计算机科学2004Vol.31№.9A

文本自动分类技术和算法研究综述

吕 琳 刘玉树

(北京理工大学信息科学技术学院 北京100081)

摘 要 文本自动分类技术是面向Internet搜索引擎的重要研究方向和关键技术。它是指在给定的分类体系下，根据文本的内容自动确定文本关联类别的过程。本文总结和跟踪了各种常用和最新的文本分类的技术、算法及其适用范围，对评价与比较分类器性能的定理和方法进行了分析，并简述了独立于算法的机器学习。

关键词 文本分类，训练，度量，非度量，机器学习

Summary of Research on Techniques and Algorithms of Text Categorization

LV Lin LIU Yu-Shu

(School of Information Science & Technology, Beijing Institute of Technology, Beijing 100081)

Abstract Text Categorization is the task of increasing importance and the key technique of Internet search engine. It refers to the automated assigning of natural language texts to predefined categories based on their contents. This paper summarizes and traces all kinds of usual and the up-to-date techniques and algorithms about Text Categorization, accompanied by their application ranges. It also analyses the theorems and methods about evaluation and comparison of sort performance, and briefly describes machine learning independent of algorithm.

Keywords Text categorization, Training, Metric, Non-metric, Machine learning

1 引言

近年来，随着Internet的迅猛发展，人们迫切需要能够从Web资源中快速准确地发现知识的工具。搜索引擎就是Web上最常见的信息发现工具，但由于查准率不高，效果远不能令人满意。目前急需解决的关键技术之一就是将文本自动分类技术应用到搜索引擎上，通过文本分类器自动地将检索结果快速分类，以分类目录树的方式来显示检索结果，这样可大大减少浏览的数量，方便用户快速查询到相关的信息。

文本分类系统的任务是在给定的分类体系下，根据文本的内容自动判别文本的类别。本文主要对各种常用和最新的文本分类的技术、算法进行了总结和跟踪，对评价与比较分类器性能的定理和方法进行了分析，并简述了独立于算法的机器学习。

2 文本分类的技术和算法研究

训练与分类的技术、算法是文本分类系统的核心部分。目前存在多种基于机器学习的训练和分类方法。按照是否含有向量间距离度量的概念，可分为度量方法和非度量方法[3]。本部分将对每种方法涉及的关键技术、具体算法步骤及其适用范围逐一加以研究探讨。

2.1 度量方法

度量方法适用于所有涉及向量间距离度量概念的分类，包括简单向量距离分类法[2]、贝叶斯法[3]、K近邻规则(KNN)[2]、线性判别函数法[3]、多层神经网络法[3]、随机方法[3]、最大熵方法[4]等，这些方法研究的问题有以下共同特点：特征向量是实数，且有距离的概念。

1. *简单向量距离分类法* 该方法首先为每类文本集找到一个代表该类的中心向量，然后在新文本到来时，确定新文本向量，计算该向量与每类中心向量间的相似度，最后判定文本属于与文本距离最近的那个类。下面给出一个具体的算法，描述如下：

① 计算每类文本集的中心向量，计算方法为对所有训练文本向量进行简单的算术平均；

② 新文本到来后，预处理，特征词提取，将文本表示为特征向量；

③ 计算新文本特征向量与每类中心向量间的相似度，考虑两个特征矢量之间的夹角余弦[1]，公式为：

$$\mathrm{sim}(d_i,d_j)=\frac{V(d_i)\cdot V(d_j)}{|V(d_i)|\times|V(d_j)|}=\frac{\sum_{k=1}^{M}W_{ik}\times W_{jk}}{\sqrt{(\sum_{k=1}^{M}W_{ik}^2)(\sum_{k=1}^{M}W_{jk}^2)}}$$

