计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2021.20200521
58(9): 2040-2051, 2021

一种双层贝叶斯模型:随机森林朴素贝叶斯

张文钧[1] 蒋良孝[1,2] 张 欢[1] 陈 龙[1]
[1](中国地质大学计算机学院 武汉 430074)
[2](智能地学信息处理湖北省重点实验室(中国地质大学) 武汉 430074)
(wjzhang@cug.edu.cn)

A Two-Layer Bayes Model: Random Forest Naive Bayes

Zhang Wenjun[1], Jiang Liangxiao[1,2], Zhang Huan[1], and Chen Long[1]
[1](*School of Computer Science, China University of Geosciences, Wuhan* 430074)
[2](*Hubei Key Laboratory of Intelligent Geo-Information Processing (China University of Geosciences), Wuhan* 430074)

Abstract Text classification is an essential task in natural language processing. The high dimension and sparsity of text data bring many problems and challenges to text classification. Naive Bayes (NB) is widely used in text classification due to its simplicity, efficiency and comprehensibility, but its attribute conditional independence assumption is rarely met in real-world text data and thus affects its classification performance. In order to weaken the attribute conditional independence assumption required by NB, scholars have proposed a variety of improved approaches, mainly including structure extension, instance selection, instance weighting, feature selection, and feature weighting. However, all these approaches construct NB classification models based on the independent term features, which restricts their classification performance to a certain extent. In this paper, we try to improve the naive Bayes text classification model by feature learning and thus propose a two-layer Bayes model called random forest naive Bayes (RFNB). RFNB is divided into two layers. In the first layer, random forest (RF) is used to learn high-level features of term combinations from original term features. Then the learned new features are input into the second layer, which is used to construct a Bernoulli naive Bayes model after one-hot encoding. The experimental results on a large number of widely used text datasets show that the proposed RFNB significantly outperforms the existing state-of-the-art naive Bayes text classification models and other classical text classification models.

Key words naive Bayes (NB); random forest; feature learning; feature representation; text classification

摘 要 文本分类是自然语言处理领域的一项基础工作.文本数据的高维性和稀疏性,给文本分类带来了许多问题和挑战.朴素贝叶斯模型因其简单、高效、易理解的特点被广泛应用于文本分类任务,但其属性条件独立假设在现实的文本数据中很难满足,从而影响了它的分类性能.为了削弱朴素贝叶斯的属性条件独立假设,学者们提出了许多改进方法,主要包括结构扩展、实例选择、实例加权、特征选择、特征加权等.然而,所有这些方法都是基于独立的单词特征来构建朴素贝叶斯分类模型,在一定程度上限制了它们的分类性能.为此,尝试用特征学习的方法来改进朴素贝叶斯文本分类模型,提出了一种双层贝叶斯模型:随机森林朴素贝叶斯(random forest naive Bayes, RFNB).RFNB分为2层,第1层利用随机森林

收稿日期:2020-06-29;**修回日期**:2020-10-20
基金项目:国家自然科学基金联合基金重点项目(U1711267);中央高校基本科研业务费专项资金项目(CUGGC03)
The work was supported by the Joint Fund Key Projects of the National Natural Science Foundation of China (U1711267) and the Fundamental Research Funds for the Central Universities (CUGGC03).
通信作者:蒋良孝(ljiang@cug.edu.cn)

从原始的单词特征中学习单词组合的高层特征.然后将学习到的新特征输入第 2 层,经过一位有效编码后用于构建伯努利朴素贝叶斯模型.在大量广泛使用的文本数据集上的实验结果表明,提出的 RFNB 模型明显优于现有的最先进的朴素贝叶斯文本分类模型和其他经典的文本分类模型.

关键词 朴素贝叶斯;随机森林;特征学习;特征表示;文本分类

中图法分类号 TP391

近年来,随着互联网信息的指数式增长,如何将海量的文本数据按照指定的类别自动归类,已成为自然语言处理的一项重要研究任务[1].为解决复杂的文本数据带来的独特挑战[2],已有许多经典的机器学习方法被用来解决文本分类问题,比如朴素贝叶斯、支持向量机、K 近邻、决策树等.

朴素贝叶斯模型因其简单、高效、易理解的特点被广泛应用于文本分类任务.在处理文本分类任务时,文档的特征通常用其中出现的单词来表示.按照单词在文档中是否出现,可将每个单词看作一个布尔变量,由此构建的朴素贝叶斯模型称为多变量伯努利朴素贝叶斯模型(Bernoulli naive Bayes, BNB)[3].BNB 只考虑单词在文档中是否出现,而完全忽略了单词在文档中出现的频次信息.为了弥补 BNB 所面临的这一不足,McCallum 和 Nigam[4] 提出了多项式朴素贝叶斯分类模型(multinomial naive Bayes, MNB),在分类过程通过捕获单词在文档中的出现频次来获得更好的分类性能,因此 MNB 较之于 BNB 有更加广阔的应用前景.

不过 MNB 与 BNB 都存在一个共同的不足:其属性条件独立假设在面对复杂的文本数据时往往很难得到满足.因此,削弱朴素贝叶斯文本分类模型所要求的属性条件独立假设成为改进朴素贝叶斯文本分类模型的重要途径.为此,学者们提出了许多改进方法,概括起来主要包括结构扩展[5]、实例选择[6-7]、实例加权[8]、特征选择[9-10]、特征加权[11-15]等.相关的实验研究也验证了这些方法的有效性.

然而,所有这些方法都是基于原始的单词特征来构建分类模型,由于朴素贝叶斯模型假定单词之间相互独立,因此在分类中很难学习到多个单词组合在一起时对分类的影响.然而现实情景中单个的单词特征携带的信息有限,单词组合在一起才可以表达更多的语义信息.基于当前朴素贝叶斯文本分类模型的这一局限,本文尝试用特征学习的方法,从原始的单词特征中学习到由多个单词特征组合在一起表示的高层特征,再基于学到的高层特征训练朴素贝叶斯模型.基于这个想法,本文提出了一种双层贝叶斯模型:随机森林朴素贝叶斯(random forest naive Bayes, RFNB),以改进朴素贝叶斯文本分类模型的分类性能.RFNB 分为 2 层,在第 1 层采用随机森林做特征学习,利用 Bagging 集成的优势以及随机决策树的随机性从原始的单词特征中学习多个单词特征组合在一起表示的高层特征,构成新的特征表示.学到的新特征输入第 2 层,经过一位有效编码(one-hot encoding)后用于构建伯努利朴素贝叶斯模型.

