计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(6):135−140 [doi: 10.15888/j.cnki.csa.006888]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于二维信息增益加权的朴素贝叶斯分类算法①

任世超, 黄子良

(成都信息工程大学 通信工程学院, 成都 610225)
通讯作者: 任世超, E-mail: 972199342@qq.com

摘　要: 由于朴素贝叶斯算法的特征独立性假设以及传统 TFIDF 加权算法仅仅考虑了特征在整个训练集的分布情况, 忽略了特征与类别和文档之间关系, 造成传统方法赋予特征的权重并不能代表其准确性. 针对以上问题, 提出了二维信息增益加权的朴素贝叶斯分类算法, 进一步考虑到了特征的二维信息增益即特征类别信息增益和特征文档信息增益对分类效果的影响, 并设计实验与传统的加权朴素贝叶斯算法相比, 该算法在查准率、召回率、*F*1 值指标性能上能提升 6% 左右.

关键词: 朴素贝叶斯; 文本分类; 特征加权; 二维信息增益; 加权算法

引用格式: 任世超,黄子良.基于二维信息增益加权的朴素贝叶斯分类算法.计算机系统应用,2019,28(6):135–140. http://www.c-s-a.org.cn/1003-3254/6888.html

Naive Bayes Classification Algorithm of Feature Weighting Based on Two-Dimensional Information Gain

REN Shi-Chao, HUANG Zi-Liang

(School of Communication Engineering, Chengdu University of Information Engineering, Chengdu 610225, China)

Abstract: Naive Bayes algorithm is based on feature-independence assumption and the traditional TF-IDF weighting algorithm, and only considers the distribution of features in the whole training set, but ignores the relationship between feature and categories or documents, so the weights given by traditional method cannot represent its performance. To solve the above problems, this study proposes a naive Bayes classification algorithm of feature weighting based on two-dimensional information gain. It considers the effects of two-dimensional information gain of features, which are the information gain of category and the information gain of documents. Compared with the traditional naive Bayesian algorithm of feature weighting, the proposed algorithm can improve about 6% in the precision, recall, *F*1 value performance.

Key words: naive Bayes; text classification; feature weighting; two-dimensional information gain; weighting algorithm

引言

随着互联网技术的快速发展, 海量的文本信息及其多样化使得文本分类任务越来越受到研究界的关注. 文本分类在信息检索方面能够加速检索过程, 提高检索性能. 同时, 文本分类在新闻专线过滤、专利分类和网页分类方面都发挥了重要的作用. 文本分类中数据往往具有的高维、稀疏、多标号等特点, 这些往往是机器学习需要解决的问题. 因此文本分类在机器学习方面具有重要的价值.

目前虽然有许多算法可以实现文本分类, 如 SVM, KNN, 神经网络、深度学习等, 但是在简洁性和有效性方面朴素贝叶斯算法都要优于其他算法[1–3]. 朴素贝叶斯算法发源于古典数学理论, 有着坚实的数学基础, 以及稳定的分类效率. 在数据较少的情况下仍然有效, 它

① 收稿时间: 2018-11-02; 修改时间: 2018-11-23, 2018-12-11; 采用时间: 2018-12-17; csa 在线出版时间: 2019-05-25

是在贝叶斯定理的基础上提出了一个属性条件独立性假设, 即对于已知类别, 假设所有属性相互独立, 对于分类结果互不影响[4], 所以朴素贝叶斯可以有效的用于文本多分类任务. 因此我们决定使用朴素贝叶斯算法用来文本分类的研究. 由于朴素贝叶斯算法是在条件独立性假设的基础上提出来的, 即所有属性的权值都为 1, 但实际上每个属性对于文档的分类的重要性是不同的, 也就是权值的取值不同. 特征提取是文本分类的关键步骤, 由于不同的加权算法对应权值计算, 直接会对我们的特征的提取以及最终分类的结果造成比较大的影响, 所以研究者们提出不同的权值计算方法来进行改进加权. 如文献[5]中把特征信息增益加到 TF-IDF 算法中相应改善算法性能后, 之后文献[6]又把信息增益与信息熵结合, 文献[7]中提出了根据特征在类间的词频和文档频率重新计算反文档频率, 文献[8]中把各特征相对于类别的互信息作为权重. 但是这些算法没有全面考虑影响特征权重的因素.

