

分布权值调节概率标准差的文本分类方法

焦庆争^{1,2}, 蔚承建¹

(1. 南京工业大学 电子与信息工程学院, 南京 210009; 2. 安徽师范大学 信息管理中心, 安徽 芜湖 241000)

(qzjiao@mail.ahnu.edu.cn)

摘要:针对文本分类问题,基于特征分布评估权值调节特征概率标准差设计了一种无须特征选择的高效的线性文本分类器。该算法的基本思路是使用特征概率标准差量化特征在文档类中的离散度,并作为特征的基础权重,同时以后验概率的 Beta 分布函数为基础,运用概率确定性密度函数,评估特征在类别中的分布信息得到特征分布权值,将其调节基础权重得到特征权重,实现了线性文本分类器。在 20Newsgroup、复旦中文分类语料、Reuters-21578 三个语料集进行了比较实验,实验结果表明,新算法分类性能相对传统算法优势显著,且稳定、高效、实用,适于大规模文本分类任务。

关键词:文本分类;特征概率标准差;特征离散度;特征分布;Beta 概率密度函数;自然语言处理

中图分类号: TP18 **文献标志码:** A

Text categorization approach based on probability standard deviation with evaluation of distribution information

JIAO Qing-zheng^{1,2}, WEI Cheng-jian¹

(1. College of Electronic and Information Engineering, Nanjing University of Technology, Nanjing Jiangsu 210009, China;

2. Information Management Center, Anhui Normal University, Wuhu Anhui 241000, China)

Abstract: For text categorization, an approach was introduced to construct the simplest linear classifier, in which the feature weight was computed by probability standard deviation of features as a base line weight regulated with features distributed parameters. In the assessment process of weighting, the probability standard deviation was considered as feature base weighting to quantify dispersion degree of feature, while distributed parameters were evaluated by using beta probability density functions to measure feature distributed information. In the experiments, 20Newsgroup, Fudan Chinese evaluation data collection and Reuters-21578 were used to evaluate the effectiveness of the techniques proposed in this paper, respectively. The experimental results show the method can improve significantly the performance for text categorization, and is simple, stable and suitable for large-scale text categorization.

Key words: text categorization; probability standard deviation of feature; dispersion degree of feature; feature distribution; Beta probability density function; natural language processing

0 引言

自动文本分类是一种有监督的学习任务^[1],即根据已分类的训练文档集合,对未分类文档分配类标签。近年来,越来越多的统计理论和机器学习方法被运用到文本自动分类,如贝叶斯概率方法、K 近邻(K-Nearest Neighbor, KNN)分类法、支持向量机方法、LLSF 算法、Boosting 算法等,使文本分类研究日趋成熟,成为一项实用技术。关于传统算法的性能比较,文献[1]中做了详细论述。一个设计优良的文本分类算法除了有较高的分类性能外还必须解决特征高维性、语料集的不均匀性、算法执行效率等问题。传统统计分类方法各有优缺点,朴素贝叶斯方法(Naive Bayes, NB)的基本思想是利用单词和类之间的联合概率来估计给定文档属于类的概率,NB 分类器不需要计算单词之间的联合分布概率,使得其速度远快于非朴素贝叶斯算法(达到指数级复杂度)。但 NB 在不均匀的语料下分类性能不能令人满足,限制了其普及。SVM 和 KNN 具有较高分类性能,但它们都存在执行效率上的问题。

一直以来,SVM 性能超过了其他算法,但它需要 cn^2 (n 为训练样本数, c 为算法依赖的常量)训练时间。KNN 不需要训练,但一个测试文档却要扫描所有训练样本。

近两年来,国内外研究主要是对传统方法的改进和算法优化,例如通过特征选择提高算法的性能和执行效率。也有较少研究提出新的分类方法,但多数实验不仅语料集较少,而且语料集类别数较少、类别之间文档数分布较均匀,结果不具代表性。文献[2]基于假设边界提出一种新型中心文本分类器,并在 20Newsgroup、Reuters-21578、WebKB 等国内外常用文本分类评测的语料集上进行了实验,取得了较好的分类性能和效率。

