一种新的基于软集合理论的文本分类方法

袁鼎荣[1,2], 谢扬才[2], 陆广泉[2], 刘 星[2]

(1. 北京工业大学 计算机学院, 北京 100124; 2. 广西师范大学 计算机科学与信息工程学院, 广西 桂林 541004)

摘 要: 文本分类技术是文本信息处理的核心技术之一, 主要包括文本的向量模型表示、文本特征选择和分类器训练三大过程。本文提出了一种混合(EIBA+ DHChi2) 特征选择算法, 并将所获取的特征作为软集合理论中的参数集进行文本分类, 从而建立了一种新的基于软集合理论的文本分类技术。实验表明查准率与查全率比原有算法都有所提高, 说明新的基于软集合理论的文本分类算法是有效的。

关键词: 文本分类; 特征选择; Chi2 假设检验; 独立度; 模糊软集合

中图分类号: TP391 文献标识码: A 文章编号: 1001-6600(2011)01-0129-04

文本分类技术是文本信息处理的核心技术之一, 主要包含文本的向量模型、特征选择和分类器训练等 3 个过程, 其任务是指依据文本的内容, 将文本判分为预先定义好的类别。虽然, 已经建立了许多可用的文本分类系统, 并取得了一定的成果, 但仍需继续完善。比如: 文本特征的选择和抽取技术不完善, 导致文本分类结果不理想。

特征选择是从确定的特征空间中选取能够充分代表文档内容的特征子集的过程, 是文本分类中的关键, 目前已经存在许多相关工作, 如: 文献[1] 基于独立性理论、文献[2-4] 基于贝叶斯粗糙集方法、文献[5] 基于粗糙集和灰色关联度的综合、文献[6] 结合优化的文档频和 PA 方法进行文档特征选择。

软集合理论是 1999 年提出的处理模糊对象的数学工具, 近年来许多学者对其理论和应用进行研究。如文献[7] 提出一种基于软集合文本分类方法。

本文在文本特征选择阶段采用一种新的基于独立度和齐性 Chi2 假设检验的特征选择方法, 将所获取的特征作为软集合理论中的参数集进行文本分类, 从而建立一种新的文本分类技术。

1 相关理论基础

1.1 齐性 Chi2 假设检验

齐性 Chi2 假设检验的目的是检验随机变量在 m 次试验中的独立性假设。对于 n 个 m 项式试验可以用 $n\times m$ 的联表表示(表 1)。其中: m 表示试验次数, n 表示每次试验中随机变量的个数。

表 1 m 个 n 项式实验的 $m\times n$ 列联表[1]

Tab. 1 $m\times n$ table

变量	C_1	C_2	…	C_m	总数
t_1	O_{11}	O_{12}	…	O_{1m}	$O_{1.}$
t_2	O_{21}	O_{22}	…	O_{2m}	$O_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮
t_n	O_{n1}	O_{n2}	…	O_{nm}	$O_{n.}$
总数	$O_{.1}$	$O_{.2}$	…	$O_{.m}$	$O_{..}$

1. O_{ij} 表示第 j 个试验的第 i 个观察值; $O_{i.}$ 表示第 i 个随机变量在所有 m 次试验中的观察值之和; $O_{.j}$ 表示第 j 次试验中所有 n 个观察值之和; $O_{..}$ 表示所有试验中的所有观察值之和。

零假设 H_0 为: 随机变量 t_i 与试验无关, 有 $O_{i1}= O_{i2}= \cdots = O_{im}$。它的检验统计量表示为:

收稿日期: 2010-12-20

基金项目: 国家自然科学基金重大研究计划培育项目(90718020); 澳大利亚 ARC 项目(DP0667060)

通讯联系人: 袁鼎荣(1967-), 男, 广西全州人, 广西师范大学副教授, 硕士。E-mail: dryuan@mailbox.gxnu.edu.cn

$$x^2 = \sum_{j=1}^{m}\sum_{i=1}^{n}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}。 \tag{1}$$

当 H_0 为真时, 第 j 个多项式实验的第 i 个观察值的期望可表示为:

$$E_{ij} = \frac{O_{i\cdot}O_{\cdot j}}{O_{\cdot\cdot}}, \tag{2}$$

因此式(1)近似服从自由度为$(m-1)(n-1)$的 Chi2 分布。式(1)的值越大, 相关性越高。

1.2 随机事件的独立度

我们定义随机事件独立度如下: 设样本空间中的 2 个事件 A 和 B, 我们称$|P(A,B) - P(A)P(B)|$的值为随机变量 A 和 B 之间的独立程度(依赖度)。