吕 琳 博士，主要从事人工智能和自然语言理解的研究。刘玉树 教授，博导，主要研究方向：人工智能、计算机图形图像处理。

其中，d_i 为新文本的特征向量，d_j 为第 j 类的中心向量，V 为特征向量，M 为特征向量的维数，W_k 为向量的第 k 维；

④ 比较每类中心向量与新文本的相似度，将文本划分到相似度最大的类中。

简单向量距离分类法适用于文本表示简单，向量维数较低的情况。

2. 贝叶斯决策论　贝叶斯决策论是解决文本分类的一种基本统计途径。为了最小化分类问题中的误差概率，它总是选择那些使后验概率最大的类别。其理论基础是贝叶斯公式。以下给出一个具体算法：

① 计算特征词属于各个类别的概率向量(w_1，w_2，w_3，…，w_n)。其中，

$$w_k=P(W_k|C_j)=\frac{1+\sum_{i=1}^{|D|}N(W_k,d_i)}{|V|+\sum_{r=1}^{|V|}\sum_{i=1}^{|D|}N(W_r,d_i)}$$

② 新文本到来后，预处理，特征词提取，再按以下公式计算该文本 d_i 属于类 C_j 的概率[2]：

$$P(C_j|d_i;\hat{\theta})=\frac{P(C_j|\hat{\theta})\prod_{k=1}^{n}P(W_k|C_j;\hat{\theta})^{N(W_k,d_i)}}{\sum_{r=1}^{|C|}P(C_r|\hat{\theta}|)\prod_{k=1}^{n}P(W_k|C_r;\hat{\theta})^{N(W_k,d_i)}}$$

其中，$|C|$ 为类的总数。

$$P(C_j|\hat{\theta})=\frac{C_j\text{ 训练文档数}}{\text{总训练文档数}}$$

$P(C_r|\hat{\theta})$ 为相似含义，$N(W_k,d_i)$ 为 W_k 在 d_i 中的词频，n 为特征词总数。

③ 比较新文本属于所有类的概率，将文本划分到概率最大的类中。

贝叶斯法适用于以下情况：决策问题可用概率的形式描述，且所有有关的概率结构均已知。

3. K 近邻规则(KNN)　该规则从测试样本点 x 开始生成，不断扩大区域，直到包含进 K 个训练样本点为止，并且把测试样本点 x 的类别归为这最近的 K 个训练样本点出现频率最大的类别。具体算法如下：

① 将各个训练文本表示为特征向量；

② 在新文本到来后，预处理，提取特征词，确定新文本的向量表示；

③ 在训练文本集中选出与新文本最相似的 K 个文本，$\mathrm{Sim}(d_i,d_j)$ 的计算公式及各符号含义与简单向量距离分类法中的完全相同，同样是考虑矢量之间的夹角余弦。目前，K 值的确定没有很好的方法，一般采用先定一个初值，然后根据实验测试的结果调整 K 值，通常初值定在几百到几千之间；

④ 在新文本的 K 个邻居中，依次计算每类的权重，计算公式为：

$$p(\tilde{x},C_j)=\sum_{\tilde{d}_j\in KNN}Sim(\tilde{x},\tilde{d}_i)y(\tilde{d}_i,C_j)$$

其中，$\tilde{x}$ 为新文本的特征向量，$Sim(\tilde{x},\tilde{d}_i)$ 同上一步的相似度计算公式，$y(\tilde{d}_i,C_j)$ 为类别属性函数，若 $\tilde{d}_i$ 属于 C_j 类，则值为1，否则为 0；

⑤ 比较类的权重，将文本划分到权重最大的类中。

KNN 算法具有很好的分类性能，但这种方法计算复杂度较高，不适用于文本的实时处理，且数据维较高时，分类质量明显下降。

4. 线性判别函数法　此方法直接假定判别函数的参数形式已知，而用训练的方法来估计判别函数的参数值。松驰算法、最小平方误差方法、线性规划算法和支持向量机算法等等均属于线性判别函数法。这里我们重点讨论支持向量机(SVM)算法。SVM 法不仅可处理线性情况，对于非线性，可通过非线性变换转化为某高维空间的线性问题。它通过适当地到一个足够高维的非线性映射 $\Phi(\cdot)$，数据(属于两类)总能被一个超平面分割。得到的判别函数虽然不是·的线性函数，但却是关于 Φ 的线性函数，从此意义上讲，它属于广义线性判别函数。其具体算法描述如下：

