文章编号:1001 - 9081(2007)01 - 0205 - 02

一种基于聚类的文本特征选择方法

张文良,黄亚楼,倪维健
(南开大学 软件学院,天津 300071)
(waynesoft@ mail. nankai. edu. cn)

摘 要:传统的文本特征选择方法存在一个共性,即通过某种评价函数分别计算单个特征对类别的区分能力,由于没有考虑特征间的关联性,这些方法选择的特征集往往存在着冗余。针对这一问题,提出了一种基于聚类的特征选择方法,先使用聚类的方法对特征间的冗余性进行裁减,然后使用信息增益的方法选取类别区分能力强的特征。实验结果表明,这种基于聚类的特征选择方法使得文本分类的正确性得到了有效的提高。

关键词:特征选择;聚类;文本分类;信息增益
中图分类号: TP311.13 **文献标识码**:A

Clustering-based feature selection in text categorization

ZHANG Wen-liang, HUANG Ya-lou, NI Wei-jian
(*College of Software, Nankai University, Tianjin* 300071, *China*)

Abstract: Traditional text feature selection methods have a common characteristic that they evaluate features with some evaluation function individually. Because they do not consider the relationship among features, the selected feature subsets are redundant sometimes. This paper introduced a new feature selection method, which first used clustering to reduce redundancy among features and then used Information Gain to choose good features. The experimental results show that the new approach significantly improves classification accuracy.

Key words: feature selection; clustering; text categorization; information gain

0 引言

文本分类是指在给定分类体系下,根据文本的内容将其分到相应的预定义的类别的过程[1]。文本分类最初是应信息检索系统的要求出现的,随着信息网络的普及,海量的电子化文本信息的出现迫切要求由机器来自动地进行文本分类。文本自动分类可节约大量人力和财力,避免人工分类带来的周期长、费用高和效率低等诸多缺陷。

文本分类过程的一般步骤:1)预处理。将文本信息表示成计算机可以处理的结构化信息;2)特征选择。运用特征选择算法在特征集中选择最能体现类别信息的特征,从而得出最佳的特征子集;3)分类器训练及分类运算。特征选择的目的在于减小文本的特征向量维数,保留有区分能力的特征。同时,具有区分能力的特征可以提高系统的效率和精度,有效防止过渡拟合问题[2],所以特征选择在文本分类过程中是至关重要的。

文本分类中,特征选择的基本思想都是构造一个评价函数,对特征集的每个特征进行评估。这样每个特征都获得一个评估分,然后对所有的特征按照其评估分的大小进行排序,选取预定数目的最佳特征作为特征子集。但是由于没有考虑特征间的关联性,这些方法选择的特征子集在类别区分能力上往往存在着冗余。本文提出的基于聚类的特征选择方法,能够有效地减少特征集的冗余性。

1 特征冗余性

在数据集中,某些特征可能总是同时出现,那么这些特征在类别区分能力上是非常相似的,也就是说在类别区分能力方面,这些特征间存在着冗余。

一个简单的例子:假设训练集有 100 个文档,两个类别 $C1$ 和 $C2$,属于 $C1$ 的文档数为60,属于 $C2$ 的文档数也为40,如果有两个特征 A、B:

1) $C1$ 类中的包含 A 的文档为58个,不包含 A 的文档为两个;$C2$ 类中包含 A 的文档为 1 个,不包含 A 的文档为 39 个。

2) $C1$ 类中的包含 B 的文档为59个,不包含 B 的文档为1个;$C2$ 类中包含 B 的文档为 2 个,不包含 B 的文档为 38 个。

在训练集中,A 与 B 几乎同时出现,特征 A 和特征 B 对分类提供的指导信息是非常相似的。如果待分类文档中包含 A 或 B,则它很有可能属于 $C1$ 类,如果待分类文档中既不包含 A 也不包含 B,则它很有可能属于 $C2$ 类。显然,特征 A 与 B 在类别区分能力上存在冗余。

既然 A 和 B 在类别区分能力非常相似,那么在特征选择时,只要选择其中之一即可。在利用特征选择方法对 A 和 B 进行评估时,A 和 B 会得到非常近似的分数,因此很有可能会被同时选进特征子集中,从而导致最终的特征子集也存在冗余。后面的实验证明,在文本分类中,特征冗余性是普遍存在的。