本文通过对现有文献中文本分类算法的研究, 提出了基于二维信息增益的加权算法 IGC-IGD (Information Gain of ocument and Category), 从类别信息增益和文档信息增益两个方面综合考虑特征词对分类效果的影响, 并设计实验进一步验证了 IGC-IGD 算法在各个评价指标上都要优于其他算法.

1 朴素贝叶斯及其改进加权算法

1.1 朴素贝叶斯算法

本文采用了文献[9]所给出的贝叶斯模型为多项式模式, 算法思想是: 首先计算出各个类别的先验概率, 再利用贝叶斯定理计算出各特征属于某个类别的后验概率, 通过选出具有最大后验概率 (maximum a posteriori, MAP) 估计值的类别即为最终的类别[9].

算法描述:

设文档类别为$C=\{C_1,C_2,\cdots,C_j\}$, $j=1,2,3,\cdots,V$, 则每个类的先验概率为$P(C_j)$. 设D_i为任意一篇文档, 其包含的特征词为$D_i=\{t_1,t_2,\cdots,t_m\}$, 把$D_i$归为哪一类的概率就是后验概率$P(C_j|D_i)$.

$$P(C_j|D_i)=\frac{P(D_i|C_j)P(C_j)}{P(D_i)} \tag{1}$$

贝叶斯分类的过程就是求解$P(C_j|D_i)$最大值的过程, 显然对于给定的训练文档$P(D_i)$是个常数. 所以求解过程可转化成求解:

$$\max_{C_j\in C}P(C_j|D_t)=\max_{C_j\in C}P(D_t|C_j)P(C_j) \tag{2}$$

因为$D_i=\{t_1,t_2,\cdots,t_m\}$, 根据朴素贝叶斯假设, $\{t_1,t_2,\cdots,t_m\}$各特征相互独立, 所以式 (2) 可等效成:

$$C_{map}=\max_{C_j\in C}P(C_j|D_i)=\max_{C_j\in C}P(C_j)\prod_{i=1}^{m}P(t_i|C_j) \tag{3}$$

其中, $P(C_j)$表示C_j类出现的概率, $P(t_i|C_j)$出现t_i属于C_j类的概率.

1.2 基于 TF-IDF 加权朴素贝叶斯算法

由于朴素贝叶斯算法没有考虑到不同特征对分类效果造成的影响, 通常采用 TFIDF 算法[10]对特征进行特征加权. 加权朴素贝叶斯模型:

$$C_{map}=\max_{C_j\in C}P(C_j|D_t)=\max_{C_j\in C}P(C_j)\prod_{i=1}^{m}P(t_i|C_j)^{W_i} \tag{4}$$

由于每次计算的概率可能会比较小, 为了避免出现下溢的情况, 通常采用对决策规则取对数的形式:

$$\begin{aligned}C_{map}&=\max_{C_j\in C}P(C_j|D_i)\\&=\max_{C_j\in C}[\ln P(C_j)+\sum_{i=1}^{m}\ln P(t_i|C_j)*W_i]\end{aligned} \tag{5}$$

TF-IDF 算法的思想: 特征单词虽然在整个文本集中出现的频率比较低, 但是在某特定文本中出现的频数越大, 则对于该文本的分类作用越大, 反之, 特征单词在大多数文档中出现的频数越大, 对于文本的分类作用越小[6,11]. TF-IDF 算法将词频和反文档频率结合作为特征的权重, 归一化计算方法:

$$IDF(t_i)=lb(\frac{N}{n(t_i)}+0.01) \tag{6}$$

$$W_i=TF(t_i)*IDF(t_i)=\frac{TF(t_i)*IDF(t_i)}{\sqrt{\sum_{i=1}^{m}(TF(t_i)*IDF(t_i))^2}} \tag{7}$$

其中, $TF(t_i)$为特征t_i在训练集中出现的频数, $IDF(t_i)$是反文档频率, N表示训练集的总文档数, $n(t_i)$表示出现特征t_i的文档数.

