自动文本分类特征选择方法研究

张海龙， 王莲芝
（中国农业大学 信息与电气工程学院，北京 100083）

摘 要：文本分类是指根据文本的内容将大量的文本归到一个或多个类别的过程，文本表示技术是文本分类的核心技术之一，而特征选择又是文本表示技术的关键技术之一，对分类效果至关重要。文本特征选择是最大程度地识别和去除冗余信息，提高训练数据集质量的过程。对文本分类的特征选择方法，包括信息增益、互信息、X^2统计量、文档频率、低损降维和频率差法等做了详细介绍、分析、比较研究。

关键词：文本分类； 特征选择； 信息增益； 互信息； X^2统计量法； 文档频率； 低损降维； 频率差
中图法分类号：TP391 **文献标识码**：A **文章编号**：1000-7024 (2006) 20-3838-04

Automatic text categorization feature selection methods research

ZHANG Hai-long, WANG Lian-zhi
(College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China)

Abstract: Text categorization is a task of classifying text documents into predefined set of categories based on their content, text representation is one of kernel technology of text categorization, and feature selection is one of key technology of text representation, it is very important to text categorization effect. Text feature selection is a process of recognizing and deleting redundant information and enhancing training documents cluster quality. The text feature selection methods are introduced, analysed and researched, including information gain, mutual information, X^2 statistics, document frequency, low loss dimensionality reduction, relative frequency difference.

Key words: text categorization; feature selection; information gain; mutual information; X^2 statistics; document frequency; low loss dimensionality reduction; relative frequency difference

0 引 言

文本分类是指根据文本的内容或属性，将大量的文本归到一个或多个类别的过程，文本分类包括文本归类和文本聚类。

文本归类技术通过分析待分类对象，提取待分类对象的特征，比较待分类对象和系统预定义类别对象的特征，将待分类对象划归为特征最近的一类，并赋予相应的分类号。

文本聚类在对文本进行自动标引的基础上，构造文本的形式化表示——文本特征向量，然后根据一定的聚类方法，计算出文本与文本之间的相似度，并把相似度较高的文本集中在一起，形成一个个的文本类。但由于现有的聚类算法、时间复杂度、次序独立性、重叠度等不满足要求，并且由于存在文本的多样性和复杂性，聚类算法大多限于理论上的探讨，很少投入实际应用。因此目前的文本分类研究仍以文本归类为主，在本文后所指的文本分类主要指的是文本归类。

诱导学习的特征选择是来自于机器学习的中的概念，特征选择也可以看作是求最优解的问题。根据 John 等人的研究，在机器学习中有两类特征选择方法：筛选(filtering)方法和复选(wrapper)方法。通过特征筛选方式得到的最优特征子集仅依赖与训练样本的统计特性和判别函数本身，而与分类器所采用的学习算法无关，相反，特征复选的结果不仅与训练样本的统计特性有关，而且与测试样本的统计特性和学习算法密切相关，因而比特征筛选方法要复杂很多。所有在解决实际的模式识别问题如文本分类问题时，人们往往选择使用特征筛选方式。

机器学习中的多数特征选择方法并不能适应大量特征的情况。文本学习的特征选择方法需要处理训练文本集的每个词，产生大量的特征。所以文本分类中的特征选择方法是有别于机器学习特征选择方法的。

1 文本分类的核心技术

文本自动分类核心技术包括文本表示技术和文本分类算法技术。

1.1 文本表示技术

文本表示技术是文本自动分类的基础。文本自动分类系统是针对文本数据来进行设计的，而未经过处理的文本型数据显然不能直接用于分类，因此必须将文本表示成能参与分类计算的数据类型。主要工作包括分词和特征选择。

收稿日期：2005-08-19。
作者简介：张海龙（1976－），男，辽宁盘锦人，硕士研究生，研究方向为数据库与数据挖掘技术； 王莲芝（1956－），女，副教授，研究方向为人工智能与智能信息处理技术。

1.1.1 分　词

针对文本，目前的一般做法是先对文本进行分词，然后进行词频统计，根据词频统计结果以及停用词表，采用一定的策略对汉语词语及短语进行过滤筛选，形成文本特征向量，同时对预定义类别中的文本也作如上操作，形成类别特征向量，用于以后的文本分类。

