改进的文本特征选取算法研究

朱世玲,郑　彦
(南京邮电大学 计算机软件学院,江苏 南京 210023)

摘　要:特征选取的好坏决定了文本分类的准确度。文本特征选取通常有文档频率、互信息、信息增益、卡方统计量等方法。文中讨论了文档频率和互信息在特征选取时的缺点,基于这些缺点,提出了一种混合文档频率和互信息的改进算法。文档频率进行特征选取时会偏向选择高频词,而没有考虑到该词是否在类别间有区分度,所以提出通过计算词的文档频率的类别方差作为文档频率的权重来进行特征选取。互信息偏向选择低频词,也忽略了互信息值为负的那些特征作用,有些互信息为负的词反而包含更多的类别信息。所以对互信息的值取了绝对值来加强互信息为负的词的作用。通过对比 DF、MI 和改进的 DFMI 的实验结果,发现该算法在精度、召回率和 F_1 度量上都有所提高,验证了该方法的有效性。

关键词:特征选取;互信息;文档频率;文本分类;改进互信息;改进文档频率
中图分类号:TP301.6　　**文献标识码**:A　　**文章编号**:1673-629X(2019)05-0066-04
doi:10.3969/j.issn.1673-629X.2019.05.014

Research on Improved Text Feature Selection Algorithm

ZHU Shi-ling, ZHENG Yan
(School of Computer Software, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: The quality of feature selection determines the accuracy of text classification. Text feature selection usually includes document frequency, mutual information, information gain, Chi-square statistics and so on. We discuss the shortcomings of document frequency and mutual information in feature selection, and on the basis propose an improved algorithm for hybrid document frequency and mutual information. In the feature selection of document frequency, high-frequency words are preferred without considering whether there is a degree of discrimination between categories. Therefore, we propose to take the category variance of word frequency as the weight of document frequency for feature selection. The mutual information tends to select low-frequency words, but also ignores those features with negative mutual information value. Some words with negative mutual information contain more category information. Therefore, the absolute value of mutual information is taken to strengthen the role of words with negative mutual information. The experimental comparison of DF, MI and improved DFMI indicates that the proposed algorithm improves in accuracy, recall rate and F_1 measure, which verifies its effectiveness.

Key words: feature selection; mutual information; document frequency; text classification; mutual information improved; document frequency improved

0　引　言

随着计算机信息技术的快速发展,网络上各种各样的文本数据极速增长。对这些文本数据的快速处理成为了重要的研究课题,文本分类也因此得到了快速发展。文本分类是在给定一些特定的文本类别下,根据文本的内容将文本自动划分到一个或多个类别中[1-2]。在文本分类时,通常需要将文本信息用向量空间模型或词频矩阵来表示[3]。如果直接用文本向量来表示,则向量空间维数会很大,而且会包含很多无用的属性,所以需要对文本数据进行预处理,去除无关属性,降低文本向量空间的维数以及排除一些无关信息对分类的干扰。预处理通常包括去除停用词、特征选取等方法[4],而特征选取是文本分类预处理中的重要一步,也是一直以来很多学者研究的重点问题[5-7]。目前,文本分类中常用的特征选取算法有文档频率(document frequency, DF)、卡方统计量(CHI-square statistic, CHI)、信息增益(information gain, IG)、互信息(mutual information, MI)[8]等。文档频率就是设置

收稿日期:2018-06-07　　**修回日期**:2018-10-09　　**网络出版时间**:2018-12-21
基金项目:国家"863"高技术发展计划项目(2006AA01Z201)
作者简介:朱世玲(1993-),女,硕士研究生,研究方向为机器学习、数据挖掘;郑　彦,教授,硕导,研究方向为数据挖掘、机器学习、人工智能。

一个阈值,只要在训练集中包含该词的文本数大于这个阈值就选取作为特征词。在文本分类特征选取中,互信息衡量的是一个特征和类别之间的相关程度,互信息值越大,所包含的类别信息就越多,对分类影响就越大。近年来很多学者都对互信息进行了改进[9-12]。在此基础上,文中分别讨论了文档频率和互信息在进行特征选取时的缺点,提出了一种混合文档频率和互信息的改进算法,并通过实验对其有效性进行验证。

1　传统的特征选取算法

1.1　传统的文档频率算法

文档频率算法是文本分类中最简单、复杂度最低的特征选取算法。它是指在训练集中包含某个词条的文本数。将得到的每个词条的 DF 值和预先设定的阈值进行比较,如果大于这个阈值,就表示这个词条属于高频词对文本分类有价值,就保留作为特征词,如果小于这个阈值,就认为该词条属于低频词对分类没有贡献并删除。这种方法简单,计算快速,能够胜任大规模的文本分类任务。

