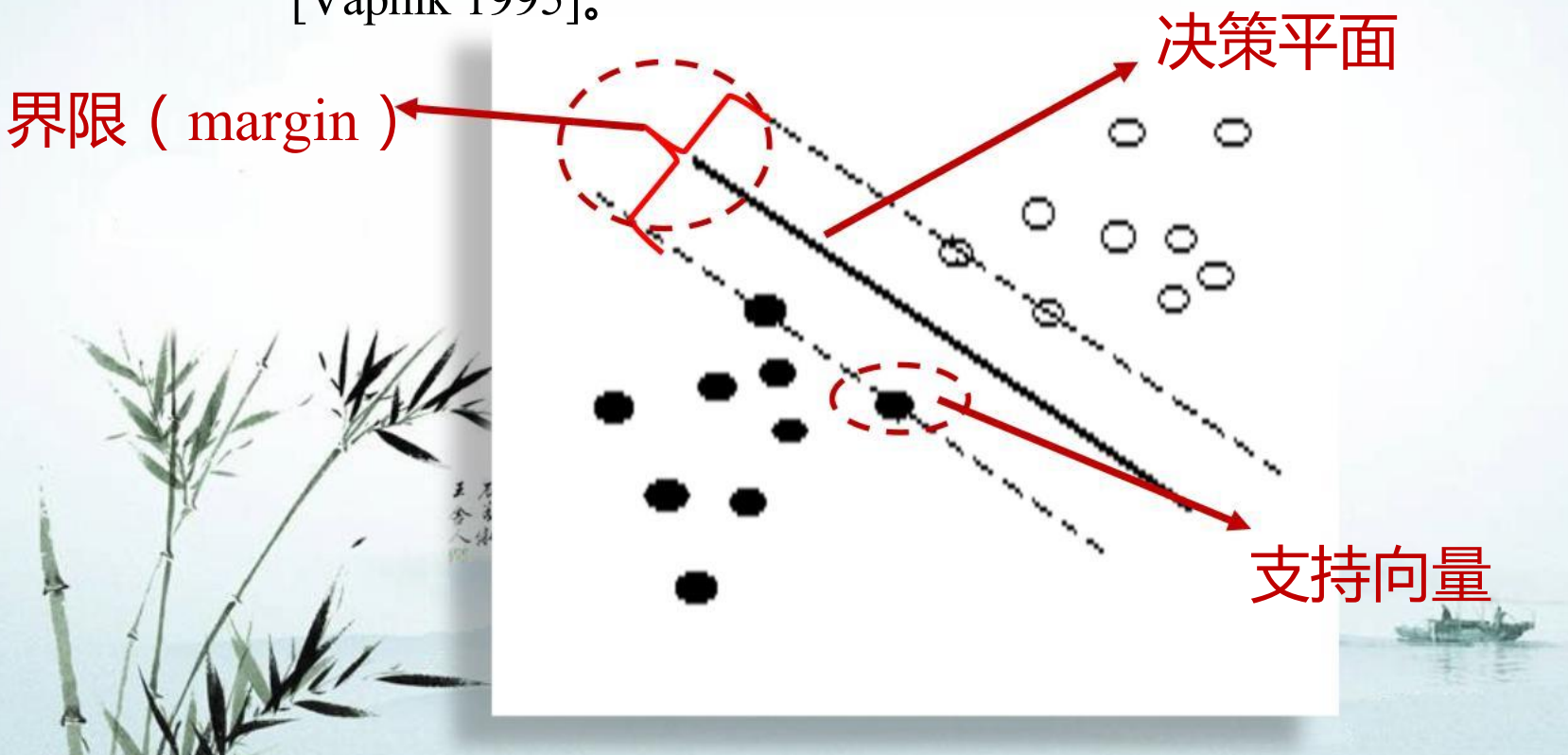
基于监督学习的情感分类

- 朴素贝叶斯法（Naïve Bayesian，NB）
- ✓ 基本思想：利用特征项和分类的联合概率来估计给定文档的分类概率
- ✓ 基本假设：文本是基于词的unigram模型，即文本中词的出现依赖于文本类别，但不依赖于其他词及文本的长度，也就是说，词与词之间是独立的

$$P(c_i \mid Doc) = \frac{P(c_i)\prod_{t_j \in V} P(Doc(t_j) \mid c_i)}{\sum_i P(c_i)\prod_{t_j \in V} P(Doc(t_j) \mid c_i)}$$

基于监督学习的情感分类

- **支持向量机法（Support Vector Machines，SVM）**
  - ✓ 基本思想：是在向量空间中找到一个决策平面（Decision surface），这个平面能 最好"地分割两个分类中的数据点[Vapnik 1995]。

