

基于多特征选择的中文文本分类

董梅,胡学钢

(合肥工业大学 计算机与信息学院,安徽 合肥 230009)

摘要:自动文本分类就是在给定的分类体系下,让计算机根据文本的内容确定与它相关联的类别。特征选择作为文本分类中的关键,困难之一是特征空间的高维性,因此寻求一种有效的特征选择方法,降低特征空间的维数,成为文本分类中的重要问题。在分析已有的文本分类特征选择方法的基础上,实现了一种组合不同特征选择方法的多特征选择方法,应用于KNN文本分类算法,实验表明,多特征选择方法分类效果比单一的特征选择方法分类效果有明显的提高。

关键词:文本分类;特征选择;多特征选择

中图分类号:TP18

文献标识码:A

文章编号:1673-629X(2007)07-0117-03

Text Categorization Based on Multiple Features Selection

DONG Mei, HU Xue-gang

(School of Computer & Information, Hefei University of Technology, Hefei 230009, China)

Abstract: Automatic text categorization is the assigning of pre-defined category to a new text based on its content. Feature selection is the key of text categorization. Feature space's high dimension is one of difficulties of it. So to find an effective feature selection method and to reduce feature space's dimension has become the important problem of text categorization. Based on analyzing most known text categorization's feature selection methods and a new multiple feature selection method that combined different feature selection methods was given. Experiments were done using KNN algorithm. The results show that the new multiple features selection method had better efficiency than single feature selection method.

Key words: text categorization; feature selection; multiple features selection

0 引言

自动文本分类就是对大量的自然语言文本按照一定的主题类别进行自动分类。它能根据用户的信息需求,在动态的信息流中,搜索用户感兴趣的信息,屏蔽其它无用的信息,目前主要应用于信息检索、机器翻译、自动文摘、信息过滤、邮件分类等。文本分类所面临的首要问题是如何在计算机中合理地表示文本,这种表示法既要包含足够的信息以反映文本的特征,又不至于太过庞大使学习无法处理,这就涉及到文本的特征选择。特征空间的高维性和文档表示向量的稀疏性是特征选择的困难之一,因此寻求一种有效的特征选择方法,降低特征空间的维数,提高分类的效率和精度,成为文本自动分类中需要首先面对的重要问题。

目前文本分类方法很多,比较著名的有 Bayes^[1],

KNN^[2], LLSF^[3], Nnet^[4], Boosting^[5] 及 SVM^[6] 等。近年来在文本分类中使用较多的特征选择方法^[7] 包括:文档频率(DF),它是最简单的评估函数,常把它作为评判其他评估函数的基准;信息增益(IG),考虑了未出现词对文本的影响,分类效果好,但统计花费大;互信息(MI),它考虑了低频词带有信息量的情况,低频词的互信息比常用词的互信息高,但过于倾向低频词,分类效果不好;CHI统计,分类效果好但统计花费大,对于低频词效果不好。文中提出了一种将不同的特征选择方法组合的多特征选择,这种方法综合多种特征评估函数选择特征子集,不受具体文本语料的限制,降低了“噪音”影响。实验证明,这种多特征选择方法比单一的特征选择方法的分类效果有明显的提高。

1 相关工作

1.1 文本特征表示

文本的特征表示是文本分类面临的首要问题。Salton等人提出的向量空间模型 VSM(Vector-Space-Model)^[8] 是目前应用最多且效果较好的文本表示法

收稿日期:2006-09-11

基金项目:安徽省自然科学基金资助项目(050420207)

作者简介:董梅(1977-),女,河北保定人,硕士研究生,研究方向为数据挖掘;胡学钢,教授,博士,硕士生导师,研究方向为人工智能、数据挖掘、数据结构。

之一。在 VSM 中,文本空间被看作是由一组正交词条向量组成的向量空间。文本经过分词程序分词后,首先去除停用词,然后统计词频,最终表示为向量形式。

假设所有训练文本的特征总数是 n , 则构成一个 n 维的向量空间, 每一个文本 d 被表示为一个 n 维的特征向量:

$$V(d) = (t_1, \omega_1(d); t_2, \omega_2(d); \dots; t_n, \omega_n(d)) \quad (1)$$