1.2　传统的互信息算法

在文本分类特征选取中,互信息衡量的是特征和类别之间的统计关联程度。它的理论基础是如果类别 c 中包含特征 t 的文档数占类别文档数的比重高,而包含特征 t 占文档集总数的比重低,则表明特征 t 与类别具有强相关性,不是相互独立关系,其互信息值大[13]。特征与类别之间的互信息计算公式如下:

$$\mathrm{MI}(t,c)=\log\frac{P(t,c)}{P(c)P(t)}=\log\frac{P(t\mid c)}{P(t)} \tag{1}$$

其中,$P(t,c)$ 表示在类别 c 中文本包含特征 t 的概率;$P(c)$ 表示属于类别 c 的文本占训练集文本的概率;$P(t)$ 表示在训练集中文本包含特征 t 的概率。当特征 t 和类别 c 相互独立时,$P(t\mid c)=P(t)P(c)$ 的值就等于 0。$P(t\mid c)$ 值越大,$P(t)$ 值越小,互信息值就越大,特征与类别之间的关联性就越强,特征就具有更多的分类信息。

特征 t 对于整个类别的互信息主要有两种计算方式,分别是互信息的最大值和各类互信息的平均值。两种计算公式如下:

采用平均值:

$$\mathrm{MI}(t)=\sum_{i=1}^{m}P(c_i)\mathrm{MI}(t,c_i) \tag{2}$$

采用最大值:

$$\mathrm{MI}(t)=\max\mathrm{MI}(t,c_i) \tag{3}$$

2　改进的特征选取算法

2.1　传统文档频率方法的不足与改进

文档频率算法虽然简单直白,复杂度低,但是缺点也很明显,即没有确切的理论基础,通常被认为是一种经验方法。而且考虑特征词和类别之间的关系,有的词条小于预先设定的阈值,被认为低频词而删除,但却在某个类别中集中出现,能够很好地反映该类别特征。有的词条虽然大于预先设定的阈值,但却在每个类别中均匀出现,这样的特征词对分类就没有价值[14]。基于这个缺点,文中为特征词的文档频率加入类别间的方差权重,选择词条在每个类别中文档频率方差比较大的词条。这样可以降低在每个类别中均等出现词的作用。

改进后的文档频率公式如下:

$$\mathrm{DF}(t)=\beta\times\log\mathrm{DF} \tag{4}$$

其中,$\mathrm{DF}(t)$ 表示改进后的特征 t 的文档频率;β 表示特征 t 在各个类别中的文档频率的方差权重;DF 表示特征 t 的文档频率。

β 的计算公式为:

$$\beta=\frac{1}{m}\sum_{j=1}^{m}\left[\mathrm{df}_j(t)-\frac{1}{m}\sum_{i=1}^{j}\mathrm{df}_i(t)\right] \tag{5}$$

其中,m 表示类别总数;$\mathrm{df}_j(t)$ 表示特征 t 在类别 j 中的文档数。

2.2　传统互信息方法的不足与改进

根据式 1 可知,当两个特征的 $P(t\mid c)$ 相同时,$P(t)$ 越小的特征的互信息值反而越大,所以会偏向选择低频词[15]。而且对于特征 t 和类别 c,当互信息值大于零时,$P(t\mid c)$ 越大或 $P(t)$ 越小时,互信息的值就越大,绝对值越大;当互信息值小于零时,$P(t\mid c)$ 越大或 $P(t)$ 越小时,互信息的值越小,绝对值反而越大。换句话说,当 $P(t\mid c)$ 和 $P(t)$ 越接近时,特征 t 和类别 c 的相关联度就越小,互信息的绝对值越小,反之,互信息的绝对值就越大。所以,互信息值的绝对值越大的特征越能反映特征和类别之间的关联程度。改进后的互信息公式如下:

$$\mathrm{MI}(t)=\sum_{i=1}^{m}P(c_i)\mid\mathrm{MI}(t,c_i)\mid \tag{6}$$

其中互信息的值采用平均值。

2.3　改进的混合算法

文中提出了混合 DF 和 MI 的特征选取算法,并对 DF 和 MI 各自的不足进行了分析和改进。针对 DF 方法偏向选择高频词和 MI 方法偏向选择低频词,考虑将两种方法进行混合来削弱它们的不足,使在特征选取时选择的特征词既不偏向低频词也不偏向高频词,也避免选取在类别中均等出现的特征词。混合 DF 和 MI 的特征选取公式如下:

$$\mathrm{DFMI}=\frac{1}{m}\sum_{j=1}^{m}\left[\mathrm{df}_j(t)-\frac{1}{m}\sum_{i=1}^{m}\mathrm{df}_i(t)\right]^2\times\mathrm{DF}\times\sum_{i=1}^{m}P(c_i)\mid\log\frac{P(t\mid c)}{P(t)}\mid \tag{7}$$