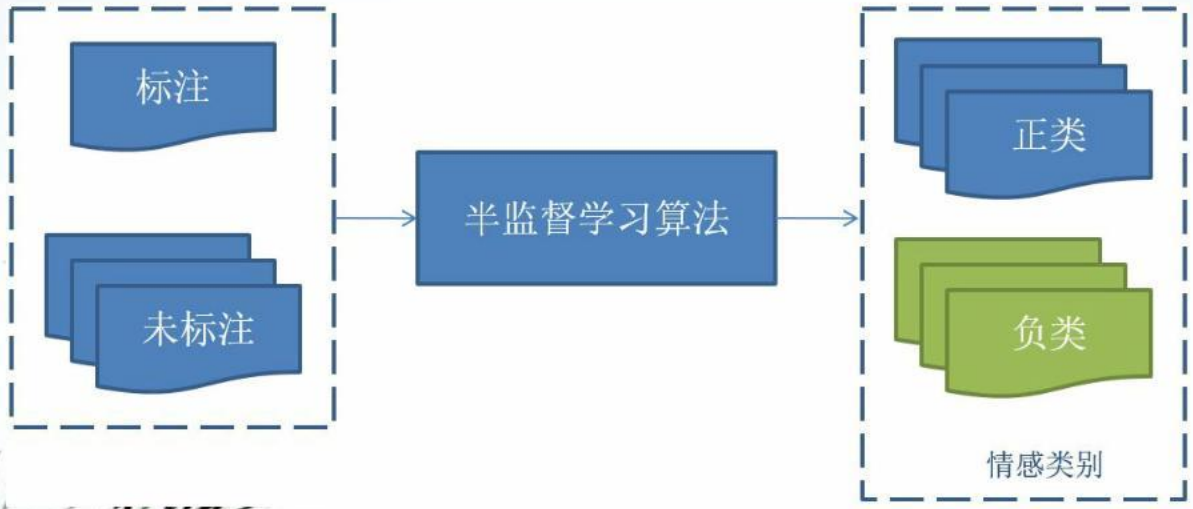


本节提纲

- 情感分析简介
- 主要任务：情感分类
- 基于监督学习的情感分类
- 基于半监督学习的情感分类 *
  - ✓ **基于个人与非个人视图的情感分类**
  - ✓ **基于不平衡数据的半监督情感分类**
  - ✓ **基于集成学习的半监督情感分类**

基于半监督学习的情感分类

- 输入：少量标注样本，大量未标注样本
- 输出：新的未标注样本的情感极性



基于半监督学习的情感分类

- 基于个人与非个人视图的情感分类
- 基于不平衡数据的半监督情感分类
- 基于集成学习的半监督情感分类

本节提纲

- 情感分析简介
- 主要任务：情感分类
- 基于监督学习的情感分类
- 基于半监督学习的情感分类
  - ✓ **基于个人与非个人视图的情感分类** *
  - ✓ **基于不平衡数据的半监督情感分类**
  - ✓ **基于集成学习的半监督情感分类**

基于个人与非个人视图的情感分类

- 我们提出两个视图，个人(personal)的和非个人(impersonal)的视图，并将它们系统地集成到半监督情感分类中
- 基于协同训练算法(co-training algorithms)学习这两个视图的信息

基于个人与非个人视图的情感分类

- 发掘个人和非个人视图
  - ✓ 个人句(Personal sentence): 整句话的主语是人
    - ✓ 我喜欢这个面包机
    - ✓ 我感到非常失望
  - ✓ 非个人句(Impersonal sentence): 整句话的主语是实体或其他成分
    - ✓ 它的屏幕非常的漂亮
    - ✓ 这个面包一点也不好吃

基于个人与非个人视图的情感分类

- 区分个人和非个人视图的算法

Input:

The training data D

Output:

All personal and impersonal sentences, i.e. sentence sets $S_{personal}$ and $S_{impersonal}$.

Procedure:

(1). Segment all documents in D to sentences S using punctuations (such as periods and interrogation marks)

(2). Apply the heuristic rules to classify the sentences S with proper pronouns into, S_{p1} *and* S_{i1}

(3). Train a binary classifier f_{p-i} *with* S_{p1} *and* S_{i1}

(4). Use f_{p-i} *to classify the remaining sentences into* S_{p2} *and* S_{i2}

(5). $S_{personal} = S_{p1} \cup S_{p2}$, $S_{impersonal} = S_{i1} \cup S_{i2}$

基于个人与非个人视图的情感分类

- 基于个人和非个人视图的半监督学习算法

Input:

The labeled data L containing personal sentence set $S_{L-personal}$ and impersonal sentence set $S_{L-impersonal}$

The unlabeled data U containing personal sentence set $S_{U-personal}$ and impersonal sentence set $S_{U-impersonal}$

Output:

New labeled data L

Procedure:

Loop for N iterations until $U = \phi$

(1). Learn the first classifier f_1 *with* $S_{L-personal}$

(2). Use f_1 *to label samples from* U *with* $S_{U-personal}$

(3). Choose n_1 *positive and* n_1 *negative most confidently predicted samples* A_1

(4). Learn the second classifier f_2 *with* $S_{L-impersonal}$

(5). Use f_2 *to label samples from* U *with* $S_{U-impersonal}$

(6). Choose n_2 *positive and* n_2 *negative most confidently predicted samples* A_2

(7). Learn the third classifier f_3 *with* L

(8). Use f_3 *to label samples from* U

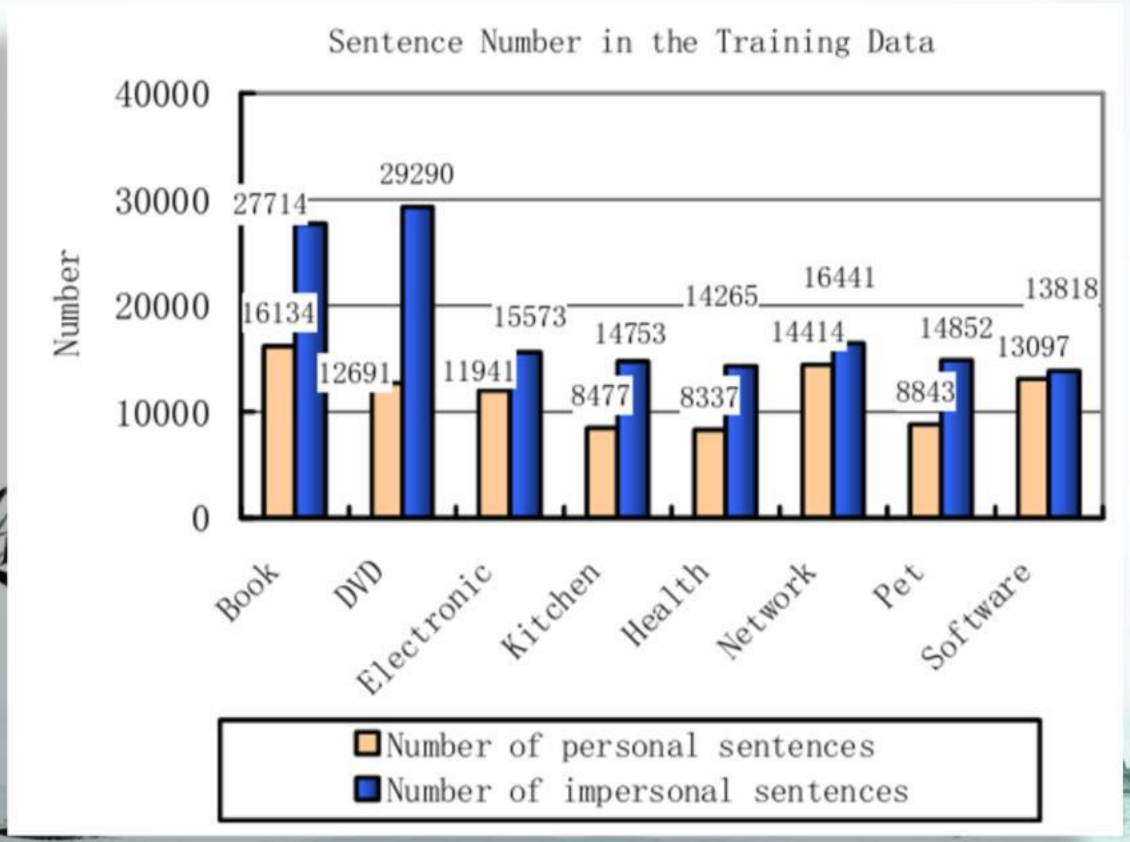
(9). Choose n_3 *positive and* n_3 *negative most confidently predicted samples* A_3

