

文本分类中基于 CHI 改进的特征选择方法

宋呈祥, 陈秀宏, 牛 强

(江南大学 数字媒体学院, 江苏 无锡 214122)

摘 要: 针对传统卡方统计量(Chi)方法在全局范围内做特征选择时忽略词的频度、词的分布等问题, 提出了一种改进的文本特征选择方法. 该方法通过定义特征词频度分布相关性系数, 选择局部出现的强相关性特征, 同时考虑特征词类间分布差异性来提升不平衡数据集的分类指标. 结果表明, 改进的方法不仅在分类效果上有明显的提高, 而且性能更加稳定.

关键词: 文本分类; 卡方统计量; 特征选择; 不平衡数据集

中图分类号: TP391.1

文献标识码: A

文章编号: 1000-7180(2018)09-0074-05

Improved Feature Selection Method Based on CHI for Text Categorization

SONG Cheng-xiang, CHEN Xiu-hong, NIU Qiang

(School of Digital Media, Jiangnan University, Wuxi 214122, China)

Abstract: Because the traditional Chi-square method chooses the feature in the global scope and ignores the information of word frequency and distribution, this paper proposes an improved feature selection method. The method selects a number of strong features with the defined feature distribution coefficient, and takes into account feature distribution that improves the performances of Chi-square method in the unbalanced dataset. The results of the experiments show that the improved algorithm not only shows a significant improvement in classification efficiency, but also has more stable performance.

Key words: text categorization; Chi-square; feature selection; unbalanced dataset

1 引言

文本分类是指在预先设定的分类体系下, 根据待分类文本内容自动确定其类别的过程. 在文本分类领域中, 文本文档通常使用向量空间模型将其映射成特征向量, 如果将文档中所有的词作为特征词, 不仅使向量空间维度过大, 造成文本稀疏而且含有大量噪声^[1]. 因此, 合理的特征选择, 不仅可以降低文本特征维度, 还能降低分类时间复杂度, 提高分类效果. 特征选择问题主要集中在如何选取最具类别区分度的特征词、各个特征词应该赋予多高的权重等. CHI作为一种常用的特征选择方法, 具备更低的时间复杂度和应用便利性等优点^[2], 但是CHI方

法只统计特征词在文本中是否出现, 没有考虑词频等信息, 文献[3-5]研究表明, 利用特征词的频度和其在类内、类间的分布信息, 能够有效提升传统CHI方法的分类效果.

针对传统CHI方法在不平衡数据集中做特征选择时, 未考虑特征词频度和特征词类间分布差异性问题, 提出一种改进的特征选择方法, 该方法首先考虑全局特征词频度信息, 然后分析不平衡数据集中特征词类间分布差异性对CHI方法的影响, 最后选择支持向量机(SVM)对文本分类. 实验结果表明, 该方法在均衡数据集和不均衡数据集的分类效果都优于传统CHI方法和文献[4]的方法.

2 卡方统计量

卡方统计量(CHI)是用来衡量特征词 t_k 和类别 c_i 之间的相关程度。假设 t_k 和 c_i 之间符合具有一阶的自由度 χ^2 分布,则 t_k 与 c_i 的CHI值定义为^[2]:

$$\text{CHI}(t_k, c_i) = \frac{n [p(t_k, c_i) p(\bar{t}_k, \bar{c}_i) - p(t_k, \bar{c}_i) p(\bar{t}_k, c_i)]^2}{p(t_k) p(c_i) p(\bar{t}_k) p(\bar{c}_i)} \quad (1)$$

式中, n 表示数据集的文本数量; $p(t_k, c_i)$ 表示文本包含特征词 t_k 并且属于类别 c_i 的概率; $p(\bar{t}_k, \bar{c}_i)$ 表示文本既不包含特征词 t_k 也不属于类别 c_i 的概率; $p(\bar{t}_k, c_i)$ 表示文本不包含特征词 t_k 但属于类别 c_i 的概率; $p(t_k, \bar{c}_i)$ 表示文本包含特征词 t_k 但不属于类别 c_i 的概率。