随着 Internet 的高速发展,要处理的 Web 文档与日俱增,构建简单高效的线性文本分类方法是解决问题的理想途径。本文通过评估特征在文档类别中的分布信息,提出一种基于特征分布权值调节特征概率标准差的线性文本分类方法(DWPSD),较好地解决了文本分类中的各种问题。

收稿日期:2009-06-24;修回日期:2009-08-09。

基金项目:国家自然科学基金资助项目(60703071);安徽省高校省级自然科学基金重点项目(KJ2009A63)。

作者简介:焦庆争(1974-),男,安徽安庆人,讲师,硕士研究生,主要研究方向:自然语言处理;蔚承建(1957-),男,江苏南京人,教授,博士,主要研究方向:可信计算。

1 基于分布权值调节概率标准差分类模型

1.1 特征概率标准差

在文本分类中,特征在文档类中出现的频率越不均匀,即特征分布得越离散,往往特征对类别判定越重要,利用这一性质可以考查特征在分类中的重要程度。离散度通常可以用标准差或方差来计算,本文使用特征在文档类中的概率标准差对特征重要性进行定量描述,此特征概率标准差将作为特征的基本权重参与文本分类。

首先,我们给出特征频率的定义。传统特征频率的计算是统计特征在文档类中出现的次数,使分类某种程度上依赖于长文档。为了消除长文档对学习算法的负面影响,对单个文档中特征频率做归一化处理是必要的。将去除停用词后剩下的单词组成用于分类的特征,特征频率的计算公式可描述如下:

$$TF_{ij} = \sum_k Bit(k, i) \frac{\text{文档 } k \text{ 中特征 } j \text{ 数}}{\text{文档 } k \text{ 中特征总数}}$$

其中: i 是文档类序号, j 是特征序号, k 是训练文档序号, $Bit(k, i)$ 函数是判断文档 k 是否属于 i 类,是则返回 1, 否则返回 0。

特征 j 在类 i 中的特征概率可用如下公式计算:

$$p_{ij} = TF_{ij} / \sum_i TF_{ij}$$

由于所有类特征概率和为 1, 类别总数的倒数即是特征概率平均值, 特征概率标准差计算如下:

$$psd_j = \sqrt{\frac{1}{c} \sum_{i=1}^c [p_{ij} - (1/c)]^2}$$

其中 c 是文档类数目。

1.2 特征类分布权值

线性文本分类算法如果能较好评估文档类对特征的信任度, 将此信任度作为特征分布权值参数调节特征基础权重得到特征权重即能较好地衡量特征对各文档类的贡献。特征在文档中的出现有复杂的同现关系, 大部分特征对分类是有益的, 即使是那些对分类没有作用的特征, 算法本身也应该能自动予以甄别, 特征选择算法是以牺牲性能为代价的。

特征分布权值既要反映同一特征在不同类别中的分布又要反映不同特征在同类中的分布。本文从主观逻辑框架出发分析特征分布权值的评估问题。Jøsang 等人提出基于主观逻辑 (Subjective Logic) 的信任模型^[3-5], 引入了事实空间 (Evidence space) 和观念空间 (Opinion Space) 概念来描述和度量信任关系。Jøsang 等人以描述二项事件的后验概率的 Beta 分布函数为基础, 给出了一个由观察到的肯定事件数和否定事件数来确定概率确定性密度函数 $pcdf$, 并以此为基础计算实体事件概率的可信度。设 θ 为事件概率, r 和 s 分别表示观测到肯定事件和否定事件数, 则 $pcdf$ 公式表述为:

$$f(\theta | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}$$

函数中的 θ, α 和 β 三个变量定义如下:

$$\alpha = r + 1$$

$$\beta = s + 1$$

$$\theta = \alpha / (\alpha + \beta)$$

Jøsang 模型是基于二项事件的, 无法直接应用于文本分类。为了解决多类事件的信任评估问题, 可以根据各文档类

构造多个二项事件, 每个二项事件针对某个特定文档类。例如针对类 i , 二项事件肯定事件为特征在类 i 中的特征频率, 否定事件为在非类 i 中的特征频率。但在实验中我们发现, 根据以上事件计算出来的概率确定性密度作为特征分布权值时存在两个问题: 第一是特征在类 i 中的特征频率由于否定事件的范围过大而不能有效调节分布权值, 这在文档类数很多时尤其明显; 第二是概率确定性密度是非规范化的结果^[6], 分类时, 某些对文档贡献大的特征由于过高的密度值而掩盖了其他特征的密度, 不能发挥整体特征的作用。因此, 在实际模型构造中引入两类二项事件空间。第一类事件空间是显著事件空间, 对类 i , 二项事件肯定事件为特征在类 i 中的特征频率, 否定事件为 0; 第二类事件空间是平均事件空间, 即对类 i , 以所有类的特征频率的平均值为事件总数, 根据显著空间中肯定与否定事件概率计算二项事件的肯定与否定事件数。

首先, 为了消除训练文档集的先验文档频率对分类性能的影响, 我们利用它对特征频率进行调节, 公式如下:

$$\widetilde{TF}_{ij} = TF_{ij} / c_i$$

其中 c_i 为类 i 的文档数。

在显著事件空间, 事件的概率确定性密度定义如下:

$$pcdf_{ij} = f(\theta | \widetilde{TF}_{ij} + 1, 1)$$

其中 $\theta = \widetilde{TF}_{ij} + (1 / \widetilde{TF}_{ij}) + 2$ 。

在平均事件空间, 事件的概率确定性密度定义如下:

$$\widetilde{pcdf}_{ij} = f(\theta | (\widetilde{TF}_j + 2)\theta, (\widetilde{TF}_j + 2)(1 - \theta))$$

其中 $\widetilde{TF}_j = \frac{1}{c} \sum_{i=1}^c \widetilde{TF}_{ij}$ 。

不同文档类对特征的信任度的规范化度量, 也即特征的分布权值可用如下公式计算:

$$dw_{ij} = (pcdf_{ij} - 1) / (\widetilde{pcdf}_{ij} - 1)$$

1.3 线性分类器及时间复杂度

利用特征分布权值为参数调节概率标准差可以得到用于分类的特征权重, 计算如下:

$$W_{ij} = psd_j \times dw_{ij}$$

判定测试文档类别时, 根据特征出现数目, 采用特征权重加权求和即可求得测试文档的类别分类值, 计算如下:

$$DC_i = \sum_j n_j W_{ij}$$

其中 n_j 为测试文档中特征 j 出现的个数。

根据单标签分类、多标签分类的不同, 选取文档分类值较大的类作为测试类别:

文本分类时间复杂度分析包括训练阶段与测试阶段两个方面。在训练阶段, 本文方法需要两次遍历: 一次是遍历所有训练文档, 计算特征频率; 另一次是遍历所有特征, 计算特征权重。时间复杂度为 $O(n + m)$, 其中 n 为训练文档总数, m 为特征总数。在测试阶段, 算法只需根据测试文档中出现的特征对特征权重求和, 每个文档测试的时间复杂度为 $O(k)$, 其中 k 为测试文档特征数。

2 实验分析

2.1 实验方法

为了较全面测试算法对不同分类语料集的适应性, 本文

选择三个具有典型特征的语料集: Reuters-20Newsgroup (20Newsgroup)^[7]、复旦大学计算机信息与技术系国际数据库中心提供的中文分类语料集 (Fudan) 和 Reuters-21578。这 3 个分类语料被国内外学者广泛用于评测文本分类性能,便于评测数据的比较。这 3 个语料又各具特色,20Newsgroup 为单标签平衡英语语料,Fudan 为单标签不均匀中文语料,Reuters-21578 为多标签不均匀英语语料。Fudan、Reuters-21578^[8] 中文档数分布极不均匀,Fudan 语料为 20 个类别,最大类 1600 篇,最小类 25 篇;Reuters-21578 为 90 个类别,每个文档平均对应 1.3 个类,最大类有 2 877 个文档,82% 的类只有不到 100 个训练文档,33% 的类的文档数甚至少于 10。本文使用的语料均从相应官方网站下载,并使用了语料提供的所有文档类别及所有文档。所有语料均去除了停用词,英语语料做了词干处理,中文语料使用 Stanford 自然语言处理实验室的中文分词系统进行分词。