1.3 基于 TF-IDF*IGC 加权改进算法

虽然 TF-IDF 算法一定程度上能提高分类的精确度, 但效果并不是很明显. 因为该算法只考虑了特征词在训练集中的总体分布情况, 而忽视了特征词在类别中的分布情况对其权重造成的影响. 针对这个问题, 文献[5]的工作主要是把信息论中信息增益应用到文本集合的类别层次上. 提出了一种改进的权重公式 TF-

IDF*IGC, 首先计算出各个类别的信息熵, 然后计算各特征词在每个类别中的条件信息熵, 利用两者的差值计算出单词在各个类别中的信息增益, 把该信息增益反映在权重中, 计算公式:

$$W_i = TF(t_i) * IDF(t_i) * IG(t_i, C_j) \tag{8}$$

$$IG_c(C,t) = E(C) - E(C_j|t) = \sum_{j=1}^{V} P(C_j|t) * lb(P(C_j|t)) - \sum_{j=1}^{V} P(C_j) * lb(P(C_j)) \tag{9}$$

其中, C 为文档的类别集合, $P(C_j)$为类别C_j在训练集中的概率, $P(C_j|t)$为每个特征词 t 在类别C_j中出现的概率, 表示类别总数.

利用 TF-IDF*IGC 算法能够将特征在类别中的信息反应出来, 并同时能够对每个特征权重做一定的修正. 当特征词 t 在某个类别中分布很多, 而在其他类别中分布很少时, 利用信息增益计算公式就能得到很高的信息增益值, 这样就能很好的反应出特征词的分布对分类的影响, 反之就能得到较小的信息增益值[5], 所以在一定程度提高了算法的精确度.

2 基于 IGC-IGD 特征加权朴素贝叶斯算法

由于 1.3 节给出的 TF-IDF*IGC 算法只考虑了特征词在类列间的分布情况并没有考虑到特征词在每个类别文档中的出现情况, 因此会对对权重造成的影响. 以进一步提高算法精度为目标, 针对 TF-IDF*IGC 算法的缺陷, 本节定义一个新的权重计算函数: IGC-IGD 函数. 由于信息增益是描述某个属性对分类效果提升作用的指标, 信息增益越大, 意味着特征属性对文档分类提升越大[4]. 二维信息增益的定义即为同时从特征词关于文档的信息增益和特征词关于类别的信息增益这两个维度进行考虑, 有效的结合了特征在两个方面的性能来刻画特征类别和特征文档对分类作用的提升程度, 这里定义了新的方法求特征类别概率:

$$P(t,C_j) = \frac{tf(t,C_j) + L}{\sum_{j=1}^{V} tf(t,C_j) + V * L} \tag{10}$$

其中, $tf(D_t,C_j)$表示各特征词在类 C_j 中的频数, 所以 $P(t,C_j)$就表示类 C_j 中出现的特征词在训练集该特征词总数中出现的概率. 式中的 L 是为了抑制概率为 0 的情况所加入的平滑因子, 本文中取 L=0.01, V 表示类别数. 同样的方法得到各类别中特征词出现的文档数在训练集中对应特征词所出现的总文档数中出现的概率:

$$P(D_t,C_j) = \frac{tf(D_t,C_j) + L}{\sum_{j=1}^{V} tf(D_t,C_j) + V * L} \tag{11}$$

其中, $tf(D_t,C_j)$表示在类 C_j 中 t 含有特征词的文档数, L=0.01 为平滑因子, V 表示类别数.

传统的求特征文档信息增益的方法仅仅考虑了特征与文档的关系[11], 而忽略了文档与文档类别的关系, 所以这里定义新的求特征文档信息增益的公式把特征与文档的关系同时把文档与文档类别的关系结合在一起, 由此可以得到新的特征类别信息增益和特征文档信息增益:

$$IGC = E(C_j) - E(C_j|t) = \sum_{j=1}^{V} P(t,C_j) * lb(P(t,C_j)) - \sum_{j=1}^{V} P(C_j) * lb(P(C_j)) \tag{12}$$

$$IGD = E(C_j) - E(C_j|D_t) = \sum_{j=1}^{V} P(D_t,C_j) * lb(P(D_t,C_j)) - \sum_{j=1}^{V} P(C_j) * lb(P(C_j)) \tag{13}$$

其中, IGC 表示特征类别信息增益, 刻画特征与类别的关系; IGD 表示特征文档信息增益, 刻画特征与文档的关系; $P(D_t,C_j)$和$P(t,C_j)$分别表示上文提出的求特征类别概率和特征文档概率. 这样得到的两组信息增益能够准确的反应出每个特征词对每个类别的影响力以及每个特征词对每类文档的影响力. 同时把特征词类别信息增益和文档信息增益结合起来, 并采用归一化方法进行处理, 得到权重表示为:

$$W_{IGC-IGD} = \frac{IGC(t_i) * IGD(t_i)}{\sum_{i=1}^{m} (IGC(t_i) * IGD(t_i))^2} \tag{14}$$

其中, $IGC(t_i)$、$IGD(t_i)$ 分别表示数据集的某一个特征的类别信息增益和文档信息增. 式 (14) 首先计算 $IGC(t_i)$×$IGD(t_i)$ 的值获得原始数据, 然后再进行归一化. 归一化的目的是为了使数据等比例的变化, 这样不会影响整体的权重调整. 这也就是二维信息增益的具体定义.