2 基于聚类的特征选择方法

针对特征在类别区分方面的冗余性,本文提出了一种基于聚类的特征选择方法(Clustering based Information Gain, CBIG),该方法分为两步:冗余性过滤和依据对类别的区分能力进行特征选择。主要思想是根据特征间的相似度,对特征进行聚类,在每个簇中选择一个特征代表整个簇,将簇中的其他特征从候选特征集中剔除,这样特征集中的冗余性就大大减小;然后对剩余的特征使用信息增益方法进行特征选择。

收稿日期:2006 - 06 - 21;**修订日期**:2006 - 08 - 27

作者简介:张文良(1981 -),男,浙江杭州人,硕士研究生,主要研究方向:知识工程、数据挖掘; 黄亚楼(1964 -),男,河北沧州人,教授,博士生导师,主要研究方向:智能信息处理、智能机器人系统; 倪维健(1981 -),男,山东临沂人,博士研究生,主要研究方向:文本挖掘、信息检索.

具体流程如图 1 所示。

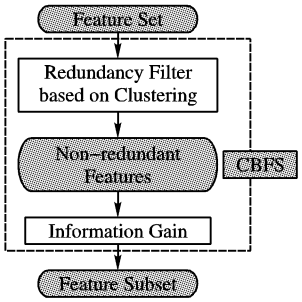


图 1 基于聚类的特征选择方法流程

基于聚类的特征选择方法的具体步骤如下：

第 1 步 使用 TF-IDF 公式对训练集文本进行特征权值计算：

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

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

第 2 步 对特征进行聚类：

每个特征作为聚类的一个样本，将特征 T 在训练集上每个文档中的权值作为一维，则聚类的输入向量 T_{ij} 表示在文档 D_j 中，特征 T_i 的权重（$1 \leqslant i \leqslant v, 1 \leqslant j \leqslant n$，其中 v 为训练集中的特征数，n 为训练集中的文档数）。

定义特征 A、B 间的相似性度量为：

$$\mathrm{sim}(A,B) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

可见 $0 \leqslant \mathrm{sim}(A,B) \leqslant 1$，当两个特征越相似，$\mathrm{sim}(A,B)$ 的值越大。

其中采用的聚类方法如下：

1）随机选取一个特征作为第一个簇的中心。

2）取得下一个特征，依次计算该特征与已有簇中心的相似度。

3）如果该特征与所有簇中心的相似度都小于阈值，则以该特征作为中心建立一个新簇；否则将该特征加入到相似度最大的簇中（本文实验中的相似度阈值设定为 0.95）。

4）重复 2、3，直到所有的特征都被处理。

第 3 步 保留每个簇的中心，将簇中的其他特征剔除。

第 4 步 采用信息增益（Information Gain，IG）对剩余的特征进行排序，选择分数最高的若干特征作为最终的特征子集。IG 是信息论中的一个重要概念，它表示了某一个特征项的存在与否对类别预测的影响，定义为某一特征项在文本中出现前后的信息熵之差，即：

$$IG(t) = P(t)\sum_i P(C_i|t)\log\frac{P(C_i|t)}{P(C_i)} + P(\bar{t})\sum_i P(C_i|\bar{t})\log\frac{P(C_i|\bar{t})}{P(C_i)}$$

其中，$i = 1,2,\cdots,M$，M 为类别数。

$P(C_i)$ 表示类别 C_i 在训练集中出现的概率，$P(t)$ 表示训练集中包含特征项 t 的文本的概率，$P(C_i|t)$ 表示文本包含特征项 t 时属于 C_i 类的条件概率，$P(\bar{t})$ 表示训练集中不包含特征项 t 的文本的概率，$P(C_i|\bar{t})$ 表示文本不包含特征项 t 时属于 C_i 类的条件概率。显然，某个特征项的信息增益值越大，类别区分能力就越强。

通过将特征聚类，簇内特征间都非常相似，这些特征在类别区分能力上往往也是类似的。因此，我们取每个簇的中心代表整个簇，将簇中的其他特征过滤掉，这样特征集的冗余性就大大降低。

3 实验分析

目前常用的文本特征选择方法有：文档频率、信息增益、互信息、χ^2 统计量、期望交叉熵、文本证据权和几率比等[3,4]。其中信息增益较选择的特征子集质量较其他方法更优[5]。因此在本文的实验中，对比了基于聚类的特征选择方法与 IG 方法所选取的特征在分类时的效果。