这里, t_i 为词条项, $\omega_i(d)$ 为 t_i 在文本 d 中的权值, 一般采用 TF · IDF 向量表示法^[8]:

$$\omega_i(d) = \frac{tf_i(d) \times \log(N/n_i)}{\sqrt{\sum_j (tf_j(d) \times \log(N/n_j))^2}} \quad (2)$$

公式(2)中, $tf_i(d)$ 为词条 t_i 在文档 d 中出现的词频, N 为所有文档的数目, n_i 为出现了词条 t_i 的文档的数目。

1.2 特征选择

文本分类中,特征选择的基本思想通常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结果的特征子集。下面介绍几种常用的特征选择方法:

1) 文档频率(DF, Document Frequency)。

文档频率可表示为:

$$DF(t) = \frac{\text{出现特征 } t \text{ 的文档数}}{\text{训练集的总文档数}}$$

DF 是指在训练语料中出现该词条的文档数。文档频率方法提取文档频率较高的特征,它的目的是去掉在训练集上出现次数过少的特征,保留具有一定影响力的特征。在各个特征选择方法中,DF 方法是最简单的。

2) 信息增益(IG, Information Gain)。

IG 是在机器学习领域应用较为广泛的特征选择方法,它考虑了未出现词对文本的影响。对于词条 t 和文档类别 c , IG 考察 c 中出现和不出现 t 的文档频数来衡量 t 对于 c 的信息增益, $IG(t)$ 定义如下:

$$IG(t) = - \sum_{i=1}^m P(c_i) \log P(c_i) + P(t) \sum_{i=1}^m P(c_i | t) \log P(c_i | t) + P(\bar{t}) \sum_{i=1}^m P(c_i | \bar{t}) \log P(c_i | \bar{t})$$

其中 $P(c_i)$ 表示 c_i 类文档在语料中出现的概率, $P(t)$ 表示语料中包含词条 t 的文档的概率, $P(c_i | t)$ 表示文档包含词条 t 时属于 c_i 类的条件概率, $P(\bar{t})$ 表示语料中不包含词条 t 的文档的概率, $P(c_i | \bar{t})$ 表示文档

不包含词条 t 时属于 c_i 的条件概率, m 表示类别数。

3) 互信息(MI, Mutual Information)。

MI 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,特征 t 和类别 C 的互信息体现了特征与类别的相关程度。在某个类别 C 中出现的概率高,而在其它类别中出现的概率低的特征 t 将获得较高的互信息。MI 可表示为:

$$MI(t) = \sum_i P(C_i) \log \frac{P(t | C_i)}{P(t)}$$

4) X^2 统计量 CHI (X^2 Statistic)。

$$CHI(F) = \sum_i P(C_i) X^2(t, C_i) =$$

$$\sum_i P(C_i) \frac{N(AD - BC)^2}{(A + C)(B + D)(A + B)(C + D)}$$

式中, A 是特征 t 和第 i 类文档共同出现的频度; B 是特征 t 出现而第 i 类文档不出现的频度; C 是第 i 类文档出现而特征 t 不出现的频度; D 是第 i 类文档和特征 t 都不出现的频度; N 为总共的文本数。 X^2 方法认为特征 t 与文本类别 C_i 之间的非独立关系类似于具有一维自由度的 X^2 分布。

2 多特征选择文本分类器构造

2.1 多特征选择(MFS, Multiple Feature Selection)

每一种特征选择方法对原始特征空间进行评估,通过选取合适的阈值,都可以得到一个特征子集,不同的特征选择方法得到的特征子集也不一样。基于多特征方法思想是:先利用不同的特征选择方法对原始特征空间进行筛选,得到多个特征子集。再对这些特征子集求并集,作为文本分类的特征空间。一般地,设原始特征空间为 X , 最终选择的特征空间为 S , t 种特征选择方法分别为 $\theta_1, \theta_2, \dots, \theta_t$, 其对应的阈值为 $\sigma_1, \sigma_2, \dots, \sigma_t$ 。经过这 t 种特征选择后,得到的每个特征子集为 S_1, S_2, \dots, S_t , 则 $S = \bigcup_{i=1}^t S_i$ 。