3 实 验

3.1 数据集及开发工具

实验数据集采用搜狗数据集,总共 9 个类别,分别为财经、IT、健康、体育、旅游、教育、招聘、文化、军事。每个类别 300 篇文章,共 2 700 篇文章,其中每个类别的 200 篇文章用于训练,100 篇文章用于测试分类结果。为了验证该算法的有效性,将传统的 DF 方法和传统的 MI 方法与提出的混合 DFMI 方法进行比较。分类器选择实现简单,分类效果良好的朴素贝叶斯,用 Java 语言实现,开发工具为 Eclipse。

3.2 分类效果评估

一篇文本的分类情况可以分为四种:真正例(true position)、假正例(false position)、真反例(true negative)、假反例(false negative),如表 1 所示。

表 1 文本分类结果

真实情况	预测结果	
	正例	反例
正例	TP(真正例)	FN(真反例)
反例	FP(假正例)	TN(假反例()

评价算法好坏的度量指标采用精度(precision,又称查准率)、召回率(recall,又称查全率)、F_1 度量。

精度(P)可以看作精确性的度量,即标记为正类的元组实际为正类所占的百分比,公式如下:

$$P = \frac{TP}{TP + FP} \tag{8}$$

召回率(R)是完全性度量,即正元组标记为正的百分比,公式如下:

$$R = \frac{TP}{TP + FN} \tag{9}$$

F_1 度量是把精度和召回率组合到一起的度量方法,公式如下:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{10}$$

3.3 实验结果及分析

在 Eclipse 上用 Java 语言实现朴素贝叶斯分类,来验证不同特征选取方法对分类结果的影响。先利用中科大 ICTCLAS 分词系统对所有文本进行分词,根据分词后的结果,再选取名词性和既有名词性和动词性的词语,得到预处理后的特征集合。使用不同特征选取方法进行特征选取,特征词都是 1 000 个。将所有文本向量化,最后利用朴素贝叶斯分类器对文本进行分类,实验结果如表 2 所示。

表 2 DF、MI、DFMI 方法在精度、召回率和 F_1 上的比较 %

方法	指标	财经	IT	健康	体育	旅游	教育	招聘	文化	军事
DF	P	82.29	69.72	83.75	100	78.26	68.18	65	86.54	75.57
	R	87	76	67	78	72	80	91	45	99
	F_1	86.14	72.72	74.44	88.52	74.99	74.77	75.83	59.21	85.34
MI	P	40.26	29.21	47.96	21.89	27.61	26.26	38.37	28.33	35.71
	R	31	26	47	14	45	26	33	34	35
	F_1	35.03	27.51	47.48	17.08	34.22	54.06	35.48	30.91	35.35
DFMI	P	86.14	77.78	87.34	100	82.52	70.8	68.94	91.3	78.33
	R	87	77	69	84	85	83	91	42	99
	F_1	88.57	77.39	77.1	91.3	83.75	75.11	78.45	60.53	84.26

从表中可以看出,改进的混合 DFMI 方法明显比 MI 方法好很多,无论在精度、召回率还是 F_1 度量上都明显提高,和 DF 相比也均有提升,从而验证了混合 DFMI 方法的有效性。

4 结束语

MI 方法简单,应用广泛,但倾向选择低频词,忽略了互信息绝对值较大的特征也具有较好的类别区别能力,因此通过对互信息取绝对值后再取平均值排序进行特征选择。DF 方法虽然简单直白,但有的特征虽然出现的频率很好,但在类别中均等出现这样的特征也没有区别能力,所以考虑加入文档频率类别方差。基于两种改进后的方法,提出一种混合的 DFMI 特征选取算法。实验结果表明,该算法在精度、召回率和 F_1 度量上均有所提高。

现有的特征选取算法都是从不同的角度进行特征选取,都有各自的优缺点,因此将不同的特征选取算法进行混合,使之从多个角度进行考虑,兼顾多个方面,是一个值得研究的方向。

参考文献:

[1] 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(9):1848-1859.

[2] YANG Yiming. An evaluation of statistical approaches to text categorization[J]. Information Retrieval,1999(1):69-90.