下面举例说明新权重的合理性, 假设训练集包含 3 个类别, 每个类别中有 3 篇文档, 特征词集合为$\{t_1,t_2,t_3\}$分布情况如表 1 所示.

表 1 特征词分布

特征词	类别 1			类别 2			类别 3		
	1	2	3	1	2	3	1	2	3
t_1	5	5	5	0	0	0	0	0	0
t_2	8	0	0	0	8	0	0	0	8
t_3	0	0	0	0	6	0	0	10	0

由表 1 知 t_1 在三个文本中都出现过, t_1 只是在类别 1 中出现过, 说明 t_1 能够准确的代表类别 1 的信息, 应当给予较大的权重, t_3 在三个类别中都出现相同的次数, 说明不具有分类能力, 应当给予较小的权重, 大部分出现在类别 3 中, t_2 所以分类能力要比 t_1 好, 但是比 t_2 要差, 所以权重值应当介于 t_1 和 t_2 之间, 使用以上三个算法得到的权重结果如表 2 所示.

表 2 特征加权算法结果比较

加权算法	各特征单词权重		
	t_1	t_2	t_3
TF-IDF	0.419	0.671	0.612
TFIDF*IGC	0.865	0.000	0.502
IGC-IGD	1.000	0.000	0.142

由表 2 中的结果可以看出, TF-IDF 算法因为针对的是整个训练集中的特征, 所以词频越大的特征被分配的权重越大, 导致结果与实际情况有点截然相反. TF-IDF*IGC 算法不仅考虑了在整个训练集中的情况, 还考虑了特征词与类别间的关系, 所以权重分配比较更合理一些, 但因为仍然与反文档频率相结合导致 t_1 与 t_3 的权重相差很小, 这种时候可能会影响到分类效果, 相比之下 IGC-IGD 算法不仅让没有分类能力的特征词 t_2 权重消零, 而且让 t_1 与 t_3 的权重拉开了差距, 这样能让各个特征起到决定分类作用的效果.

3 实验设计与结果分析

3.1 实验数据及其预处理

实验数据采用国际通用的 20_NewsGroup 数据集进行验证. 该数据集为英文数据集, 从 qwone.com/~jason/20Newsgroups 官网下载, 一共包含 20 个类, 从中选出 6 个类: alt.atheismcomp.graphics, misc.forsale, rec.autos, sci.crypt, talk.politics.guns, 从每个类别中各抽出 100 篇文档, 一共 600 篇文档, 采用交叉验证法[12], 随机选出 60%(360 篇) 作为训练集, 40%(240 篇) 作为测试集.

实验数据预处理: 去除掉标点符号, 停用词, 数字以及一些特殊符号, 为了降低空间复杂度和分类计算的时间, 把词频特作为选择单词的标准, 对每次选取 500 特征进行实验, 重复实验十次求平均值来验证算法的准确性.

3.2 实验结果分析

本实验中使用上文介绍的三个加权算法与朴素贝叶斯算法结合进行实验. 采用查准率 (P)[12,13], 召回率 (R)[14,15], $F1$ 值 ($F1$) 和宏 $F1$ 值 ($Macro\_F1$) 四个指标[16–20] 算公式如下:

查准率: $P = \dfrac{TP}{TP+FP}$

宏查准率: $Macro\_P = \dfrac{1}{V}\sum_{i=1}^{V} P$, V 表示类别数

召回率: $R = \dfrac{TP}{TP+FN}$

宏召回率: $Macro\_R = \dfrac{1}{V}\sum_{i=1}^{V} R$

$F1$ 值: $F1 = \dfrac{2*P*R}{P+R}$

宏 $F1$ 值: $Macro\_F1 = \dfrac{2*Macro\_P*Macro\_R}{Macro\_P+Macro\_R}$

TP 表示正确的标记为正, FP 错误的标记为正, FN 错误的标记为负, TN 正确的标记为负[4], 如表 3 所示.

表 3 参数含义

真实情况	预测情况	
	正例	反例
正例	TP	FN
反例	FP	TN

实验结果如表 4 所示.