本文的主要贡献包括 2 个方面:

1) 对朴素贝叶斯文本分类模型进行较为全面的调查研究,总结了朴素贝叶斯文本分类模型的 5 类改进方法;创新性地提出了改进朴素贝叶斯文本分类模型的特征学习新方法,构建了一种双层贝叶斯模型——随机森林朴素贝叶斯(RFNB),设计并实现了学习 RFNB 模型的新算法.

2) 针对朴素贝叶斯文本分类模型假定单词特征之间完全独立的不足,本文提出的特征学习新方法可以从原始单词特征表示中学习到有益于分类的新特征表示;在大量广泛使用的文本数据集上的实验结果表明,相较于现有改进模型和其他经典的机器学习文本分类模型,RFNB 模型取得了较好的实验效果.

1 相关工作

在众多朴素贝叶斯文本分类模型中,首先被提出的是多变量伯努利朴素贝叶斯文本分类模型(BNB).BNB 假定文档由二进制特征向量表示,该向量表示哪些单词在文档中出现,哪些单词在文档中不出现.用 $\boldsymbol{d}$ 来表示 1 篇待分类的文档,其特征向量 $(w_1, w_2, \cdots, w_i, \cdots, w_m)$ 表示构成文档的单词向量.BNB 运用式(1)来分类文档 $\boldsymbol{d}$:

$$\hat{c}(\boldsymbol{d}) = \arg\max_{c \in C} \left[\ln P(c) + \sum_{i=1}^{m} \ln(B_i P(w_i \mid c) + (1 - B_i)(1 - P(w_i \mid c))) \right], \quad (1)$$

其中,C 为文档所属类别 c 的集合,m 是词库中不同

单词的数目,即词库的大小,w_i 表示单词向量中的第 i 个单词,B_i 表示第 i 个单词在文档 $\boldsymbol{d}$ 中是否出现,出现为 1,不出现为 0.先验概率 $P(c)$和条件概率 $P(w_i|c)$分别运用式(2)(3)来估计:

$$P(c)=\frac{\sum_{j=1}^{n}\delta(c_j,c)+1}{n+l}, \tag{2}$$

$$P(w_i|c)=\frac{\sum_{j=1}^{n}w_{ji}\delta(c_j,c)+1}{\sum_{j=1}^{n}\delta(c_j,c)+2}, \tag{3}$$

其中,n 是训练文档的数目,l 表示文档的类别数目,c_j 是第 j 篇训练文档的类标记,w_{ji}表示第 j 篇训练文档中的第 i 个单词,其值为布尔类型,若第 i 个单词在第 j 篇文档中出现则 $w_{ji}=1$,否则 $w_{ji}=0$. $\delta(\alpha,\beta)$是一个二元函数,当 $\alpha=\beta$ 时 $\delta(\alpha,\beta)=1$,否则 $\delta(\alpha,\beta)=0$.

为了克服 BNB 忽略单词出现的频次信息这一不足,McCallum 和 Nigam[4] 通过捕获文档中单词出现的频次信息,提出了一种多项式朴素贝叶斯文本分类模型.MNB 运用式(4)来分类待分类文档 $\boldsymbol{d}$:

$$\hat{c}(\boldsymbol{d})=\arg\max_{c\in C}\left[\ln P(c)+\sum_{i=1}^{m}f_i\ln P(w_i|c)\right], \tag{4}$$

其中 f_i 是单词 w_i 在文档 $\boldsymbol{d}$ 中出现的频次,先验概率 $P(c)$与 BNB 中计算方法一致,仍然运用式(2)来估计.条件概率 $P(w_i|c)$则运用式(5)来估计:

$$P(w_i|c)=\frac{\sum_{j=1}^{n}f_{ji}\delta(c_j,c)+1}{\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ji}\delta(c_j,c)+m}, \tag{5}$$

其中 f_{ji}是单词 w_i 在第 j 篇文档中出现的频次.

MNB 较 BNB 具有更加广泛的应用,许多研究工作都以 MNB 为基准进行改进,但是 MNB 要求的属性条件独立假设在现实应用中往往很难得到满足.为了削弱它要求的属性条件独立假设,学者们提出了许多改进方法,蒋良孝和李超群[16] 对这些改进方法做了详细总结,概括起来主要包括 5 类:结构扩展、实例选择、实例加权、特征选择、特征加权.

1.1 结构扩展

结构扩展方法通过在存在相互依存关系的特征之间添加有向边,由此来学习结构扩展的 MNB 模型.给定 1 篇待分类的文档 $\boldsymbol{d}$,结构扩展的 MNB 运用式(6)来分类文档 $\boldsymbol{d}$:

$$\hat{c}(\boldsymbol{d})=\arg\max_{c\in C}\left[\ln P(c)+\sum_{i=1}^{m}f_i\ln P(w_i|S_{w_i},c)\right], \tag{6}$$

其中 S_{w_i}表示贝叶斯网络中 w_i 的父特征集.先验概率 $P(c)$仍运用式(2)来估计,但条件概率 $P(w_i|S_{w_i},c)$在估计时要首先通过结构学习确定每个特征的父特征集,这类似于学习一个最优的贝叶斯网络,被证明是一个 NP-hard 问题.

为了构建一种不需要进行结构学习但仍然可以在某种程度上考虑到特征之间依存关系的贝叶斯网络模型,Jiang 等人[5] 提出了一种结构扩展的 MNB 模型(structure extended multinomial naive Bayes, SEMNB).SEMNB 提供了一种简单有效的学习方法,通过加权平均所有的一依赖多项式估计来削弱 MNB 模型要求的属性条件独立假设.SEMNB 模型无需复杂的结构学习过程,保持了 MNB 模型的结构简单性.

1.2 实例选择

实例选择方法首先利用局部学习的思想从整个训练文档集中选择部分实例来组建待分类文档的邻域,然后在组建的邻域上构建 MNB.给定 1 篇待分类的文档 $\boldsymbol{d}$,实例选择的 MNB 仍然运用式(4)来分类文档 $\boldsymbol{d}$,但先验概率 $P(c)$和条件概率 $P(w_i|c)$分别运用式(7)(8)来估计:

$$P(c)=\frac{\sum_{j=1}^{k}\delta(c_j,c)+1}{k+l}, \tag{7}$$

$$P(w_i|c)=\frac{\sum_{j=1}^{k}f_{ji}\delta(c_j,c)+1}{\sum_{i=1}^{m}\sum_{j=1}^{k}f_{ji}\delta(c_j,c)+m}, \tag{8}$$

其中 k 表示文档 $\boldsymbol{d}$ 的邻域中训练文档的数目.

