

基于两种特征影响度的特征选择

胡 强

(川渝中烟工业公司 长城雪茄烟厂, 四川 什邡 618400)

摘 要: 定义了两种特征影响度: 一种是特征对类间文档分散程度的影响度, 该影响度越大越好; 另一种是特征对类内文档分散程度的影响度, 该影响度越小越好. 然后把这两种特征影响度有机地结合起来设计了一个新的特征选择方法. 该方法能够对所选特征进行综合考虑, 从而使获得的特征集具有较好的代表性. 仿真实验表明, 所提特征选择方法在一定程度上能够提高文本分类性能.

关键词: 特征选择; 文本分类; 特征影响度; 文档分散程度

中图分类号: TP301

文献标识码: A

文章编号: 1000-7180(2010)12-0065-04

Feature Selection Based on Two Kinds of Feature Influence Degree

HU Qiang

(GreatWall Cigar Factory, China Tobacco Chuanyu Industrial Corporation, Shifang 618400, China)

Abstract: Two kinds of feature influence degree were defined; one was the feature influence degree of document dispersion degree amongst categories, the contribution that was larger was better. Another was the feature influence degree of document dispersion degree in category, the contribution that was smaller was better. And then, the two kinds of feature influence degree were integrated organically and a new feature selection method was designed. The method can inspect selected feature synthetically so that the feature set that is more representative is obtained. Simulation experiments show that, to a certain extent, the feature selection method is able to improve performance of text categorization.

Key words: feature selection; text categorization; feature influence degree; document dispersion degree

1 引言

文本自动分类作为一项有效处理大量文本信息的核心技术而受到高度关注. 文本自动分类根据文档内容以及其属性给文档自动分配一个或多个类别^[1]. 目前, 在文本分类模型中一般用空间向量模型来对文档进行表示^[2]. 在这个表示模型中, 训练文档通常由大量特征词组成, 每个特征词对分类都起着一定的作用, 但这些大量的特征中也肯定包含许多相关的、冗余的特征, 这些冗余特征的存在不仅增加了系统的时空开销, 而且还大大限制了分类算法的选择、降低了分类算法的性能. 因此, 在实行文本自动分类之前必须去掉冗余特征以减少运算开销、提高分类的效率和精度^[3].

在消除冗余特征方面, 特征选择是一种比较有效的方法, 它从原始特征集中选择出一部分有代表性的、对分类类别贡献比较大的词条组成特征子集. 其一般过程就是使用某种评估函数对每个词条进行打分, 然后把词条按照分值高低排队, 取分值最高的一些词条构成作为文本特征子集.

目前比较成熟的特征选择方法有文档频、互信息、交叉熵、信息增益、 χ^2 统计量等^[4], 这些方法在评价时只是偏重于特征的某个方面, 从而使得选择的特征子集不具有代表性^[5]. 为此, 文中提出了一个新的特征选择方法, 该方法从特征对类间文档分散程度的影响度和特征对类内文档分散程度的影响度两个方面来考虑, 能够较全面地对所选特征进行评价, 从而使得选择的特征具有较高的代表性, 能较好

地提高分类系统的性能.

2 相关基础定义

在文本分类中,文本常用空间向量模型来表示,其中所牵涉到的基本定义如下^[6]:

定义 1 特征词 文本通常是由一些基本语言单位,例如字、词、词组或短语等所组成的集合,这些基本的语言单位统称为特征词,通常把特征词简称为特征.对文本 T 可以用特征集表示,即 $T(T_1, T_2, \dots, T_n)$, 其中 T_k 是特征词, $1 \leq k \leq n$.

定义 2 特征权重 对于含有 n 个特征的文本 $T(T_1, T_2, \dots, T_n)$, 其中 T_k 是特征, $1 \leq k \leq n$, T_k 常常被赋予一定的权重 W_k , 表示它在文本中的重要程度, 即 $T(T_1, W_1; T_2, W_2; \dots; T_n, W_n)$. 有时在特征确定时, 文本常简记为 $T(W_1, W_2, \dots, W_n)$.

定义 3 向量空间模型 (Vector Space Model, VSM) 给定一文本 $T(T_1, W_1; T_2, W_2; \dots; T_n, W_n)$, 通常特征 $T_k (1 \leq k \leq n)$ 在文本中是可以重复出现多次的, 并且按照一定的次序出现的, 这就大大增加了对本分析的难度. 为了便于简化分析, 可以暂时不考虑 T_k 在文本中出现的次序, 并且还认为各个特征是互不相同的, 这样就可以把 $T_1, T_2, \dots, T_n$ 看成一个 n 维坐标系, 而 $W_1, W_2, \dots, W_n$ 为相应的坐标值, 因而 $T(W_1, W_2, \dots, W_n)$ 被看成是 n 维空间中的一个向量, 称 $T(W_1, W_2, \dots, W_n)$ 为文本 T 的向量表示.