在 20Newsgroup、Fudan 语料集上,采用 NB 与 KNN 方法做比照实验,为了提高 NB 分类性能,本文采用了最大 LTC-TFIDF 加权处理,准确率比原始算法提高近 3 个百分点。

同时为了与更多的分类算法进行性能比较,本文引用了文献[1]在 Reuters-21578 语料上分类实验结果(文献[1]的实验涵盖目前主流文本分类方法,实验结果被广泛引用于文本分类性能)。

本文分类方法是无须特征选择的分类算法,但比照算法需要利用特征选择来提高分类性能及算法执行速度,因此本文使用信息增益 (Information Gain, IG)^[9] 统计量来进行特征选择,比较算法在不同特征规模下的性能表现。IG 统计量是根据某一特征在不考虑该特征的熵和考虑该特征后的熵的差值来选择预测能力较强的特征。计算公式如下:

$$\begin{aligned} Gain(t) = & - \sum_{i=1}^M P(c_i) \log P(c_i) - \\ & P(t) \left\{ - \sum_{i=1}^M P(c_i | t) \log P(c_i | t) \right\} - \\ & P(\bar{t}) \left\{ - \sum_{i=1}^M P(c_i | \bar{t}) \log P(c_i | \bar{t}) \right\} \end{aligned}$$

2.2 实验分析

本文分别采用微平均 F1 指标和宏平均 F1 指标评价算法的性能^[10]。

1) 20Newsgroup 语料的实验结果。

由于三个算法的微平均 F1 值和宏平均 F1 值十分接近,为了图表清晰直观,我们只绘出微平均 F1 值。图 1 显示, DWPSD 与 NB 性能明显高出 KNN。本文目标是构建无特征选择文本分类算法,因此我们以所有特征参与 DWPSD 分类获得的性能与 NB 与 KNN 方法各自取得最佳性能时的指标进行比较,结果, DWPSD 比 NB 高 1.4%, 比 KNN 高 7.8%。在此语料上,各算法在大规模特征参与分类时,分类准确率均未出现明显下降。

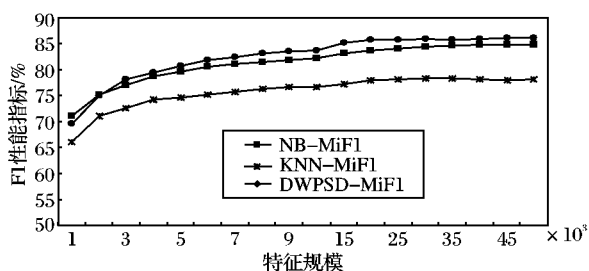


图 1 20NewsGroup 分类语料分类结果

2) 复旦中文分类语料的实验结果。

与 20Newsgroup 语料相比, NB 与 KNN 性能出现明显分化,图 2 显示,在各自取得最佳性能时, KNN 比 NB 微平均 F1 值高出 7.8%, 宏平均 F1 值高 18.3%, NB 严重忽视了小类的识别,导致宏平均 F1 指标过低。DWPSD 在所有特征参与分类时,微平均 F1 指标与宏平均 F1 指标分别比 NB 高 12.2%、36.1%, 比 KNN 高 4.4%、17.5%。DWPSD 在特征规模不断扩大的过程中,其性能一直上升趋势。

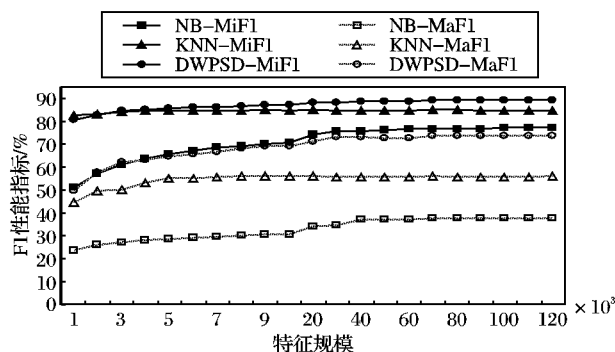


图 2 复旦中文分类语料分类结果

3) Reuters-21578 语料的实验结果。

此语料实验中,文档类标签确定采用阈值法,阈值确定使用的算法是最优截尾法 (score-based local optimization threshold)。图 3 显示了 DWPSD 在此语料上的分类性能,从图上我们看到,算法在较小规模特征时就达到了最佳性能,随着特征规模的不断扩大,无论是微平均 F1 指标还是宏平均 F1 指标,性能并没有出现明显下滑,基本上保持在最佳分类状态。