为了组建待分类文档 $\boldsymbol{d}$ 的邻域,Jiang 等人[6] 提出了局部加权 MNB 模型(locally weighted multinomial naive Bayes, LWMNB).LWMNB 采用 K 近邻算法来搜索待分类文档 $\boldsymbol{d}$ 的邻域,是一种典型的消极学习的模型.此外,Wang 等人[7] 受 NBTree 的启发,利用决策树算法来搜索待分类文档 $\boldsymbol{d}$ 的邻域,然后在决策树的叶子节点上构建 MNB 模型,从而提出了多项式朴素贝叶斯树模型(multinomial naive Bayes tree, MNBTree).大量实验表明,MNBTree 具有良好的改进效果.

1.3 实例加权

实例加权方法首先为不同的训练文档学习不同

的权值,然后在加权之后的训练文档集上构建MNB.给定1篇待分类的文档 $\boldsymbol{d}$,实例加权的MNB仍然运用式(4)来分类文档 $\boldsymbol{d}$,但先验概率 $P(c)$ 和条件概率 $P(w_i|c)$ 分别运用式(9)(10)来估计:

$$P(c)=\frac{\sum_{j=1}^{n}W_j\delta(c_j,c)+1}{\sum_{j=1}^{n}W_j+l}, \tag{9}$$

$$P(w_i|c)=\frac{\sum_{j=1}^{n}W_jf_{ji}\delta(c_j,c)+1}{\sum_{i=1}^{m}\sum_{j=1}^{n}W_jf_{ji}\delta(c_j,c)+m}, \tag{10}$$

其中 W_j 表示第 j 篇训练文档的权值.

为了学习得到不同训练文档的权值,Jiang等人[8]提出一种判别加权的MNB模型(discriminatively weighted multinomial naive Bayes, DWMNB).DWMNB首先用原始训练文档集构建一个MNB模型,然后在每次迭代中,根据MNB估计的条件概率损失为不同的训练文档分配不同的权值.最后用学到的实例权值更新所有训练文档,并在更新得到的训练文档集上训练新的MNB模型作为最终学到的模型.

1.4 特征选择

特征选择方法首先利用数据归约的思想从整个特征空间选择出一个最佳特征子集,然后在选择的最佳特征子集上构建MNB.给定1篇待分类的文档 $\boldsymbol{d}$,特征选择的MNB运用式(11)来分类文档 $\boldsymbol{d}$:

$$\hat{c}(\boldsymbol{d})=\arg\max_{c\in C}\left[\ln P(c)+\sum_{i=1}^{q}f_i\ln P(w_i|c)\right], \tag{11}$$

其中 q 表示被选择的特征数目,先验概率 $P(c)$ 仍然运用式(2)来估计,但条件概率 $P(w_i|c)$ 运用式(12)估计:

$$P(w_i|c)=\frac{\sum_{j=1}^{n}f_{ji}\delta(c_j,c)+1}{\sum_{i=1}^{q}\sum_{j=1}^{n}f_{ji}\delta(c_j,c)+q}. \tag{12}$$

为了从整个特征空间选择出一个最佳的特征子集,学者们提出了许多特征选择方法.其中Yang和Pedersen[9]对文本分类的特征选择方法进行了比较研究,总结并评估了基于文档频率、信息增益、互信息、卡方统计量、单词强度这5种特征选择方法.他们的实验结果表明,在不损失分类精度的前提下,信息增益和卡方统计量的效果最好.不过在特征选择的过程中,基于信息增益的选择方法没有考虑属性值个数对结果的影响,对属性值较多的单词有所偏好.作为对信息增益的补充和改进,Zhang等人[10]提出了一种基于信息增益率做特征选择的MNB模型(gain ratio-based selective multinomial naive Bayes, GRSMNB),广泛的实验研究表明GRSMNB取得了很好的分类效果.

1.5 特征加权

特征加权方法首先为不同的特征学习不同的权值,然后在加权之后的训练文档集上构建MNB.给定1篇待分类的文档 $\boldsymbol{d}$,特征加权的MNB运用式(13)来分类文档 $\boldsymbol{d}$:

$$\hat{c}(\boldsymbol{d})=\arg\max_{c\in C}\left[\ln P(c)+\sum_{i=1}^{m}W_if_i\ln P(w_i|c)\right], \tag{13}$$

其中 W_i 表示单词 w_i 的权值,先验概率 $P(c)$ 仍然运用式(2)来估计,但条件概率 $P(w_i|c)$ 运用式(14)估计:

$$P(w_i|c)=\frac{\sum_{j=1}^{n}W_if_{ji}\delta(c_j,c)+1}{\sum_{i=1}^{m}\sum_{j=1}^{n}W_if_{ji}\delta(c_j,c)+m}. \tag{14}$$

为了学习特征权值,Jiang等人[11]率先提出了一种相关性深度特征加权的MNB模型(deep feature weighted multinomial naive Bayes, DFWMNB).之后,Zhang等人[12]提出了一种基于决策树做特征加权的MNB模型(decision tree-based feature weighted multinomial naive Bayes, DTWMNB).2020年Ruan等人[13]在深度特征加权的基础上引入类依赖的思想,有区别地为不同类别上的每个特征学习不同的权值,由此改进MNB模型.此外,Kim等人[14]基于信息增益、卡方统计量以及一种风险概率的扩展版本提出了3种特征加权方法.Li等人[15]提出了另一种基于卡方统计量的特征加权方法,通过在训练阶段准确测量特征和类别之间的正向依赖来计算特征的权值.