定义 4 文本特征向量 在空间向量模型中, 每一个文本都用一个向量来表示, 其元素是由特征词的权重组成, 该向量就称之为该文本特征向量. 文本特征向量是文本的一个特征表示, 在某种意义上可以完全代表该文档, 例如, 有 n 个特征词的文本 T , 其特征向量表示为 $T(W_1, W_2, \dots, W_n)$.

定义 5 中心向量 如果一个文本集有 m 篇文本, 其中 $T_i (i = 1, 2, \dots, m)$ 代表第 i 篇文本, 文本集共有 n 个特征词、 p 个类别, 则整个文本集的中心向量为:

$$C = (1/m) \sum_{i=1}^m T_i \quad (1)$$

第 k 个类别的向量中心为:

$$C_k = (1/m_k) \sum_{j=1}^{m_k} T_j \quad (2)$$

其中 m_k 代表属于类别 k 的文本数. 这两个中心向量

有如下关系: $C = (1/m) \sum_{i=1}^p m_i \times C_i$.

3 两种特征影响度及其特征重要性

特征选择的目的是要找到一个较具代表性的特征子集, 让该特征子集来表示原来的文档集. 在经典的线性判别分析中, 类间分布矩阵和类内分布矩阵是两个常用的目标函数, 其中类间分布矩阵表示的是各类别文档之间的离散程度, 类内分布矩阵表示类内文档之间的离散程度. 通常来说, 如果各类别文档之间的离散程度越高、类内文档越密集, 那对分类也就越有利. 在这种情况下, 特征选择就是要找到这样一个特征子集: 如果用这个特征子集来表示训练文本集, 那么该文本集中各类别文档之间的离散程度较大, 而类内文档之间的离散度较小也即集中度较大. 此时, 一个特征对类间文档分散程度的影响度越大越好, 而它对类内文档分散程度的影响度是越小越好. 基于这个思想, 论文定义了两种特征影响度.

定义 6 特征对类间文档分散程度的影响度, 该程度越大, 表明该特征区分类别的能力也就越强. 其公式表示如下:

$$\text{Influence-degree}_1(i) = (1/p) \sum_{j=1}^p (C(i) - C_j(i))^2 \quad (3)$$

式中 p 表示类别个数, $C(i)$ 表示第 i 个特征在整个文本集中心向量中的权重, $C_j(i)$ 表示第 i 个特征在类别 j 的中心向量中的权重.

定义 7 特征对类内文档分散程度的影响度, 该程度越小, 表明该特征的对类的代表性也就越强. 其公式表示如下:

$$\text{Influence-degree}_2(i) = (1/m) \sum_{j=1}^p \sum_{k=1}^{m_j} (C(j) - T_{jk}(i))^2 \quad (4)$$

式中 m 表示文本集中文档总数, p 表示类别个数, $C_j(i)$ 表示第 i 个特征在类别 j 的中心向量中的权重, $T_{jk}(i)$ 表示第 i 个特征在类别 j 的文本集中的第 k 个文本向量中的权重.

在上述两个特征影响度的基础上, 可以得到特征的总体重要性, 其表示公式如下:

$$\text{Importance-degree}(i) = \frac{\text{Influence-degree}_1(i)}{\text{Influence-degree}_2(i)} \quad (5)$$

4 新的特征选择方法

依据式(3)可以设计一个新的特征选择方法, 其简单描述如下:

Input: 原始特征向量和类别集合

Output: 一个特征子集 Q

Step1: 根据式(1)计算文本集的中心向量 C ;

Step2: 根据式(2)计算各类别的中心向量 C_k ;

Step3: 根据式(3)计算各特征对类间文档分散程度的影响度 $\text{Influence-degree}_1(i)$;

Step4: 根据式(4)计算各特征对类内文档分散程度的影响度 $\text{Influence-degree}_2(i)$;

Step5: 根据式(5)计算各特征的总体重要性 $\text{Importance-degree}(i)$;