1.3 软集合相关理论

定义 **1** 设 U 是给定的论域, E 是一个参数集, 一个集合对(F,E)被称为域 U 上的一个软集合(soft set), 当且仅当 F 是 E 到所有 U 子集中某集合的映射, 如 $F: E \to P(U)$, 其中, $P(U)$ 是 U 的幂集。软集合是 U 的子集的一个参数族。该参数族中每个集合 $F(e)$ $(e \in E)$ 可以看成软集合(F,E)的 e 个元素的集合, 或者是软集合的 e 个相似元素的集合。

定义**2** $P(U)$ 为 U 上所有模糊集, E 为一参数集, $A_i \in E$, 集合对(F_i, A_i)被称为 U 上的一个模糊软集合, 当且仅当 F_i 是 A_i 到 $P(U)$ 的一个映射, 如 $F_i: A_i \to P(U)$。

2 基于 DHChi2 与 EIBA 的特征选择算法

2.1 DHChi2(Distributed Homogeneous Chi2) 特征选择方法

按照分布式思想, 将假设检验的零假设分解成多个零假设, 使表 1 中每一行对应的观察值代表一个零假设, 那么各行对应的观察值之间可以通过其对多个试验的独立度(依赖度) 进行比较筛选确定。其方法描述如下:

设 t_i 为特征空间中的第 i 个特征, $c_1, c_2, \cdots, c_i, \cdots, c_m$ 对应于 m 个文档类, 假设 t_i 文档类无关, 则通过式(3) 计算统计量, 统计量越大则 t_i 对文档依赖性越强。

$$x^2(t_i) = \sum_{j=1}^{m}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{3}$$

其中

$$E_{ij} = \frac{O_{i\cdot}}{m}。 \tag{4}$$

2.2 EIBA(Event IndependenceBased Approach) 特征选择方法

对于特征 t_i 和文档类 c_j, $P(t_i) \neq 0$ 且 $P(c_j) \neq 0$, t_i 与 c_j 相互独立的充分必要条件为 $P(t_i, c_j) = P(t_i)P(c_j)$。则 t_i 相对 c_j 的独立程度或依赖度表示为式(5), 评价函数表示为式(6):

$$\mathrm{EIBA}(t_i, c_j) = |P(t_i, c_j) - P(t_i)P(c_j)|, \tag{5}$$

$$\mathrm{EIBA}_{\max}(t_i) = \max_{j=1}^{m}\{\mathrm{EIBA}(t_i, c_j)\}, \tag{6}$$

评价函数值越大, t_i 对文本 c_j 具有越好的代表性。

2.3 混合(**EIBA+ DHChi2**)特征选择算法

我们给出 EIBA 和 DHChi2 的组合方法, 记为 EIBA + DHChi2。

算法 1 EIBA + DHChi2。

输入: n 个初始特征和 m 个文档类

输出: k 个特征属性

步骤: ①基于 EIBA, 利用式(5)、(6) 先计算每个特征对各个文档类的依赖度;

②计算表 1 中的观察值 O_{ij}, 累计行、列的观察值之和;

③基于 DHChi2, 利用式(3) 作为评价函数选择特征子集;

④当 $x^2(t_i) \geq k$, 选取符合条件的 k 个特征子集。

3　改进的基于软集合的文本分类方法

3.1　文本的软集合表示

文本分类时, 选择的文本特征$\{C_1, C_2, \cdots, C_k\}$, $\{T_1, T_2, \cdots, T_k\}$为待分类的文本样例。构造文本软集合如表 2。

表 2　软集合表

Tab. 2　Soft set table

U	C_1	C_2	…	C_k
T_1	0.6	0.8	…	0.9
T_2	0.3	0.4	…	0.8
⋮	⋮	⋮	⋮	⋮
T_k	0.4	0.3	…	0.6

表 3　软集合对照表

Tab. 3　Soft set compared table

U	T_1	T_2	…	T_k
T_1	4	4	…	3
T_2	0	4	…	2
⋮	⋮	⋮	⋮	⋮
T_k	1	2	…	4

3.2　构造软集合$(\boldsymbol{F}, \boldsymbol{E})$对照表

一个含有 n 个对象的软集合(F, E), 其对照表由 n 行 n 列构成, 其中, n 为软集合中对象个数, 单元格记为 C_{ij}, 其取值为: 在软集合(F, E)中对象 x_i 的取值大于等于 x_j 值所对应的参数个数, 则表 2 所对应的对照表如表 3 所示。

3.3　文本分类算法

算法 2　新的基于软集合理论的分类算法。

输入: 算法 1 选择的 k 个特征属性与待选择的文本特征向量

输出: 待选择的文本所属类别

步骤　①由 C 个分类类别 M 个特征属性构造一个 $C \times M$ 软集合图表;