目前，在信息处理方向上，文本的表示主要采用向量空间模型(vector space model，VSM)。向量空间模型的基本思想是以向量来表示文本 $(W_1,W_2,\cdots,W_n)$，其中 W_i 为第 i 个特征项的权重，普遍认为选取词作为特征项要优于字和词组，因此，要将文本表示为向量空间中的一个向量，就首先要将文本分词，由这些词的频率作为向量的维数来表示文本，词频分为绝对词频和相对词频，绝对词频，即使用词在文本中出现的频率表示文本，相对词频为归一化的词频，其计算方法主要运用 TF-IDF 公式，目前存在多种 TF-IDF 公式，一种比较普遍的 TF-IDF 公式为

$$W(t,\vec{d})=\frac{tf(t,\vec{d})\times\log(N/n_t+0.01)}{\sqrt{\sum_{t\in\vec{d}}[tf(t,\vec{d})\times\log(N/n_t+0.01)]^2}} \quad (1)$$

其中：$W(t,\vec{d})$——词 t 在文本 $\vec{d}$ 中的权重，$tf(t,\vec{d})$——词 t 在文本 $\vec{d}$ 中的词频，N——训练文本的总数，n_t——训练文本集中出现 t 的文本数，分母为归一化因子。

1.1.2 特征选择

这是本文要重点讨论的问题。文本分类的错误率主要是由于特征词没有适当选取和表示造成的，不管分类算法如何改进都无法降低这个错误率，这是进行特征选择的原因之一；构成文本的词汇数量是相当大的，因此，表示文本的向量空间的维数也相当大，可以达到几万维，为了提高程序的效率，提高运行速度，因此需要进行维数压缩的工作，这是进行特载选择第 2 个原因。

文本特征特征选择的方法如图 1 所示。

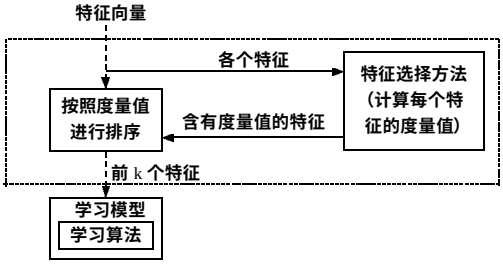

图 1　文本分类特征选择方法

图 1 中的 k 就是特征选择中抽取特征项的数量(阈值)，具体 k 值取多少，目前无很好的解决方法，一般采用先定初始值，然后根据实验测试和统计结果确定最佳值，一般初始值定在几千左右。

1.2 分类算法技术

分类算法是中文文本自动分类的关键。文本自动分类的核心内容为分类函数的选取，即根据一定的算法，比较文本特征向量与特征向量的相似度，把文本归类于最相似的那个预定义类别。

常见的分类算法有简单向量距离、朴素贝叶斯、支持向量机、神经网络和 K 最近邻居算法等，但不是本文论述的重点，所以不在此详述。

2 特征选择的方法

为了提高分类精度，对于每一类，我们应去除那些表现力不强的词汇，筛选出针对该类的特征项集合，存在多种筛选特征项的算法，如：信息增益、互信息、X^2 统计量、文档频率、低损降维法和频率差法。每一种算法都有着自己的优点和不足，如何更加全面、有效的结合这些方法，值得我们思考。

下面，为了便于我们对文本分类中常用的特征选择方法进行讨论，先引入如下符合：

A—属于类别 c 且包含特征 x 的训练文本个数；

B —不属于类别 c(或属于类别 $\bar{c}$)且包含特征 x 的训练文本个数；

C—属于类别 c 且不包含特征 x 的训练文本个数；

D—不属于类别 c 且不包含特征 x 的训练文本个数；

M—属于类别 c 文本个数；

N—训练文本总数。

显然有 $A+C=M$ 和 $A+B+C+D=N$。

2.1 信息增益法

信息增益(information gain，IG)是机器学习中的概念，用在决策树中来计算特征的权值，信息增益被定义为类特征向量的平均值。

特征 x 与类 c 的信息增益 $IG(x,c)$ 计算公式如下

$$IGain(x)=IG(c,x)=H(c)-H(c|x)=H(x)-H(x|c)=IG(x,c) \quad (2)$$

类向量的熵如下

$$H(c)=\Sigma_i P(c_i)\log P(c_i)$$

条件熵如下

$$H(c|x)=-\Sigma_j P(x_j)\Sigma_i P(c_i|x_j)\log P(c_i|x_j)$$

将上式中各个事件的概率用其相应的频率来代替(例如，概率 $p(x,c)$ 用属于类别 c 且包含特征 x 的训练文本个数占训练文本总数的比率来代替)，条件概率 $P(c|x)=P(x|c)/P(x)$，信息增益 $IG(x,c)$ 可以用下式来近似计算

$$IG(c,x)\approx\frac{A+C}{N}\log\frac{A+C}{N}+\frac{1}{N}[(A+D)\cdot\frac{A}{M}\cdot\frac{1}{A+B}\cdot\log(\frac{A}{M}\cdot\frac{1}{A+B})] \quad (3)$$