算法一:多特征选择算法。

设原始特征空间 X , t 种特征选择方法 $\theta_1, \theta_2, \dots, \theta_t$ 对应的阈值 $\sigma_1, \sigma_2, \dots, \sigma_t$ 。特征选择后空间为 S 。

步骤 1:任意的特征选择方法 θ_i , 根据其评估函数,计算原始特征空间中每个特征的评估值,按从大到小的顺序排列,评估值不小于阈值 σ_i 的特征构成特征子集 S_i 。

步骤 2:对每一种特征选择方法重复步骤 1, 得到特征子集空间集合 $\{S_1, S_2, \dots, S_t\}$ 。

步骤 3: $S = \bigcup_{i=1}^t S_i$, 对特征子集求并集。

2.2 文本分类器构造

KNN 是一种简单、有效基于实例的文本分类算法,在文本分类中得到广泛使用,并且取得了很好的效

果。相似度可以通过欧几里德距离或向量间夹角来度量。文中采用 KNN 方法。

KNN 算法的基本思路是:在给定新文本后,考虑在训练文本集中与该新文本距离最近(最相似)的 K 篇文本,根据这 K 篇文本所属的类别判定新文本所属的类别。

算法二:文本分类器构造算法。

步骤 1:根据算法一得到的特征项集合重新描述训练文本向量。

步骤 2:在新文本到达后,根据特征词分词新文本,确定新文本的向量表示。

步骤 3:在训练文本集中选出与新文本最相似的 K 个文本,计算公式为:

$$\text{Sim}(d_i, d_j) = \frac{\sum_{k=1}^M W_{ik} \times W_{jk}}{\sqrt{(\sum_{k=1}^M W_{ik}^2)(\sum_{k=1}^M W_{jk}^2)}}$$

步骤 4:在新文本的 K 个邻居中,依次计算每类的权重,计算公式如下:

$$p(x, C_j) = \sum_{d_i \in \text{KNN}} \text{Sim}(x, d_i) y(d_i, C_j)$$

其中, x 为新文本的特征向量, $\text{Sim}(x, d_i)$ 为相似度计算公式,与上一步骤的计算公式相同,而 $y(d_i, C_j)$ 为类别属性函数,即,如果 d_i 属于类 C_j , 那么函数值为 1, 否则为 0。

步骤 5:比较类的权重,将文本分到权重最大的那个类别中。

3 实验结果与分析

3.1 测试集及性能评价指标

文中的实验数据为中文自然语言处理开放平台上用于文本分类的语料库。表 1 列出了这些语料的类别及其包含的文档数目。

进行实验时,任意选取其中 1883 篇作为训练集,934 篇作为测试集。预处理采用分词工具 ICTCLAS 对文档进行分词,并去掉停用词。采取普遍接受的宏平均 F1 值来评价文档分类的性能^[9]。

表 1 测试文档库

类名	环境	计算机	交通	教育	经济	军事	体育	医药	艺术	政治	合计
文档数	201	201	214	220	325	249	450	204	248	505	2817

3.2 实验结果与分析

在 CPU 赛扬 1.7GHz, 内存 256MB, VC++ 6.0 环境下,首先对 IG, CHI 和 MI 三种单一的特征选择方法进行了实验,然后将 IG 和 CHI、MI 和 IG、MI 和 CHI 互相结合即多特征选择 MFS(multiple feature selection)方法进行实验,文本分类算法采用 KNN,其中

k 近邻值取 30,选取的特征维数分别从 800 到 3000。比较试验结果,IG 和 CHI 相结合分类效果最好。图 1 所示为 IG 和 CHI 及其结合的 MFS 三种特征选择方法,随着特征数目变化的宏 F1 值比较图。

