模式识别

文本分类综述

A Survey on Text Categorization

（中国科学院自动化研究所模式识别国家重点实验室，北京 100080） 靳小波

靳小波（1979－）
男，湖北随州人，中国科学院自动化研究所在读博士，研究方向为文本挖掘和模式识别理论。

1 文本分类的背景和意义

上世纪九十年代以来，因特网以惊人的速度发展起来，它容纳了海量的各种类型的数据和信息，包括文本、声音、图像等。文本数据与声音和图像数据相比，占用网络资源少，更容易上传和下载，这使得网络资源中的大部分是以文本（超文本）形式出现的。如何从这些浩瀚的文本中发现有价值的信息是信息处理的一大目标。基于机器学习的文本分类系统能够在给定的分类模型下，根据文本的内容自动对文本分门别类，从而更好地帮助人们组织文本、挖掘文本信息，因此得到日益广泛的关注，成为信息处理领域最重要的研究方向之一。

2 文本分类的研究历史和现状

文本分类的研究可以追溯到上世纪六十年代，早期的文本分类主要是基于知识工程（Knowledge Engineering），通过手工定义一些规则来对文本进行分类，这种方法费时费力，且必须对某一领域有足够的了解，才能写出合适的规则。

到上世纪九十年代，随着网上在线文本的大量涌现和机器学习的兴起，大规模的文本（包括网页）分类和检索重新引起研究者的兴趣。文本分类系统首先通过在预先分类好的文本集上训练，建立一个判别规则或分类器，从而对未知类别的新样本进行自动归类。大量的结果表明它的分类精度比得上专家手工分类的结果，并且它的学习不需要专家干预，能适用于任何领域的学习，使得它成为目前文本分类的主流方法。

1971 年，Rocchio 提出了在用户查询中不断通过用户的反馈来修正类权重向量，来构成简单的线性分类器。Mark van Uden、Mun 等给出了其他的一些修改权重的方法。1979 年，van Rijsbergen 对信息检索领域的研究做了系统的总结，里面关于信息检索的一些概念，如向量空间模型(Vector Space Model)和评估标准如准确率(Precision)、回召率(Recall)，后来被陆续地引入文本分类中，文中还重点地讨论了信息检索的概率模型，而后来的文本分类研究大多数是建立在概率模型的基础上。

1992 年，Lewis 在他的博士论文《Representation and Learning in Information Retrieval》中系统地介绍了文本分类系统实现方法的各个细节，并且在自己建立的数据集 Reuters22173（后来去掉一些重复的文本修订为 Reuters21578 数据集）上进行了测试。这篇博士论文是文本分类领域的经典之作。后来的研究者在特征的降维和分类器的设计方面作了大量的工作，Yiming Yang 对各种特征选择方法，包括信息增益(Information Gain)、互信息(Mutual Information)、χ^2统计量等，从实验上进行了分析和比较。她在 1997 年还对文献上报告的几乎所有的文本分类方法进行了一次大阅兵，在公开数据集 Reuters21578 和 OHSUMED 上比较了各个分类器的性能，对后来的研究起到了重要的参考作用。

1995 年，Vipnik 基于统计理论提出了支持矢量机(Support Vector Machine)方法，基本思想是寻找最优的高维分类超平面。由于它以成熟的小样本统计理论作为基石，因而在机器学习领域受到广泛的重视。Thorsten Joachims 第一次将线性核函数的支持矢量机用于文本分类，与传统的算法相比，支持矢量机在分类性能上有了非常大的提高，并且在不同的数据集上显示了算法的鲁棒性。至今，支持矢量机的理论和应用仍是研究的热点。

在支持矢量机出现的同时，1995 年及其后，以 Yoav Freund 和 Robert E. Schapire 发表的关于 AdaBoost 的论文为标

志，机器学习算法的研究出现了另一个高峰。Robert E. Schapire 从理论和试验上给出 AdaBoost 算法框架的合理性。其后的研究者在这个框架下给出了许多的类似的 Boosting 算法，比较有代表性的有 Real AdaBoost，Gentle Boost，LogitBoost 等。这些 Boosting 算法均已被应用到文本分类的研究中，并且取得和支持矢量机一样好的效果。

相比于英文文本分类，中文文本分类的一个重要的差别在于预处理阶段：中文文本的读取需要分词，不像英文文本的单词那样有空格来区分。从简单的查词典的方法，到后来的基于统计语言模型的分词方法，中文分词的技术已趋于成熟。比较有影响力的当属中国科学院计算所开发的汉语词法分析系统 ICTCLAS，现已公开发布供中文文本分类的研究使用。