由公式(1)可知,特征词 t_k 和类别 c_i 的卡方统计量值越高,表示 t_k 与 c_i 之间的相关性越大, t_k 携带 c_i 类的鉴别信息就越多。当CHI值等于0时,表明特征词 t_k 与 c_i 类相互独立。传统的CHI方法是把训练集合的文本数据进行文本预处理后,形成候选特征集合,在根据式(1)计算候选特征词相对于各个类别的CHI值,取其中的最大值(见公式(2))作为该特征词相对于整个训练集合的最后CHI值。然后将这些CHI值进行降序排列,选取前 K 个CHI值对应的特征词作为最后的特征集合^[6]。

$$\text{CHI}_{\max}(t_k) = \max_{1 \leq i \leq m} \{\text{CHI}(t_k, c_i)\} \quad (2)$$

式中, m 表示类别数目。

3 卡方统计量的优化

3.1 基于特征词频度的改进

Yang等人^[2]的研究表明,CHI相对于其他特征选择效果较好,但是仍然存在一些问题,卡方统计量只统计特征是否出现,而并未考虑特征词在文本中出现的次数,从而会受到低频词的影响,分类效果会很差。由公式(1)知,当 $p(t_k, c_i) \rightarrow 0, p(\bar{t}_k, \bar{c}_i) \rightarrow 0$ 时,即 t_k 分布在少量数据集时,CHI值约等于0,这样如根据CHI值排序进行特征选择,CHI值低的特征词 t_k 很容易被过滤掉。比如,当少量数据集中大量出现“篮球”、“足球”时,这些特征词CHI值约等于0。由于CHI值较低,将无法选择出这些特征词,而这些特征词最能代表“体育”这个类别。当特征词 t_k 出现次数较少且均匀分布在各个类别数据集时,特征词 t_k 含较低的类别区分度,应当使 t_k 获得较小的权值。

针对传统CHI全局特征选择以及未考虑词频信息等问题,定义一种新的特征分布系数(Feature Distribution Coefficient, FDC):

$$\text{FDC}(t_k) = \log\left(\frac{\mu \times n_{\mu t}}{\text{df}(t_k)}\right) (\mu > 1) \quad (3)$$

式中, $n_{\mu t}$ 表示 t_k 在所有数据集中出现的总次数; n 表示数据集文本总数量; μ 为调节因子,对于不同的数据集取值不同; $\text{df}(t_k)$ 表示数据集中出现特征词 t_k 的文本数目。当 $\frac{n_{\mu t}}{\text{df}(t_k)} \rightarrow 1$ 时, $\log \frac{n_{\mu t}}{\text{df}(t_k)} \rightarrow 0$,此

时 t_k 获得较小的权值。而当 $\frac{n_{\mu t}}{\text{df}(t_k)} > 1$ 时,表示特征词 t_k 大量分布于某些文本集合中, t_k 携带类别信息就越多;FDC值越大, t_k 就可能被保留下来。

3.2 基于不均衡数据集的改进

一个有分类价值的特征词应该在指定类分布较多,而在其他类别中很少出现。基于此,文献[7]提出一种特征词在指定类的文本频数与所有类文本频数平均值间的差值来衡量特征词分布的方法,然而方法只考虑了各个类内包含特征词的文本数目,忽略了类内数据不均衡的影响。数据不均衡表现为特征词在数据集中各类别间的文本数量可能存在数量级的差距(偏斜)。在不均衡数据集上计算特征词在某个类内的集中程度,也常常使用包含该词的文本类概率代替文本频数来衡量特征词出现的频繁程度。于是,针对不均衡数据集中文本数目偏斜问题,在文献[7]基础上,将各个类别的文本数目考虑进来,提出一种新的特征词文本数目偏离程度的差值(Document Probability Variance, DPV):

$$\text{DPV}(t_k, c_i) = \left(\frac{\text{df}(t_k, c_i)}{\text{df}(c_i)} - \frac{1}{m} \sum_{j=1}^m \frac{\text{df}(t_k, c_j)}{\text{df}(c_j)} \right)^2 \quad (4)$$