① 假设每个文本 x_k 变换到 $y_k=\Phi(x_k)$，选择将输入数据映射到更高维空间的非线性 Φ 函数。

对 n 个文本中的每一个，$k=1,2,\cdots,n$，根据文本属于 ω_1 或 ω_2，我们分别令 $z_k=\pm1$，增广空间 y 上的判别函数就是

$$g(y)=a^t y \qquad (1)$$

这里的权向量和变换后的文本向量都是增广的(相应地取 $a_0=\omega_0$，$y_0=1$)

② 一个分隔超平面保证：

$$z_k\,g(y_k)\geqslant 1,k=1,\cdots,n \qquad (2)$$

支持向量是使此式等号成立的(变换后的)文本向量，即支持向量是最接近超平面的。

③ 间隔 b 是到判定超平面的任何正的距离。从超平面到变换后的文本 y 的距离是 $|g(y)|\ /\ \|a\|$，如果正的间隔 b 存在的话，由(2)式推出

$$z_k\,g(y_k)/\ \|a\|\geqslant b,k=1,\cdots,n \qquad (3)$$

我们的目标是找到一个使 b 最大化的权向量 a。当然，解向量可以任意伸缩，同时保持超平面不变，这样就保证了我们加上的限制条件 $b\|a\|=1$；即表达式(1)、(2)的解是 $\|a\|^2$ 的极小值。

若能找到以下形式的判别函数：它们或是样本点 x 的各个分量的线性函数，或是以 x 为自变量的某些函数(一个适当的非线性映射)的线性函数，就可使用线性判别函数法进行分类。这样，在所有两类样本集的情况下这些判别都能确定一个判定超平面。对于多类问题，可以把 C 类问题转化为 C 个两类问题来处理。该方法适用于模型相对简单且较低维的情况。其中，SVM 法是在统计学习基础上发展起来的一种新的机器学习方法，它能有效地解决过

学习问题，具有良好的推广性和较好的分类精确性，其重要的优点之一是算法复杂度可用支持向量的个数刻划，而与变换空间的样本维数无关。

5. *多层神经网络算法* 多层神经网络算法寻求的是一种在训练线性判别函数的同时学习其非线性程度的方法。它基本上执行的仍是线性判决，只是该执行过程是在输入信号的非线性映射空间中进行的。它包括前向神经网络、径向基函数网络、卷积网络算法[5]等等。这里我们重点研究一种很流行的训练多层前向神经网络的方法，即误差反向传播算法(BP)，它通过梯度下降法来进行训练，以实现由网络拓扑所定义的模型中权值的最大似然估计。下面给出三层网络随机反向传播的算法，它可以很容易地推广到更一般的多层网络。其中，n_H 为隐单元的个数，w 为网络里所有的权值，θ 为某预设值，η 为学习率，ω_{ji}为输入层单元 i 到 隐含层单元 j 的权值，ω_{kj}为隐含层单元 j 到输出层 k 的权值，δ_k为单元 k 的敏感度，x_i 为每个输入单元对应的分量，y_j 为每个隐单元的输出对应的分量，$\nabla J(w)$为准则函数 $J(w)$的变化量。具体算法描述如下：

① *begin initialize* n_H，w，准则 θ，η，$m \leftarrow 0$
② do $m \leftarrow m+1$
③ $x^m \leftarrow$随机选择文本
④ $\omega_{ji} \leftarrow \omega_{ji}+\eta\delta_j x_i$；$\omega_{kj} \leftarrow \omega_{kj}+\eta\delta_k y_j$
⑤ until $\|\nabla J(w)\| < \theta$
⑥ return w
⑦ end

该网络的主要优点是学习算法的简便性、模型选择的简易性以及容易嵌入各种启发式信息和约束条件，它能较容易地预测非线性系统，包容噪音和错误数据。非线性激活函数 f(net)的参数、文本的预处理、目标值以及权值初始化都可以通过统计学原理求出。因此这类模型非常强大，具有很好的理论基础，可应用到大量的实际问题中。较线性判别函数法而言，适用于模型相对复杂一些的情况。

