关于文本分类中特征降维方式的研究

伍建军，康耀红

（海南大学 信息科学技术学院， 海南 海口 570228）

摘 要：文本分类的一个关键点就是如何对高维的特征集进行降维。而特征降维的常用方法就是特征选择、特征抽取等。首先介绍了几种常见的特征选择和特征抽取方法，并对这些特征降维方法的优缺点进行了分析。最后结合 K-近邻分类算法对四种特征选择方法（文档频率、互信息、CHI 统计法、相关系数法）进行了分类测试，同时通过测试分析，提出了一些改进的、可行的互信息评价函数，实验结果表明,采用这种方法,在同等训练语料的情况下,分类效果比基于传统的互信息评价函数的分类效果要好。

关键词：文本分类；特征降维；特征选择；互信息；

Research about feature dimension reduction in text categorization

WU Jian-jun, KANG Yao-hong

（College of Information Science & Technology, Hainan University, Hainan, Haikou 570228,China）

Abstract: The key point of text categorization is how to reduce the high dimension of the feature vectors. Feature selection and feature extraction is the main methods of feature reduction. This paper introduces five methods of feature selection and feature extraction. And then k-nearest neighbor is selected as the evaluating classifier to compare the performance of the four feature selection methods(including Document Frequency、Mutual Information、Chi-square、correlation coefficient) in text categorization. From the test result, a new improved method of FS is presented based on Mutual Information, and is proved to be effective by experiment.

Key words: text categorization; Feature reduction; Feature selection; Mutual information;

1 引言

随着网络技术的高速发展，网络上的电子文档也迅速增长，如何有效地、更好地帮助用户查找、过滤、管理这些海量数据显得越来越重要，因此文本分类（基于文档内容，把一篇新文档分类到预定义好的类别中）在许多信息组织和管理领域中的应用越来越广泛。

文本分类大致可以分为三个步骤：文本的向量模型表示、特征抽取、分类器训练。目前大多数使用向量空间模型对文本表示成为向量形式，而向量的属性则有可能涉及到中文中的所有词汇，其向量的维数是非常巨大的，同时考虑到一篇文章只不过包含极少数词语（比如，一篇文档只由几百个词语组成），可知文档表示向量的稀疏性。这样高维的特征空间对文本分类的运算时间和空间复杂性是很不利的，因此在进行文本分类之前需要对文本进行特征降维，以最大程度的提高文本分类的精度，同时高效的特征降维能够节省更多的存储空间、提高分类速度。特征降维的主要方法有特征选择、特征抽取，下面分别对几种不同的特征降维方法进行介绍。

作者简介：伍建军（1982-），男，湖南祁阳人，硕士研究生，研究方向为 Internet 信息检索、数据挖掘

E-mail： happier5281@yahoo.com.cn ；

康耀红（1963-），男，陕西韩城人，教授，博士生导师，研究方向为 Internet 信息检索等

2 特征选择

特征选择指的是从特征总集中挑选出一部分有用的、对分类类别有贡献的词条组成特征子集，其一般的方法是使用某种评估函数独立地对每个特征词打分，然后把特征词按照分值高低排队，取最高分的一些特征词作为文本特征子集。特征选择并没有改变原始特征空间的性质，只是从原始特征空间中选择了一部分重要的特征，组成一个新的低维空间[6]。下面就四种比较常用的特征选择方法，包括文档频率（Document Frequency）、互信息（Mutual Information）、X^2统计量(Chi-square)、相关系数（correlation coefficient）等方法做进一步介绍。

2.1 文档频率（Document Frequency DF）

词条的文档频率(DF)就是指在训练样本集中出现该词条的文档数。在进行特征抽取时，将 DF 高于某个特定阈值的词条提取出来，低于这个阈值的词条给予滤除。

DF 评估函数的理论假设是稀有词条不含有有用信息，或含有的信息太少不足以对分类产生影响，而应当被去除。然而这种假设与一般的信息抽取观念有点冲突，因为在信息抽取中，有些稀有词条（如类别特征词）却恰恰比那些中频词更能反映类别的特征而不应该被滤除，因此单独使用 DF 评估函数进行特征选择未免太武断了。同时在文献[1]中的实验证明：在 CHI 等统计方法的计算“费用”太高而变得不可用时，DF 可以替代它们被使用。

