ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2011/05(04)-0313-11
DOI: 10.3778/j.issn.1673-9418.2011.04.003

E-mail: fcst@vip.163.com
http://www.ceaj.org
Tel: +86-10-51616056

半监督文本分类综述*

牛 罡, 罗爱宝, 商 琳+
南京大学 计算机软件新技术国家重点实验室, 南京 210093

A Survey of Semi-supervised Text Categorization*

NIU Gang, LUO Aibao, SHANG Lin+
State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China
+ Corresponding author: E-mail: shanglin@nju.edu.cn

NIU Gang, LUO Aibao, SHANG Lin. A survey of semi-supervised text categorization. Journal of Frontiers of Computer Science and Technology, 2011, 5(4): 313-323.

Abstract: Text categorization is a regular problem in people daily work and an interesting research area of machine learning. Semi-supervised learning algorithms, which consider both labeled and unlabeled data, can improve learning effectiveness significantly. This paper gives the definition and characteristic of text categorization and introduces the traditional supervised learning algorithms and evaluation indicators. Then it analyzes the characteristic and basic theory of semi-supervised text categorization, and discusses some algorithms on semi-supervised text categorization, such as Bayesian method and regularization method.
Key words: text categorization; semi-supervised learning; naïve Bayesian; manifold and spectralgraph

摘 要：文本分类是人们日常工作中经常遇到的问题, 也是机器学习的重要研究内容。半监督学习算法同时考虑有标记和无标记数据, 能显著提升学习效果。给出了文本分类的定义和特点, 介绍了传统的监督学习分类算法和评价指标, 对半监督文本分类的特点和基础理论进行了分析, 并具体介绍了一些半监督文本分类算法, 如贝叶斯方法和正则化方法。
关键词：文本分类; 半监督学习; 朴素贝叶斯; 流形和谱图
文献标识码：A **中图分类号：**TP181

*The National Natural Science Foundation of China under Grant No. 60775046 (国家自然科学基金).
Received 2010-04, Accepted 2010-07.

1 引言

传统的监督学习中，学习器通过大量的有标记样本建立模型，用于预测未见样本的标记。随着数据收集和存储技术的飞速发展，收集大量无标记样本变得相当容易，然而获取大量有标记样本则相对困难，对收集到的样本进行标定通常要花费很多时间和精力。事实上，在现实世界的应用问题中，通常存在大量的无标记样本，但有标记样本则相对较少，尤其在文本分类任务中，如果仅使用有标记样本，问题规模将受到很大的限制。

在统计学界，从有标记和无标记的混合数据中学习的思想可以追溯到几十年前，早在 1969 年 Day 就提出了针对无标记数据的类似期望最大化算法(expectation maximization, EM)的迭代算法[1]，后来 Dempster 等人整理了前人估计数据缺失值的技术，于 1977 年建立了 EM 算法的理论框架[2]。此后 EM 算法不断发展，到 20 世纪 90 年代，为了估计不能显示求解的对数似然函数，期望最大化算法即 EM 算法被引入机器学习领域[3−5]。

显然，如果仅使用少量的有标记样本，利用它们训练出的学习系统往往很难具有强泛化能力，同时，大量廉价的无标记样本弃之不用，是对数据资源的极大浪费。如何利用大量的无标记数据来改善学习性能已成为当前机器学习研究中最受关注的问题之一[6−17]。这样的学习方法被称为半监督学习和直推式学习。

半监督学习的核心思想在于光滑性假设:

(1) 距离越近的点越倾向于拥有相同的标记;

(2) 数据的分布具有某种内在结构(簇或流形)。

上述假设前者是局部性质，后者是全局性质，它们与设计学习算法的一些原则，例如低密度分离原则是一致的。事实上，传统的监督学习也是依赖光滑性假设的，但这些算法一般只考虑局部的光滑性假设。

传统的文本分类使用最多的分类器是朴素贝叶斯(naïve Bayesian, NB)和支持向量机(support vector machine, SVM)。在监督学习框架下，NB 考虑如何估计后验概率 $P(y|x)$，这是一个条件概率，所以估计数据本身的概率分布 $P(x)$ 对于估计 $P(y|x)$ 基本没有作用。SVM 也是一样，忽略了类别内信息，只考虑类别间信息。但是 $P(y|x)$ 是数据和标记的联合概率 $P(x,y)$ 的条件分布，而 $P(x)$ 是 $P(x,y)$ 的边缘分布，巧妙地利用无标记样本提供的信息将有助于捕捉数据的真实内在结构，从而提高学习器的泛化能力，如图 1 和图 2。

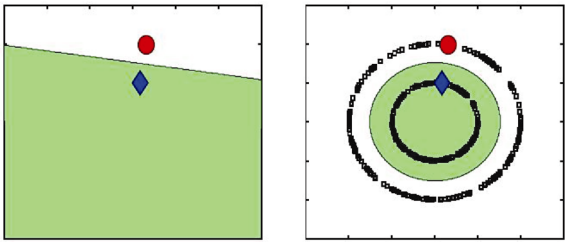

Fig.1 Unlabeled data and prior assumptions

图 **1** 无标记数据和先验假设

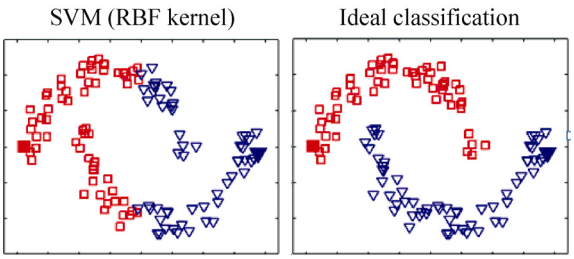

Fig.2 Standard SVM and the ideal classifier

图 **2** 标准的 **SVM** 和理想的分类器

理论分析和实验表明，文本分类时有充分的无标记数据，只要它们和有标记数据产生自同样的概率分布，考虑这些无标记的数据就能建立更适合的模型，使用这些模型的分类器也就具有更好的泛化能力[4,6]。