式中, $\text{df}(t_k, c_i)$ 表示类别 c_i 出现特征 t_k 的文本数; $\text{df}(c_i)$ 表示类 c_i 的文本总数, m 表示类别数。如果各个类内出现特征词 t_k 的文本类概率越接近,DPV值就越小,表明特征词 t_k 的类别区分度越小;如果各个类内出现特征词 t_k 的文本类概率相差越大,DPV值就越大,表明该特征词 t_k 具有很好的分类辨别力。

在不均衡数据集中,如何选择出具有高区分度稀有类别特征词,对于提高稀有类别的分类性能不可忽视。进一步分析公式(1)可知,令 $\beta = p(t_k, c_i) p(\bar{t}_k, \bar{c}_i) - p(t_k, \bar{c}_i) p(\bar{t}_k, c_i)$,当 $\beta < 0$ 时,CHI特征选择表现为特征词与类别呈现负相关性。特别在不均衡数据集中,稀有类别特征词往往表现为一定的负相关性。文献[4]提出一种类别倾向因

子 $\alpha = p(t_k, c_i) / (p(t_k, c_i) + p(\overline{t_k}, \overline{c_i}))$ 来调整特征词的正负相关性。但是,类别倾向因子 α 只是考虑了 $p(t_k, c_i), p(\overline{t_k}, \overline{c_i})$ 两种情况的文本数目的影响,忽略了不平衡数据集中特征词类间分布差异性。于是,将公式(4)引入,定义一种新的特征词类间分布因子(Category Distribution Factor, CDF),公式如下:

$$CDF(t_k, c_i) = \begin{cases} 0, & \beta < 0 \text{ and } \alpha < 0.5 \\ DPV, & other \end{cases} \quad (5)$$

可知,当 CDF 值越大,特征词 t_k 越集中在指定类 c_i 中,对分类贡献就越大。最后给出以下改进的特征选择公式 PCHI(Polished CHI, PCHI) 为:

$$PCHI(t_k, c_i) = CHI(t_k, c_i) \times FDC(t_k) \times CDF(t_k, c_i) \quad (6)$$

公式(6)表明,如特征词 t_k 出现次数越多,类间分布越有差异性,特征词的分类区分度就越大,分类效果也就越好。

综上所述,得到以下改进的特征选择的文本分类算法流程:

(1)文本预处理. 文本预处理包括读入用户词典并分词、去除特殊符号以及停用词、词性标注;只保留名词、动词和形容词等重要词语,获取文本词语集合。

(2)特征选择. 使用本文算法计算训练集文本词语集合的 FDC 值和每个类别的 CHI, CDF 值,得到每个词和对应类别的 PCHI 值;对于重复的词,取最大值作为该词最终的 PCHI 值。将每个词按 PCHI 值降序排序,选取 TOP-K 作为整个数据集的特征词集合,并用于文本向量表示。

(3)权重计算. 对于每篇文本的词语集合,若步骤(2)的特征词集合含有该词,使用 TF-IDF^[1] 权重计算公式对该词进行赋权,为了消除文本长度的影响,将权值进行归一化处理,构造文本特征向量。

(4)分类器训练. 利用步骤(3)得到训练集文本特征向量,并训练分类器。

(5)测试分析. 将测试集分别进行步骤(1)、(3)处理获取测试集文本特征向量,并对步骤(4)得到的分类器测试评估,输出实验结果。

4 实验设计与结果

4.1 实验准备

实验数据利用复旦大学中文数据集、网易新闻数据集和搜狗实验室数据集,其中复旦大学中文数据集和搜狗实验室数据集分别选取 3 156 篇、2 100 篇文本组成训练集和测试集,选取数量如表 1、表 2;

