动态置信度的序列选择增量学习方法①

李 念[1,2], 廖闻剑[2], 彭艳兵[2]

[1](武汉邮电科学研究院, 武汉 430074)
[2](烽火通信科技股份有限公司 研发部, 南京 210019)

摘 要: 贝叶斯在训练样本不完备的情况下, 对未知类别新增训练集进行增量学习时, 会将分类错误的训练样本过早地加入到分类器中而降低其性能, 另外增量学习采用固定的置信度评估参数会使其效率低下, 泛化性能不稳定. 为解决上述问题, 提出一种动态置信度的序列选择增量学习方法. 首先, 在现有的分类器基础上选出分类正确的文本组成新增训练子集. 其次, 利用置信度动态监控分类器性能来对新增训练子集进行批量实例选择. 最后, 通过选择合理的学习序列来强化完备数据的积极影响, 弱化噪声数据的消极影响, 并实现对测试文本的分类. 实验结果表明, 本文提出的方法在有效提高分类精度的同时也能明显改善增量学习效率.

关键词: 贝叶斯分类器; 增量学习; 置信度; 序列选择

Incremental Learning Method of Dynamic Confidence Level and Sequence Selectable

LI Nian[1,2], LIAO Wen-Jian[2], PENG Yan-Bing[2]

[1](Wuhan Research Institute of Posts and Telecommunications, Wuhan 430074, China)
[2](FiberHome Communications Science&Technology Development Co., Ltd., Nanjing 210019, China)

Abstract: Under the condition of insufficiency of the training sets, Bayesian will easily make the classification of the new incremental and unlabeled training texts incorrectly. If these incorrectly labeled texts are added to the Bayesian classifier early, it will reduce the performance of Bayesian classifier. In addition, incremental learning with fixed confidence level parameter will cause low learning efficiency and instable generalization ability. In order to solve the above problems, this paper proposes an incremental learning method of dynamic confidence level and sequence selectable. Firstly, the new incremental training subsets are made up of these texts which are classified by current Bayesian classifier correctly. Secondly, it uses confidence level to dynamically monitor the performance of classifier, and then chooses texts from the new incremental training subsets. Finally, strengthen the positive impact of the more mature data, weaken the negative impact of the noise data, and complete the text classification of the test sets by choosing reasonable learning sequence. The experimental results show that the classification efficiency and precision are both advanced by using the method this paper proposes.

Key words: Bayesian classification; incremental learning; confidence level; sequence selectable

作为信息处理领域的重要研究方向之一, 文本分类是指分析文本内容, 通过预先构造的分类器来给文本分配一个合适的类别以缩小信息检索的范围[1], 提高文本检索效率. 较为著名的文本分类技术有朴素贝叶斯(NB)[2]、K-最邻近(KNN)[3]、支持向量机(SVM)[4]、决策树(Decision Tree)[5,6]、线性最小二乘法拟合和神经网络[7]等方法, 朴素贝叶斯由于其计算高效、精确度高, 并且具有坚实的理论基础而得到广泛的应用[8].

目前大多数文本分类系统存在训练样本不完备且分类体系复杂易变更, 致使训练初期分类器分类效果不够理想, 朴素贝叶斯由于其健壮性且能够充分利用先验知识和样本信息, 使其成为增量式分类的自然选

① 项目基金:国家高技术研究发展计划(863)(2012AA013002)
收稿时间:2015-05-20;收到修改稿时间:2015-06-15

择. 对于贝叶斯增量学习, 很多研究者提出了方法的改进, 龚秀军等[9,10]提出的基于 0-1 分类损失[11]增量学习算法的问题, 用分类损失来确定新增训练集中文本加入到原始训练集中的顺序, 但是没有考虑到原始样本数少, 知识储备不足等因素造成对新增训练集的分类产生很多错误的分类结果, 如果这些错误的分类结果过早地选择加入到原始训练集中, 会使噪声数据一直传播下去, 进而影响整体的分类精度; 罗福星等[12]讨论了一种基于贝叶斯加权的增量学习算法, 通过设置可信度来确定测试集文本加入到训练集样本中, 类可信度的动态调整可以加快分类器的增量学习过程, 但是直接将测试集样本加入到训练集样本中会使分类器出现过拟合, 导致其泛化能力不强; 马后锋等[13]提出了一种改进的增量贝叶斯分类算法, 通过选择合理的学习序列来强化较完备数据对分类的积极影响, 但是需要每次迭代寻出最优个体, 极大地弱化了增量学习的效率.