本文第 2 章首先介绍在监督学习框架下如何对文本进行分类，给出文本分类的定义，介绍常见的分类器和评价指标。第 3 章介绍使用精巧的产生式模型的 NB，它虽然简单，但对文本分类很有效，模型先验概率和精度具有强的正相关性。第 4 章介绍流形正则化框架下的归纳式文本分类算法以及基于邻接图的直推式文本分类算法，注意到谱图论把图上的 Laplace 算子和流形上的 Laplace 算子联系起来，使得 SVM 通过嵌入流形信息扩展为半监督学习器。

此外还有一类算法，即协同训练[18]，在协同训练框架下，任何适用于文本分类的传统学习器比较容易扩展为针对混合数据的学习器，鉴于篇幅所限，本文暂不涉及。

2　传统的文本分类方法

在研究半监督文本分类之前，首先简单介绍传统的文本分类方法。文本分类任务有其自身特点，监督学习框架下的很多用于文本分类的经典算法都是针对这些特点设计的，了解这些算法对于继续深入探讨半监督文本分类是十分有益的。关于传统的文本分类方法可参考文献[19–21]。

2.1　文本分类的定义

文本分类问题的形式化定义为：使用一个分类器，通过学习，对一个未知函数 $\Phi: D\times C\to\{T,F\}$ 进行估计，其中 $C=\{c_1,c_2,...,c_M\}$ 为预先定义好的类别集，D 为文档集。如果 $\Phi(d_i,c_j)=T$，则 d_i 是 c_j 的一个正例(或成员)，如果 $\Phi(d_i,c_j)=F$，则 d_i 是 c_j 的一个反例。

文档的类别仅仅是符号化的标记(即自然数)，并没有额外的过程性或说明性的知识，也没有元数据(如中图法分类号)，只能从文档的内容本身提取信息来对文档进行分类。这是文本分类研究的一般设定，当然，对于特定的应用，使用外部知识或元数据可提高分类器的泛化能力。

一般认为，文本分类是一个主观任务，即便两位资深的人类专家的分类结果也经常会不一致。因此，文本分类中的机器学习技术与其说提供了这种不一致性的黄金标准，不如说通过学习重现了专家判断的主观性。

根据不同的应用，文本分类可以是单标记的(每个文档 d_i 对应一个类别 $c_j\in C$)，也可以是多标记的(可以有 $0\leqslant n_i\leqslant M$ 个类别被赋予一个文档)。单标记的一种特殊情形是二元文本分类，此时，任意的 $d_i\in D$ 要么属于一个类别 c_j，要么属于该类别的补 $\bar{c}_j$，类别对应目标函数 $\Phi_j: D\to\{T,F\}$。理论上，二元分类的情况比多标记(也比单标记)更为一般，因为一个二元文本分类算法同样可以进行多标记分类，反之则不行。常见的做法是把原问题转化为 M 个二类任务 $\{c_j,\bar{c}_j\}, j=1,2,...,M$，这种转化要求子任务是统计独立的，即 $\Phi(d_i,c')$ 不依赖 $\Phi(d_i,c'')$。另外，从机器学习的角度，训练二元分类器要比多类分类器更容易实现(如SVM)，所以若无特别说明，文本分类通常指代二元的或单标记的文本分类。

2.2　索引化(向量化)

文档索引技术属于知识表示的范畴，它把文档从字符流映射到一个度量空间(如欧氏空间)，使文档向量化，向量化的文档可以作为未训练的分类器的输入来对分类器进行训练，或者作为已训练的分类器的输入来预测类别。目前使用的索引技术几乎全部来自文本信息检索领域，原始文本流 d_i 被表示成 $x_i=\langle w_{1i},w_{2i},...,w_{|T|i}\rangle$，其中 T 是一个字典，包含了至少在 k 个文档中出现过的项(或称为特征)，$0\leqslant w_{ki}\leqslant 1$ 为项 t_k 在文档 d_i 中的语义重要程度。注意，在不引起混淆的情况下，本文 d_i 和 x_i 是通用的。

为文档做索引有两个要点，一是构造项集，二是计算某个项的重要度。一个常用的方法是，对单词的词干在全部文档中出现的次数计数，排除一些常见的高频词，次数大于某阈值的词即成为项。项的重要度一般为实数，常用的方法是计算归一化的 $tf*idf$ [22]，其中 tf 为项的频繁度，idf 为逆文档频繁度，注意用频率估计概率时要做平滑处理。

如果分类器不能处理大规模数据，使用常规的信息检索技术把文档向量化之后还要降维，可以使用特征选择或特征提取技术[23]。

举例来说，可以使用 CMU(Carnegie Mellon University)的彩虹软件包[24]，首先把文档输入到Porter提取词干，然后用SMART过滤掉高频的无用词，再去掉出现次数小于给定阈值的单词词干，最后计算 $tf*idf$ 并归一化。

2.3　训练分类器

现在需要解决的问题是如何从向量表示的文档集中归纳出一个数学模型，使得该模型可以对未见文档样本进行分类。关于机器学习的一般问题可以

参考 Mitchell 的教材[25]。

常见的学习器都可以胜任此任务，例如 NB、SVM、决策树、人工神经网络、遗传算法、隐 Markov 模型等。这里介绍 NB 和 SVM，它们是第 3 章和第 4 章的基础。