(10). Add samples $A_1 \cup A_2 \cup A_3$ *with the corresponding labels into* L

(11). Update $S_{L-personal}$ *and* $S_{L-impersonal}$

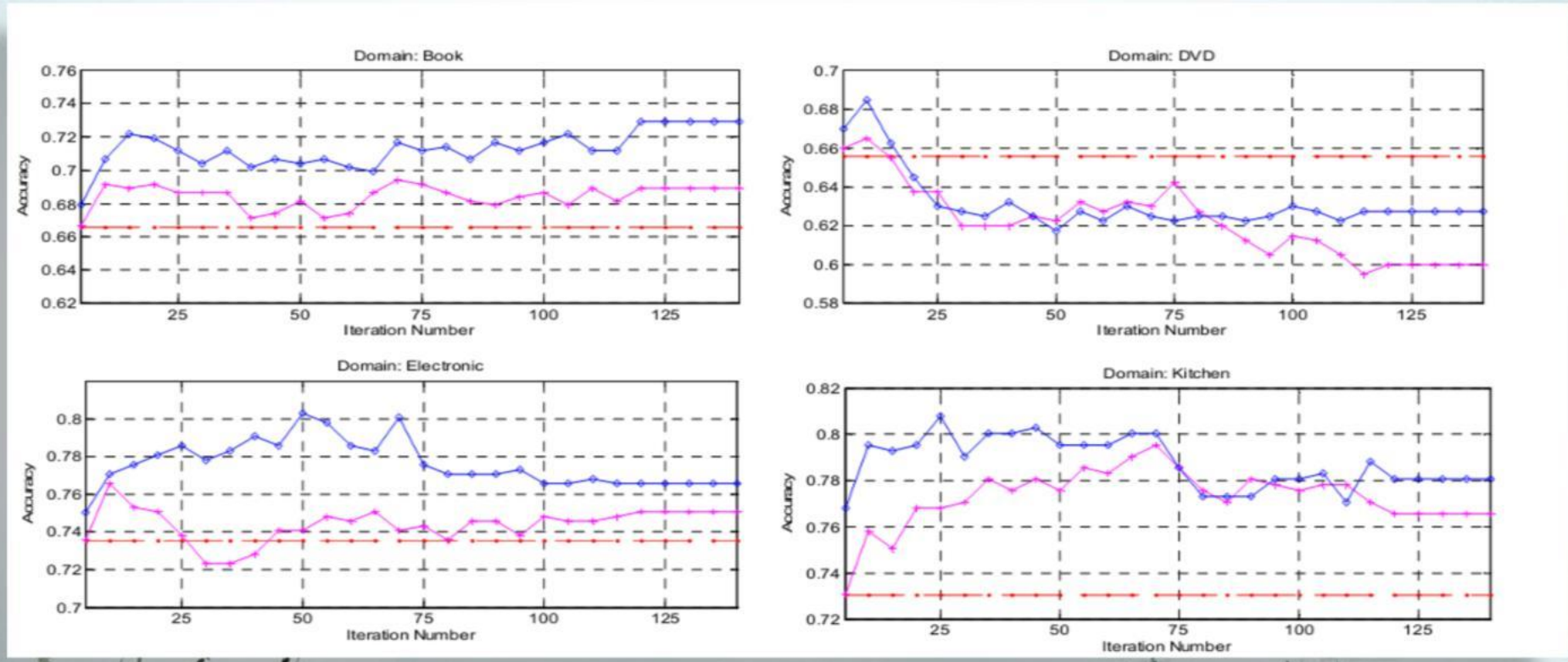
基于个人与非个人视图的情感分类

- 实验结果（一）
  - ✓ 在不同领域，个人和非个人视图句子的分布情况



基于个人与非个人视图的情感分类

- 实验结果（二）
  - ✓ 在不同领域上的实验结果



本节提纲

- 情感分析简介
- 主要任务：情感分类
- 基于监督学习的情感分类
- 基于半监督学习的情感分类
  - ✓ **基于个人与非个人视图的情感分类**
  - ✓ **基于不平衡数据的半监督情感分类** *
  - ✓ **基于集成学习的半监督情感分类**

基于不平衡数据的半监督情感分类

- 已有的半监督学习方法认为正负类样本是平衡的
  - ✓ 实际上，情感分类中，正类样本是远多于负类的

Domain	N_+	N_-	N_+/N_-
Book	425159	58315	7.29
DVD	69175	11383	6.08
Electronic	15397	4316	3.57
Kitchen	14290	3784	3.78

基于不平衡数据的半监督情感分类

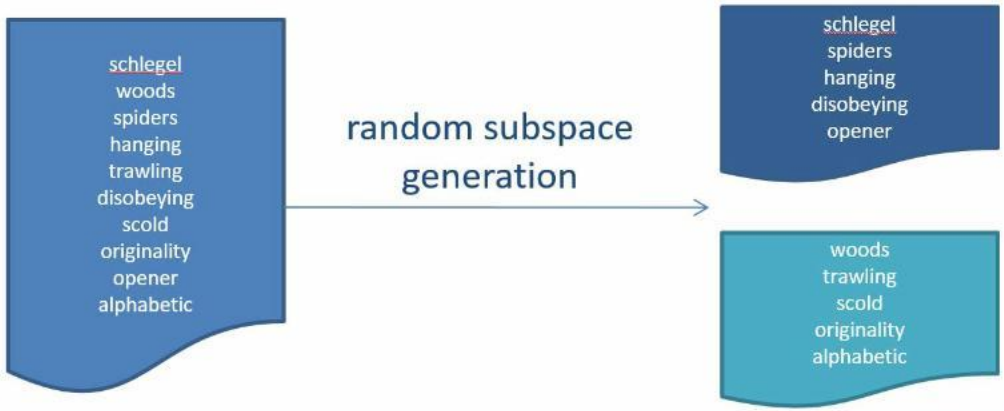
- 不平衡的挑战
  - ✓ 处理不平衡的标注数据
    - 如何完全使用所有的标注样本
  - ✓ 处理不平衡的未标注数据
    - 如何从不平衡的未标注数据中获取信息