3.1 数据集

本文中的实验采用的数据集为 PU1 Corpus[6]。PU1 Corpus 包含 1 099 个样本，其中 481 个为垃圾邮件，618 个为合法邮件，垃圾邮件率为 43.77%。邮件中的邮件头和 HTML 标签都已被删除，只剩下主题和正文内容。为了维护用户的隐私，邮件中的单词都被映射为一个唯一的整数。PU1 数据集有四个版本：bare、lemm、emm_stop 和 stop，本文中的实验采用了 lemm_stop 版本。此版本的邮件内容单词已经过优化处理，排除了“-ed，-ing”等词型派生变化情况，去掉了“a，and，for”等频繁使用但提供信息量较小的单词。我们对整个数据集进行分析，发现其中共包含 21 705 个不同特征（单词）。

整个数据集分成大小相等的 10 个样本子集，依次为 Part1，Part2，…，Part10，每个部分的样本数大约为 110 个，每个样本子集中合法邮件与垃圾邮件的比例大致为 6∶5。本文中所有实验都采用了 10-folder 交叉验证。

3.2 分类器与参数设置

朴素贝叶斯分类器（Naive Bayes Classifier）被广泛应用于文本分类中，经常被用作与其他算法比较的基准。因此本文的实验采用了朴素贝叶斯分类器。

在分类器建立之后，就可以对邮件 $\vec{d}$ 进行分类，如果后验概率 $\frac{P(spam|\vec{d})}{P(legit|\vec{d})} > \lambda$[6]，那么该信息就被分类为垃圾邮件。在实验中，令误分合法邮件为垃圾邮件的代价是误分垃圾邮件为合法邮件代价的 9 倍，即令 $\lambda = 9$。

3.3 实验结果评估指标

准确率（Precision）和查全率（Recall）是目前常用的两种分类质量评价指标。准确率和查全率的计算公式分别为：

$$p = \frac{n_{S\to S}}{n_{S\to S} + n_{L\to S}}$$

$$r = \frac{n_{S\to S}}{n_{S\to S} + n_{S\to L}}$$

$n_{S\to S}$ 表示将垃圾邮件正确分类的个数，$n_{S\to L}$ 表示将垃圾邮件错误分类的个数，$n_{L\to S}$ 表示将合法邮件错误分类的个数。

准确率和查全率反映了分类质量的两个不同方面，两者必须综合考虑。$F1$ 值是综合考虑这两者的一个评价指标，其数学公式如下：

$$F1 = \frac{2pr}{p + r}$$

本文的实验结果评价采用了 Precision、Recall 和 F1 值三种指标。

3.4 实验结果与分析

在下面的实验中，我们分别使用 IG 和 CBIG 两种特征选择方法在前文所示的数据集上选取前 1 000、2 000、3 000、4 000、5 000 个特征训练朴素贝叶斯分类器对垃圾邮件进行过滤，分类结果如图 2 所示。 （下转第 209 页）

从图 3 可以看出,在文档集合篇数相同时,局部类别分析算法的查准率要高于其他两种方法。影响潜语义标引算法查准率的主要原因是加入扩展词后产生的主题偏离问题;局部上下文分析算法由于只对首次查询返回的结果集进行分析,有时只能得到较少的相关信息,特别是在结果集包含过多无关文档时会加入很多无关词语,严重影响查准率;而新算法由于获得了初次检出文档以外更多的相关信息,从而获得了更高的查准率。在图 3 中,这三种方法的查准率都在文档篇数为200左右时最高,之后随着文档篇数的增加而降低,这主要是因为随着文档集合的增大,无关信息的干扰也会增大,导致查准率随之降低。

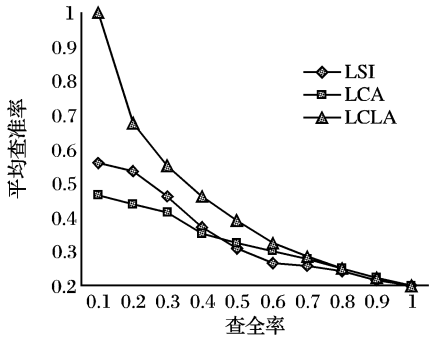


图 2　三种算法在不同查全率时的平均查准率

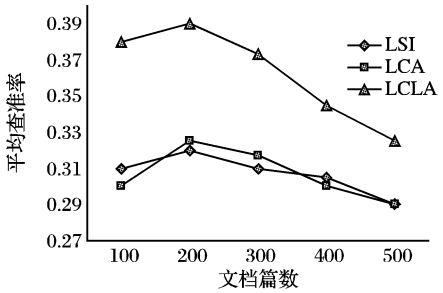


图 3　三种算法在不同文档数时的平均查准率

通过以上的实验和分析可以看出,在与两种常用的自动查询扩展算法的比较中,局部类别分析算法的时间开销较小,查准率较高,因此具有更好的检索性能。

3　结语

在信息检索系统中引入查询扩展,容易产生与原查询主题偏离的问题,严重影响系统的检索性能。本文提出一种基于局部类别分析的查询扩展算法,通过分析与用户查询相关的文档类别,并利用相关类别中词语的共现关系来选取扩展词,避免加入与原查询不相关的词,以缓解主题偏离的问题,提高检索系统的查准率。实验表明新算法取得了较好的效果。

参考文献:

[1] CROUCH CJ, YANG B. Experiments in automatic statistical thesaurus construction[A]. Proceedings of the ACM-SIGIR Conference on Research and Development in Information Retrieval[C]. 1992.77 - 88.