2.3.1 朴素贝叶斯分类器

贝叶斯方法使用贝叶斯定理计算文档 d_i 属于类别 c_j 的后验概率：

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

其中，$P(d_i)$ 为随机选取一个文档时该文档具有向量形式 $\boldsymbol{x}_i=\langle w_{1i},w_{2i},...,w_{|T|i}\rangle$ 的概率，而 $P(c_j)$ 为随机选取一个文档时该文档属于类别 c_j 的概率。

计算 $P(d_i)$ 或者 $P(d_i \mid c_j)$ 的代价非常巨大，所以经常假定 x_i 的项 t_k 之间是没有关系的，即 w_{ki} 是统计独立的，在这个假设下，

$$P(x_i \mid c_j)=\prod_{k=1}^{|T|}P(w_{ki} \mid c_j)$$

这个分类器即是 NB。最著名的 NB 是二元独立分类器[26]。令 p_{kj}、$p_{k\bar{j}}$ 分别表示 $P(w_{ki}=1 \mid c_j)$ 和 $P(w_{ki}=1 \mid \bar{c}_j)$，则

$$\log\frac{P(c_j \mid x_i)}{1-P(c_j \mid x_i)}=\log\frac{P(c_j)}{1-P(c_j)}+\sum_{k=1}^{|T|}w_{ki}\log\frac{p_{kj}(1-p_{k\bar{j}})}{p_{k\bar{j}}(1-p_{kj})}+\sum_{k=1}^{|T|}\log\frac{1-p_{kj}}{1-p_{k\bar{j}}}$$

上式中，等号前是 $P(c_j \mid x_i)$ 的单调增函数，可以当作后验概率，而等号后第一项和第三项对所有的文档都是常数，因而可以忽略，所以只需估计 $2|T|$ 个参数 $\{p_{1j},p_{1\bar{j}},p_{2j},p_{2\bar{j}},...,p_{|T|j},p_{|T|\bar{j}}\}$。事实上，由于文本分类问题中项固有的稀疏性，大部分 w_{ki} 为零，只要估计非零的 w_{ki} 对应的 p_{ij}、$p_{i\bar{j}}$ 就足够了。

用于文本分类的 NB 及其改进在文献[27]中有详细描述。

2.3.2 支持向量机

SVM 是以结构风险最小化原则为基础，建立在正则化框架下的基于核方法的通用分类器[28−29]。SVM 把数据从输入空间映射到核诱导特征空间，并在核诱导特征空间中建立线性分离超平面，引入松弛变量得到软边界，分离超平面使正例和反例在该平面的法向上尽可能分开。换句话说，SVM 在核诱导特征空间中，试图最大化样本到分离超平面的最小距离，这等价于最小化泛化误差。

训练一个 SVM 的复杂性和样本所在的特征空间的维度无关，只和样本的个数有关，这意味着在为文档做索引时可以提取大量的特征，之后却只需付出很少的计算代价。

最经典的 SVM 近似算法为序列最小优化算法(sequential minimal optimization, SMO)[30]，经验时间复杂度为 $O(\ell^{2.3})$，而最快的近似算法核向量机(core vector machines, CVM)被证明达到了 $O(\ell)$，即线性的时间复杂度[31]，其中 ℓ 为用于训练的文档集中包含的文档个数。

Joachims 在文献[32]中指出了 SVM 适合做文本分类的四点理由，文档向量具有以下特征：

(1) 高维的输入空间;

(2) 很少的无关特征;

(3) 项的稀疏性;

(4) 多数问题线性可分或近似线性可分。

历史上 SVM 逐渐进入人们的视野，一个重要原因正是 SVM 适用于文本分类，SVM 适用于文本分类的理由也正是文本分类问题自身的特点。

网上有很多功能齐全的开源 SVM 包，常用的有 LIBSVM[33](a library for support vector machines)和 SVM Light[34]。

2.4 评价指标

训练效率、分类效率和分类的有效性是评价分类器的三个指标(这里的效率一般指时间)。

理论研究通常只考虑分类结果的有效性，即分类器的泛化能力，但在实际应用中，不能不考虑训练和使用分类器的效率，必须在三个指标中寻求平衡，例如经常与用户交互的系统就不能使用分类效率很低的分类器。当然随着硬件计算速度的提升，分类结果有效性的地位变得越来越重要。

最常用的有效性指标是正确率(或等价的，错

误率), 是指被正确分类的文档比例。当类别的先验概率非常不平衡时, 就要考虑代价是否敏感, 否则简单地把文档预测为先验概率较大的类别将获得较高的正确率。一种改进是引入召回率的概念, 用精度和召回率取代正确率, 此时精度 P 表示在分类器认为属于一个类别的文档中确实属于该类别的文档所占的比例, 召回率 R 表示在一个类别的所有文档中被分类器正确预测为该类的文档所占的比例。有多种方法可以合并精度和召回率, 令 $F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P+R}$, 则最常用的指标就是 $F_1 = \frac{2PR}{P+R}$, 它是精度 P 和召回率 R 的调和平均值。另一个常用指标是 AUC(area under the ROC curve)值, 它是 ROC(receiver operating characteristic)曲线与 x 轴和直线 $x=1$ 围成的面积。

3 贝叶斯方法

下面将讨论如何有效利用无标记数据来建立 NB 分类器, 主要方法是使用 EM 算法(或模拟退火算法)对贝叶斯方法中的未知参数进行估计。虽然仅使用简单的产生式模型和简单的学习器, 模型先验概率和分类正确率还是高度的正相关。因为引入了无标记数据, 对数似然函数中含有对数项, 简单的极大似然估计不再适用, 所以使用 EM 算法来估计模型中的未知参数, 模拟退火算法是 EM 的一个变种, 用于有效地规避局部极值点。