表 4 算法测试结果比较

类别	基于二维信息增益 IGC-IGD 加权的朴素贝叶斯文本分类算法 (IGDCNB)			基于 TFIDF*IGC 特征加权的朴素贝叶斯文本分类算法 (TFIDF*IGCNB)			基于传统的 TFIDF 特征加权的朴素贝叶斯文本分类 (TFIDFNB)		
	P	R	$F1$	P	R	$F1$	P	R	$F1$
alt.atheism	0.9740	0.9238	0.9450	0.9682	0.9062	0.9313	0.9329	0.8989	0.9115
comp.graphics	0.9440	0.9405	0.9392	0.9199	0.9609	0.9329	0.8541	0.9666	0.9014
misc.forsale	0.9093	0.9838	0.9446	0.8884	0.9947	0.9357	0.8454	0.9947	0.9109
rec.autos	0.9808	0.8640	0.9093	0.9853	0.7954	0.8556	0.9228	0.8442	0.8749
sci.crypt	0.9329	0.9977	0.9673	0.9198	0.9805	0.9488	0.9775	0.8369	0.9002
talk.politics.guns	0.9590	0.9775	0.9674	0.9513	0.9607	0.9541	0.9722	0.9331	0.9504
平均值	0.9500	0.9478	0.9455	0.9388	0.9331	0.9264	0.9175	0.9124	0.9082

由表 4 可以看出, 当使用基于二维信息增益 IGC-IGD 加权的朴素贝叶斯文本分类算法时, 查准率, 召回率和 $F1$ 值这三个指标总体上都有明显的提高. 具体的, IGDCNB 算法对所有类别的查准率都要高于其他两种算法, 除了 comp.graphics, misc.forsale 两个类别的召回率略低于 TF-IDF*IGC 特征加权的朴素贝叶斯文本分类算法, 对于 $F1$ 值, IGDCNB 算法也都要领先于其他两种算法, 个别类别比如 sci.crypt 和 talk.politics.guns, 虽然传统的基于 TF-IDF 特征加权的朴素贝叶斯文本分类算法具有较高的查准率, 但其召回率却远远低于 IGDCNB 算法, 这也不是研究者希望出现的. 总体上看, 与 TF-IDF*IGC 加权算法相比, 三个指标都能提高 3% 到 4%; 与 TF-IDF 加权算法相比三个指标都能提高 4% 到 6%, 这充分证明了 IGC-IGD 加算法的有效性. 查准率对应实际被分类的比例, 召回率对应应该被分类的比例. 由于查准率较高时召回率不一定能够很高, 所以本文最后采用比较三种算法的宏 $F1$ 值的方法来验证 IGC-IGD 加权的朴素贝叶斯文本分类算法的有效性. 本文通过选择不同的特征数来验证算法的准确性, 实验结果如图 1 所示.

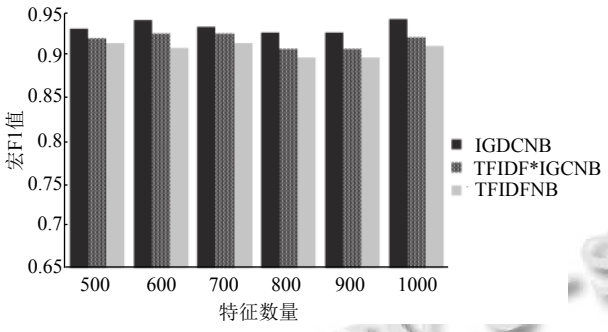

图 1 3 种算法宏 $F1$ 值比较

$F1$ 值对应查准率 P 和召回率 R 的调和均值. 宏 $F1$ 值是所有类别对应的 $F1$ 值得平均, 能够反应各加权算法整体性能 (查准率、召回率、$F1$ 值) 的指标. 由图 1 可以看出, 当特征数量从 500 增加到 1000 时, IGC-IGD 加权的朴素贝叶斯分类算法的宏 $F1$ 值要高于其他两个算法, 根据 3.2 宏 $F1$ 值的计算公式计算可以得到该算法相比传统的加权算法宏 $F1$ 值提升将近 6% 左右, 而且该算法本身不会因为特征数量的变化出现较大的波动, 说明在给定一定量有价值的特征时, 二维信息增益 IGC-IGD 加权的朴素贝叶斯分类算法能够有效的对文本进行分类.