2 随机森林朴素贝叶斯

如第1节所述,已有许多方法被用于削弱朴素贝叶斯的属性条件独立假设.大量的实验研究表明,这些改进方法可以提升朴素贝叶斯文本分类模型的分类性能.然而现有改进方法都是基于原始的单词特征来构建朴素贝叶斯模型,属性条件独立假设使得朴素贝叶斯模型假定单词之间完全独立,因此

忽略了多个单词组合在一起时对分类的影响.由于实际情景中单词组合携带的语义信息往往比单个单词更加充分,基于这一问题,本文尝试用特征学习的方法来改进朴素贝叶斯文本分类模型,为此我们提出了一种双层贝叶斯模型:随机森林朴素贝叶斯(RFNB).模型的具体结构如图 1 所示,图 1 中 $document$ 表示 1 篇文档,$term_m$ 表示第 m 个单词在 $document$ 中的出现次数,$\boldsymbol{v}_1$ 表示由所有单词特征在 $document$ 中的出现次数构成的特征向量,特征向量中矩形的密集程度表示了特征的维度.第 1 层由随机森林构成,随机森林中的第 T 个基学习器用 $RandomTree_T$ 表示,每个基学习器为 1 棵随机决策树.第 1 层利用随机森林从特征向量 $\boldsymbol{v}_1$ 中捕获 T 维单词组合的高层特征,构成特征向量 $\boldsymbol{v}_2$.由于 $\boldsymbol{v}_2$ 的维度等于随机森林中基学习器的个数 T,而 T 往往远小于单词维度 m,因此特征向量 $\boldsymbol{v}_2$ 的维度远远低于原始的特征向量 $\boldsymbol{v}_1$ 的维度.第 2 层由 BNB 构成,由于 BNB 接受二进制特征,所以将特征向量 $\boldsymbol{v}_2$ 进行一位有效编码(one-hot encoding)后转化为二进制特征向量 $\boldsymbol{v}_3$.特征向量 $\boldsymbol{v}_3$ 的维度为 $T\times l$,是特征向量 $\boldsymbol{v}_2$ 的维度的 l 倍.最后在 $\boldsymbol{v}_3$ 上构建 BNB 模型,并由 BNB 预测待测文档的类别 $class$.

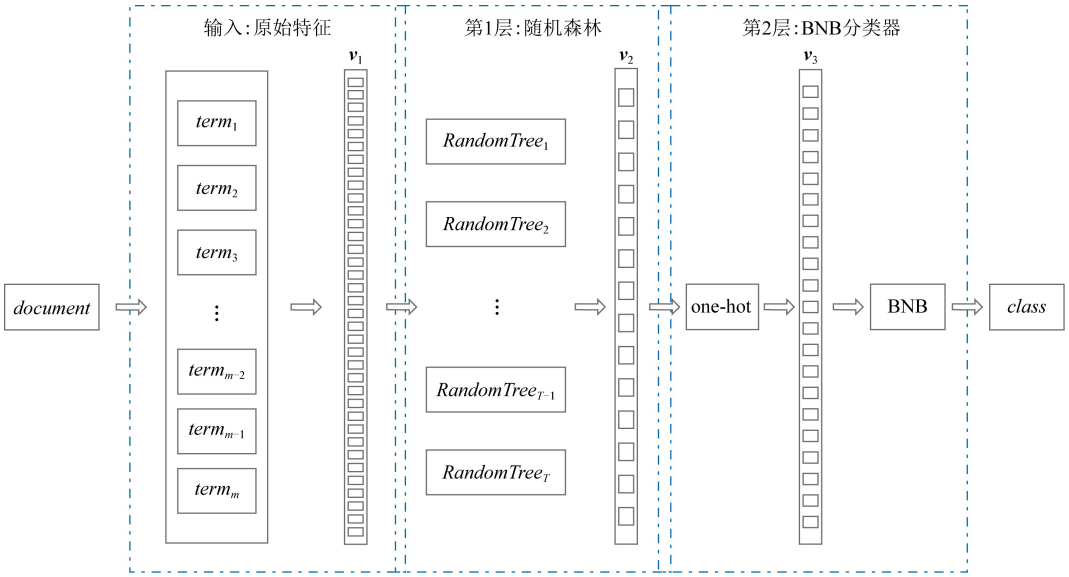

Fig. 1 Structure framework of RFNB

图 1 RFNB 的结构框架

可见,怎样从原始的单词特征中学习到单词组合的高层特征,以及怎样利用新特征构建高分类性能的朴素贝叶斯分类模型,是学习 RFNB 要解决的 2 个核心问题.下面的 2.1 节和 2.2 节将详细讨论如何解决这 2 个问题.

2.1 第 1 层:用随机森林学习新的特征表示

原始文本数据由原始的单词特征构成,特征值为单词在文档中的出现频率,单词空间的高维性导致了文本数据集的高维稀疏的特点.朴素贝叶斯的属性条件独立假设要求在给定类标记的前提下各特征之间完全相互独立,因而难以捕获多个单词组合的高层特征.RFNB 第 1 层的任务就是从高维稀疏的单词特征中捕获单词组合的高层特征.

随机森林[17]作为 Bagging 的扩展变体,采用随机决策树作为其基学习器.Bagging 有效防止在集成学习的过程中陷入过拟合,同时保证了随机森林各基学习器之间具有较高的独立性,随机决策树增加了训练基学习器时的随机性,进一步保证了基学习器之间的独立多样性.使用文本数据集训练随机森林,图 2 所示的树形结构表示 1 棵训练好的随机决策树,其中圆形节点表示分裂节点,三角形表示叶子

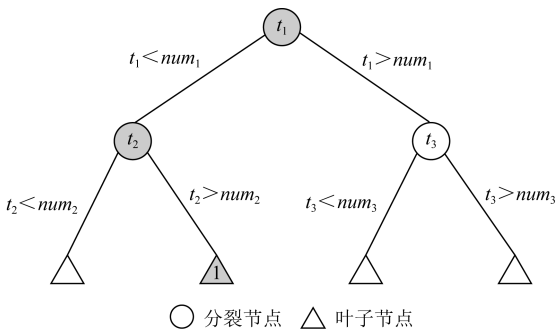

Fig. 2 Structure diagram of random decision tree

图 2 随机决策树结构图

节点,t_i 表示单词空间中的第 i 个单词特征,特征值为第 i 个单词在文档中的出现次数,用在分裂节点中表示使用该单词作为所在分裂节点的划分单词特征,num_i 表示第 i 个单词特征的最佳划分值.假定 1 篇文档经过这棵随机决策树的路径如阴影标记所示,最后落入一个叶子节点,并根据叶子节点的分类信息给出最终预测类别下标为 1.

由于预测类别下标 1 的得出同时参考了单词 t_1 和单词 t_2 的频率信息,这表示在这棵树中当文档的单词 t_1 的频率值小于 num_1 且单词 t_2 的频率值大于 num_2 时,2 个单词在这组具体特征值下的高层特征就可以用类别下标 1 来表示.类似上述情形,文档从输入随机决策树根结点到落入一个叶子节点的过程中,经过了一系列的中间节点,每个中间节点按照各自节点划分单词的特征值将文档送入对应的孩子节点,因此随机决策树的每个叶子节点上的类别下标输出都对应 1 组单词特征在不同的单词频率下的高层特征.因此 1 篇文档被随机森林中的每个基学习器分类的过程可以看作提取不同单词组合的高层特征的过程,且随机森林的 Bagging 和随机决策树的长树规则,保证我们学习这些不同的单词组合的高层特征更加容易.因此我们在第 1 层选择随机森林来进行特征学习,具体的算法步骤如下.