3.1 最简单的模型

最简单的产生式模型中隐含了三个假设：

(1) 数据由一个混合模型产生;

(2) 混合成分和类别一一对应;

(3) 混合成分是项的多项式分布[27]。

假定每个文档由一个混合模型生成, 为了评价模型, 不妨设每个模型被一个参数 θ 唯一刻画。根据全概率公式, 有

$$P(x_i \mid \theta) = \sum_{j=1}^{M} P(c_j \mid \theta) P(x_i \mid c_j; \theta)$$

文档表示成向量形式 $\boldsymbol{x}_i = \langle x_{i1}, x_{i2}, ..., x_{i|T|} \rangle$, 若向量的元素是单词出现的次数, 则 $|x_i| = \sum_{k=1}^{|T|} x_{ik}$, 由标准的 NB 假设得

$$P(x_i \mid c_j; \theta) \propto P(|x_i|) \prod_{k=1}^{|T|} P(t_k \mid c_j; \theta)^{x_{ik}}$$

每个独立的混合成分的参数 $\theta_{c_j} \equiv P(c_j \mid \theta)$ 都定义了一个项的多项式分布 $\theta_{t_k|c_j} \equiv P(t_k \mid c_j; \theta)$。不失一般性, 可以假定文档的长度符合均匀分布, 这样 $P(|x_i|)$ 可以不予考虑。

只考虑有标记数据, 项和类别的极大似然估计是

$$\hat{\theta}_{t_k|c_j} \equiv P(t_k \mid c_j; \hat{\theta}) = \frac{1 + \sum_{i=1}^{|D|} \delta_{ij} x_{ik}}{|T| + \sum_{h=1}^{|T|} \sum_{i=1}^{|D|} \delta_{ij} x_{ih}}$$
$$\hat{\theta}_{c_j} \equiv P(c_j \mid \hat{\theta}) = \frac{1 + \sum_{i=1}^{|D|} \delta_{ij}}{M + |D|} \tag{1}$$

其中, δ_{ij} 是类别的指示函数, 当 x_i 对应的类别 $y_i = c_j$ 时 δ_{ij} 为 1, 否则为 0。

当考虑无标记数据时, 极大似然估计(1)的封闭形式不存在, 需要使用 EM 算法。首先, 在有限的有标记样本上建立标准的 NB, 然后使用刚才建立的 NB 对无标记样本做预测, 注意不要输出概率最大的类别, 而是为每个类别赋予一个概率, 这样每个类别都把一个无标记的样本当作一部分来对待。在新得到的混合数据上建立 NB 给无标记样本赋予概率, 重复这个过程直至收敛[4]。

形式化地说, 令 D_l、D_u 分别表示有标记和无标记的样本, 此时需求解如下的对数似然函数(忽略常数项)：

$$l(\theta \mid D, Y) = \log(P(\theta)) + \sum_{x_i \in X_u} \log \sum_{j=1}^{M} P(c_j \mid \theta) P(x_i \mid c_j; \theta) + \sum_{x_i \in X_l} \log(P(y_i = c_j \mid \theta) P(x_i \mid y_i = c_j; \theta)) \tag{2}$$

上式中等号后第二项导致不能使用简单的极大似然法, 而 EM 算法[2]使用迭代的类似爬山的策略在参数空间中搜索局部极大值, 在 E 步使用当前的模型给无标记样本赋予概率, 在M步使用式(1)估计模

型参数，当式(2)中两次迭代似然函数的改变量小于一个阈值时，算法停止。实验表明，使用相同的产生式模型，考虑无标记数据会提高 NB 的精度。

3.2 稍复杂的模型

现实世界的应用中，最简单的产生式模型的第二个隐含假设，即混合成分和类别一一对应，并不总是成立，考虑垃圾邮件分类，不管垃圾邮件还是正常邮件都可能包含许多主题，也就不可能用一个多项式分布来描述它们。为了解决这个问题，可以使用更加复杂的产生式模型，混合成分和类别的对应关系改为多对一，每个类别下可以有不同的子主题，每个子主题对应一个多项式分布[35]。

此时对于无标记的样本，现在有两个缺失值，类别和子主题，对于有标记的样本，子主题也是未知的。现在用 $\{s_1,s_2,...,s_N\}$ 表示混合成分(子主题)，用 $\{c_1,c_2,...,c_M\}$ 表示类别，若 s_j 属于 c_a，则 q_{aj} 为 1，否则为 0。一个文档 x_i 的类别和子主题表示为 y_i、z_i。

如果只考虑有标记数据，那么同前面类似，即

$$\hat{\theta}_{t_k|s_j} \equiv P(t_k \mid s_j;\hat{\theta}) = \frac{1+\sum_{i=1}^{|D|}\delta_{ij}x_{ik}}{|T|+\sum_{h=1}^{|T|}\sum_{i=1}^{|D|}\delta_{ij}x_{ih}}$$

$$\hat{\theta}_{c_a} \equiv P(c_a \mid \hat{\theta}) = \frac{1+\sum_{i=1}^{|D|}\delta_{ia}}{M+|D|} \tag{3}$$

$$\hat{\theta}_{s_j|c_a} \equiv P(s_j \mid c_a;\hat{\theta}) = \frac{1+\sum_{i=1}^{|D|}\delta_{ij}\delta_{ia}}{\sum_{j=1}^{N}q_{aj}+\sum_{i=1}^{|D|}\delta_{ia}}$$