[3] 袁　轶,王新房. 一种基于方差的文本特征选择算法[J]. 计算机工程,2012,38(12):155-157.
[4] KO Y,SEO J. Automatic text categorization by unsupervised learning[C]//Proceedings of 18th conference on computational linguistics. Saarbrücken, Germany: ACM, 2000: 453-459.
[5] BIDI N,ELBERRICHI Z. Feature selection for text classification using genetic algorithms[C]//International conference on modelling, identification and control. Algiers, Algeria: IEEE,2017:806-810.
[6] 林艳峰. 中文文本分类特征选择方法的研究与实现[D]. 西安:西安电子科技大学,2014.
[7] 李军怀,付静飞,蒋文杰,等. 基于 MRMR 的文本分类特征选择方法[J]. 计算机科学,2016,43(10):225-228.
[8] 郑　伟. 文本分类特征选取技术研究[D]. 呼和浩特:内蒙古大学,2008.
[9] QIU Liqing, ZHAO Ruyi, ZHOU Gang, et al. An extensive empirical study of feature selection for text categorization [C]//IEEE/ACIS international conference on computer and information science. Portland, OR, USA: IEEE, 2008: 312-315.
[10] 林少波. 中文文本分类特征提取方法的研究与实现[D]. 重庆:重庆大学,2011.
[11] 康岚兰,董丹丹. 一种改进的互信息特征选择方法[J]. 电脑知识与技术,2009,5(35):9889-9890.
[12] 卢新国,林亚平,陈治平. 一种改进的互信息特征选取预处理算法[J]. 湖南大学学报:自然科学版,2005,32(1):104-107.
[13] 蒋　健. 文本分类中特征提取和特征加权方法研究[D]. 重庆:重庆大学,2010.
[14] DING Xiaoming, TANG Yan. Improved mutual information method for text feature selection[C]//International conference on computer science & education. Colombo, Sri Lanka: IEEE,2013:163-166.
[15] JIANG Xiaoyu, JIN Shui. An improved mutual information-based feature selection algorithm for text classification[C]//5th international conference on intelligent human-machine systems and cybernetics. Hangzhou, China: IEEE, 2013: 126-129.

(上接第 65 页)

[2] BIRKHOFF G. Lattice theory[M]. 3rd ed. [s. l.]:[s. n.], 1973.
[3] LI Yongming. Quantitative model checking of linear-time properties based on generalized possibility measures[J]. Fuzzy Sets and Systems,2015,320:17-39.
[4] HOLZMANNN G J. The SPIN model checker: prinmer and reference manual[M]. Reading: Addison-Wesley,2004.
[5] CIMATTI A, CLARKE E, GIUNCHIGLIA F, et al. NuSMV: a new symbolic model checker[J]. International Journal on Software Tools for Technology Transfer,2000,2(4):410-425.
[6] LI Yongming, MA Zhanyou. Quantitative computation tree logic model checking based on generalized possibility measures[J]. IEEE Transactions on Fuzzy Systems,2015,23(6):2034-2047.
[7] 陈云霁,马　麟,沈海华,等. 龙芯 2 号微处理器浮点除法功能部件的形式验证[J]. 计算机研究与发展,2006,43(10):1835-1841.
[8] 张　良,易江芳,佟　冬,等. 使用局部建模的微处理器测试程序自动生成方法[J]. 电子学报,2011,39(7):1639-1644.
[9] 马占有,李永明. 基于决策过程的广义可能性计算树逻辑模型检测[J]. 中国科学:信息科学,2016,46(11):1591-1607.
[10] 邓楠轶. 基于广义可能性测度的模型检测器 GPoCheck 的设计与实现[D]. 西安:陕西师范大学,2015.
[11] 范艳焕,李永明,潘海玉. 不确定型模糊 Kripke 结构的计算树逻辑模型检测[J]. 电子学报,2018,46(1):152-159.
[12] LI Yongming, LI Yali, MA Zhanyou. Computation tree logic model checking based on possibility measures[J]. Fuzzy Sets and Systems,2015,262:44-59.
[13] 周从华,邢支虎,刘志锋,等. 马尔可夫决策过程的限界模型检测[J]. 计算机学报,2013,36(12):2587-2600.
[14] 林惠民,张文辉. 模型检测:理论、方法与应用[J]. 电子学报,2002,36(12A):1907-1912.
[15] 刘　阳,李宣东,马　艳,等. 随机模型检测研究[J]. 计算机学报,2015,38(11):2145-2162.
[16] 周从华,刘志锋,王昌达. 概率计算树逻辑的限界模型检测[J]. 软件学报,2012,23(7):1656-1668.
[17] 李永明,李　平. 模糊计算理论[M]. 北京:科学出版社,2016.
[18] 严蔚敏,吴伟民. 数据结构[M]. 北京:清华大学出版社,2012.