当给定文档集中文档的数目为 n,随机森林中基学习器的数目为 T 时,针对给定文档集,采用装袋方式处理:先从文档集中随机选择 1 篇文档作为采样文档放入采样集,再把该采样文档放回初始的文档集,使得下次采样时该文档仍有机会被选中;经过 n 次采样,得到包含 n 篇采样文档的采样集;将采样过程重复 T 次,得到 T 个分别含有 n 篇采样文档的采样集.采样集之间相互独立,分别针对每个采样集构建随机森林中的一个基学习器.为了保证基学习器之间的独立性,文档数目 n 应远大于基学习器数目 T.

本文所构建的随机森林中的基学习器是 1 棵随机决策树,在随机决策树的长树阶段,每次从给定的分裂节点中随机选取出 k_w 个单词构成候选分裂特征集合 A.从集合 A 中求出最佳分裂单词的最佳分割点,进行分割.k_w 的取值影响基学习器的分类性能和基学习器之间的差异性,经过实验分析,最终确定 k_w 值:

$$k_w=\sqrt{m}, \tag{15}$$

其中 m 值为所有文档中不同单词的数目.

给定随机决策树当前分裂节点上的文档集 D,在计算单词 w 的分割点时,假定单词 w 在文档集 D 上出现了 s 个不同的频次值,即 s 个不同的特征值,将这些频次值按照从小到大进行排序,记为 $\{w^1,w^2,\cdots,w^s\}$;设基于划分点 t,将文档集 D 分为文档集 D_t^- 和文档集 D_t^+,其中文档集 D_t^- 为单词 w 出现频次不大于 t 的文档,文档集 D_t^+ 为单词 w 出现频次大于 t 的文档;对于单词 w,包含 $s-1$ 个元素的候选划分点集合 T_w 为

$$T_w=\left\{\frac{w^i+w^{i+1}}{2}\,|\,1\leqslant i\leqslant s-1\right\}. \tag{16}$$

分类回归树(classification and regression tree,CART)[18]中采用基尼指数作为数值型特征的划分标准,分裂时仅将当前节点的数据集合分成 2 部分,分别进入左右子树,取得了良好的效果.同样地,RFNB 采用基尼指数划分数值型单词特征.在本文中划分前后基尼指数的差值用基尼增益表示,单词 w 划分文档集 D 的最佳基尼增益值为

$$Gini\_gain(D,w)=\max_{t\in T_w}\left\{Gini(D)-\sum_{\lambda\in\{-,+\}}\frac{|D_t^\lambda|}{|D|}Gini(D_t^\lambda)\right\}, \tag{17}$$

其中 $|D|$ 为文档集 D 中的文档数目.

假定文档集 D 中文档的类别数目为 l,文档集 D 中第 j 类文档所占的比例为 $p_j(j=1,2,\cdots,l)$,文档集 D 的基尼值 $Gini(D)$ 为

$$Gini(D)=1-\sum_{j=1}^{l}p_j^2. \tag{18}$$

根据以上方法,分别求出候选分裂特征集合 A 中每个单词的最佳基尼增益划分点,然后计算当前节点的最佳分裂单词 w_* 为

$$w_*=\arg\max_{w\in A}Gini\_gain(D,w). \tag{19}$$

若候选分裂特征集合 A 内每个单词对应的最佳基尼增益均不大于 0,则从剩下的分裂特征集合中继续随机不重复地选择单词并求其最佳基尼增益,直到第 1 个最佳基尼增益大于 0 的单词出现为止,将这个单词设置为分裂单词 w_*;若特征空间中所有单词的最佳基尼增益均不大于 0,则将 w_* 标记为空.若当前分裂点的最佳划分单词 w_* 为空,表示当前节点没有适合用来分裂的单词,则当前节点为叶子节点,叶子节点的类别为该节点所含文档最多的类别;否则按最佳划分单词 w_* 长树,将在单词 w_* 上出现频次不大于 t 的文档集 D_t^- 放入当前分裂节点的左子节点;将在单词 w_* 上出现频次大于 t 的文档集 D_t^+ 放入当前分裂节点的右子节点,完成

当前节点的分裂.从根节点出发,若所有节点均不可继续分裂,则长树完成.

我们用构建好的随机森林对训练集中的每篇文档进行分类,然后用每篇文档在所有随机决策树上的预测类别下标构成一个类别下标向量,以此作为从原始文档中学到的新的特征表示.

综上所述,RFNB 的第 1 层用随机森林将每篇原始文档从原始的单词特征表示转化为高层特征表示,并输入到 RFNB 的下一层.2.2 节将详细介绍如何在新特征表示上构建朴素贝叶斯文本分类模型.

2.2 第 2 层:在新特征表示上构建 BNB

通过 RFNB 第 1 层学习得到的新特征的特征值为预测类别的下标,虽然也是整型数值,但并不能作为单词在文档中出现的频次来使用,因此不能直接用于文本分类模型的学习,为了使新特征可以用于朴素贝叶斯文本分类模型的学习,本文对第 1 层学习得到的新特征做了一位有效编码(one-hot encoding).经过一位有效编码后的特征值变成了布尔类型,即只有 0 和 1 这 2 个值,刚好可以视作单词在文档中出现的频次(1 表示单词在文档中出现,0 表示单词在文档中未出现),此时才可以用于朴素贝叶斯文本分类模型的学习.当文档集 D 中文档类别数目为 l,随机森林中基学习器的数目为 T 时,经过一位有效编码产生的二进制特征表示的维度为 $T\times l$.

接下来的问题是,RFNB 的第 2 层应该选择哪一种朴素贝叶斯文本分类模型.MNB 相比于 BNB 的优势在于捕获了单词在文档中出现的频次信息.而新特征经过一位有效编码后的特征值是布尔类型,单词的频次只有 0 和 1 这 2 个值,因此 MNB 相比于 BNB 的优势消失.同时,MNB 按照式(4)分类时,当单词的频次 $f_i=0$ 时,与之对应的累加项被彻底抛弃,这相当于完全忽略了未出现单词对文档分类的影响.不同于 MNB,BNB 依据式(1)分类不仅考虑了出现单词对文档分类的影响,还考虑了未出现单词对文档分类的影响.并且 BNB 忽略单词频次信息的不足刚好被一位有效编码得到的二进制特征表示所弥补,因此 BNB 被选为最终的朴素贝叶斯文本分类模型.