2.2 互信息（Mutual Information MI）

互信息本来是信息论中的一个概念，用于表示信息之间的关系，而将这个概念引入到特征选择来表示词条与类别之间的关系。使用互信息理论进行特征抽取是基于如下假设：在某个特定类别出现频率高、但在其它类别出现频率比较低的词条与该类的互信息比较大。

对于某个类别C_j与特定词条W的互信息计算公式如下：$MI(C_j,W)=\log\left(\frac{P(W|C_j)}{P(W)}\right)=\log(P(W|C_j))-\log(P(W))$L (1)，其中$P(W|C_j)$表示为词条 W 在类别$C_j$出现的频率，$P(W)$表示词条 W 在整个训练文档中出现的频率，$MI(C_j,W)$表示词条 W 和类别$C_j$的互信息，即特征词与类别之间的相关程度。当特征词的出现只依赖于某一类别时，特征与该类别的互信息很大；当特征与类别相互独立时，互信息为 0；当特征很少在该类别文本出现时，它们之间的互信息为负数，即负相关。由公式（1）可以看出，对于频度比较小的特征，$\log(P(W))$的变化比$\log(P(W|C_j))$快，是互信息的主要部分，这就使得低频特征具有较大的互信息。

如果使用了 m 个类，则对于每个词条 W 有 m 个值，取它们的最大值作为每个词条的全局互信息量，然后将这些值进行排序，设定一个恰当的阈值，并保留高于阈值的词条作为文本的特征。

2.3 X^2统计法(Chi-square CHI)

CHI 统计方法度量词条与文档类别之间的相关程度，并假设词条与类别之间符合具有一阶自由度的X^2分布。词条对于某个类别的 统计量越高，表明它与该类之间的相关性越大，所携带的类别信息也就越多。令 A 表示属于 C 类且包含词条 W 的文档频率，B 表示不属于 C 类但包含词条 W 的文档频率，C 表示属于 C 类但不包含 W 的文档频率，D 表示既不属于 C 类也不包含 W 的文档频率，则词条 W 对于类别 C 的 CHI 统计值由下列式子计算：

$$X^2(C,W)=\frac{(AD-CB)^2\times(A+B+C+D)}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

利用 CHI 统计方法来进行特征抽取是基于如下假设：在指定类别文本中出现频率高的词条与在其它类别文本中出现频率比较高的词条，对判定文档是否属于该类别都是很有帮助的。词条的 CHI 统计值比较了词条对一个类

别的贡献和其它类别的贡献，以及词条和其它词条对分类的影响，当词条 W 与类别 C 互相独立时，CHI 值为 0；若 $AD-CB>0$，说明 W 与 C 正相关，即词条的出现说明某个类别也可能出现；反之 $AD-CB<0$，说明 W 与 C 负相关，即词条出现说明某个类别很可能不出现。为了将 CHI 统计量应用到所有的类别中，与互信息的处理类似，取每个类别对应的词条的 CHI 统计量最大值作为该词条的全局统计量，然后对这些统计量进行排序，设置一个阈值，保留高于阈值以上的词条组成特征子集。

2.4 相关系数（correlation coefficient CC）

在公式（2）中，分子取平方使得特征与类别的正相关能力与负相关能力被同等对待，而文献[4]指出：对于文本分类而言，特征的重要性主要由特征词与类别的正相关能力决定，所以提出了一种相关系数方法，其认为仅仅使用本类别文本中出现高的词语作为特征词项，能取得更好的分类效果。虽然在其它类中出现频率高的词语能够对判别文本不属于类别 C 有很好的提示作用，但是这个作用对分类效果的影响并不明显，于是得到了 CHI 统计法的改进形式，其数学表达式为：

$$CC(C,W)=\frac{(AD-CB)\times\sqrt{(A+B+C+D)}}{\sqrt{(A+C)\times(B+D)\times(A+B)\times(C+D)}}。$$

3 其它特征降维方式

不管使用哪种特征选择方法，基于特征选择的降维方式只是从原始特征空间中选择了一部分重要的特征，组成一个新的低维空间，并没有改变原始特征空间的性质。而特征抽取方法则可以看作从测量空间到特征空间的一种映射或变换，一般是通过构造一个特征评分函数，把测量空间的数据投影到特征空间，得到在特征空间的值，然后根据特征空间中的值抽取最高的若干个特征。