对未见文档样本进行分类时，首先计算该文档属于一个子主题的概率，再把一个类别中的子主题相加，即

$$P(z_i=s_j \mid x_i;\hat{\theta}) = \frac{\sum_{a=1}^{M}q_{aj}\hat{\theta}_{c_a}\hat{\theta}_{s_j|c_a}\prod_{k=1}^{|D|}(\hat{\theta}_{t_k|s_j})^{x_{ik}}}{\sum_{r=1}^{N}\sum_{b=1}^{M}q_{bj}\hat{\theta}_{c_b}\hat{\theta}_{s_j|c_b}\prod_{k=1}^{|D|}(\hat{\theta}_{t_k|s_j})^{x_{ik}}} \tag{4}$$

$$P(y_i=c_a \mid x_i;\hat{\theta}) = \sum_{j=1}^{N}q_{aj}P(z_i=s_j \mid x_i;\hat{\theta})$$

当考虑无标记数据时，需要使用 EM 算法，同之前一样，在 M 步使用式(3)估计模型参数，在 E 步使用式(4)计算无标记样本属于子主题和类别的概率。此外，还需要计算有标记样本属于其所在类别包含的子主题的概率，因为样本属于其他类别的子主题的概率为零，所以只要计算样本对应类别的子主题的概率 $P(z_i=s_j \mid x_i;\hat{\theta})$ 并归一化即可。

当没有子主题与类别的对应关系，甚至没有任何关于子主题的可用知识时，就需要进行模型选择，可以使用常见的标准，如 AIC(Akaike information criterion)或 BIC(Bayesian information criterion)，通过 10 折交叉验证等技术为每个类别设置一定数量的子主题。此时模型的复杂性和数据的稀疏性导致了一个问题：如果子主题的数目等于文档的数目，那么可以完美地拟合训练数据，但此时模型的泛化能力将会非常差。事实上当子主题的数目很多时就存在这个问题，因为这时需要估计很多的参数，而关于每个子主题却只有很少的数据，并不能准确地估计真实概率。当子主题的数目很少时，相反的问题就会出现，虽然参数可以完美地估计，但由于模型自身的限制，依然没有好的泛化能力。如何避免过拟合和欠拟合是一个悬而未决的问题[35]。

4 正则化方法

4.1 理论背景

正则化方法是 Tikhonov 于 1963 年提出的,主要用于解决不适定问题[36–37]，思想是在不适定算子方程的对应泛函上施加若干正则项。不失一般性，设泛函 $W(f)$，只要满足以下条件，则这个不适定问题可以转化为一个适定问题[28]：

(1) 算子方程的解属于 $W(f)$ 的定义域；

(2) 在定义域上 $W(f)$ 非负；

(3) 所有 $\{f:W(F)\leqslant c\},c\geqslant 0$ 都是紧的。

流形正则化试图捕捉数据概率分布的几何特征，作为一个正则项加入目标泛函。此时，目标泛函包含两个正则项，一个正则项得自类别间信息，在样本邻域空间控制分类器的复杂性，另一个正则项得自类别内信息，使用数据的几何特征控制分类

器的复杂性。

任取一个 Mercer 核函数 $K: X \times X \to \mathbb{R}$，存在与之对应的再生核 Hilbert 空间(reproducing kernel Hilbert space, RKHS) H_K 和空间上的范数 $\| \|_K$。给定有标记样本 $(x_i, y_i), i = 1,2,...,l$，采用传统的正则化方法求解：

$$f^* = \arg\min_{f \in H_K} \frac{1}{l}\sum_{i=1}^{l} V(x_i, y_i, f) + \gamma \|f\|_K^2 \tag{5}$$

其中，V 是选定的损失函数，γ 是正则化因子。惩罚 RKHS 范数能在凸优化的可行解之上附加光滑性条件。经典的表示定理指出，优化问题(5)的解具有如下形式：

$$f^*(x) = \sum_{i=1}^{l} \alpha_i K(x_i, x)$$

这样，问题转化为由系数 α_i 张成的有限维空间上的优化问题，这是 SVM 和岭回归等算法的基础。

在半监督学习框架下，考虑到数据的内在几何结构，可以求解

$$f^* = \arg\min_{f \in H_K} \frac{1}{l}\sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2 \tag{6}$$

其中，$\|f\|_I^2$ 是选定的关于数据的边缘概率分布 P_X 的惩罚项。若 P_X 已知，则式(6)具有如下形式：

$$f^*(x) = \sum_{i=1}^{l} \alpha_i K(x_i, x) + \int_M \alpha(z) K(x, z) \mathrm{d}P_X(z)$$

其中，$M = supp\{P_X\}$ 是 P_X 的支持集。若 P_X 未知，则必须对 P_X 和相应的 $\| \|_I$ 做经验估计。根据引言中的光滑性假设，当 $supp\{P_X\}$ 是一个紧流形时，一个自然的选择是所有样本的邻接图 Laplace 算子，附加一定的条件，它等价于流形上的 Laplace-Beltrami 算子 Δ_M。

给定有标记样本 $\{(x_i, y_i)\}_{i=1}^{l}$ 和无标记样本 $\{x_i\}_{i=l+1}^{l+u}$，考虑如下问题：

$$f^* = \arg\min_{f \in H_K} \frac{1}{l}\sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_K^2 + \frac{\gamma_I}{(u+l)^2} \sum_{1 \le i < j \le l+u} (f(x_i) - f(x_j))^2 W_{ij} \tag{7}$$