网易新闻数据集包含六个类别,随机选取每个类别的 300 篇文本,以 2:1 的比例组成训练集和测试集。

表 1 复旦大学中文数据集训练集和测试集的选取情况

类别	农业	艺术	环境	政治	空间	医药	体育	经济
训练集	286	453	212	164	267	322	134	265
测试集	143	227	106	82	134	161	67	133

表 2 搜狗实验室数据集训练集和测试集的选取情况

类别	汽车	财经	IT	健康	教育	招聘
训练集	66	133	200	266	333	400
测试集	34	67	100	134	167	200

实验中,使用中科院 NLPPIR^[8] 工具对文本进行预处理,利用台湾大学的林智仁教授开发的 SVM 分类器^[9] 对文本进行分类;实验性能评估使用宏 F1 值 (macro_F1) 来度量所有类别的总体分类指标:

$$macro_P = \frac{1}{m} \sum_{i=1}^m P_i, macro_R = \frac{1}{m} \sum_{i=1}^m R_i,$$
$$macro_F1 = \frac{2 \times macro_P \times macro_R}{macro_P + macro_R}, \quad (7)$$

式中, m 表示类别个数; P_i, R_i 分别表示 c_i 类的查准率 (Precision, P) 和查全率 (Recall, R); $macro_P$ 表示宏查准率; $macro_R$ 表示宏查全率。

实验中各个实验方法表示为: E1 表示传统 CHI 方法实验; E2 表示文献[4]提出的 CHI 方法实验; E3 表示本文提出的 PCHI 方法实验。

4.2 实验结果与分析

4.2.1 在复旦大学中文不平衡数据集的分类实验
在复旦大学数据集中得到分类实验如图 1 所示。

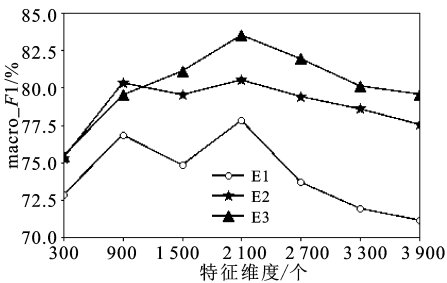


图 1 不同特征维度的分类宏 F1 值 (%)

由图 1 可见,当特征集合维度增大时,宏 F1 值也随着变大,当维数继续增大,由于冗余特征的增多致使宏 F1 值变小。E3 综合考虑特征词频度和特征词类别之间的分布差异性等因素,性能优于 E1, E2。在 2100 维时, E3 达到 83.53% 的宏 F1 值,比 E1, E2 宏 F1 值分别高出 5.72%、3.02%,此时分析

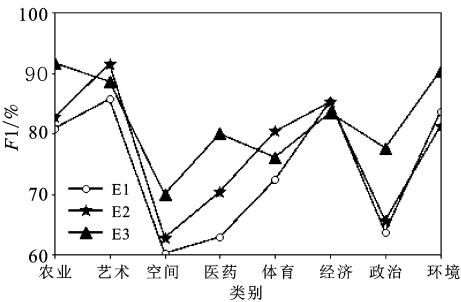


图 2 特征维度 2100 时分类 F1 值(%)

各个类别的 F1 值,如图 2. 各个类别 F1 值差别明显,原因是如果文本长度较短,含有很多空值,使向量稀疏,造成分类结果较低;若文本含有词数较多,并且具有一些类别区分度特征词,此时宏 F1 值较大.

4.2.2 在网易新闻均衡数据集的分类实验

为了验证本文提出的方法在均衡数据集上的有效性,在网易新闻数据集上的实验如图 3 所示.

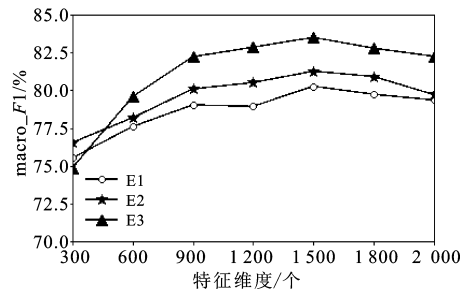


图 3 不同特征维度的分类宏 F1 值 (%)

由图 3 可知,因各个类别文本数目分布均匀, E1 也能保持一个良好的性能,在特征维度 1 500 维时,E1,E2,E3 分别达到最优值,此时分析各个类别的 F1 值,如图 4 所示.