综上所述,RFNB 分为 2 层,第 1 层采用随机森林做特征学习,然后将新特征输入第 2 层,经过一位有效编码后用于构建 BNB.因此,RFNB 的整个学习过程可以分为训练过程和分类过程,本文分别用算法 1 和算法 2 进行详细的描述.

算法 1. RFNB-Training.

输入:训练文档集 D;

输出:训练好的随机森林、$P(c)$、$P(w_i|c)$.

① 使用训练文档集 D,根据 2.1 节描述的长树方法构建随机森林;

② 建立新的样本集合 D',D'初始为空;

③ 运用构建的随机森林分类 D 中的每篇训练文档 $\boldsymbol{x}$ 得到新的特征向量 $\boldsymbol{x}'$,并将新特征向量 $\boldsymbol{x}'$存入样本集合 D'中;

④ 分别对 D'中的每一个样本 $\boldsymbol{x}'$做一位有效编码,编码后的样本表示为 $\boldsymbol{x}'_{\text{one-hot}}$,更新后的样本集合表示为 $D'_{\text{one-hot}}$;

⑤ 用$D'_{\text{one-hot}}$作为训练集,运用式(2)(3)计算 BNB 的先验概率 $P(c)$和条件概率 $P(w_i|c)$;

⑥ 返回训练好的随机森林、$P(c)$、$P(w_i|c)$.

算法 2. RFNB-Classification.

输入:随机森林、$P(c)$、$P(w_i|c)$、待分类文档 $\boldsymbol{d}$;

输出:待分类文档 $\boldsymbol{d}$ 的预测类别.

① 将待分类文档 $\boldsymbol{d}$ 输入随机森林,用基学习器预测的类别下标为 $\boldsymbol{d}$ 构建新特征向量 $\boldsymbol{d}'$;

② 对$\boldsymbol{d}'$做一位有效编码,经过编码后的二进制特征表示用 $\boldsymbol{d}'_{\text{one-hot}}$表示;

③ 运用式(1)分类 $\boldsymbol{d}'_{\text{one-hot}}$,得到 RFNB 模型的最终预测结果 $\hat{c}(\boldsymbol{d})$;

④ 返回待分类文档 $\boldsymbol{d}$ 的预测类别 $\hat{c}(\boldsymbol{d})$.

3 实验与结果

3.1 实验设置与实验数据

为了验证本文所提模型的有效性,本文从每一类现有改进方法中分别挑选了一种经典模型作为 RFNB 的比较对象,具体模型包括 SEMNB[5],MNBTree[7],DWMNB[8],GRSMNB[10],DTWMNB[12].这些改进模型已经被证明具有很好的改进效果.除此之外我们还增加了支持向量机(support vector machine, SVM)[19] 和随机森林(random forest, RF)[17]两种经典的文本分类模型作为比较对象.下面是所有这些比较对象的全称和缩写.

BNB:伯努利朴素贝叶斯[3];

MNB:多项式朴素贝叶斯[4];

SEMNB:结构扩展的 MNB[5];

MNBTree:多项式朴素贝叶斯树[7];

DWMNB:判别加权的 MNB[8];

GRSMNB:信息增益率特征选择的 MNB[10];

DTWMNB:基于决策树特征加权的 MNB[12];

SVM:支持向量机[19];

RF:随机森林[17].

我们在国际数据挖掘实验平台 WEKA[20]上实现本文提出的 RFNB,并设置随机森林中基学习器的数目 $T=200$.实验数据为 WEKA 平台发布的 15 个广泛使用的标准文本数据集,代表了不同的文本分类场景和特征,具体信息如表 1 所示:

Table 1 The Detailed Description of 15 Datasets

表 1 15 个数据集的详细描述

数据集	文档数目	单词数目	类别数目
fbis	2 463	2 000	17
oh0	1 003	3 182	10
oh10	1 050	3 238	10
oh15	913	3 100	10
oh5	918	3 012	10
re0	1 657	3 758	25
re1	1 504	2 886	13
tr11	414	6 429	9
tr12	313	5 804	8
tr21	336	7 902	6
tr23	204	5 832	6
tr31	927	10 128	7
tr41	878	7 454	10
tr45	690	8 261	10
wap	1 560	8 460	20

3.2 实验评价指标

本文采用 2 种评价指标衡量算法的分类性能,分别是分类精度(classification accuracy)和加权平均 F_1 值(weighted average of F_1, F_{wa}).分类精度是指算法分类正确样本占总样本的比例,是最常用的评价指标.F_{wa}用于衡量模型在多类问题中针对不同类别分类性能的整体表现,其具体计算为

$$F_{wa}=\frac{1}{n}\sum_{j=1}^{l}F_j\,count_j, \tag{20}$$

其中,$count_j$ 表示第 j 类文档样本个数,F_j 表示当用第 j 类样本作为正类样本、其余样本作为负类样本时的 F_1 值,F_j 计算为

$$F_j=\frac{2\times p_j\times r_j}{p_j+r_j}, \tag{21}$$

其中 p_j 和 r_j 分别表示对应情景下的精确率和召回率.

3.3 实验结果与分析

表 2 展示了本文提出的 RFNB 及其比较对象在每个数据集上通过 10 次 10 折交叉验证获得的平均分类精度,表 3 展示了实验获得各个算法的 F_{wa}.表 2 和表 3 中每行加粗的数字表示在该数据集上获得的最高的分类精度或 F_{wa}.作为各模型相对性能的总体指标,本文将每个模型在 15 个数据集上的平均分类精度和平均 F_{wa}汇总在表格底部.