②给定一个待分类的文本, 计算出其符合条件的 k 个特征向量 $\boldsymbol{V}_f$;

③由 $\boldsymbol{V}_f$ 与软集合表构成新的表格, 单元格 $\boldsymbol{V}_{ij} = 1 - \dfrac{|\boldsymbol{V}_{ij} - \boldsymbol{V}_f|}{\max(\boldsymbol{V}_{ij})}$;

④计算单元格的权重 $S_i = r_i - r_j$ 其中 r_i、r_j 为行列的累加和;

⑤则 $C = \max S_i$ 就是待分类文本的所属类别。

4　实验结果分析

我们采用新浪网站提供的 1 000 篇文档作为训练集和测试集, 从中提取知识并对分类算法进行评价, 训练集 800 篇文档, 测试集 200 篇文档, 其中的文本涉及 8 个主题: 新闻、体育、财经、娱乐、科技、汽车、房产、生活。评价分类性能[8-9]的 2 种常用指标是查准率和查全率。为了评估算法在整个数据集上的性能, 分别与 KNN 文本分类算法和基于软集合文本分类算法的性能进行比较(表 4)。

表 4　3 种算法性能比较

Tab. 4　Three algorithms compare

特征数	KNN		软集合方法		新的软集合方法	
	查全率/%	查准率/%	查全率/%	查准率/%	查全率/%	查准率/%
100	88.84	86.58	89.87	87.36	90.13	88.56
400	90.12	87.62	91.98	89.19	92.36	90.45
1 000	93.78	90.41	93.35	91.21	94.56	92.01
1 500	93.89	91.66	93.95	92.02	94.78	93.52
2 000	93.12	91.35	93.89	91.89	94.21	93.09

从表 4 实验数据可以得出以下结论:

①新的基于软集合理论的文本分类算法与 KNN 算法、基于软集合理论算法相比, 查准率和查全率有所提高。

②当特征数达到一定数量时, 3 种算法的性能均有所下降。

5 结 语

本文提出一种混合(EIBA + DHChi2)特征选择算法, 将所获取的特征作为软集合理论中的参数集进行文本分类, 从而建立一种新的基于软集合理论的文本分类技术。实验表明其查准率与查全率比原来算法都有所提高, 说明本文提出的新的基于软集合理论文本分类算法是有效的。特征数在什么情况下使得算法查准率和查全率最大, 这将是我们下一步要做的研究。

参 考 文 献:

[1] 冯霞, 刘志辉, 田继存. 基于独立性理论的文本分类特征选择方法[J]. 计算机工程, 2010, 36(12): 22-27.
[2] 朱颢东, 钟勇. 基于贝叶斯粗糙集的文本特征选择方法[J]. 河南师范大学学报: 自然科学版, 2009, 37(4): 31-35.
[3] 袁野, 封化民. 基于 Vague 集的 Web 内容安全文本分类[J]. 广西师范大学学报: 自然科学版, 2010, 28(1): 147-152.
[4] 吕小勇, 石洪波. 基于粗糙集的多标签文本分类算法[J]. 广西师范大学学报: 自然科学版, 2009, 27(3): 150-153.
[5] 黄玉龙, 王翰虎, 陈梅. 基于粗糙集理论的 KNN 分类[J]. 广西师范大学学报: 自然科学版, 2007, 25(4): 75-79.
[6] 朱颢东, 钟勇. 结合优化的文档频和 PA 的特征选择方法[J]. 计算机应用研究, 2010, 27(1): 36-38.
[7] 洪智勇, 秦克云. 基于模糊软集合理论的文本分类方法[J]. 计算机工程, 2010, 36(13): 90-92.
[8] 柴玉梅, 朱国重, 咎红英. 基于质心的文本分类算法[J]. 计算机工程, 2009, 35(20): 83-85.
[9] 苏金树, 张博锋, 徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报, 2006, 17(5): 1848-1859.

A New Text Classification Approach Based on Fuzzy Soft Set Theory

YUAN Ding-rong[1,2], XIE Yang-cai[2], LU Guang-quan[2], LIU Xing[2]

(1. College of Computer Science and Technology, Beijing University of Technology, Beijing 100124, China;
2. College of Computer Science and Information Technology, Guangxi Normal University, Guilin Guangxi 541004, China)

Abstract: Text classification is one of the key techniques in text information process, which includes how to establish vector model, select feature and train classifier. EIBA and DHCHi2 are integrated to select the features of text and the features are used as parameters in a fuzzy soft set theory. Then a new technique of text classification is established based on a fuzzy soft set. Experiments show that the technique is effective, and the ratios of accuracy and recall are improved comparing with other methods.

Key words: text classification; feature select; Chi2 hapothesis testing; independent degree; fuzzy soft set

(责任编辑 王龙杰)