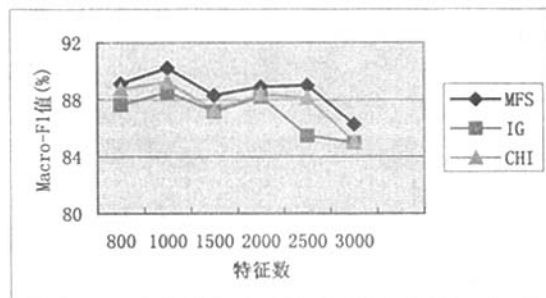


图 1 三种特征选择方法宏 F1 值比较图

实验可以得出如下结论:

1)特征维数 1000 时,分类效果最好。随着特征维数的升高和下降,文本分类的准确率也下降。这说明,通过特征选择,在特征维数 1000 时,能够比较好地反映训练文本信息。

2)将不同的特征选择方法结合起来的 MFS 方法,文本分类准确率有明显的提高。

4 结 论

文中提出将不同的特征选择方法结合起来的特征选择方法,并在中文语料上验证了算法思想。实验表明,多特征选择比单一的特征选择分类效果有明显提高。下一步,将研究多特征选择方法在不同语料集的分类情况,并将其结合到具体应用中。

参考文献:

- [1] Lewis D D. Naive(Bayes)at forty: the independence assumption in information retrieval[C]// In: The 10th European Conference on Machine Learning. New York: Springer, 1998:4-15.
- [2] Yang Y, Liu X. A re-examination of text categorization methods[C]// In: The 22nd Annual International ACM SIGIR Conference on Research and Development in the Information Retrieval. New York: ACM Press, 1999.
- [3] Yang Y, Chute C G. An example-based mapping method for text categorization and retrieval[J]. ACM Transaction on Information Systems, 1994,12(3):252-277.
- [4] Wiener E. A neural network approach to topic spotting[C]// In: The 4th Annual Symposium on Document Analysis and Information Retrieval. Las Vegas, NV:[s. n.],1995.
- [5] Schapire R E, Singer Y. Improved boosting algorithms using confidence-rated predications[C]// In: The 11th Annual

有效性。

由于拉格朗日插值系数是每个实体都知道的,则 $P_i \in B$ 可以通过 $SS_{(i,j)}$ 直接推导出 d_j ,但是除了自身拥有者对任何其它实体都应该是保密的,因此存在一些不足,文中通过采用一种混合因子算法来弥补这个缺陷。

在这个算法中,每个参与产生密钥份额的实体 $P_j \in A$ 都随机产生一个小数 Δ ,参与产生密钥份额的实体之间都要互相交换它们的小数 Δ ,拥有大 ID 号的实体把这个数看作是正数,其它的实体把这个数看作是负数。那么每个成员都有 $t-1$ 个这样的小数,每个成员计算这 $t-1$ 个的和以及 $SS_{(i,j)}$,然后把这个具有混淆部分的 $SS_{(i,j)} = SS_{(i,j)} + \sum \Delta$ 发送给 $P_i \in B$ 。因为每个 Δ 都有正负两个,所有的 Δ 相加仍为零,所以可以验证 $P_i \in B$ 得到了相同的 d_i 。

3.4 秘密重构

先验式秘密共享机制是在保证秘密信息的秘密性、完整性和可用性的前提下,共享秘密信息的机制。因此对于 (t, n) 门限秘密共享方案来说,只要有 t 台正常的主机,根据拉格朗日多项式在任何条件下都可以重构出秘密信息。

每对 $(x_i, d_i) (1 \leq i \leq n)$ 是“曲线” $f(x)$ 上的一个点。因为 t 个点可以唯一确定一个 $t-1$ 次多项式,所以 $f(x)$ 可以从 t 个共享密钥中重构出来。但是任何少于 t 个共享密钥无法确定 $f(x)$ 或者 d 。