3.1 特征抽取—潜在语义索引

常用的特征抽取方法主要有主成分分析、潜在语义索引、非负矩阵分解等，下面就潜在语义索引[2]作进一步的介绍。

由于文本中存在同义词和多义词现象，导致特征向量构造的空间存在“斜交”的特点，也就是说，特征向量的各个分量之间存在一定的相关性。潜在语义索引（Latent Semantic Indexing LSI）通过挖掘文本与特征之间潜在的语义结构，将文本特征矩阵分解为一个低维的正交矩阵，实现了特征空间的降维。LSI 的过程实际上是将高维空间中的文档向量（词条向量）投影到低维的潜在语义空间中，使得原来没有任何共同项的两个文档（词汇），得到 LSI 处理后有可能找到它们彼此间的比较有意义的关联性，体现了文档（词汇）间的语义。

在 LSI 模型中，一个文本集可以表示为一个 m×n 的词—文档矩阵 A，其中 m 表示文本集中包含的所有不同的词条个数，n 表示文本集中的文本个数，即每个文档对应着矩阵 A 的一列，每个不同的词对应 A 的一行。A 表示为 $A=\left[a_{ij}\right]$，其中 a_{ij} 表示第 i 个词在第 j 篇文档中出现的权重。词—文档矩阵 A 建立以后，就可以利用奇异值分解计算 A 的 K 维近似矩阵 $\mathrm{A_k}$，其中 $k=\min(m,n)$。通过奇异值分解，矩阵 A 表示为 3 个正交矩阵的乘积：$\mathrm{A_{m\times n}=TSD^T}$，其中 $T_{m\times r}=(t_1,t_2\mathrm{L\ L}\ t_r)$，$t_1,t_2\mathrm{L\ L}\ t_r$ 为 T 的左奇异向量，并且为 AA^T 的特征向量；$S_{r\times r}=(\sigma_1,\sigma_2\mathrm{L\ L}\ \sigma_r)$ 为对角矩阵 S，$\sigma_1,\sigma_2\mathrm{L\ L}\ \sigma_r$ 为 A 的所有奇异值，同时也是 AA^T 所有特征值的平方根，并且满足关系 $\sigma_1\geq\sigma_2\geq\mathrm{L\ L}\geq\sigma_r\geq0$；$D_{n\times r}=(d_1,d_2\mathrm{L\ L}\ d_r)$ 为正交矩阵，$d_1,d_2\mathrm{L\ L}\ d_r$ 为 A 的右奇异向量，并且是 AA^T 的特征向量。

并将矩阵 A 的奇异值递减排序，取前 K 个奇异值构成对角阵 S_k，同时取 T 和 D 最前面的 k 列分别得到 T_k，D_k，然后进行反运算得到 A 的近似矩阵 A_k，其计算公式为：$A_k=T_kS_kD_k^T$，其中 T_k 和 D_k 列向量都是正交向量，分别作为词向量和文本向量，在此基础上进行文本分类或其它各种文档处理。

LSI 利用潜在的语义结构来表示词条和文本，将词条和文本投影到同一个 K 维的语义空间中去，词条和文本所表示的向量中的元素不再是反映词条出现的频率和分布关系，而是反映了语义关系。它在保持了原始的大部分信息的同时，克服了文本内容表示方法中存在的同义词、多义词等现象。

3.2 基于特征词聚合的降维方式

特征选择的主要做法是保留对分类有用的词条，过滤掉对分类无用特征，而没有抓住词条与词条之间的关联性，虽然潜在语义索引将特征空间里的信息进行重组，很少丢失原始特征空间里的信息，同时体现出了词条之间的语义关系，但是计算复杂度比较高，而且在大规模数据集上进行奇异值分解非常困难[2]。

在文献[3]中，提出了一种基于特征词聚合的朴素贝叶斯分类法。它把词在文本中的出现看成一个事件，先通过搜索算法，计算每一个特征词的分布，合并对分类有相似作用的特征词，然后建立一种新的、降维的事件空间，然后在这个事件空间中进行处理，测试结果表明，基于特征词聚类的文本分类精度更高，分类速度更快。需要说明的是，基于特征词聚合的降维方式并不象特征选择那样将那些无用信息过滤掉，而是使用某种聚类方式将其包含的信息附加到其它的词条身上，因此聚类处理后的特征信息更能体现原文档信息。