4 结束语

本文通过有机的结合特征类别信息增益 (IGC) 和特征文档信息增益 (IGD), 提出了二维信息增益加权的朴素贝叶斯分类算法, 充分利用了文本中特征的二维信息, 克服了传统朴素贝叶斯算法分类性能上的缺陷, 通过实验进一步验证了该算法的有效性. 为了进一步提升朴素贝叶斯文本分类算法的性能, 以得到更为精确迅速的分类方法. 下一步的工作将会从朴素贝叶斯算法中的条件概率计算方法这个方面进行改进.

参考文献

1 邸鹏, 段利国. 一种新型朴素贝叶斯文本分类算法. 数据采集与处理, 2014, 29(1): 71–75. [doi: 10.3969/j.issn.1004-9037.2014.01.010]
2 Han JW, Kamber M. 数据挖掘: 概念与技术. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2007.
3 李忠波, 杨建华, 刘文琦. 基于数据填补和连续属性的朴素贝叶斯算法. 计算机工程与应用, 2016, 52(1): 133–140. [doi: 10.3778/j.issn.1002-8331.1401-0232]
4 周志华. 机器学习. 北京: 清华大学出版社, 2016.
5 张玉芳, 陈小莉, 熊忠阳. 基于信息增益的特征词权重调整算法研究. 计算机工程与应用, 2007, 43(35): 159–161. [doi: 10.3321/j.issn:1002-8331.2007.35.048]
6 李学明, 李海瑞, 薛亮, 等. 基于信息增益与信息熵的 TFIDF 算法. 计算机工程, 2012, 38(8): 37–40. [doi: 10.3778/j.issn.1002-8331.2012.08.011]
7 饶丽丽, 刘雄辉, 张东站. 基于特征相关的改进加权朴素贝叶斯分类算法. 厦门大学学报 (自然科学版), 2012, 51(4): 682–685.
8 武建军, 李昌兵. 基于互信息的加权朴素贝叶斯文本分类算法. 计算机系统应用, 2017, 26(7): 178–182.
9 贺鸣, 孙建军, 成颖. 基于朴素贝叶斯的文本分类研究综述. 情报科学, 2016, 34(7): 147–154.
10 Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing & Management, 1988, 24(5): 513–523.
11 李凯齐, 刁兴春, 曹建军. 基于信息增益的文本特征权重改进算法. 计算机工程, 2011, 37(1): 16–18, 21. [doi: 10.3969/j.issn.1000-3428.2011.01.006]
12 Jiang LX, Li CQ, Wang SS, *et al*. Deep feature weighting for naive Bayes and its application to text classification. Engineering Applications of Artificial Intelligence, 2016, 52: 26–39. [doi: 10.1016/j.engappai.2016.02.002]

13 Zhang LG, Jiang LX, Li CQ, *et al*. Two feature weighting approaches for naive Bayes text classifiers. Knowledge-Based Systems, 2016, 100: 137–144. [doi: 10.1016/j.knosys.2016.02.017]

14 Song Y, Kołcz A, Lee Giles C. Better Naive Bayes classification for high-precision spam detection. Software—Practice & Experience, 2009, 39(11): 1003–1024.

15 He W, Zhang Y, Yu SJ, *et al*. Deep feature weighting with a novel information gain for naive Bayes text classification. JIHMSP, 2019, 10(1).

16 Wu J, Cai ZH, Zhu XQ. Self-adaptive probability estimation for naive Bayes classification. Proceedings of 2013 International Joint Conference on Neural Networks. Dallas, TX, USA. 2013. 1–8.

17 Li L, Li C. Research and improvement of a spam filter based on naive Bayes. Proceedings of the 7th International Conference on Intelligent Human-Machine Systems and Cybernetics. Hangzhou, China. 2015. 361–364.

18 Jiang QW, Wang W, Han X, *et al*. Deep feature weighting in Naive Bayes for Chinese text classification. Proceedings of the 2016 4th International Conference on Cloud Computing and Intelligence Systems. Beijing, China. 2016. 160–164.

19 Arar ÖF, Ayan K. A feature dependent naive Bayes approach and its application to the software defect prediction problem. Applied Soft Computing, 2017, 59: 197–209. [doi: 10.1016/j.asoc.2017.05.043]

20 Chen JN, Huang HK, Tian SF, *et al*. Feature selection for text classification with naïve Bayes. Expert Systems with Applications, 2009, 36(3): 5432–5435. [doi: 10.1016/j.eswa.2008.06.054]