基于以上分析, 本文提出一种动态置信度的序列选择增量学习方法, 通过最大化分类置信度, 从新增训练集中选出完备训练样本加入到原始训练集中, 利用微平均 $F1$ 均值实时监控并评估分类器性能, 动态调整 α 阈值(置信度因子)来改善分类器学习效率, 有效兼顾增量学习样本数据的完备性和增量学习的高效性. 最后通过实验验证该方法有效提高分类精度的同时也能明显改善增量学习效率.

1　加权朴素贝叶斯分类模型

分类模型, 在贝叶斯原理的基础上, 假定各条件属性对决策属性的作用是相互独立的, 通过类先验概率和文本的类条件概率计算后验概率的模式识别方法, 文本 d_j 归属某一类别 C_i $(1 \le i \le n)$ 的概率可用下面公式求的:

$$C_{NB}(d_j) = \arg_i[\max \prod_{k=1}^{n} P(t_k \mid C_i) \cdot P(C_i)] \qquad (1)$$

其中 $P(C_i)$ 为类 C_i 条件概率, 其值为属于类别 C_i 的文本数与总的训练文本数的比值, $P(t_k \mid C_i)$ 为特征项 t_k 在类别中 C_i 出现的概率, 这里综合考虑特征项词频和文档词频, 利用郑霖等[14]提出的改进 $tf\text{-}idf$ 权重计算公式:

$$w_{kj} = \frac{\log(tf_{kj}+1) \times idf'}{\sqrt{\sum_{t_k \in d_j} [\log(tf_{kj}+1) \times idf']^2}} \qquad (2)$$

改进后的 idf' 的公式如下:

$$idf' = \log(\frac{\frac{m}{m+j}}{\frac{m}{m+j}+\frac{k}{k+p}} \times N) \qquad (3)$$

其中, m 为类 C_i 内包含特征项 t_k 的文档数, j 为类 C_i 内不含特征项 t_k 的文档数, k 为非类 C_i 包含特征项 t_k 的文档数, p 为非类 C_i 不包含特征项 t_k 的文档数. $\frac{m}{m+j}$ 表示包含特征项 t_k 的文档在类 C_i 文档集内的比例, $\frac{k}{k+p}$ 表示包含特征项 t_k 的文档在非类 C_i 文档集中的比例, 改进后的 idf' 有效解决类别之间分布均匀对类别区分度不大的特征项赋予很高的权值和一个类别内部只集中在某几个文本的特征项赋予很高的权值的问题, 其将用于本文的文本分类预处理.

2　增量贝叶斯模型

朴素贝叶斯算法依据训练样本的类先验概率和类条件概率来预测后验概率, 得出测试样本的类别标签, 如果训练样本没有良好的数据完备性, 使得预测的样本类别标签可能不准确; 或是遭遇信息如潮涌般地涌入, 不可能一次将训练集全部放入到内存中, 一般的分类算法会表现得手足无措, 通过引进增量学习, 可以很好地解决上述问题, 朴素贝叶斯充分利用样本信息的各个特点使其成为增量学习的最佳模型.