4 基于特征选择方法的 K-近邻文本分类及其测试分析

4.1 分类器设计

文本分类一般由预处理、特征选择、文本的向量模型表示、分类算法等几部分组成。

首先对训练文档集进行预处理，主要包括中文分词、停用词过滤等，得到初始的特征子集 A，然后分别使用上面介绍到的几种特征选择方法（文档频率、互信息法、CHI 统计法、相关系数法）对特征子集 A 中的每个特征词打分，并设定合适的阈值，保留高于该阈值的特征词，虑去低于该阈值的特征词，从而构成用于分类所使用的特征空间。再根据该特征空间里的特征项对训练文本进行向量表示，同时将所有的训练文本向量进行训练得到一定的分类规则，然后对需要分类的文本进行文本向量表示，使用 K-近邻分类算法[5]即可得到分类结果，最后对分类结果进行评估。

4.2 分类测试实验

本实验中所采用的语料库来自于“中文自然语言处理开放平台(http://www.nlp.org.cn)”提供的复旦大学收集的文本分类语料库，该语料库包括了测试集（共 9833 篇文档）与训练集（共 9804 篇文档），一共分为 20 个类别，包括计算机、经济、政治、军事、教育、环境、体育等。本实验选用了其中的 5 个类别，分别为教育、计算机、交通、医药、环境类等，每个类别随机选取 50 篇文章作为训练样本，在这 5 个类别各自随机抽取了 50 篇文章作为测试样本。由于本实验中使用了中国科学院的免费汉语词法分析系统，在一定程度上制约了用于分词的文档规模，所以一共随机选择了 250 篇作为训练样本集，250 篇作为测试样本集。考虑到分类速度的问题和免费分词系统的问题，在一定程度上限制了训练样本集和测试样本集的规模，但这并不影响本文要求的分类精度比较的验证。

其中本实验使用了中国科学院的汉语词法分析系统 ICTCLAS 对文本进行分词处理，该系统的功能有：中文分词；词性标注；未登录词识别，其中分词正确率高达 97.58%。而停用词是指那些不包含类别信息的词汇（主要包括介词、虚词等），如：的、是、啊等词，在本实验中使用的停用词词表的规模为 586 个。同时还考虑到单字词所表达的信息不够完整，所以分词后在单字词在本实验中也被滤除。

经过分词及停用词过滤以后得到的词条总数为 10910 条。分别选用 DF、MI、CHI、CC 这四种方法进行特征选择，并进行文本向量表示，使用 KNN 分类算法进行待分类文本的分类测试。其中 DF 法设定的阈值为 2，即文档

频率小于 2 的词条过滤掉，而其它特征选择方法的阈值设为 4000（保留排序在前 4000 的词条构成特征子集），且 K-近邻分类器中 K 取 25，仅仅保留相似度最大的 25 篇训练样本，分类后得到的具体结果如表 1。

表 1 分类测试结果

Tab.1 The result of text categorization experiment

特征选择 / 精度 / 分类类别	7 8 DF	MI	CHI	CC
教育	94%	90%	88%	94%
计算机	98%	96%	100%	100%
交通	86%	56%	78%	82%
医药	82%	84%	74%	80%
环境	74%	52%	64%	76%
平均精度	86.8%	75.6%	80.8%	86%

从表 1 可以看出，总体而言，在相同的测试条件下，使用文档频率 DF 评估函数进行特征降维，得到的分类精度最好。不管从评估函数公式还是实验中可以看出，其它特征选择方法，如 MI、CHI 统计法、CC 统计法，都要进行大量的计算才能统计出各个词条的打分，而 DF 只需要计算每个词条的文档频率就能对其进行打分，排序输出特征词，所以在分类过程中，DF 所使用的分类时间是最短的。

CC 统计法相对于 CHI 统计法而言，它只考虑本类别文本中出现高的词条作为特征词项的几率更大，一般能取得更好的分类效果，在表 1 中可以看出，对于所有的类别，使用 CC 统计法进行特征降维比 CHI 统计法的分类精度要高，而平均精度高出 5.2%。

从表 1 中还可以看出，使用互信息评估函数进行特征降维，对分类的效果最差，这个测试结果恰恰和文献[1]中的测试结果相一致。同时在文献[1]中，CHI 统计法在英文文本分类问题中表现良好，优越于 DF，而在我们的实验中，基于 DF 方法的分类效果比基于 CHI 统计法的分类效果要好一点，造成这种差别可能是因为使用类别信息的特征抽取方法 CHI 统计法对低频词的倚重和中文相对于英文具有更高的特征空间维数等。