在很长一段时间内，中文文本分类的研究没有公开的数据集，使得分类算法难以比较。现在一般采用的中文测试集有：北京大学建立的人民日报语料库、清华大学建立的现代汉语语料库等。

其实一旦经过预处理将中文文本变成了样本矢量的数据矩阵，那么随后的文本分类过程和英文文本分类相同，也就是随后的文本分类过程独立于语种。因此，当前的中文文本分类主要集中在如何利用中文本身的一些特征来更好地表示文本样本。

总而言之，尽管机器学习理论对于文本分类的研究起了不可低估的作用，在这之前文本分类的研究曾一度处于低潮，但是文本分类的实际应用和它自身的固有的特性给机器学习提出新的挑战，这使得文本分类的研究仍是信息处理领域一个开放的、重要的研究方向。

3 文本分类系统概述

文本分类的任务是在给定的分类体系下，对一未知类别标号的文本，根据其内容进行归类，它可以归为多类，也可以不属于任何类（对给定的类集合而言）。下面来看一个简单的文本分类系统是如何工作的，图 1 是整个文本分类系统的简单的架构图，其中实线表示的是分类器建立过程中的数据流，虚线表示的是分类器测试过程中的数据流。

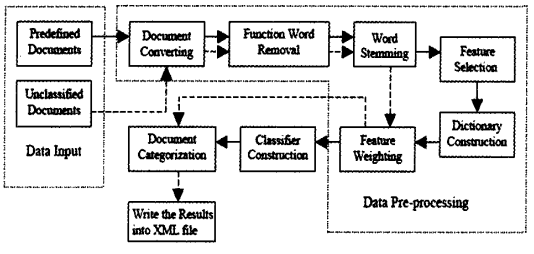

图 1 文本分类系统架构图

3.1 文本的预处理

机器学习算法要求将文本的单词序列转换为数值表示的一个矢量。首先假定每个文本是单词的无序集合(Set)，不考虑单词之间的相互位置，这样一个单词代表哪一维是随机的，文本不再是一个序列，而是属性—值的集合。在给出文本中每个属性（单词）的值之前，需要对文本集进行预处理：

- 去掉一些低频词，比如某些单词只在一两个文本中出现过，这些词留在集合中会导致大部分文本样本的该属性值为 0；
- 去掉停止词(Stop word)，一般这种词几乎不携带任何信息，如中文中“的”，“啊”，英文中的“the”，“a”；
- 去掉一些标记信息，这主要针对网页文本或其他的标记语言文本，比如SMGL(Standard Generalized Markup Language)标记的文本；
- 去掉单词的前后缀（Word Stemming），该步骤主要用在英文文本分类中，将同词根的单词映射到一个特征词属性，可以大大地减少特征词的个数。

经过以上处理后，就可以用特征词权重函数来度量文本中每个单词的属性值。常用的权重函数是 TF-IDF(Term frequency–Inverse Document frequency)。文本中某个特征词的属性值的计算与下列因素有关。一个是特征词在该文本中出现的频率，特征词出现的频率越高，则权重值越大。另一个是特征词的文档频率，即包含该特征词的文本的个数，含有特征词的文本数目越多，则该特征词越普通，它对类别的区分作用越小，给它分配的权重也就越小。最后，考虑到文本的长短不一，要将文本长度归一化，这样同一个特征词在不同长度的文本中的权重才具有可比性。基于以上因素考虑，特征词的权重函数经常采用如下方式计算：

$$a_{ij}=\frac{tf(w_j,d_i)\times\log(\frac{N}{df(w_j)})}{\sqrt{\sum_{k=1}^{M}\left[tf(w_k,d_i)\times\log(\frac{N}{df(w_k)})\right]^2}} \quad (1)$$

其中 tf(ω j，di)表示特征词 ω j 在文本 di 中出现的频率，df(ω j)表示了特征词的文档频率，N 表示文档总数，M 表示特征总数，aij 表示了第 i 个样本矢量的第 j 个分量值。这样，文本的每个特征词都对应一个属性值，从而实现了从特征词到数值的转换。

3.2 特征的降维方法

经过上一节的预处理之后，特征矢量的维数仍然很高，需要在尽量不损失分类信息的情况下生成一个低维的特征矢量。一种方法是从原有的特征词集合中挑出最有效的一些特征构成新的特征矢量，这个过程叫特征选择(Feature Selection)；还有一种方法是将高维特征矢量映射（一般为线性映射）到低维空间，这个过程叫特征提取(Feature Extraction)。