在增量贝叶斯模型中, 设事件发生的先验概率用参数 C 来表示, $P(C \mid T_0)$ 表示其概率密度函数, 其中 T_0 表示先验认识, 依据贝叶斯原理, 当有新的样本 S 进入到训练集时, 根据先验知识 $P(C \mid T_0)$, 经计算得出后验知识 $P(C \mid S, T_0)$ 的公式为:

$$P(C \mid S, T_0) = \frac{P(S \mid C, T_0) P(C \mid T_0)}{P(S \mid T_0)} \qquad (4)$$

其中:

$$P(S \mid T_0) = \int [P(S \mid C, T_0) P(C \mid T_0)] dC \qquad (5)$$

通过图 1 可以看出, 贝叶斯增量学习模型全面综合了先验知识和 T_0 与新增的样本信息 S, 来得到最终的后验知识 T_1, 即:

后验知识(T_1)=先验信息(T_0)+新增信息(S)　(6)

这样一旦有新增样本抵达, 之前积累的后验知识立马会转换成此刻的先验知识, 所以贝叶斯增量学习

的实质是一个不断修正完善自身的动态过程, 即不断地使用新增信息来完善自身当前信息的过程, 这里有一个重要的前提就是后验知识必须与先验知识同分布[15], 否则是不能使用当前的后验知识作为下一次的先验知识.

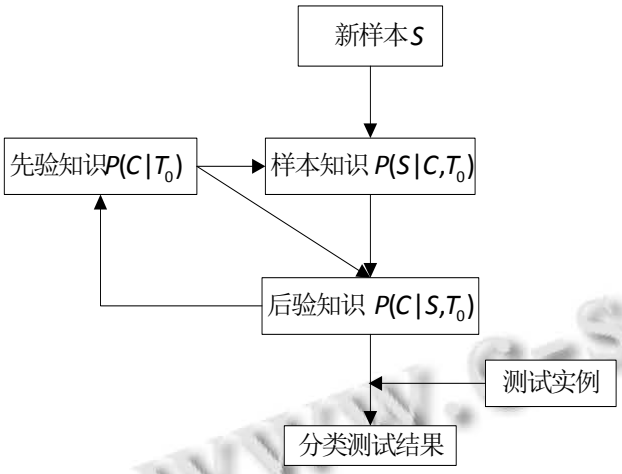

图 1 增量贝叶斯模型

3 动态置信度的序列选择增量学习方法

3.1 方法思想

为了下面叙述的方便, 对文中出现的符号做出约定: 原始训练集 D, 分类器 C, 未标注类别的新增训练集 S, 测试集 O, 置信度因子 α, 微平均 $F1$ ($Micro\_F1$).

利用改进的互信息公式和 *tf-idf* 加权公式对原始训练集进行特征选择和特征加权, 构造出当前分类器 C, 分类器 C 对未标注类别新增训练集 S 进行分类, 然后选出分类正确的文本组成新增训练子集 S'. 在最大化置信度的前提下, 计算新增训练子集 S' 中置信度最大的文本和新增训练子集 S' 中每个文本加入到分类器后对测试集 O 的微平均 $F1$, 根据微平均 $F1$ 的均值, 调整置信度因子 α 的阈值, 依据 α 值从新增训练子集 S' 中选出完备样本加入到训练集 D 中. 改进后的方法不仅使分类错误的文本加入到训练集 D 中的机率大大降低, 弱化噪声文本的影响, 同时随着分类器知识储备的充实, α 的动态调整将加快增量学习的学习进度, 使分类器在最短时间达到最优化.

该方法从两个方面进行了优化, 考虑到初始时分类器性能较弱, 通过序列选择让完备数据加入到分类器中, 有效降低了初始时噪声数据加入到分类器中的几率. 在增量学习的中后期, 随着分类器知识储备的完善, 通过调整置信度因子的阈值, 加快增量学习进度, 达到分类精度和增量学习效率的兼优.

整个动态置信度序列选择增量学习方法可以分为三步: 第一步通过当前分类器 C 过滤掉噪声数据, 第二步利用新增训练子集 S' 对测试集 O 微平均 $F1$ 的均值来检测分类器 C 的性能, 判断是否调整置信度因子 α 的阈值, 第三步在第二步的基础上利用置信度进行增量实例选择.