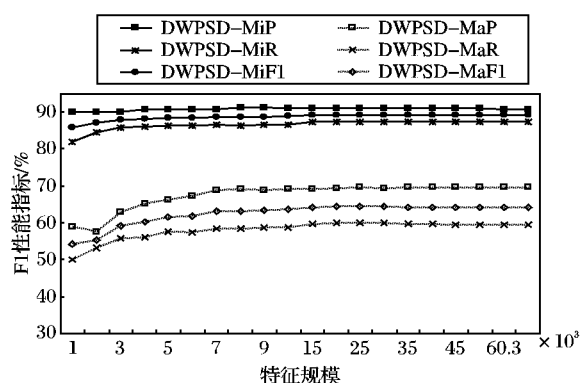


图 3 Reuters-21578 文本分类结果

文献[1]使用 Reuters-21578 语料对主流文本分类方法进行实验评价,并列出了 SVM、KNN、LLSF、NNET、NB 等算法性能的实验数据,为了与更多分类方法进行性能比较,本文引用文献[1]实验数据列于表 1 中。DWPSD 性能标为所有特征均参与分类时的指标,特征规模为 10 830。DWPSD 除微平均精度略低于 SVM 外,其他各项指标均优于其他方法。

表 1 Reuters-21578 分类性能比较

方法	MiR	MiP	MiF1	MaF1
DWPSD	0.8752	0.9074	0.8910	0.6420
SVM *	0.8120	0.9137	0.8599	0.5251
KNN *	0.8339	0.8837	0.8567	0.5242
LLSF *	0.8507	0.8489	0.8498	0.5008
NNET *	0.7842	0.8785	0.8287	0.3765
NB *	0.7688	0.8245	0.7956	0.3886

注:“*”数据来源于文献[1]。

本文还使用了 11-point average precision 评测标准对 DWPSD 分类结果进行评测,在所有特征均参与分类时,分类精度为 0.9132。

综合实验结果可以总结出本文提出的 DWPSD 分类方法的三点优势:

1) 对不同语料有较好的适应性。DWPSD 在三个具有典型特征的语料上均表现出较高分类性能,显示了对不同特色语料分类的稳定性,克服了一些分类方法挑剔语料的问题。

2) DWPSD 分类方法为线性分类方法,性能明显优于朴素贝叶素等线性分类方法,而且相对 KNN、SVM 等非线性分类方法亦能显示出性能优势,是高性能且高效率的文本分类方法。

3) DWPSD 分类方法无须特征选择,能较好地甄别所有特征对分类的贡献,避免因特征选择而造成分类信息的丢失。

3 结语

与以往基于统计理论的文本分类方法大多从机器学习理论寻求解决途径不同,本文基于文档类别与特征之间同现事件的确定性角度,发展了一种分布权值调节概率标准差的文本分类方法,实质上是将文本分类问题理解为特征的信任分析问题,算法直观、简洁、易于理解,为机器学习吸收其他研究领域理论提供了有益尝试。实验分析表明,算法适于不同语种、单一多标签、不同文档分布平衡性等语料的分类,具有较高稳定性和适应性,性能相对传统分类优势较明显。同时算法为线性分类方法,因此适于大规模文本分类,有很强的实用性。在未来的研究工作中,我们将在更广泛的语料上测试系统性能,深入研究可信计算与机器学习之间的关系,为自然语言处理领域的问题寻求更好的解决途径。

(上接第 3292 页)

由于篇幅的关系,本文仅列出了几个关键点的推导。

3.2 结论分析

目前,模式匹配工作的完成都是在半自动化工具的基础上完成。构造源模式到目标模式间映射的常规方案是直接将源模式中的元素与目标模式中元素进行一一匹配,使用工具对模式进行半自动化匹配时,需要不断的人机交互,才能保证源模式与目标模式匹配的准确性。这种方式在应用过程中人机交互过于频繁,极大地影响到了匹配的效率和,在映射过程中出现误差和错误匹配几率很大。本文提出的映射过程是在分析源模式与目标模式的基础上,对源模式中的对象集和关系集进行充分扩充,再使用关系代数对扩充后的模式进行去重、合并等简化处理。这样在映射的过程中,使用半自动化工具时可以将人工交互置于后半部分的化简工作中去,避免出现频繁人机交互存在的问题。