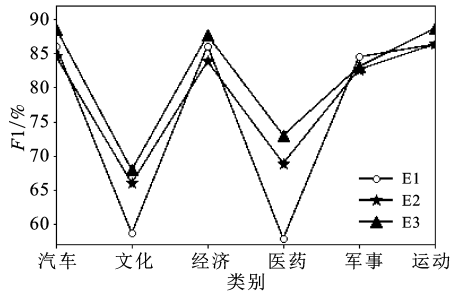


图 4 特征维度 1 500 时的分类 F1 值(%)

4.2.3 在搜狗实验室不均衡数据集的分类实验

在搜狗实验室不均衡数据集中特征维度达到 1 200 维时,分析各个类别的查准率 P、查全率 R 和 F1 值,如表 3.

表 3 在特征维度 1 200 时各个方法的分类效果(%)

		汽车	财经	IT	健康	教育	招聘	均值
E1	R	38.34	55.22	65.00	97.01	80.24	96.50	72.05
	P	100.00	97.68	87.84	65.66	72.83	86.16	85.03
	F1	55.32	70.48	74.71	78.31	76.35	91.04	74.37
E2	R	55.88	77.61	68.00	94.78	89.22	94.00	79.92
	P	100.00	96.30	87.18	85.81	72.68	94.95	87.82
	F1	71.70	85.95	76.40	90.07	80.11	94.47	83.12
E3	R	76.47	89.55	65.00	97.01	92.81	94.50	85.89
	P	100.00	95.24	94.20	85.53	77.11	98.95	91.84
	F1	86.66	92.31	76.92	90.91	84.24	96.68	87.95

由表 3 可得,在 1200 维时,E1,E2,E3 的宏 F1 值分别为 78.00%、83.68%、88.77%;可知,E3 对一些较低的分类指标值有较好的效果.

5 结束语

特征选择在文本分类过程中具有重要作用. 本文提出了一种改进的考虑特征词频度 CHI 统计特征选择方法,同时定义特征词类间分布因子解决 CHI 方法在不均衡数据集分类差问题,该方法能有效改善传统 CHI 方法的缺陷,选择出具有类别区分度的特征词,使得分类效果更好,性能更稳定. 在后续工作中,将考虑特征词的语义关系,进一步进行特征降维,在减少算法时间复杂度的同时提升分类效果.

参考文献:

[1] 周爱武,马那那,刘慧婷. 基于卡方统计的情感文本分类[J]. 微电子学与计算机,2017,34(8):57-61.

[2] Yang Y, Pedersen JO. A Comparative Study on Feature Selection in Text Categorization[C]// Fourteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 1997:412-420.

[3] Junejo K N, Karim A, Hassan M T, et al. Terms-based discriminative information space for robust text classification[J]. Information Sciences, 2016, 372(c): 518-538.

[4] 谢娜娜,房斌,吴磊. 不均衡数据集上文本分类方法研究[J]. 计算机工程与应用,2013,49(20):118-121.

[5] Jin C, Ma T, Hou R, et al. Chi-square Statistics Feature Selection Based on Term Frequency and Distribution for Text Categorization[J]. Iete Journal of Research, 2015, 61(4):351-362.

[6] 叶敏,汤世平,牛振东. 一种基于多特征因子改进的中文文本分类算法[J]. 中文信息学报,2017,31(4):

132-137.

[7] 邱云飞, 王威, 刘大有, 等. 基于方差的 CHI 特征选择方法[J]. 计算机应用研究, 2012, 29(4): 1304-1306.

[8] 中国科学院计算技术研究所. NLPir[EB/OL]. [2017-03-14]. <http://ictclas.nlpir.org/>.