特征 x 与类别 c 的信息增益越大，说明特征 x 中包含的与类别 c 有关的鉴别信息就越多。

2.2 互信息法

互信息(mutual information，MI)是信息论中的概念，它用于度量一个消息中两个信号之间的相互依赖程度。在特征选择领域中人们经常利用它来计算特征 x 与类别 c 之间依赖程度，将特征 x 与各个类的互信息融合起来作为特征的权重。特征 x 与类 c 的互信息 $MI(x,c)$ 计算公式如下

$$MI(x,c)=p(x,c)\cdot\log\frac{p(x,c)}{p(x)\cdot p(c)}+p(x,\bar{c})\cdot\log\frac{p(x,\bar{c})}{p(x)\cdot p(\bar{c})}+p(\bar{x},c)\cdot\log\frac{p(\bar{x},c)}{p(\bar{x})\cdot p(c)}+p(\bar{x},\bar{c})\cdot\log\frac{p(\bar{x},\bar{c})}{p(\bar{x})\cdot p(\bar{c})} \quad (4)$$

将上式中各个事件的概率用其相应的频率来代替，互信息 $MI(x,c)$ 可以用下式来近似计算

$$MI(x,c)\approx\frac{1}{N}\cdot[A\cdot\log A+B\cdot\log B+C\cdot\log C+D\cdot\log D-(A+B)\cdot\log(A+B)-(C+D)\cdot\log(C+D)-(A+C)\cdot\log(A+C)-(B+D)\cdot\log(B+D)+N\cdot\log N] \quad (5)$$

由于 N 是一个常数，对于给定的类别c，$A+C(=M)$和$B+D(=N-M)$也是常数，因此上式可以简化为

$$MI(x,c)\approx A\cdot\log(A)+B\cdot\log(B)+C\cdot\log(C)+D\cdot\log(D)-(A+B)\cdot\log(A+B)-(C+D)\cdot\log(C+D) \quad (6)$$

特征x与类别c的互信息越大，说明特征x中包含的与类别c有关的鉴别信息就越多。

2.3 X^2统计量法

X^2统计量(chi-square statistic，CHI)特征选择方法又被称作开方拟合检验(CHI，X^2-test)，这个概念来自列联表检验(contingency table test)，它可以用来衡量特征x与类别c之间的统计相关性。其计算公式如下

$$CHI(x,c)=\frac{N\cdot[p(x,c)\cdot p(\bar{x},\bar{c})-p(x,\bar{c})\cdot p(\bar{x},c)]^2}{p(x)\cdot p(\bar{x})\cdot p(c)\cdot p(\bar{c})} \quad (7)$$

用各个事件的频率代替其相应的概率，X^2统计量$CHI(x,c)$可以用下式来近似计算

$$CHI(x,c)\approx\frac{N\cdot(A\cdot D-B\cdot C)^2}{(A+C)\cdot(B+D)\cdot(A+B)\cdot(C+D)} \quad (8)$$

考虑到N，$A+C$和$B+D$均是常数，上式可以进一步简化为

$$CHI(x,c)\approx\frac{(A\cdot D-B\cdot C)^2}{(A+B)\cdot(C+D)} \quad (9)$$

当特征x与类别c相互独立时，$CHI(x,c)=0$，此时特征x不包含与类别c有关的鉴别信息。$CHI(x,c)$的值就越大，此时特征x包含的与类别c有关的鉴别信息就越多。

Hwee Tou Ng 等人认为仅使用本类别文本中出现频率高的词语作为特征项，能取得更好得分类效果。诚然，在其它类中出现频率高得词语能够对人们判别文本不属于类别C有很好得提示作用，但是这个作用对分类效果得影响并不明显。于是他们对X^2统计量进行了改进，得到Correlation Coefficient法，其数学表达式为

$$CC(x,c)=\frac{\sqrt{N}\cdot[P(x,c)\cdot P(\bar{x},\bar{c})-P(x,\bar{c})\cdot P(\bar{x},c)]}{\sqrt{P(x)\cdot P(\bar{x})\cdot P(c)\cdot P(\bar{c})}} \quad (10)$$