3.2 增量方法描述与置信度

在无信息的 *Dirchlet* 先验假设条件下, 根据先验知识和样本信息可以得到样本条件概率 $P(C_i)$ 和类条件概率 $P(t_k|C_i)$ 的计算公式, 公式如下:

$$P(C_i)=\frac{1+|D_{C_i}|}{|C|+|D|} \tag{7}$$

$$P(t_k|C_i)=\frac{1+m}{2+|D_{C_i}|} \tag{8}$$

式中 $|D_{C_i}|$ 为训练集样本类别为 C_i 样本数, $|D|$ 为训练样本总数, $|C|$ 为类别数目, m 为类别 C_i 内包含特征项 t_k 的文档数.

当有新的增量样本 s_p 加入到训练集中时 ($D+\{<s_p,C_p>\}$), 后验知识可由先验知识和新增文本信息共同得到, $P(C_i)$ 与 $P(t_k|C_i)$ 的更新公式如下:

$$P^*(C_i)=\begin{cases}\dfrac{|C|+|D|}{1+|C|+|D|}*P(C_i), 当C_p\neq C_i时\\ \dfrac{|C|+|D|}{1+|C|+|D|}*P(C_i)+\dfrac{1}{1+|C|+|D|}, 当C_p=C_i时\end{cases} \tag{9}$$

$$P^*(t_k|C_i)=\begin{cases}\dfrac{1+|D_{C_i}|}{2+|D_{C_i}|}*P(t_k|C_i), 当C_p=C_i且t_k\notin s_p时\\ \dfrac{1+|D_{C_i}|}{2+|D_{C_i}|}*P(t_k|C_i)+\dfrac{1}{2+|D_{C_i}|}, 当C_p=C_i且t_k\in s_p时\\ P(t_k|C_i), 当C_p\neq C_i时\end{cases} \tag{10}$$

通过分析上述公式变化可知, s_p 加入训练集 D 后, 与它相关项的估计变化较大, 而与它无关项变化较小, 为了降低测试时算法的复杂度, 计算只在与测试文本 s_p 中相同的特征项进行, 忽略与 s_p 无关项的影响, 采用龚秀军[9]提出的下式求解:

$$P^*(C_q|o_q)=\frac{P(C_q|o_q)\prod\limits_{t_k\in s_p\wedge o_q}P^*(t_k|C_q)}{\prod\limits_{t_k\in s_p\wedge o_q}P(t_k|C_q)} \tag{11}$$

式中 $P(t_k|C_q)$ 是训练集 D 训练出来的类条件概率，$P^*(t_k|C_q)$ 是训练集 D 增加了文本 s_p，即 $D+\{<s_p,C_p>\}$ 之后得到类条件概率.

衡量分类器精度标准是它的分类效果，在最大化置信度的前提下，选择 $s_p \in S$，在当前分类器下获得类标签 C_p，可由以下公式计算在 $D+\{<s_p,C_p>\}$ 下估计测试集 O 的分类置信度:

$$K(D^*)=\frac{\alpha}{|O|+1}\sum_{o\in O}(\max P^{D^*}(C_i|o)) \qquad (12)$$

式中 $|O|$ 代表测试集 O 中的文本数，D^* 是指训练集 D 中增加的新文本 s_p，$P^{D^*}(C_i|o)$ 可由(11)式可以计算. α 为置信度因子，其值越大，系统对于参与增量学习的文本要求越高，根据经验得出 $\alpha \in [0.75,1]$ 系统都能得到较好的效果.