本文提出的方法适用于对象为现实中客观存在的实体,且源模式与目标模式之间存在有较多重复、离散待合并元素以及布尔值作为元素属性的模式间的匹配问题。

4 结语

下一步工作,是要在这种映射模式的基础上,结合现有半自动化模式匹配工具,例如 Altova 工具集,将映射过程进一步智能化,实现半自动化的批处理操作,使得半自动化的模式映射过程更加方便,进一步的提高集成映射效率,保证映射过程

参考文献:

- [1] YANG Y, LIU X. A re-examination of text categorization methods [C]// Proceedings of the ACM SIGIR '99. New York: ACM Press, 1999: 42 - 49.
- [2] TAN S, CHENG X. Using hypothesis margin to boost centroid text classifier [C]// Proceedings of the ACM SAC '07. New York: ACM Press, 2007: 398 - 403.
- [3] JØSANG A. Probabilistic logic under uncertainty [C]// CATS '07: Proceedings of Computing: The Australian Theory Symposium. Bal-larat, Victoria, Australia: [s. n.], 2007: 101 - 110.
- [4] JØSANG A, KNAPSKOG S J. A metric for trusted systems [EB/OL]. [2007 - 11 - 08]. <http://csrc.nist.gov/nissc/1998/proceedings/paperA2.pdf>.
- [5] JØSANG A. Artificial reasoning with subjective logic [C]// Proceedings of the 2nd Australian Workshop on Commonsense Reasoning. Perth: [s. n.], 1997.
- [6] WANG YONG-HONG, SINGH MUNINDAR P. Formal trust model for multiagent systems [EB/OL]. [2009 - 03 - 12]. <http://www.aaai.org/Papers/IJCAI/2007/IJCAI07-250.pdf>.
- [7] LANG K. Newswreeder: Learning to filter netnews [C]// Proceedings of ICML 95. Tahoe City, CA: [s. n.], 1995: 331 - 339.
- [8] LEWIS D. Reuters-21578 text categorization test collection distribution 1.0 [EB/OL]. [2008 - 10 - 09]. <http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt>.
- [9] BRANK J, GROBELNIK M, MILIC-FRAYLING N, *et al.* Interaction of feature selection methods and linear classification models [EB/OL]. [2009 - 04 - 10]. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.6031>.
- [10] YANG Y. An evaluation of statistical approaches to text categorization [J]. Journal of Information Retrieval, 1999(1): 67 - 88.

中数据集和关系集匹配的完备性与准确性。

衡量一个映射,需要从不同的角度来说明,如集成能力、查询返回能力、映射推理能力和映射合成能力等。今后集成映射研究的重点是如何能够自动地完成集成映射与模式匹配工作,这将会对数据集成研究产生巨大的推动。

参考文献:

- [1] CALI A, CALVANESE D, de GIACOMO G, *et al.* On the expressive power of data integration systems [C]// ER2002: Proceedings of 21st International Conference on Conceptual Modeling. Berlin: Springer-Verlag, 2002: 338 - 350.
- [2] POTTINGER R, BERNSTEIN P A. Schema merging and mapping creation for relational sources [C]// ACM International Conference Proceeding Series. New York: ACM Press, 2008: 73 - 84.
- [3] McBRIEN P, POULOVASSILIS A. Data integration by bi-directional schema transformation rules [C]// ICDE 2003: Proceedings of 19th International Conference on Data Engineering. Washington, DC: IEEE, 2003: 227 - 238.
- [4] MELNIK S, BERNSTEIN P A, HALEVY A, *et al.* Supporting executable mappings in model management [C]// Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2005: 167 - 178.
- [5] BISKUP J, EMBLEY D W. Extracting information from heterogeneous information sources using ontologically specified target views[J]. Information Systems, 2003, 28(3): 169 - 212.