Luigi Galavotti 等人在此基础上，提出了一种更为简化的X^2统计量(simplified chi-square statistic，SCHI)

$$SCHI(x,c)=p(x,c)\cdot P(\bar{x},\bar{c})-P(x,\bar{c})\cdot P(\bar{x},c) \quad (11)$$

用各个事件的频率代替其相应的概率，简化的X^2统计量$SCHI(x,c)$可以用下式来近似计算

$$SCHI(x,c)\approx A\cdot D-B\cdot C \quad (12)$$

2.4 文档频率法

某个特征x的文档频率(document frequency，DF)一般定义为包含该特征的训练文档个数，即$DF(x)=A+B$。因此，文档频率通常用于全局特征选择。如果将文档频率用于局部特征选择，也可以将其定义为：$DF(x)=A$，即属于类别c且包含特征x的训练文档个数。

在运用文档频率进行特征选择时，首先要计算各个特征词的文档频率，然后将那些文档频率高于某个阈值的所有特征词挑选出来。其依据为：文档频率低的特征不包含对分类有用的鉴别信息，因而对分类结果没有什么影响。

文档频率是最简单的一种文本特征选择方法，同时也是最有效的文本特征选择方法之一。

2.5 低损降维法

基于低损降维(low loss dimensionality reduction，LLDR)的特征相关性计算公式为

$$LLDR(x,c)=\max\{p(x|c),p(x|\bar{c})\} \quad (13)$$

它可以用下式进行计算

$$LLDR(x,c)=\max\{\frac{A}{M},\frac{B}{N-M}\} \quad (14)$$

$LLDR(x,c)$的值越小，特征x包含的与类别c相关的信息就越少。

需要指出的是，那些几乎在所有正例训练样本和几乎所有反例训练样本中出现的特征也近似是冗余特征。我们之所以没有在公式(13)中考虑这种情况，是因为文本分类中很少有这种特征存在。

2.6 频率差法

一个特征成为某个类别的代表性特征，如果该类别的大多数样本均具有这一特征；一个特征称为某个类别的鉴别性特征，如果其它类别的大多数样本均不具有这一特征。显然，我们应该选择那些即具有代表性又具有鉴别性的特征。

条件概率$p(x|c)$可以作为特征x的代表性度量，而$-p(x|\bar{c})$可以作为特征x的鉴别性度量。因而可以用频率差(relative frequency difference，RFD)作为特征的相关性度量

$$RFD(x,c)=(p(x|c)-p(x|\bar{c}))^2 \quad (15)$$

它可以近似为

$$RFD(x,c)\approx(\frac{A}{M}-\frac{B}{N-M})^2=\frac{(A\cdot D-B\cdot C)^2}{M^2\cdot(N-M)^2} \quad (16)$$

由于N、$N-M$均为常数，上式可以进一步简化为

$$RFD(x,c)\approx(A\cdot D-B\cdot C)^2 \quad (17)$$

$RFD(x,c)$越大，特征x包含的与类别c有关的鉴别信息就越多。

3 不同特征选择方法的比较

3.1 实验设计

实验数据集为国际标准语料库 Reuters-21578 和 20 newsgroups，在 Reuters-21578 数据集上，按照 ModApte 方式将其划分成训练集和测试集两部分，其中训练集含有 9 603 篇文档，测试集含有 3 299 篇文档。训练样本的类别数为 118，测试演变的类别数为 93，至少有 1 个训练样本和 1 个测试样本的类别数为 90。在对某个特定的类别进行特征选择时，属于该类别的所有训练样本为正例样本，所有其它训练样本为反例样本。文本特征词典根据 9 603 篇训练文档基于标识符“ ”和“ ”之间的文本生成，不区别标题和正文，原始特征维数高达 27 942。

在 20 Newsgroups 数据集上，取每个新闻组的前 800 条消息构成训练样本集，剩下的样本构成测试数据集。文本特征词典根据 16 000 篇训练文档的正文(忽略所有的报头)生成。原始特征维数高达 98 225。

考虑到用于文本分类器有很多分类算法，如朴素贝叶斯、K最近邻、神经网络以及支持向量机等，其中以支持向量机最为有效。因而本实验选用支持向量机作为基准分类器。这里，我们用 Junshui Ma 等开发的 Matlab 支持向量机工具箱中的函数 LinearSVC。