3.3 方法步骤及整体增量学习流程

输入: 训练集: $D=\{d_1,d_2,...,d_N\}$

新增训练集(不含类别标签):

$S=\{s_1,s_2,...,s_m\}$

测试集: $O=\{o_1,o_2,...,o_m\}$

输出: 分类器 C

输入: 训练集: $D=\{d_1,d_2,...,d_N\}$

新增训练集(不含类别标签):

$S=\{s_1,s_2,...,s_m\}$

测试集: $O=\{o_1,o_2,...,o_m\}$

输出: 分类器 C

第一步: 利用互信息公式[16]和文中给出的(2)式来对训练集 D 进行特征选择与特征加权，构造当前分类器 C;

第二步: 若 $S=\Phi$，返回初始分类器 C，方法结束; 否则继续;

第三步: 利用当前的分类器 C，对新增的训练集 S 中每一个文本 s_p 进行分类，获得其类别标注 C_p; 选出当前分类器 C 分类正确的文本组成新增训练子集 $S^{'} \subset S$;

第四步: 令 k=0，α=1，对 $S^{'}$ 中的每一个文本 $s_p \in S^{'}$，重复:

①新文本 s_p 加入到训练集 D ($D+\{<s_p,C_p>\}$)中，并更新分类器，再对测试集 O 中的每一个文本分类，计算分类置信度 K 和微平均 $Micro\_F1$ 值;

②if $K>k$ then $k=K$;

第五步: 根据 $\overline{F1}$，调整 α，对 $S^{'}$ 中的每一个文本 $s_p \in S^{'}$ 重复: if $K_{s_p} \in [\alpha k,k]$ then $C.append(C_p)$, $temp.append(s_p)$;

第六步: $D+\{<temp,C>\}$，更新分类器，$S=S-temp$，然后返回第一步继续执行.

注: (1)第五步中 $\overline{F1}=\frac{1}{N_{S^{'}}}\sum_{p=1}^{N_{S^{'}}}Micro\_F1_{s_p}$，其中 $N_{S^{'}}$ 表示 $S^{'}$ 中文本数，$Micro\_F1_{s_p}$ 为 s_p 的 $Micro\_F1$ 值; (2)第六步中 $\{<temp,C>\}$ 是新增训练集组，即包含多个训练实例.

整体增量学习流程图 2 所示.

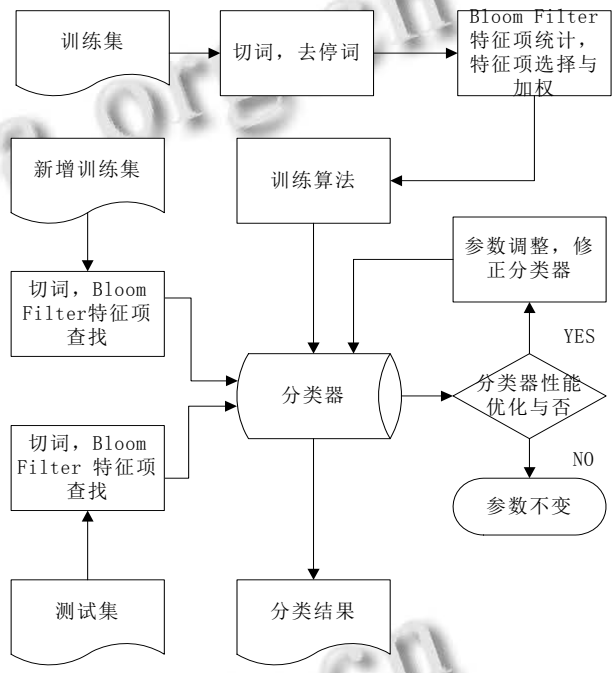

图 2　增量学习整体分类流程

4　实验测试

4.1 实验性能评估指标

对文本分类器性能的评估方法选择传统的评价标准: 微平均正确率 $Micro\_P$、微平均召回率 $Micro\_R$ 和微平均 $F1$ 值，计算公式如下:

$$Micro\_P=\frac{\sum_{i=1}^{k}T_i}{\sum_{i=1}^{k}M_i} \qquad (13)$$

$$Micro\_R=\frac{\sum_{i=1}^{k}T_i}{\sum_{i=1}^{k}N_i} \qquad (14)$$

$$Micro\_F1=\frac{2\times Micro\_P\times Micro\_R}{Micro\_P+Micro\_R} \qquad (15)$$

其中 T_i、M_i、N_i 表示第 i 类的结果中正确的文本个数、结果中出现的个数和实际包含的文本个数, 这里只进行单类别分类(一个文档只给出一个预测类别).