[9] Chih-Jen Lin, LIBSVM[EB/OL]. [2016-11-22]. <https://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html>.

(上接第 73 页)

小,平均功耗误差为-21.58%。偏小的原因在于 CGRA 部分内部结构细节无法获取,在建模时略去了对该部分结构的建模。从仿真数据误差来看,本文提出的粗粒度可重构处理器功耗模型具有一致性,可以反映出系统的整体功耗趋势,因此可以作为粗粒度系统结构初期设计功耗的评估工具。

6 结束语

系统级功耗模型可以为系统结构设计的初期提供功耗评估,本文提出的粗粒度可重构结构的系统级功耗模型,兼顾了系统级的功耗评估速度和门级结构的仿真精度。实验结果显示,预测功耗和板级实测功耗相比,平均误差在-21.58%,说明了该系统级功耗模型的有效性。因此,本文提出的功耗模型可以帮助系统结构设计者根据功耗要求,准确高效地定位出结构问题,同时对编译器的调度和映射起到指导作用。

本文提出的功耗模型目前只考虑了 CGRA 中的 RPU 部分,没有包括 ARM 主控、内存访问以及互连结构的功耗。在深亚微米工艺时代,互连功耗在集成电路中占比越来越高,本文对这部分的功耗建模尚未涉及。因此,后续的工作将在此基础上进行更加完整的功耗建模,为体系结构设计人员增加更大的设计探索空间。

参考文献:

[1] Zhongyuan Zhao, Weiguang Sheng, Naifeng Jing, et al. Resource-saving compile flow for coarse-grained reconfigurable architectures [C] // ReConFigurable Computing and FPGAs (ReConFig), 2015 International Conference on. Riviera Maya, Mexico, 2015.

[2] Sheng Li, Jung Ho Ahn, Richard D, et al. McPAT;

作者简介:

宋呈祥 男,(1989-),硕士研究生.研究方向为机器学习与自然语言处理. E-mail:scx929@163.com.

陈秀宏 男,(1964-),博士(后),教授.研究方向为模式识别,图像处理及人工智能等.

牛强 男,(1992-),硕士研究生.研究方向为机器学习与图像处理.

an integrated power, area, and timing modeling framework for multicore and manycore architectures [C]//Microarchitecture, 2009. MICRO-42. 42nd Annual IEEE/ACM International Symposium on. New York, NY, USA, 2009.

[3] International technology roadmap for semiconductors[EB/OL]. [2017-10-10]. <https://www.itrsgroup.com>.

[4] Mukherjee R, Mondal S, Memik S O. Thermal sensor allocation and placement for reconfigurable systems [C] // Proceedings of the 2006 IEEE/ACM international conference on Computer-aided design. Washington, USA, 2006; 437-442.

[5] An integrated cache and memory access time, cycle time, area, leakage, and dynamic power model[EB/OL]. [2017-10-13]. <http://www.hpl.hp.com/research/cacti>".

[6] Weste N, Harris D. CMOS VLSI design: a circuits and systems perspective[M]. 3rd ed. USA: New Jersey, Addison Wesley, 2004.

[7] Horowitz M, Ho R, Mai K. The future of wires [EB/OL]. (1999-10-01). <http://velox.stanford.edu/>.

[8] Guthaus T, R Ringenberg M, Ernst J S, et al. MiBench: a free, commercially representative embedded benchmark suite[C] // IEEE International Workshop on Workload Characterization. [s. l.]. 2001; 3-14.

作者简介:

穆昌根 男,(1991-),硕士研究生.研究方向为粗粒度可重构处理器系统级功耗建模. E-mail:alfred_mu@sju.edu.cn.

赵仲元 男,(1989-),博士研究生.研究方向为可重构计算编译架构.

绳伟光 男,(1977-),助理研究员.研究方向为高可靠集成电路设计.

毛志刚 男,(1962-),教授.研究方向为系统级芯片设计.