此外，随机方法和最大熵方法也是近年来应用较广的文本分类方法。当涉及离散模型或有过高的复杂度，而常规的解析方法或梯度下降算法都无能为力时，则可尝试采用随机搜索方法，随着计算费用的持续降低，这种计算密集型的算法将变得越来越流行；而最大熵方法则是近几年提出的一种概率分布的估算技术，在一定条件下可以比其他方法更有效的算法。其原则是：在没有外部知识的情况下，应该倾向于尽可能一致的分布。

2.2 非度量方法

以上我们研究了基于连续实数或离散数值的特征向量距离的度量方法。此处，我们的注意力从以实向量形式表示的文本，转向以非度量的语义属性来表示的文本。非度量方法研究的是“语义数据”，它包括 CART、ID3和 C5.0等基于决策树的方法、串匹配算法、基于文法和基于规则的方法等[3]。决策树方法是最具代表性的非度量方法，而其中的分类和回归树算法(CART)是仅有的一种通用的树生长算法，它提供了一种通用的框架，利用它可实例化为各种不同的判定树。对相当大范围的应用来说，树分类器要比前面讨论过的分类器，如 KNN 分类器和神经网络分类器等，能生成更精确的分类结果。对于非度量数据，树分类器特别有用。

3 评价与比较分类器性能的定理和方法

以上我们总结和跟踪了各种常用和最新的文本分类的技术和算法。面对各种各样的方法，每个人可能都会产生疑问：到底有没有一种算法是“最好”的？

3.1 没有免费的午餐定理(NFL)

NFL 定理的结论是：对于上述问题及其相关问题的回答是：“没有”。此定理说明，不存在与问题领域无关的“最优”的学习算法。简言之，没有天生优越的分类器，分类本质上属于一门实验科学，必须深入地考察恰当的特征以及数据与算法的“匹配”程度。

3.2 偏差和方差

“偏差和方差”是两种度量，用于测量学习算法和给定分类问题的“匹配”程度。“偏差”代表的是期望值和真实值之间的差异，一个高的偏差意味着很差的匹配；而一个小的“方差”意味着函数的估计并不随训练集的波动而发生较大改变，所以，一个高的方差往往意味着较弱的匹配。“偏差-方差两难问题”是说如果增加一个学习算法的灵活性，能够自适应地“匹配”训练数据，那么它将具有更小的偏差，但会有更大的方差。对分类问题而言，偏差和方差之间存在某种非线性关系，且在分类中为了得到所需要的较小的推广误差，小的差要比小的偏差更为重要。可以利用增大训练集 n 的方法来降低偏差和方差。

3.3 最小描述长度原理(MDL)

“最小描述长度原理”给出了选择某种分类器而非另外一种的理由。假设我们要用训练集 D 来设计一个分类器，MDL 指出：我们必须使模型的算法复杂度及与该模型相适应的训练数据的描述长度的和最小，即 $K(h,D)=K(h)+k(D|h)$。若分类器可用二进制串的形式来表达，则 MDL 说明，最优的模型就是那个具有最短的模型描述和训练数据的模型。只有了解了问题的具体类型、先验分布等信息，才能运用 MDL 确定哪种形式的分类器将提供最好的性能。

3.4 分类器的评价与比较方法

至少有两个理由使得我们希望知道某个分类器对给定问题的推广程度。其一是评价该分类器性能，看它是否足够好，足够适合给定的问题；其二是为了比较不同的分类器，选出更好的设计方案。目前有许

(下转第37页)