[2] QIU Y, FREI. Concept based query expansion[A]. Proceedings of the 16th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '93, Pittsburgh, PA)[C]. KORFHAGE R, RASMUSSEN E, WILLETT P, eds. New York: ACM Press, 1993. 160 - 169.

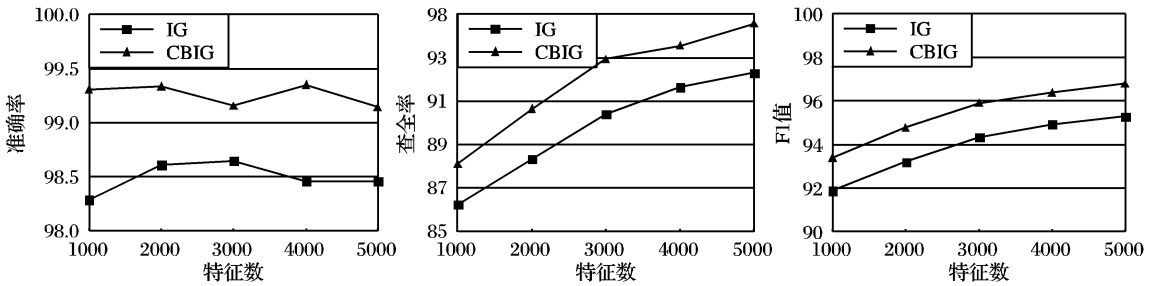
[3] BUCKLEY C, SALTON G, ALLAN J, *et al.* Automatic query expansion using SMART[A]. TREC-3: Overview of the Third Text Retrieval Conference (TREC-3)[C]. 1995.69 - 80.

[4] XU JX, CROFT WB. Improving the effectiveness of information retrieval with local context analysis[A]. ACM Transactions on Information Stems[C]. 2000.79 - 112.

[5] CLOUGH P, SANDERSON M. Measuring Pseudo Relevance Feedback & CLIR[A]. SIGIR'04[C]. 2004.484 - 485.

[6] TAO T, ZHAI CX. A Two-stage Mixture Model for Pseudo Feedback[A]. SIGIR'04[C]. 2004.486 - 487.

(上接第 206 页)



(a) 准确率比较　(b) 查全率比较　(c) F1值比较

图 2　分类结果比较

从图 2 可以看出,在特征数为 1 000、2 000、3 000、4 000、5000 时,Cluster-based IG 方法较 IG 方法在 Precision、Recall、F1 值三个指标上均有一定的提高。

通过实验可以说明,特征间的冗余性在本文数据中是普遍存在的,这正说明在特征集中确实存在着冗余性。而本文提出的基于聚类的特征方法有效地过滤了冗余特征,提高了特征子集的质量,改进了分类器的性能。

4　结语

本文设计了一种基于聚类的特征选择方法,通过对特征进行聚类,有效地过滤了冗余特征,提高了最终特征子集的质量。实验结果表明基于聚类的特征选择方法可以有效地改进特征子集的质量,从而提高分类器的性能。

参考文献:

[1] 胡佳妮, 徐蔚然, 等. 中文文本分类中的特征选择算法研究[J]. 光通信研究, 2005(3): 44 - 46.

[2] FLEURET F. Binary Feature Selection with Conditional Mutual Information[R]. Technical Report RR-941, 2003.

[3] YANG Y, PEDEREN JO. A Comparative Study On Feature Selection in Text Categorization[A]. The 14th International Conference on Machine Learning[C]. 1997.

[4] ROGATI M, YANG YM. High-performing Feature Selection for Text Categorization[A]. CIKM'02[C]. 2002.

[5] MITCHELL T. Machine Learning[M]. McGraw Hill, 1996.

[6] ANDROUTSOPOULOS I, KOUTSIAS J, CHANDRINOS KV, *et al.* An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages[A]. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR-2000)[C]. 2000.