4.3 一种改进的互信息评估函数的提出

同时在本实验的特征选择分阶段中发现，使用互信息评估函数进行特征选择时，许多词条的互信息量是完全一样的，而且同一分值的词条可能有很几百个，因此在进行特征选择时，只能随机的删除那些分值与前面的相同但是排在后面的词条，而造成大量有用信息的损失。基于这类情况的考虑，需要对互信息评估函数进行改进，以达到更好的分类效果。

从互信息计算公式（1）可以看出，它并没有考虑到关键词在文档中出现的频率，因此会造成不同频率词互信息量的相近或大量相同。例如对于词条 a 和 b 而言，如果 a 在所有文档中出现的次数 20，而在类别 C 中出现的次数为 10；b 在所有文档中出现的次数为 8，而在类别 C 中出现的次数为 4，由公式（1）计算得到的互信息量是相同的，但是从直观上来看，词条 a 对类别 C 的贡献要大。同时从公式（1）还可以看出，对同一个类，不同的词词条，在相同的 $P(W \mid C_j)$ 情况下，相对稀有的 W 会得到较大的值。基于以上考虑，改进的互信息评估函数为：$MI(C_j,W)=\log\left(\frac{P(W \mid C_j)}{P(W)}\times TF(W,C_j)\right)$。

使用相同的训练样本集和测试文档进行分类测试，结果如表 2 所示：

表 2 用 MI 及其改进的互信息评估函数作为特征选择的分类测试结果

Tab.2 The result of text categorization experiment based on MI and improved MI

分类类别 / 精度 / 特征选择	教育	计算机	交通	环境	医药	平均精度
MI	90%	96%	56%	52%	84%	75.6%

MI_{ch}	94%	100%	88%	80%	78%	88%

从表 2 可以看出，使用改进的互信息评估函数进行特征选择，得到的分类精度有明显的提高，其中环境类的分类精度提高了 28%，虽然医药类的分类精度有所下降，但平均精度提高了 12.4%。主要原因是由于在相同的传统互信息量的条件下，考虑了词频的作用，因而增加了高频词的互信息量，而减少了低频词的互信息量，在进行特征选择的时候，保留了含有类别信息的高频词，过滤了一定数量的低频词。

5 结论

考虑到文本向量巨大的维数和文本分类的运算速度，在进行文本分类之前需要进行特征降维处理。本文首先对几种特征降维方式（如特征选择、特征抽取、特征词聚合等方法）分别进行了介绍，并通过实验主要考察了基于文档频率、互信息、CHI 统计法和相关系数法等特征选择方法的文本分类，并从实验分析中提出了一种改进的互信息评估函数，最后通过测试表明，使用基于改进的互信息评估函数进行特征降维，对文本分类的精度有所提高。不同的特征降维方式都有各自的优点，下一步的工作将集中在如何选择不同的特征降维方式进行组合，从而提高分类精度。

参考文献

[1] Yiming Yang. A Comparative Study on Feature Selection in Text Categorization[A] .Proceeding of the Fourteenth International Conference on Machine Learning (ICMLp97) [C]，1997. 412-420

[2] Deerwester S , Dumais S , Furnas D . Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science , 1990 , 41(6) :391-407

[3] L. Douglas Baker, Andrew Kachites McCallum. Distributional clustering of words for text classification[A]. Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval[C].1998

[4] Luigi Galavotti, Fabrizio Sebastiani. Feature Selection and Negative Evidence in Automated Text Categorization[A]. In :Proceedings of the ACM KDD-00 Workshop on Text Mining[C], 2000.

[5] Yiming Yang. Expert network：Effective and efficient learning from human decisions in text categorization and retrieval [A] Proceedings of the 7 t h Annual International ACN-SIGIR Conference on Research and Development in Information Retrieval Dublin[C].2001

[6] Tao Liu , Shengping Liu , Zheng Chen. An evaluation on feature selection for text clustering[A]. In : Proceedings of the 20th International Conference on Machine Learning (ICML203) [C] ,2003. 488～495

[7] 张宁, 贾自艳, 史忠植 使用 KNN 算法的文本分类[J] 计算机工程 2005, 31(8) : 171-172