特征选择中一种重要的特征词度量方法是信息增益(Information Gain)。对于一个特征词，假定在不知道该特征词的条件下，所有的类有一个平均无条件信息熵；在知道这个特征词之后，所有的类有一个平均条件（条件就是知道该特征词）

信息熵，那么这两个信息熵的差就表示了该特征词所携带的信息量，即信息增益。用公式表述如下：

$$\begin{aligned} &I(T,w)=E(T)-E(T\mid w)\\ &=-\sum_{c\in C}\Pr(T(d)=c)\log\Pr(T(d)=c)\\ &+\sum_{c\in C}\Pr(T(d)=c,w=0)\log\Pr(T(d)=c\mid w=0)\\ &+\sum_{c\in C}\Pr(T(d)=c,w=1)\log\Pr(T(d)=c\mid w=1) \end{aligned} \tag{2}$$

其中E(T)表示所有类的平均信息熵，E(T | w)表示在知道特征词w是否在文本中出现的条件下所有类的平均条件信息熵。Pr(T(d)=c, w=0)表示没有特征词w出现的文本在类c中出现的概率。其他的定义类似。将特征词按信息增益值由大到小排列，选取前k个特征词构成最终的特征词集合，这样每个样本被转化为新的矢量，其维数为k。

3.3 分类器设计和评估

根据不同的应用，文本分类可以分为单标号分类（一个文本恰好属于一个类）和多标号分类（一个文本可以属于多个类或者不属于任何类）。在单标号条件下，研究得比较多的是两类问题：对于某个类ci，一个文本或者属于类ci或者不属于类ci。对于类标号集合有m个类的多标号问题，可以将它当作m个相互立的两类单标号问题来处理。为了简单起见，这一小节列举的分类器假定每个文本恰好属于一个类标号（即单标号分类）。

在建立分类器之前，首先得给分类器提供两个预先分类好的文档样本集，一个样本集称作训练集(Training Set)，在训练的过程中分类器要用到每个训练文本的类标号信息。另外一个样本集称作测试集(Test Set)，在评估（或测试）分类器的性能时会用到测试样本的类标号信息。有时为调整分类器的参数，还需要从训练集中分离一部分样本来执行参数的优化。如果仅仅给定一个样本集，可以人为地划分训练集和测试集。一般来说，训练集的样本数要远远大于测试集的样本数。

最近邻分类器是机器学习领域和文本分类领域研究得比较多的一个分类器。该算法的基本思想是：在给定新文本后，考虑在训练文本集中与新文本距离最近（最相似）的k篇文本，根据这k篇文本所属的类别采用距离加权的方式来计算每个类别的分值(Categorization Status Value)。将新样本分到分值最大（或者说与之距离最小）的类中。

另一种常用的分类器是Naïve Bayes分类器，它假定文本的各个特征属性是相互独立的，据此计算各个类的后验概率，最后将新文本分到后验概率最大的类中。根据贝叶斯公式，在给定文本d=(x1, x2……xM)(为简单起见，省去d的下标，xi表示文本d的第i个属性，M表示特征矢量的维数）的条件下，类ci(i=1, 2……l)的后验概率（即文档d属于类ci的概率）为：

$$\begin{aligned} &p(c_i\mid d)=\frac{p(c_i)p(x_1,x_2,\cdots x_n\mid c_i)}{p(x_1,x_2,\cdots x_n)}\\ &=\frac{p(c_i)p(x_1\mid c_i)p(x_2\mid c_i)\cdots p(x_n\mid c_i)}{p(x_1,x_2,\cdots x_n)}\\ &\propto p(c_i)p(x_1\mid c_i)p(x_2\mid c_i)\cdots p(x_n\mid c_i) \end{aligned} \tag{3}$$

由于对于同一个文本，上式的分母对所有的类相同，所以可以作上述简化，(3)式中类ci(i=1, 2……l)的先验概率和特征词的条件概率分别按下式计算（#用来表示集合的大小）：

$$p(c_i)=\frac{\#\{d\in c_i\}+1}{\#\{d\in C\}+l} \tag{4}$$

$$p(x_i\mid c_j)=\frac{\#\{x_i\in c_j\}+1}{\#\{x\in c_j\}+M} \tag{5}$$

当然还有其他的一些分类器（如第二节提到的支持矢量机和Boost方法）可以用于文本分类，具体细节这里就不做详细介绍了。

在分类器的评估阶段，测试集中每个文本按相同的预处理和特征转换步骤形成新的样本矢量，输入分类器进行测试。对于单标号而言，简单地统计被正确分类的测试样本数占测试样本集的比例，即得出分类系统的分类精度(Accuracy)。但是对于多标号而言，用它不足以衡量多标号分类器的分类性能。比如，假定每个类只有少量的正样本，有大量的负样本（不属于该类的样本统称为负样本），这种情况下，分类器如果总是将任意样本判为负样本，得到的分类精度也很高，但是不能认为这种决策策略是最好的。这需要从信息检索领域引入准确率和回召率来度量。