基于不平衡数据的半监督情感分类

- 我们的解决方法
  - ✓ 对于第一个难点
    - 多次欠采样（Multiple Under-sampling）
      - ✓ 在多类样本中进行多次欠采样
  - ✓ 对于第二个难点
    - 基于随机子空间生成（ random subspace generation ）的协同训练
    - 动态子空间生成能够提高性能
      - ✓ 不同的子分类器有很大的差异性

基于不平衡数据的半监督情感分类

- 随机子空间生成
  - ✓ 一个文档被表示为一个特征向量（词袋模型）
  - ✓ 随机选择一半的特征生成一个子空间，剩下的一半作为另外一个子空间

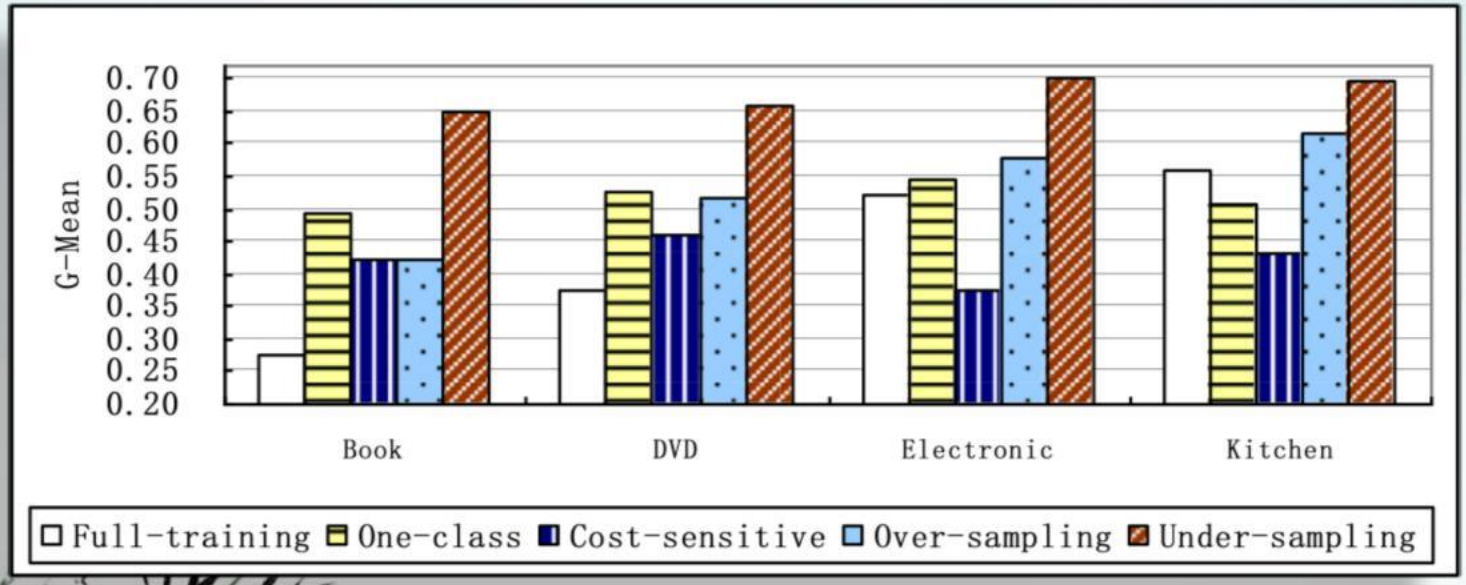


基于不平衡数据的半监督情感分类

- 基于多重采样的协同训练
  - ✓ **迭代N次** // 对于协同训练
    - – **For i =1 to K:** // 构建k个欠采样数据集
      - **生成两个随机子空间**
      - **基于第i次欠采样数据集，训练两个子空间分类器**
      - **使用每个子分类器选择最可信的样本**
      - **更新标注样本集**