Step6: 选择重要度分值最大的前 P 个特征组成特征子集 Q 并将它输出。

5 实验验证

5.1 实验语料库

文中选用的语料库是复旦大学中文文本分类语料库,其下载网址为:http://www.nlp.org.cn/categories/default.php?cat_id=16. 它包含 20 个类别,分为训练文档集和测试文档集两个部分。去除部分重复文档和损坏文档后,共保留有文档 14 378 篇,其中训练文档 8 214 篇,测试文档 6 164 篇,跨类别的重复文档没有考虑,即一篇文档只属于一个类别。该语料库中的文档的类别分布情况是不均匀的。其中,训练文档最多的类 Economy 有 1 369 篇训练文档,而训练文档最少的类 Communication 有 25 篇训练文档;同时,训练文档数少于 100 篇的稀有类别共有 11 个。训练文档集和测试文档集之间互不重叠。

5.2 实验环境

采用中科院计算所的“汉语词法分析系统 ICTCLAS”进行分词处理;使用 Weka 工具进行分类实验,该工具是纽西兰的 Waikato 大学开发的数据挖掘相关的一系列机器学习算法,它包括数据预处理、分类、回归分析、聚类、关联规则、可视化等工具,网址为 <http://www.cs.waikato.ac.nz/>,使用 MATLAB 7.0 进行数值计算。

5.3 实验所用分类器及其评价标准

实验主要用来比较本文特征选择方法与三种经典的特征选择方法:互信息(MI)、信息增益(IG)、 χ^2 统计量(CHI)对分类系统性能的影响。因此,实验中在这四种特征选择方法后使用同一分类器。由于 KNN 分类器实现简单,本实验就选择该分类器。其中分类器参数 K 设置为 10,两文本向量之间的夹角余弦作为向量间的距离。为了检测这四个方法在特征个数变化情况下的 KNN 分类性能,采用微平均 F_1 值和宏平均 F_1 值作为分类性能评价标准。

5.4 实验结果及其解释分析

在所选的语料库上,对所选择的 20 个类别的微平均 F_1 值和宏平均 F_1 值进行了计算,分别在不同的特征数目下检测分类性能,结果如图 1、图 2 所示。

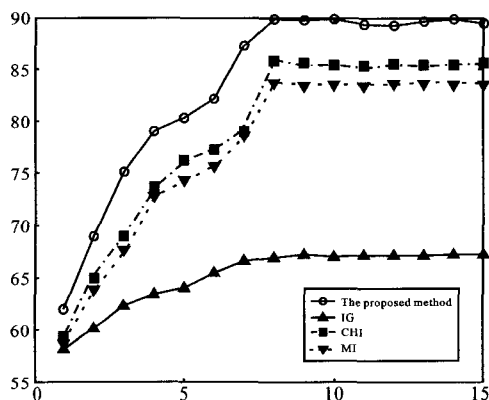


图 1 不同特征个数情况下微平均 F_1 值

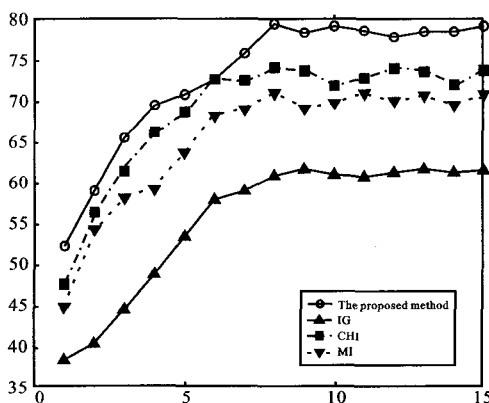


图 2 不同特征个数情况下宏平均 F_1 值

图 1 和图 2 横轴中从 1 到 15 这 15 个数字分别代表的特征个数依次为:50、100、200、500、800、900、1000、1500、2000、2500、3000、3500、4000、4500、5000。纵轴分别代表相应的微平均值 F_1 和宏平均 F_1 值,单位为%。

在特征个数变化的情况下,考察分类模型的性能变化情况,可以很好地反映一个分类模型对数据样本变化的敏感程度。从图 1 可以看出,随着特征个数的增加,微平均 F_1 值逐渐增加,并达到一个相对稳定的水平。从图 2 可以看出,宏平均 F_1 值也随着特征个数的增加而增加,但是波动相对较大,这是由于本文所选语料库类别样本分布极其不均匀所致。

由图 1 和图 2 可以看出:KNN 分类器在文中方法所选的前 1500 个特征上性能最佳,微平均 F_1 值和宏平均 F_1 值约分别为 90% 和 80%;KNN 分类器在 IG 方法所选的前 2000 个特征上性能最佳,微平均 F_1 值和宏平均 F_1 值约分别为 67% 和 62%;