4 文本分类的应用状况和应用前景

信息检索(Information Retrieval)里的许多任务都可以归结为文本分类问题，包括搜索引擎对网页的相关性排序、邮件的过滤、文档的组织。一些自然语言处理任务，像词义消歧、词性标注等等，也可以转换为文本分类的问题。

采用Naïve Bayes实现的垃圾邮件过滤系统已经进入企业产品阶段；Google公司和微软公司在网页检索方面已不再满足简单的单词匹配来给出与用户查询相关的网页，越来越多地将信息检索和文本分类的技术引入以更好地理解用户的搜索需求，提供给用户更优的信息处理服务。毫无疑问，文本分类技术作为上述应用的基础理论，必然会推动这些应用领域的发展。

网络安全问题越来越多地受到政府和大公司的重视，根据以往用户访问的历史记录，对不同用户进行分类，从而决定是否允许用户访问，这在文本分类中是一个典型的两类分类问题。

尽管文本分类在上述领域的应用取得一定的成功，但是相比于一般的机器学习应用问题，文本分类问题仍然存在如下的挑战：

(1) 文本分类的矢量矩阵一般是成千上万的稀疏矩阵，对于这样超高维数矩阵，文本分类器必须能够有效地存取和运算；

(2) 文本的特征词集合中存在多义词、同义词现象，还包含大量的噪音，中文文本中还需要恰当地分词等等，如何从文本中形成最有效的特征矢量成为文本分类中需要解决的关键问题。因为如果提取的特征不可靠，哪怕采用再好的分类器，其性能也会大打折扣；

(3) 文本分类的研究和它的实际应用在某种程度上存在脱节，比如对于信息检索来说，文本分类中采用的数据量对比

于实际需要搜索的网页的数量，仍是小巫见大巫，对于小的数据集上比较高效的算法对于大的数据集不一定成立。最近发布的新的文本分类公开测试集RCV1已经考虑到这一点，总的文档个数已达到上百万（旧的数据集Reuters21578是21 578个文档）。目前文本分类的研究中考虑更多的是算法的有效性（比如正确率、回召率、准确率），算法的运行效率也是未来文本分类研究需要考虑的一个因素。

5 总结

本文主要介绍了文本分类研究的历史和现状，简要地阐述了文本分类系统的工作流程，并给出了它的一些实际应用，讨论了文本分类的研究所遇到的一些挑战，希望能给对该领域感兴趣的读者一些有益的参考。

参考文献

[1] 许洪波, 程学旗, 王斌, 骆卫华. 文本挖掘与机器学习[N]. 信息技术快报, Vol.3, No.2, 2005.

[2] D.D. Lewis, Challenges in Machine Learning for Text Classification. The 9th Annual Conference on Computational Learning Theory. Italy, 1996.

[3] D.D. Lewis, Representation and Learning in Information Retrieval, Doctoral Thesis, 1992.

[4] F.Debole, F.Sebastiani, An Analysis of the Relative Hardness of Reuters-21578 Subsets. Journal of the American Society for Information Science and Technology, Vol. 56, No.6, 2005.

[5] F. Sebastiani, Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol. 34, No. 1, March 2002, pp.1-47.

[6] K. Aas, L. Eikvil, Text Categorisation: A Survey. http://www.nr.no/files/samba/bamg/tm_survey.ps.

[7] R.E. Schapire, Y. Singer. Improved Boosting Algorithms Using Confidence-Rated Predictions. Machine Learning, Vol.37, No.3, pp.297-336, 1999.

[8] （美）Tom M. Mitchell 著, 曾华军, 张银奎, 等译. 机器学习[M]. 机械工业出版社, 2002.

[9] Y. Yang, J. Pedersen, A comparative study on feature election in text categorization. International conference on Machine Learning (ICML), 1997. http://citeseer.ist.psu.edu/yang97comparative.html.

[10] Y. Yang, An Evaluation of Statistical Approaches to Text Categorization. Information Retrieval, Vol.1, No.1-2, 1999.

[11] C.J. van Rijsbergen, Information Retrieval. University of Glasgow, 1979.

[12] （美）Vipnik 著, 张学工译. 统计学习理论的本质[M]. 清华大学出版社, 2000.

编号：060304