其中，W_{ij} 是邻接图的边权。令对角矩阵 $\boldsymbol{D}$ 的元素 $D_{ii} = \sum_{j=1}^{l+u} W_{ij}$，则 $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ 为邻接图的 Laplace 矩阵。优化问题(7)的解有如下形式[13]：

$$f^*(x) = \sum_{i=1}^{l+u} \alpha_i K(x_i, x)$$

这是流形正则化学习算法的基础。

另一方面，基于相似图的直推式学习算法中，除了少数算法直接通过随机游走传播标记外，多数算法试图从邻接图中导出关于样本的新度量。如果图中两个节点的距离是通过 Dijkstra 算法计算的，那么这个距离可被认为是两样本在流形上的测地距离的有效近似。两种常见的选择是高斯相似度和余弦相似度，或者选用由两者归一化导出的对称 Laplace 矩阵 $\tilde{\boldsymbol{L}} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$。

4.2　半监督学习算法

直推式支持向量机(transductive support vector machines, TSVM)[6](注意 TSVM 本质是半监督算法，而非直推式算法)允许在 RKHS 中的函数集和无标记样本的标记集 $\{y_j\}_{j=l+1}^{l+u}$ 上做联合优化，它基于下式：

$$\min_{\substack{w,b,\xi_1,\xi_2,...,\xi_{l+u},\\ y_{l+1},y_{l+2},...,y_{l+u}}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i + C^*\sum_{i=l+1}^{l+u}\xi_i$$
$$\text{s.t. } y_i(\langle w, x_i \rangle + b) \geqslant 1 - \xi_i, \xi_i > 0, i = 1,2,...,l+u \tag{8}$$
$$y_j \in \{-1,1\}, j = l+1, l+2, ..., l+u$$

其中，C、C^* 是用户设定的参数，分别控制有标记样本和无标记样本上的 hinge 损失惩罚程度。由于式(8)涉及整数规划，难以直接求解，Joachims 建议使用迭代的方法。具体来说，先在有标记样本上建立 SVM 对无标记样本进行标记，然后在得到的样本集上训练新的 SVM，直到目标函数和标记集都收敛。然而这一过程只能保证收敛到局部极值点，并且收敛速度可能很慢，即使如此，TSVM 依然比只考虑有标记样本的 SVM 具有相当的优越性。

半监督支持向量机[17](semi-supervised support

vector machines, S3VM)和 TSVM 很相似，它的主要问题为(为保持一致性这里使用 2 范数)：

$$\min_{w,b,\xi_1,\xi_2,...,\xi_{l+u}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i + C^*\sum_{i=l+1}^{l+u}\xi_i$$
$$\text{s.t. } y_i(\langle w,x_i\rangle + b) \geqslant 1-\xi_i, i=1,2,...,l$$
$$|\langle w,x_i\rangle + b| \geqslant 1-\xi_i, i=l+1,l+2,...,l+u$$
$$\xi_i > 0, i=1,2,...,l+u$$

基于度量的正则化[11]使用一个基于梯度的正则项，它对高密度区域的函数变化施加大力度惩罚，给低密度区域的函数变化施加小力度惩罚，从而实现半监督学习的低密度分离原则。基于度量的正则化优化为：

$$f^* = \arg\min_{f\in F}\sum_{i=1}^{l}V(x_i,y_i,f) + \gamma\int_X \|\nabla f(x)\|^2 p(x)\mathrm{d}x$$

其中，$p(x)$ 是边缘分布 P_X 的密度。对任意的 $p(x)$，寻找一个以 $\int_X \|\nabla f(x)\|^2 p(x)\mathrm{d}x$ 为 RKHS 范数的核函数并不容易，所以 Bousquet 等建议在一组事先选定的基函数集合上进行常规优化。该方法和流形正则化非常相似，区别是范数 $\nabla f(x)$ 和 $\nabla_M f(x)$ 选择的不同，当数据近似分布在一个微分流形上时，两者的效果存在一定的差异，因为流形法向上的光滑性要求采用对学习任务没有帮助。

采用 Laplacian 支持向量机 LapSVM[13]间接地求解式(7)，首先求解 l 维二次规划：

$$\max_{\beta\in\mathbb{R}^l}\sum_{i=1}^{l}\beta_i - \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\left[\boldsymbol{YJK}\left(2\gamma_A\boldsymbol{I} + 2\frac{\gamma_I}{(u+l)^2}\boldsymbol{LK}\right)^{-1}\boldsymbol{J}^{\mathrm{T}}\boldsymbol{Y}\right]\boldsymbol{\beta}$$
$$\text{s.t. } \sum_{i=1}^{l}y_i\beta_i = 0, 0 \leqslant \beta_i \leqslant \frac{1}{l}, i=1,2,...,l \quad (9)$$

其中，$\boldsymbol{Y}$ 为对角矩阵 $Y_{ii}=y_i$，$\boldsymbol{K}$ 为所有样本的 Gram 矩阵，$\boldsymbol{L}$ 为邻接图 Laplace 矩阵，$\boldsymbol{J}$ 为 $l\times(l+u)$ 矩阵，若 $i=j$ 且 x_i 有标记，则 J_{ij} 为 1，否则为 0。设式(9)的解为 $\boldsymbol{\beta}^*$，则式(7)的解为：

$$\boldsymbol{\alpha}^* = \left(2\gamma_A\boldsymbol{I} + 2\frac{\gamma_I}{(u+l)^2}\boldsymbol{LK}\right)^{-1}\boldsymbol{J}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{\beta}^*$$