给定 t 个共享密钥 $d_i (1 \leq i \leq t)$,由拉格朗日多项式重构的 $f(x)$ 为: $f(x) = \sum_{i=1}^t d_i \prod_{\substack{j=1 \\ j \neq i}}^t \frac{x - x_j}{x_i - x_j}$, 其中,运算为 Z_q 的运算。计算出 $f(x)$ 后,通过 $d = f(0)$ 可计算出密钥 d 。

$$d = f(0) = \sum_{i=1}^t d_i \prod_{\substack{j=1 \\ j \neq i}}^t \frac{x - x_j}{x_i - x_j}, \text{ 若令 } b = \prod_{\substack{j=1 \\ j \neq i}}^t \frac{-x_j}{x_i - x_j}, \text{ 则 } d = f(0) = \sum_{i=1}^t b d_i.$$
 其中 $x_i (1 \leq i \leq n)$ 是公开的。

4 结束语

由于先应式秘密共享方法在对秘密信息进行分散保护的同时,还周期性地对秘密份额进行更新,可防止所谓的移动攻击。这种方法被广泛应用于群组通信、密钥管理、安全多方计算、电子商务等领域。先应式秘密共享机制在不泄漏秘密的前提下,可以对秘密进行拆分、更新、检测、恢复和重构,是对传统的门限密码的一个改进。先应式秘密共享机制在秘密共享的各个阶段都对秘密进行检测,在最大程度上保证了秘密的完整性、秘密性和可用性。可以说,先应式秘密共享机制是在传统的门限式秘密共享的基础上,对其安全性进行强化的结果。

参考文献:

- [1] 周全,杨华冰,黄继海,等. 先应式秘密共享方案及实现[J]. 情报指挥控制系统与仿真技术, 2005, 27(3): 57-60.
- [2] Shamir A. How to Share a Secret[J]. Communications of the ACM, 1979, 22(11): 612-613.
- [3] Pedersen T P. Distributed provers with applications to undeniable signatures[C]// In: Proc. Eurocrypt'91. Lecture Notes in Computer Science 547. New York: Springer - Verlag, 1991: 221-238.
- [4] Hankerson D, Menezes A, Vanstone S. 椭圆曲线密码学导论[M]. 张焕国, 等译. 北京: 电子工业出版社, 2005.
- [5] Nikov V, Nikova S. On Proactive Secret Sharing Schemes[J]. Lecture Notes in Computer Science, 2004, 3357: 308-311.
- [6] Asaeda H, Rahman M, Toyama Y. Structuring Proactive Secret Sharing in Mobile Ad-hoc Networks[C]// International Symposium on Wireless Pervasive Computing (ISWPC), IEEE. Phuket, Thailand: [s. n.], 2006: 1-6.
- [7] 郭渊博, 马建峰. 异步及不可靠链路环境下的先应式秘密共享[J]. 电子学报, 2004, 32(3): 399-403.
- [8] Jiejun K, Petros Z, Luo Haiyun, et al. Providing robust and ubiquitous security support for mobile ad-hoc networks[C]// IEEE Ninth International Conference on Network Protocols. Riverside, USA: [s. n.], 2001: 251-260.

(上接第 119 页)

- Conference on Computational Learning Theory. Madison: ACM Press, 1998: 80-91.
- [6] Joachims T. Text categorization with support vector machines: learning with many relevant features[C]// In: The 10th European Conference on Machine Learning. New York: Springer: [s. n.], 1998: 137-142.
 - [7] Yang Yiming, Pederson J O. A Comparative Study on Feature Selection in Text Categorization[C]// In: Proceeding of

- the Fourteenth International Conference on Machine Learning (ICML'97). Nashville: Morgan Kaufmann, 1997: 412-420.
- [8] Salton G, Wong A, Yang C S. On the specification of term values in automatic indexing[J]. Journal of Documentation, 1973, 29(4): 351-372.
 - [9] 周茜, 赵名生, 扈旻. 中文文本分类中特征选择[J]. 中文信息学报, 2004, 18(3): 17-23.