4.2 实验设计与结果

本实验采用的语料集是利用 Python 爬取了新浪博客近四个月的 XML 文件, 进行数据清洗后, 将文本类型分为科技类和非科技类, 各 2000 篇, 进行以下两组实验.

实验一: 从语料集中总共随机抽取 8 组实验数据, 每组实验数据包括: 随机抽取 100 篇科技类和 100 篇非科技类文本组成原始训练集; 随机选取 100 篇文本作为测试集; 分别选取 200、400 和 600 篇文本作为新增训练集. 其中训练集、新增训练集和测试集之间均没有交集, 利用相同的原始训练集和测试集分别对不同规模的新增训练集的进行增量学习.

实验二: 在实验一的基础上, 以 100 为等差来递增新增训练集, 计算不同数目的新增训练集完成增量学习所用的平均时间.

参数设定与方法说明: 1)实验采用的分词工具是利用 Python 直接调用 jieba 分词插件, 用 Bloom Filter 实现特征项词频和文档词频的统计; 2)本实验采用樊兴华[16]提出的改进的互信息公式, 其在特征项数目较少的情况下取得较好的分类性能, 特征项数目变化范围为 400-2000; 3)方法利用改进的词频加权公式来修正分类器, 以此凸显特征项在不同类别中的重要性; 4)实验里新增训练子集是采用当前分类器判别值的倍数来判定正确类别的, 即选取最大类别除以第二大类别值的倍数, 然后选择排序靠前的文本组成新增训练子集 S'; 5)为了提高实验效率, 采用缓冲池的方法将新增训练集加入到测试集中; 6)根据新增训练子集 S' 对测试集的 $\overline{F1}$, 相应调整 α 值, α 以 0.01 步进递减, 实现增量过程中的批量学习, 降低后期增量学习的时间开销. 实验一的 8 组实验结果如表 1 所示, 实验二的实验结果如图 3 所示, 其中方法 1 采用的是龚秀军[3]提出的增量学习算法, 方法 2 是本文提出的改进的增量学习方法.

表 1 8 组增量学习实验性能比较

数据集规模与方法	新增训练集(200) 平均值($\overline{F1}$ 值)(%)		新增训练集(400) 平均值($\overline{F1}$ 值)(%)		新增训练集(600) 平均值($\overline{F1}$ 值)(%)	
实验组别	方法 1	方法 2	方法 1	方法 2	方法 1	方法 2
1	84.32	**85.56**	86.21	**86.79**	84.86	**87.34**
2	89.81	**91.89**	87.64	**88.26**	81.34	**88.49**
3	84.98	80.32	83.26	82.17	86.22	**90.01**
4	85.73	**88.75**	83.07	**86.22**	85.61	**86.45**
5	83.64	**87.11**	88.52	**89.73**	87.96	**90.39**
6	91.12	**92.16**	89.85	**91.68**	88.63	**91.33**
7	86.21	**90.98**	88.23	87.34	90.10	88.16
8	84.90	**85.25**	85.69	**88.32**	86.23	**87.09**

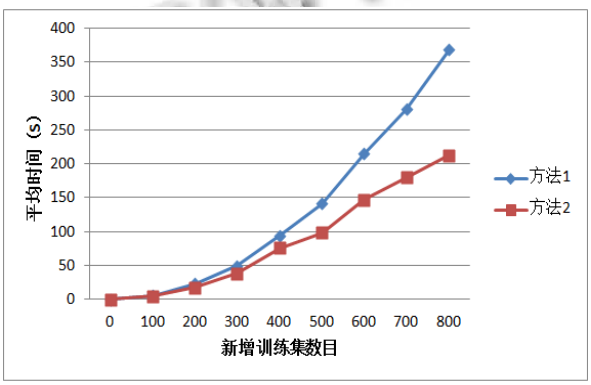

图 3 不同新增训练集完成增量学习对应的时间

4.3 实验结果分析

对表 1 和图 2 中的测试结果进行以下分析:

1)从表 1 中可以看出, 方法 2 对比方法 1 在分类精度上基本都有所提高.