Table 2 Classification Accuracy Comparison for RFNB Versus Its Competitors

表 2 RFNB 与其比较对象的分类精度比较结果 %

数据集	RFNB	BNB	MNB	SEMNB	MNBTree	DWMNB	GRSMNB	DTWMNB	SVM	RF
fbis	**86.13**	66.33	77.11	83.27	79.06	80.39	79.61	79.37	82.98	84.05
oh0	90.98	79.72	89.55	88.87	88.93	89.64	90.18	**92.28**	89.49	89.63
oh10	**85.73**	75.01	80.60	80.66	83.25	80.64	81.10	82.59	80.36	84.27
oh15	**86.90**	73.72	83.60	83.36	79.01	83.29	84.38	86.35	84.26	85.06
oh5	90.64	75.40	86.63	87.55	88.74	86.87	89.72	**90.99**	89.77	90.12
re0	**85.79**	62.77	80.02	82.73	77.3	81.81	80.56	81.18	84.91	83.39
re1	**86.57**	59.30	83.31	82.22	84.26	83.13	86.12	86.10	85.88	83.51
tr11	**90.95**	55.11	85.21	87.62	85.79	85.81	86.24	86.58	89.36	87.96
tr12	**90.06**	55.93	80.99	86.64	85.3	82.46	87.48	84.89	88.85	87.34
tr21	**94.82**	48.06	61.90	90.36	86.15	78.45	92.18	62.41	91.39	85.22
tr23	92.60	39.79	71.15	89.05	**93.04**	84.02	92.01	78.56	90.50	83.63
tr31	97.35	78.57	94.60	96.86	96.48	96.28	96.46	95.64	**98.39**	97.72
tr41	96.14	81.35	94.65	94.97	94.38	95.21	94.37	95.25	**96.59**	94.60
tr45	94.33	71.07	83.64	91.54	90.36	87.36	89.86	89.00	**94.45**	92.39
wap	84.37	67.67	81.22	80.53	75.42	81.83	80.34	82.76	**85.15**	80.87
平均值	**90.22**	65.99	82.28	87.08	85.83	85.15	87.37	84.93	88.82	87.32

注:黑体值表示最优值.

Table 3 F_{wa} Comparison for RFNB Versus Its Competitors

表 3 RFNB 与其比较对象的 F_{wa} 比较结果 %

数据集	RFNB	BNB	MNB	SEMNB	MNBTree	DWMNB	GRSMNB	DTWMNB	SVM	RF
fbis	**86.00**	66.08	77.40	82.39	78.84	80.43	79.23	79.63	82.80	82.29
oh0	90.98	77.09	89.49	88.56	88.83	89.59	90.04	**92.31**	89.29	89.50
oh10	**85.85**	71.58	80.05	79.57	83.03	80.09	80.00	82.16	79.98	82.85
oh15	**86.76**	70.79	83.25	82.89	78.63	82.84	84.01	86.06	83.97	84.67
oh5	90.62	73.61	86.55	87.27	88.57	86.78	89.50	**90.95**	89.59	89.86
re0	**85.45**	59.25	79.91	81.87	76.64	81.69	79.69	81.42	84.77	81.72
re1	84.83	51.60	81.01	78.93	82.18	81.50	83.47	84.14	**85.16**	80.13
tr11	**90.29**	50.78	83.81	85.80	84.46	84.38	84.23	85.23	88.48	84.72
tr12	**89.60**	53.47	79.76	85.78	84.46	81.62	86.51	84.57	88.24	85.72
tr21	**94.15**	49.46	63.03	87.86	83.59	78.40	90.44	63.83	90.38	80.70
tr23	92.08	31.71	72.13	87.60	**92.96**	84.67	91.22	79.78	89.71	79.01
tr31	98.12	76.88	94.42	96.59	96.37	96.14	96.28	95.52	**98.27**	97.44
tr41	**96.45**	78.79	94.44	94.61	94.18	94.96	94.11	95.05	96.32	93.59
tr45	**94.55**	68.94	82.97	90.47	90.34	86.73	88.50	88.54	94.02	91.27
wap	82.81	63.76	78.66	77.80	72.52	79.68	78.02	80.48	**84.32**	77.89
平均值	**89.90**	62.92	81.79	85.87	85.04	84.63	86.35	84.64	88.35	85.42

注:黑体值表示最优值.

基于表 2 和表 3,我们利用 KEEL(knowledge extraction based on evolutionary learning)数据挖掘软件[21]进行了系统的威尔克森符号秩检验[22],以进一步比较每一对模型之间的性能差异.威尔克森符号秩检验计算所得秩和如表 4 和表 5 所示.根据威尔克森符号秩检验的临界值表,显著性水平为 $\alpha=0.05$和 $\alpha=0.1$ 时,如果正差秩和与负差秩和中较小的一个小于或者等于 25 和 30,则认为算法显著不同.由此得到详细的统计比较结果如表 6 和表 7 所示.

从表 2~7 可以看出:

1) 直接在原始的文本数据上构建 BNB,分类精度和 F_{wa}都是最低的,在 15 个数据集上的平均分类精度仅仅只有 65.99%,平均 F_{wa}为 62.92%.

2) 相比于 BNB,MNB 因为捕获了单词在文档中出现的频次信息而获得了更好的分类性能,在 15 个数据集上的平均分类精度快速提升到 82.28%,平均 F_{wa}为 81.79%.

Table 4 Ranks of Classification Accuracy Computed by the Wilcoxon Test

表 4 分类精度的威尔克森测试的秩和值

算法	RFNB	BNB	MNB	SEMNB	MNBTree	DWMNB	GRSMNB	DTWMNB	SVM	RF
RFNB	—	120	120	120	119	120	120	114	106	119
BNB	0	—	0	0	0	0	0	0	0	0
MNB	0	120	—	18	36	8	7	0	3	5
SEMNB	0	120	102	—	87	95	49.5	64	3	32
MNBTree	1	120	84	33	—	70	24	46	13	22
DWMNB	0	120	112	25	50	—	28	40	3	13
GRSMNB	0	120	113	70.5	96	92	—	56	28	45
DTWMNB	6	120	120	56	74	80	64	—	21	33
SVM	14	120	117	117	107	117	92	99	—	96
RF	1	120	115	88	98	107	75	87	24	—

注:“—”表示空值;“—”所示的对角线上方数值表示行中模型与列中模型的正差秩和,对角线下方为负差秩和.

Table 5 Ranks of F_{wa} Computed by the Wilcoxon Test

表 5 F_{wa}的威尔克森测试的秩和值

算法	RFNB	BNB	MNB	SEMNB	MNBTree	DWMNB	GRSMNB	DTWMNB	SVM	RF
RFNB	—	136	136	136	135	136	136	131	123	136
BNB	0	—	0	0	0	0	0	0	0	0
MNB	0	136	—	25	38	4	12	0	3	9
SEMNB	0	136	111	—	92	98	51	68	0	63
MNBTree	1	136	98	44	—	78	37	53	13	42
DWMNB	0	136	132	38	58	—	36	44	3	39
GRSMNB	0	136	124	85	99	100	—	53	17	70
DTWMNB	5	136	136	68	83	92	83	—	19	69
SVM	13	136	133	136	123	133	119	117	—	120
RF	0	136	127	73	94	97	66	67	16	—

注:“—”表示空值;“—”所示的对角线上方数值表示行中模型与列中模型的正差秩和,对角线下方为负差秩和.