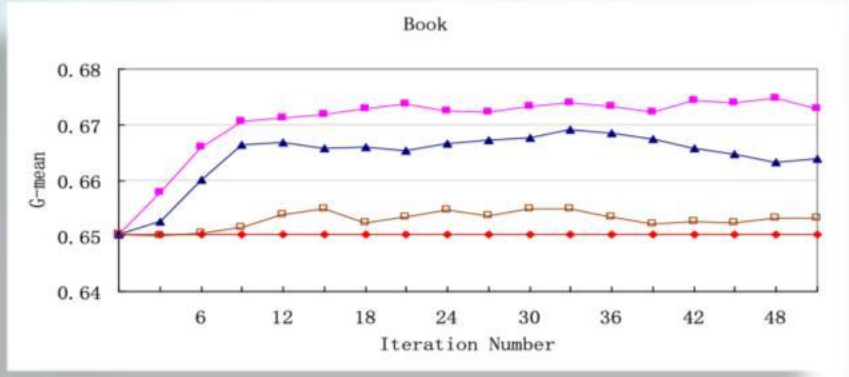
基于不平衡数据的半监督情感分类

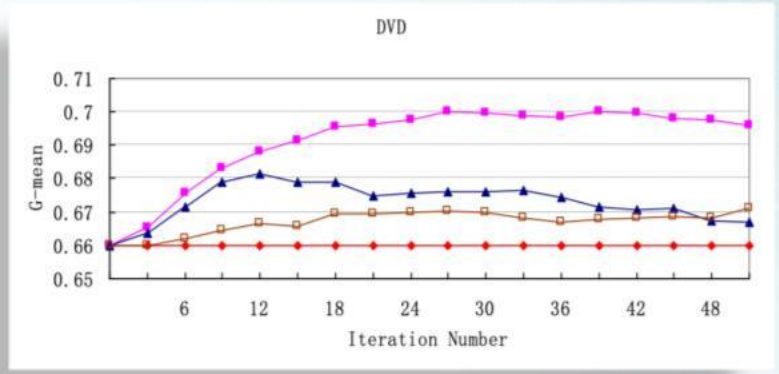
- 不同监督分类方法的比较

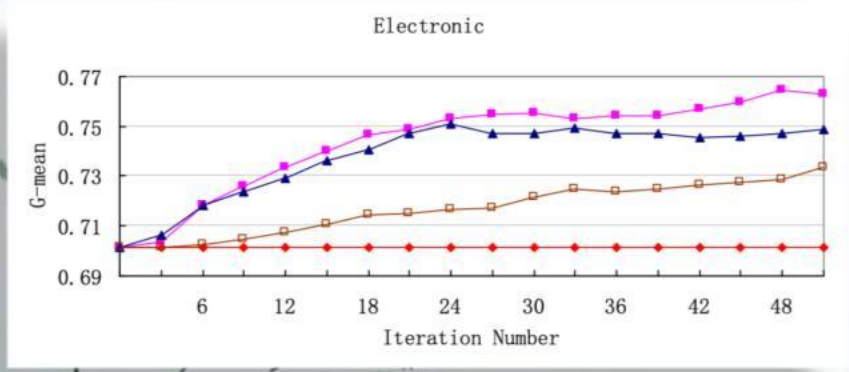


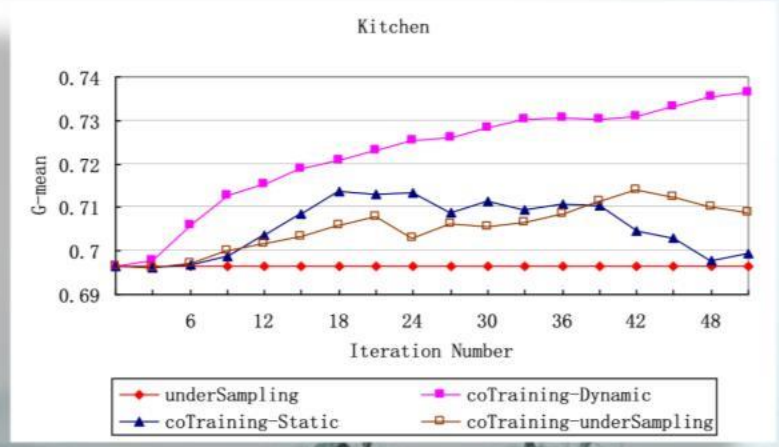
基于不平衡数据的半监督情感分类

- 不同监督分类方法的比较









本节提纲

- 情感分析简介
- 主要任务：情感分类
- 基于监督学习的情感分类
- 基于半监督学习的情感分类
  - ✓ **基于个人与非个人视图的情感分类**
  - ✓ **基于不平衡数据的半监督情感分类**
  - ✓ **基于集成学习的半监督情感分类** *

基于集成学习的半监督情感分类

- 研究动机
  - 哪种半监督学习算法表现最好？
    - ✓ 每个半监督学习算法都有其独特的特性，在特定的领域中都能获得较其他算法更好的性能。
  - 例如：
    - ✓ 协同训练算法（Co-training）：Book与Kitchen域中能获得更好的性能
    - ✓ 标签传播算法（Label Propagation）：DVD与Electronic中表现的更好
    - ✓ **结论：很难分辨出哪种算法最优！**