3.2 实验结果

6 种文本特征选择方法在数据集 Reuters-21578 上的实验结果如表 1 所示。

6 种文本特征选择方法在数据集 20 Newsgroups 上的实验结果如表 2 所示。

另外，表 3 列出了计算 Reuters-21578 数据集中 90 个类别的 IG、MI、CHI、DF、LLDR 依据 RED 值所需要的时间。

3.3 基本结论

根据理论分析和一定的对比实验，我们将 6 种主要文本特征选择方法的特点归纳如表 4 所示。

从表 4 我们不难看出，LLDR(低损降维)和 RED(频率差法)比其它的特征选择方法更适合用于文本特征的提取。

4 结束语

本文探讨了文本分类系统的关键技术之一的特征选择方法，在使用基于支持向量机分类算法的基础上比较和分析了 6 种特征选择方法，为自动文本分类选择合适的特征选择方法做了充分的研究。

将来还将继续在层次分类体系中进行文本分类系统的进一步研究。

参考文献:

[1] Dunija Mladenic,Marko Grobelnik.Feature selection on hierarchy of web documents[J].Decision Support Systems, 2003,35:45-87.

[2] Zhi-hua,Zhou KaiJiang,Ming Li.Multi-instance learning based web mining[J].Applied Intelligence, 2005,22:135-147.

[3] 李粤, 李星, 刘辉, 等. 一种改进的文本网页分类特征选择方法[J].计算机应用, 2004,24(7):119-121.

[4] 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[J].计算机应用研究, 2001,18(9):23-26.

[5] 卜东波. 聚类/分类理论研究及其在大规模文本挖掘中的应用[D].北京:中科院计算技术研究所, 2000.

[6] 刘敏. 基于支持向量机的中文文本自动分类系统的设计和实现[D].杭州:浙江大学, 2001.

[7] 黄萱菁. 大规模中文文本的检索、分类与摘要研究[D].上海:复旦大学, 1998.

[8] 宋枫溪. 自动文本分类若干基本问题研究[D].南京:南京理工大学, 2004.

[9] 寇莎莎, 魏振军. 自动文本分类中权值公式的改进[J]. 计算机工程与设计, 2005,26(6):1616-1618.

表 1 6 种文本特征选择方法对应的 Reuters90 类微平均 BEP 值

特征选择方法	选择的特征个数									
	200	400	600	800	1 000	2 000	3 000	4 000	5 000	6 000
IG	0.861	0.870	0.865	0.868	0.871	0.873	0.880	0.876	0.877	0.875
MI	0.863	0.869	0.870	0.871	0.873	0.876	0.879	0.879	0.878	0.877
CHI	0.846	0.859	0.865	0.868	0.870	0.876	0.878	0.877	0.876	0.876
DF	0.669	0.725	0.761	0.786	0.800	0.854	0.867	0.873	0.875	0.877
LLDR	0.865	0.871	0.877	0.878	0.878	0.879	0.881	0.879	0.879	0.880
RFD	0.868	0.873	0.876	0.875	0.878	0.879	0.880	0.880	0.880	0.880

表 2 6 种文本特征选择方法对应的 20 Newsgroup 的识别率

特征选择方法	选择的特征个数									
	800	1 000	2 000	3 000	4 000	6 000	8 000	10 000	20 000	30 000
IG	0.702	0.712	0.750	0.763	0.776	0.785	0.789	0.805	0.820	0.824
MI	0.707	0.715	0.745	0.766	0.777	0.790	0.800	0.809	0.821	0.827
CHI	0.705	0.723	0.754	0.763	0.773	0.786	0.797	0.803	0.824	0.827
DF	0.579	0.625	0.701	0.745	0.759	0.781	0.799	0.809	0.819	0.826
LLDR	0.616	0.637	0.713	0.754	0.766	0.781	0.801	0.807	0.821	0.826
RFD	0.643	0.664	0.728	0.759	0.771	0.790	0.801	0.810	0.821	0.826

表 3 计算 IG、MI、CHI、DF、LLDR 依据 RED 值所需要的时间

特征选择度量	IG	MI	CHI	DF	LLDR	RFD
时间(秒)	6.453	6.336	2.203	0.897	0.976	1.305

表 4 文本特征选择方法的比较

特征选择方法	IG	MI	CHI	DF	LLDR	RFD
是否有理论基础	是	是	是	否	是	是
含义是否容易理解	否	否	否	是	是	是
计算效率如何	非常低	非常低	低	高	高	较高
降维效果	很好	很好	很好	好	很好	很好