2)表 1 中有些实验结果显示方法 1 的分类精度比方法 2 要高, 主要原因是新增训练集中科技类文本的特征信息不明显而容易产生分类错误, 改进后的方法使新增训练文本选择范围减小, 受到噪声数据影响所致, 但随着新增训练集规模的增大和特征信息的完备, 分类器的性能会逐渐完善. 例如第三组数据, 新增训练集为 200 篇时分类精度较低, 但新增训练集为 400

和 600 篇时分类精度较高.

3)随着新增样本数据的增加, 分类器的知识储备不断完善, 类别区分度不断加强, 受到噪声数据的干扰也随之降低, 改进后的方法表现出来的分类性能相对比较稳定.

4)从图 3 可以看出, 随着新增数据集的增加, 方法 1 增量学习所用时间呈指数递增, 方法 2 则呈线性递增, 当新增数据集规模较大时, 方法 2 要明显优于方法 1, 大大缩减增量学习所用时间的时间开销.

5 结语

本文提出的一种动态置信度的序列选择增量学习方法, 主要通过提高分类器较强类型辨别能力和过滤噪声数据对分类器的影响等方法来选择合理的学习序列, 强化完备数据的积极影响, 弱化噪声数据的消极影响; 通过动态监控分类器性能来调整置信度因子阈值, 加快增量学习过程, 实验表明改进的方法提高了分类器分类精度的同时也提高了增量学习效率.

同时本文也存在一些问题: 增量学习过程中的动态特征选择、更为有效过滤噪声数据的方法、批量学习过程合理的参数调整机制以及增量学习性能不稳定等问题, 这些问题是未来进一步研究的主要内容.

参考文献

1 Sebastiani F. Machine learning in automated text categorization. ACM Computer Surveys, 2002, 34(1): 11–12, 32–33.

2 McCallum A, Nigam K. A comparison of event models for Naïve Bayes text classification. AAAI-98 Workshop on Learning for Text Categorization. 1998. 41–48.

3 Han EH, Karypis G, Kumar V. Text categorization using weight adjusted k-nearest neighbor classification. University of Minnesota, 1999: 11–12.

4 Joachims T. Text categorization with support vector machines: Learning with many relevant features. Machine Learning: ECML-98, Tenth European Conference on Machine Learning. 1998. 137–142.

5 Fuhr N, Buckley C. A probabilistic learning approach for document indexing. Information Systems, 1999, 9(3): 223–248.

6 Zhou ZH, Tang W. Selective ensemble of decision trees. Lecture Notes in Artificial Intelligence 2639, Berlin: Springer, 2003: 476–483.

7 Ruiz M, Srinivasan P. Hierarchical text categorization using neural networks. Information Retrieval, 2002, 5(1): 87–118.

8 Yager R. An extension of the native Bayesian classifier Information Sciences 2006 176: 577–588.

9 龚秀军,刘少辉,史忠植.一种增量贝叶斯分类模型.计算机学报,2002,25(6): 654–650.

10 姜卯生,王浩,姚宏亮.朴素贝叶斯分类器增量学习序列算法研究.计算机工程与应用, 2004,14,57–57.

11 Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 2007, 29(2-3): 103–130.

12 罗福星,刘卫国.一种朴素贝叶斯分类增量学习算法.微计算机应用,2008,29(6):107–112.

13 马后锋,樊兴华.一种改进的增量贝叶斯分类算法.仪器仪表学报,2007,28(8III):312–316.

14 郑霖,徐德华.基于改进的 TFIDF 算法的文本分类研究.计算机与现代化,2014,229:6–10.

15 Samuel K. Modern Bayesian Statistics. George Washington University Press, 2000: 109.

16 樊兴华,孙茂松.一种高性能的两种中文文本分类方法.计算机学报,2006,29(1):124–131.