基于集成学习的半监督情感分类

- 解决方案
  - 提出一种新的基于元分类器的集成学习方法
    - ✓ 通过集成多个半监督学习方法进行半监督学习
  - 核心模块：
    - ✓ 元学习（Meta-learning）
      - ✓ 重新预测未标注样本的类别标签
        - ✓ 利用多个半监督学习方法的输出结果作为样本训练元分类器

基于集成学习的半监督情感分类

- 系统框架图

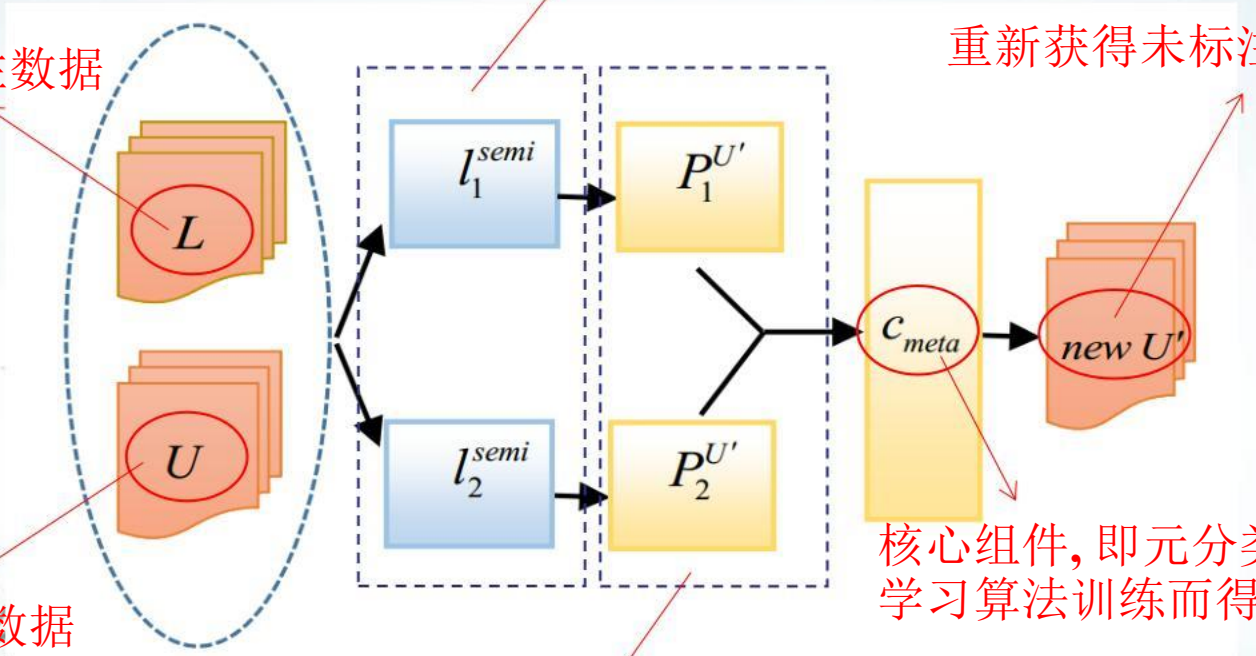


Figure 1: The framework of *semi-stacking*

基于集成学习的半监督情感分类

➢ 元学习（Meta-Learning）

- 元学习算法的训练样本不是利用传统的特征进行表示，例如：词袋特征
- 利用两个半监督算法的输出概率进行组合作为样本特征表示

元学习算法训练样本：

$$x^{meta} = < p_1(pos \mid x_k), p_1(neg \mid x_k), p_2(pos \mid x_k), p_2(neg \mid x_k) >$$