KNN 分类器在 CHI 方法所选的前 1500 个特征上性能最佳,微平均 F_1 值和宏平均 F_1 值约分别为 86% 和 74%; KNN 分类器在 MI 方法所选的前 1500 个特征上性能最佳,微平均 F_1 值和宏平均 F_1 值约分别为 84% 和 71%。因此,这四种特征选择方法的总体效果排序为:文中方法>CHI>MI>IG。产生这个结果的主要原因在于:文中方法既考虑了特征对类间文档分布的影响又考虑了特征对类内文档分布的影响,从而使得文中方法不受所选语料库类别文档分布的影响,能够对所选特征进行综合考虑,因此所选的特征具有较好的代表性;MI 方法在选择特征时仅考查了所选特征存在的情况,而 CHI 方法既考查了特征发生的概率又考查了特征没发生的概率,因此 CHI 方法要优于 MI 方法;由于 IG 方法对样本分布情况极其敏感,如果在样本分布不均匀的情况下使用该方法,那么它所选择的特征集代表性就较差,而文中所选语料库类别文本分布是极其不均匀的,所以在此情况下该方法效果最差。

6 结束语

文中定义了特征对类间文档分散程度的影响度以及特征对类内文档分散程度的影响度,并以这两个影响度为基础设计了一个新的特征选择方法。通过在复旦大学分类语料库上的实验表明,同三种经

典特征选择方法 MI 和 CHI 以及 IG 相比,文中方法所选择的特征具有较好的代表性,能够为后续的知识发现算法减少时间与空间复杂性,从而使得文中方法在文本分类中有一定的使用价值,同时也为文本特征选择提供一种思路。

参考文献:

- [1] Fabrizio S. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34 (1): 33-47.
- [2] Zhu Haodong, Zhao Xianghui, Zhong Yong. Feature selection method combined optimized document frequency with improved RBF network[C]// Proc. of 5th International Conference, ADMA 2009. China, Beijing, 2009:796-803.
- [3] 陈平,刘晓霞,李亚军. 文本分类中改进型互信息特征选择的研究[J]. 微电子学与计算机, 2008, 25(6): 194-197.
- [4] 张建兵,戴新宇,陈家骏. 文档分类之特征选择方法的实验比较[J]. 广西师范大学学报:自然科学版, 2008, 26 (3): 181-185.
- [5] 朱颢东,钟勇. 一种新的基于多启发式的特征选择算法[J]. 计算机应用, 2009, 29(3): 849-851.
- [6] Roga M, Yang Y. High performance feature selection for text categorization [C]// The CIKM - 02. USA, Mclean, 2002:34-39.

作者简介:

胡 强 男, (1974-), 高级企业信息管理师。

(上接第 64 页)

钥空间非常大,使加密者可以随意地选择密钥,这样的选择可以使算法有着几乎一次一密的安全特性。

4 结束语

文中构造了二维 Logistic 离散混沌系统,并将所生成的混沌序列应用于图像加密。该算法混沌系统生成的序列具有很好的加密效果,及更强的伪随机特性。同时具有很大的密钥空间,加强了算法的安全性和保密性。两个 x, y 实数值混沌序列非常适用于彩色图像的分层加密以及多幅图像加密,提高算法的加密效率。

参考文献:

- [1] 焦占亚,胡子濮. 一种密钥的设计和管理办法 [J]. 微电子学与计算机, 2004, 21(10): 158-160.
- [2] 丁玮,齐东旭. 数字图像变换及信息隐藏与伪装技术 [J]. 计算机学报, 1998, 21(9): 838.
- [3] 周蕊,王震,毛鹏伟,等. 一类三维混沌系统的自适应同

步[J]. 陕西科技大学学报, 2008, 26(4): 123-126.

- [4] 王震,毛鹏伟. 一类三维混沌系统的分叉及稳定性分析 [J]. 动力学与控制学报, 2008, 6(1): 16-21.
- [5] 闵乐泉,杨森,张先华. 基于广义混沌同步的数字图像隐藏方案[J]. 北京科技大学学报, 2003, 25(5): 477.
- [6] 叶永伟,杨庆华,王颖玉. 用混沌序列对数字图像进行魔方加密[J]. 浙江工业大学学报, 2003, 31(2): 173.
- [7] 易开祥,孙鑫,石教英. 一种基于混沌序列的图像加密算法 [J]. 计算机辅助设计与图形学学报, 2000, 12(9): 672.
- [8] 余香敏,焦占亚. 一种基于 DWT 与 SVD 的数字图像水印算法[J]. 微电子学与计算机, 2008, 25(10): 199-201.

作者简介:

周 蕊 女, (1979-), 硕士研究生, 讲师. 研究方向为网络安全与计算机应用。

于晓明 女, (1965-), 硕士生导师, 副教授. 研究方向为网络信息安全。

焦占亚 男, (1947-), 硕士生导师, 教授. 研究方向为网络信息安全。