用了分布式模型。在服务器数据库的页面记录中设置记录用户评价度的数据结构和对文档矩阵的支持。

3.1 客户端的实现

agent 是一种工作在客户端具有自主性，机动性，智能性，合作性的软件形式。客户端采用这种形式，以增强灵活性和智能性。

客户端以 Plug-in 的形式与浏览器系统结合。用户首次登录系统时，系统向用户询问用户基本信息并记录在本地文件中。用户提出查询时，服务器通过关键字查找数据，并生成相应的文档矩阵。服务器返回数据后，客户端计算查询与文档的相似度，根据用户信息计算用户评价度，对 URL 进行排序。在用户浏览过程中，记录用户点击动作和浏览的 URL，并把数据存入临时文件中。系统退出时把数据发送到服务器。服务器每天根据收集的数据计算网页的点击量。

3.2 客户端用户界面的设计

与传统的搜索引擎界面不同，为了用户浏览方便，网页被分三类显示：用户已经浏览的；用户不感兴趣的；用户未浏览的和新页面。这样用户在第二次查找时可以过滤记录，用户就可以更轻松地找到需要的页面。

4 实验结果及改进

我们通过互联网收集了十个主题的5000余个网页进行实验，并在 Windows Xp 和 Internet Explore6.0平台上以 com 组件的形式实现了一个模型系统。实验性的设定 $a_1=5, a_2=3, a_3=2, a_4=0$，经过对系统一个月的使用，通过对50个人的使用情况的统计得到以下数据，如表1。

表1

用户希望网页在第一条的比率	38.7%
用户希望网页在前十条的比率	83.4
用户希望网页在后十条的比率	0.3%

实验表明，系统能够体现用户的期望。虽然实验规模比较小，但仍表明系统有良好的性能。客户端排序网页的平均延时为0.05秒，所以用户不会有任何的感觉。

实验表明的问题：在计算网页的点击量时需要一定的计算量，串行计算点击量的方法使用时间较长。改进方案，可以并行计算点击量缩短计算点击量的时间。

结论 本文提出的个性化信息检索，可以根据用户的喜好，为用户提供有效的、个性化的信息。这样，不仅可以大大节省用户的时间，提高用户的效率，还可以充分地利用网络资源，提高网络的可用性。

参 考 文 献

1 Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: Bringing order to the Web. Stanford Digital Libraries Working Paper, 1998

2 Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. In: Proc. of the 7th Intl. World Wide Web Conf., Brisbane, Australia, 1998

3 Hinrich S. Automatic word sense discrimination. Computational Linguistics, 1998, 24(1): 97～123

4 Mitchell T M. 机器学习[M]. 北京：机械工业出版社，2003

（上接第26页）

多评价和比较分类器性能的方法，例如交叉验证法、刀切法、自助法、贝叶斯模型选择法等。每种方法都基于一些假定，这些假设与最终评价并无明显的关系。

4 独立于算法的机器学习

除上述的各种具体分类算法外，我们还要特别提到一种广义的算法——独立于算法的机器学习。它包括两层含义：首先，其数学基础不依赖于所采用的特定分类器和特定学习算法；其次，这些技术可与各种不同的学习算法组合使用，或者为它们提供应用指导，以改善分类器的总体性能。以上罗列的各种评价和比较分类器性能的方法（交叉验证法等）、偏差和方差法、重采样方法（Bagging、AdaBoost 法等）、组合分类器法等均属于独立于算法的机器学习[4]。

结束语 本文总结和跟踪了各种常用和最新的文本分类的技术和算法，分析了评价与比较分类器性能的定理和方法，简述了独立于算法的机器学习。今后的研究工作将着重围绕新算法的提出和旧算法的改进展开，并将结合计算语言学和本体论知识，将现有模型中常用的词空间拓展为概念空间来进行文本自动分类。

参 考 文 献

1 史忠植. 知识发现. 清华大学出版社，2002.1

2 庞剑锋，卜东波，白硕. 基于向量空间模型的文本自动分类系统的研究与实现. 计算机应用研究，2001(18)

3 Duda R O, Hart P E. Stork D G. 模式分类. 机械工业出版社，2003.9

4 黄萱菁，吴立德. 独立于语种的文本分类方法. In: 2000 Intl Conf. on Multilingual Information Processing, 2000

5 王永庆. 人工智能原理与方法. 西安交通大学出版社，1998.5