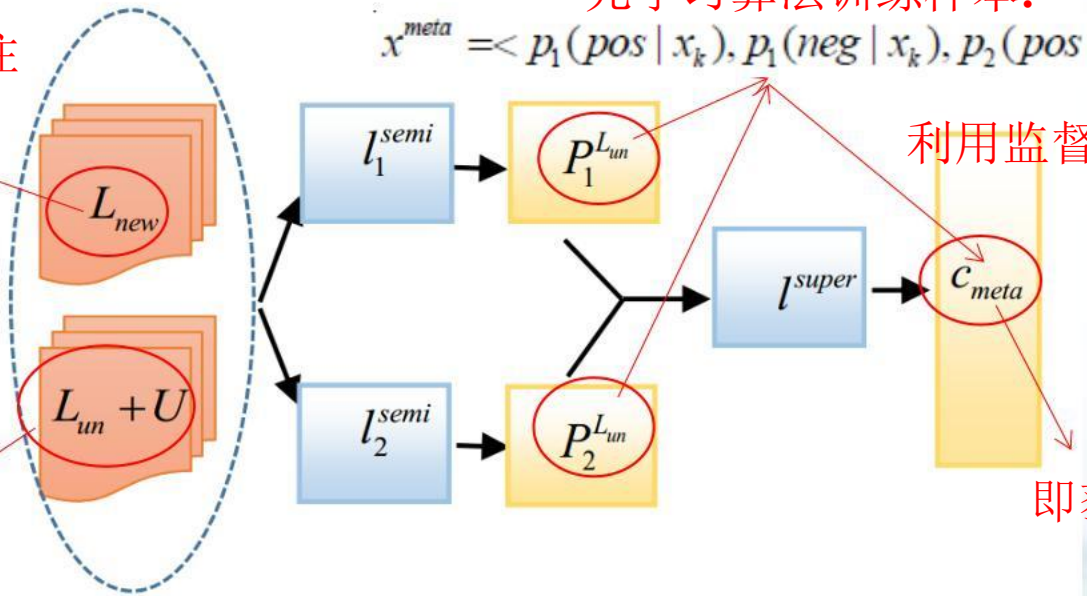


Figure 2: The framework of *meta*-learning

基于集成学习的半监督情感分类

- 实验结果

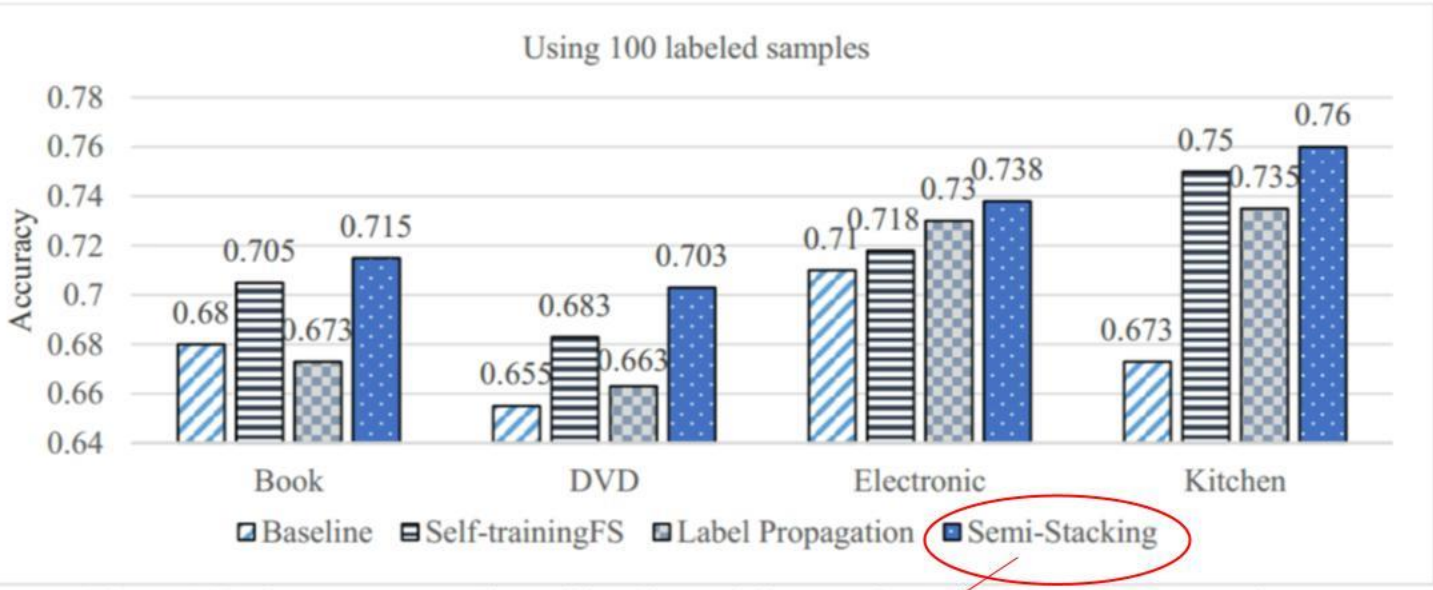


Figure 4: Performance comparison of baseline and three semi-supervised learning approaches

我们的方法

参考

- ✓ Bing Liu's homepage: http://www.cs.uic.edu/~liub/
- ✓ John Blitzer's homepage: http://john.blitzer.com/
- ✓ Movie Review Data: http://www.cs.cornell.edu/people/pabo/movie-review-data/
- ✓ MPQA: http://mpqa.cs.pitt.edu/

谢谢！

Q&A