通过这个转化，可以使用标准的 SVM 求解程序来求解 LapSVM。

4.3 直推式学习算法

作为直推式学习的代表，采用一致性学习方法[10]定义矩阵 $\boldsymbol{W}$，以高斯相似度为例，若 $i\neq j$，则 $W_{ij}=\exp(-\|x_i-x_j\|^2/2\sigma^2)$，对角线元素 $W_{ii}=0$。然后定义对角矩阵 $\boldsymbol{D}$，$D_{ii}=\sum_{j=1}^{l+u}W_{ij}$ 和归一化的矩阵 $\boldsymbol{S}=\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$。定义 $(l+u)\times M$ 矩阵 $\boldsymbol{Y}$，这里 M 为文档类别数，若 x_i 对应的标记 $y_i=j$，则 $Y_{ij}=1$，否则为 0。令 $\boldsymbol{F}(0)=\boldsymbol{Y}$，然后使用如下规则进行迭代：

$$\boldsymbol{F}(t+1)=\alpha\boldsymbol{S}\boldsymbol{F}(t)+(1-\alpha)\boldsymbol{Y}, 0<\alpha<1 \quad (10)$$

其中，α 是学习率。

事实上，$\boldsymbol{S}$ 的特征值和 α 都介于 0 和 1 之间，因此式(10)收敛于一个解析的表达式 $\boldsymbol{F}^*=(\boldsymbol{I}-\alpha\boldsymbol{S})^{-1}\boldsymbol{Y}$。Zhou 等人指出该算法同样能归于正则化框架之下[10]，学习率 α 越大 $\boldsymbol{F}^*$ 越光滑，α 越小 $\boldsymbol{F}^*$ 越接近原始标记 $\boldsymbol{Y}$，即 α 平衡了最终结果在有标记样本上的损失和光滑性。无标记数据的预测通过 $y_i=\arg\max_{1\leqslant j\leqslant M}F_{ij}^*$ 得到。

类似的直推式学习算法还有 Markov 随机游走[8]。Markov 随机游走定义在局部度量之上。具体来说，设有度量 $d(x_i,x_j)$，定义一个对称的 K 近邻图，其中对角线元素 $W_{ii}=1$，若 x_i 和 x_j 在 K 近邻图中相连，则有 $W_{ij}=\exp(-d(x_i,x_j)/\sigma)$，否则 $W_{ij}=0$。从节点 i 到 k 的单步转移概率为 $p_{ik}=W_{ik}/\sum_j W_{ij}$。令 $\boldsymbol{A}$ 为所有单步转移概率构成的矩阵，$P_{t|0}(k\,|\,i)$ 为给定起点 i 经过 t 步到达节点 k 的概率，则有

$$P_{t|0}(k\,|\,i)=[\boldsymbol{A}^t]_{ik}$$

初始概率不妨设为均匀分布，使用标准的 Markov 模型求解算法可以估计给定终点 k，它是 t 步之前从节点 i 出发的概率 $P_{0|t}(i\,|\,k)$。分类时使用下式：

$$c_k=\arg\max_{1\leqslant c\leqslant M}\sum_i P(y=c\,|\,i)P_{0|t}(i\,|\,k)$$

其中，$P(y=c\,|\,i)$ 可以使用 EM 算法估计，或者只允许有标记样本通过随机游走传播标记。

Markov 随机游走的一个关键参数是步数 t，在原始 K 近邻图连通的条件下，当 t 趋于无穷时，所有

样本将变得不可分辨，相对的，越小的 t 值越倾向于小范围内游走以传播标记。

这两个算法是直推式的，只能对无标记样本而不能对从未见过的样本进行分类。

5 结束语

本文首先给出了文本分类的定义及特点，介绍了传统的监督学习分类算法和评价指标，然后对半监督文本分类的特点和基础理论进行了分析，并具体介绍了一些半监督文本分类算法。最后指出半监督文本分类还没有完全成熟，人们会继续关注这一研究领域，并给出更多的研究成果。

References:

[1] Day N E. Estimating the components of a mixture of normal distributions[J]. Biometrika, 1969, 56(3): 463–474.

[2] Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society: Series B, 1977, 39(1): 1–38.

[3] Miller D J, Uyar H. A generalized Gaussian mixture classifier with learning based on both labelled and unlabelled data[C]//Advances in Neural Information Processing Systems 9: Proceedings of the 1996 Conference. Cambridge, MA, USA: MIT Press, 1996: 783–787.

[4] Nigam K, McCallum A, Thrun S, et al. Learning to classify text from labeled and unlabeled documents[C]// Proceedings of the 15th National/10th Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence. Menlo Park, CA, USA: AAAI Press, 1998: 792–799.

[5] Baluja S. Probabilistic modeling for face orientation discrimination: learning from labeled and unlabeled examples[C]//Advances in Neural Information Processing Systems 11: Proceedings of the 1998 Conference. Cambridge, MA, USA: MIT Press, 1998: 854–860.

[6] Joachims T. Transductive inference for text classification using support vector machines[C]//Proceedings of the 16th International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann, 1999: 200–209.

[7] Blum A, Chawla S. Learning from labeled and unlabeled data using graph mincuts[C]//Proceedings of the 18th International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann, 2001: 19–26.

[8] Szummer M, Jaakkola T. Partially labeled classification with Markov random walks[C]//Advances in Neural Information Processing Systems 14: Proceedings of the 2001 Conference. Cambridge, MA, USA: MIT Press, 2001: 945–952.

[9] Chapelle O, Weston J, Schöelkopf B. Cluster kernels for semi-supervised learning[C]//Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference. Cambridge, MA, USA: MIT Press, 2002: 585–592.

[10] Zhou D, Bousquet O, Lal T, et al. Learning with local and global consistency[C]//Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference. Cambridge, MA, USA: MIT Press, 2003: 321–328.

[11] Bousquet O, Chapelle O, Hein M. Measure based regularization[C]//Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference. Cambridge, MA, USA: MIT Press, 2003: 1221–1228.

[12] Zhu X, Lafferty J, Ghahramani Z. Semi-supervised learning using Gaussian fields and harmonic functions[C]//Proceedings of the 20th International Conference on Machine Learning. Menlo Park, CA, USA: AAAI Press, 2003: 912–919.

[13] Belkin M, Niyogi P, Sindhwani V, et al. Manifold regularization: a geometric framework for learning from examples[J]. Journal of Machine Learning Research, 2006, 7: 2399–2434.

[14] Kondor R, Lafferty J. Diffusion kernels on graphs and other discrete input spaces[C]//Proceedings of the 19th International Conference on Machine Learning, Sydney,

2002. San Francisco, CA, USA: Morgan Kaufmann, 2002: 315–322.

[15] Smola A, Kondor R. Kernels and regularization on graphs[C]//Schölkopf B, Warmuth M K. Proceedings of the 16th Annual Conference on Learning Theory (COLT). [S.l.]: Springer, 2003: 144–158.

[16] Shrager J, Hogg T, Huberman B. Observation of phase transitions in spreading activation networks[J]. Science, 1987, 236: 1092–1094.

[17] Bennett K, Demiriz A. Semi-supervised support vector machines[C]//Advance in Neural Information Processing Systems 11: Proceedings of the 1998 Conference. Cambridge, MA, USA: MIT Press, 1998: 368–374.

[18] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the 11th Annual Conference on Computational Learning Theory. New York, NY, USA: ACM, 1998: 92–100.

[19] Sebastiani F. Text categorization[M]//Text Mining and its Applications. [S.l.]: WIT Press, 2005: 109–129.

[20] Sebastiani F. Machine learning in automated text categorization[J]. ACM Computing Surveys, 2002, 34: 1–47.

[21] Sebastiani F. A tutorial on automated text categorization[C]//Proceedings of the 1st Argentinian Symposium on Artificial Intelligence (ASAI-99), Buenos Aires, AR, 1999: 7–35.

[22] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information Processing and Management, 1988, 24: 513–523.

[23] Dasgupta A, Drineas P, Harb B, et al. Feature selection methods for text classification[C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2007: 230–239.

[24] Bow: a toolkit for statistical language modeling, text retrieval, classification and clustering[EB/OL]. [2010]. http://www.cs.cmu.edu/~mccallum/bow/.

[25] Mitchell T. Machine learning[M]. [S.l.]: McGraw Hill, 1997.

[26] Robertson S, Sparck J. Relevance weighting of search terms[J]. Journal of the American Society for Information Science, 1976, 27: 129–146.

[27] Lewis D. Naive (Bayes) at forty: the independence assumption in information retrieval[C]//Lecture Notes in Computer Science 1398: Proceedings of the 10th European Conference on Machine Learning (ECML-98), Chemnitz, Germany. [S.l.]: Springer, 1998: 4–15.

[28] Vapnik V. Statistical learning theory[M]. [S.l.]: John Wiley & Sons, 1998.

[29] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods[M]. Cambridge: Cambridge University Press, 2000.

[30] Platt J. Fast training of support vector machines using sequential minimal optimization[M]//Advances in Kernel Methods – Support Vector Learning. Cambridge, MA, USA: MIT Press, 1999: 185–208.

[31] Tsang I, Kwok J, Cheung P. Core vector machines: fast SVM training on very large data sets[J]. Journal of Machine Learning Research, 2005, 6: 363–392.

[32] Joachims T. Text categorization with support vector machines: learning with many relevant features[C]//Lecture Notes in Computer Science 1398: Proceedings of the 10th European Conference on Machine Learning (ECML-98), Chemnitz, Germany. [S.l.]: Springer, 1998: 137–142.

[33] LIBSVM—a library for support vector machines[EB/OL]. [2010]. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[34] SVM—Light support vector machine[EB/OL]. [2010]. http://svmlight.joachims.org/.

[35] Chapelle O, Schölkopf B, Zien A. Semi-supervised learning[M]. Cambridge, MA, USA: MIT Press, 2006.

[36] Tikhonov A. Regularization of incorrectly posed problems[J]. Soviet Math Dokl, 1963, 4: 1624–1627.

[37] Tikhonov A. On solving incorrectly posed problems and method of regularization[J]. Dokl Acad Nauk USSR, 1963, 151: 501–504.

NIU Gang was born in 1984. He received his M.S. degree from Department of Computer Science and Technology, Nanjing University in 2010. His research interest is statistical machine learning.
牛罡(1984—), 男, 河北石家庄人, 2010 年于南京大学计算机科学与技术系获得硕士学位, 主要研究领域为统计机器学习。

LUO Aibao was born in 1986. He is a master candidate at Department of Computer Science and Technology, Nanjing University. His research interests include machine learning and data mining.
罗爱宝(1986—), 男, 江苏姜堰人, 南京大学计算机科学与技术系硕士研究生, 主要研究领域为机器学习, 数据挖掘。

SHANG Lin was born in 1973. She received her Ph.D. degree from Department of Computer Science and Technology, Nanjing University in 2004. Now she is an associate professor at Department of Computer Science and Technology, Nanjing University. Her research interests include rough sets, machine learning and data mining.
商琳(1973—), 女, 河北曲阳人, 2004 年于南京大学计算机科学与技术系获得博士学位, 现为南京大学计算机科学与技术系副教授, 主要研究领域为粗糙集, 机器学习, 数据挖掘。