3) 相比于 MNB,目前已有的 5 种改进模型 SEMNB, MNBTree, DWMNB, GRSMNB, DTWMNB 在 15 个数据集上的平均分类精度都有较大幅度的上升,分别上升到 87.08%,85.83%,85.15%,87.37%,84.93%;平均 F_{wa} 也有较大提升,上升到 85.87%,85.04%,84.63%,86.35%,84.64%.

4) 本文提出的 RFNB 在 15 个数据集上的平均分类精度达到了 90.22%,不仅比已有的 5 种改进模型都高,还高于经典的支持向量机(88.82%)和随机森林(87.32%);平均 F_{wa} 达到了 89.90%,同样高于上述文本分类模型.这说明本文提出的特征学习方法是非常有效的,新的特征表示更加有利于朴素贝叶斯文本分类模型做分类,最终学习得到的 BNB 模型分类性能达到了最佳.

5) 在原始的文本数据集上直接构建 BNB,其分类性能欠佳,甚至比最简单的 MNB 还要低大约 17 个百分点,但在本文得到的新特征表示上构建 BNB,其分类精度竟然反超 MNB 大概 8 个百分点.这表明本文提出的 RFNB,不仅可以从原始的单词特征中学习到由多个单词组合的高层特征,还可以弥补 BNB 因忽略单词频次信息造成的模型缺陷.

6) 从表 4~7 所示的威尔克森统计测试比较结果来看,本文提出的 RFNB 显著优于所有的比较对象,包括 BNB, MNB, SEMNB, MNBTree, DWMNB, GRSMNB, DTWMNB, SVM, RF.这充分证明了本文所提模型的有效性.

Table 6 Classification Accuracy Comparison Results of the Wilcoxon Tests

表 6 分类精度的威尔克森测试的比较结果

算法	RFNB	BNB	MNB	SEMNB	MNBTree	DWMNB	GRSMNB	DTWMNB	SVM	RF
RFNB	—	•	•	•	•	•	•	•	•	•
BNB	◦	—	◦	◦	◦	◦	◦	◦	◦	◦
MNB	◦	•	—	◦		◦	◦	◦	◦	◦
SEMNB	◦	•	•	—		•			◦	
MNBTree	◦	•			—		◦		◦	◦
DWMNB	◦	•	•	◦		—	◦		◦	◦
GRSMNB	◦	•	•		•		—		◦	
DTWMNB	◦	•	•					—	◦	
SVM	◦	•	•	•	•	•		•	—	•
RF	◦	•	•		•	•			◦	—

注:“—”表示空值;“◦”表示相应列中的模型显著优于相应行中的模型,而“•”表示相应行中的模型显著优于相应列中的模型;“—”所示的对角线以下的结果的显著性水平为 $\alpha=0.05$,对角线以上的结果的显著性水平为 $\alpha=0.1$.

Table 7 F_{wa} Comparison Results of the Wilcoxon Tests

表 7 F_{wa}的威尔克森测试的比较结果

算法	RFNB	BNB	MNB	SEMNB	MNBTree	DWMNB	GRSMNB	DTWMNB	SVM	RF
RFNB	—	•	•	•	•	•	•	•	•	•
BNB	◦	—	◦	◦	◦	◦	◦	◦	◦	◦
MNB	◦	•	—	◦		◦	◦	◦	◦	◦
SEMNB	◦	•	•	—					◦	
MNBTree	◦	•			—				◦	
DWMNB	◦	•	•			—			◦	
GRSMNB	◦	•	•				—		◦	
DTWMNB	◦	•	•					—	◦	
SVM	◦	•	•	•	•	•	•	•	—	•
RF	◦	•	•						◦	—

注:"—"表示空值;"◦"表示相应列中的模型显著优于相应行中的模型,而"•"表示相应行中的模型显著优于相应列中的模型;"—"所示的对角线以下的结果的显著性水平为 $\alpha=0.05$,对角线以上的结果的显著性水平为 $\alpha=0.1$.

4 总结与展望

朴素贝叶斯文本分类模型的属性条件独立假设使得模型在面对文本数据时,只考虑了单个单词特征的语义信息而忽略了不同单词组合下的高层信息.目前所有削弱其属性条件独立假设的改进方法都是基于原始的单词特征来构建文本分类模型,这在一定程度上限制了改进方法的效果.不同于现有的改进方法,本文提出用特征学习的方法来改进朴素贝叶斯文本分类模型,提出了一种双层贝叶斯模型:随机森林朴素贝叶斯(RFNB).RFNB 在第 1 层采用随机森林从原始的单词特征中学习单词组合的高层特征,然后将学到的新特征输入第 2 层,经过一位有效编码后用于构建伯努利朴素贝叶斯模型.在大量广泛使用的文本数据集上的实验结果表明,本文提出的 RFNB 模型明显优于现有的最先进的朴素贝叶斯文本分类模型和其他经典的文本分类模型.

在目前的实验中,随机森林中基学习器的数目被设置为固定值 200,固定的参数设置不利于模型在不同维度的数据集上应用,将来的研究可以尝试设计一种随数据维度自适应的参数设置方法,如设置随机森林中基学习器的数目为数据维度的开方,以此进一步增强模型的泛化能力.此外,目前的版本中随机森林采用了固定的长树方法,这在一定程度上限制了学习高层特征时的多样性.将来的另一个研究方向可以尝试用不同的方法长树,以进一步提高基学习器的多样性,持续攀升模型的分类性能、拓展其应用场景.

参 考 文 献

[1] Li Ran, Lin Zheng, Lin Hailun, et al. Textemotion analysis: A survey [J]. Journal of Computer Research and Development, 2018, 55(1): 30-52 (in Chinese)
(李然, 林政, 林海伦, 等. 文本情绪分析综述[J]. 计算机研究与发展, 2018, 55(1): 30-52)

[2] Su Jindian, Ouyang Zhifan, Yu Shanshan. Aspect-level sentiment classification for sentences based on dependency tree and distance attention [J]. Journal of Computer Research and Development, 2019, 56(8): 1731-1745 (in Chinese)
(苏锦钿, 欧阳志凡, 余珊珊. 基于依存树及距离注意力的句子属性情感分类[J]. 计算机研究与发展, 2019, 56(8): 1731-1745)

[3] Ponte J M, Croft W B, Pranckevicius T, et al. A language modeling approach to information retrieval [C] //Proc of the 21st Annual Int ACM SIGIR Conf on Research and Development in Information Retrieroceval. New York: ACM, 1998: 275-281

[4] McCallum A, Nigam K. A comparison of event models for naive Bayes text classification [C] //Proc of the AAAI/ICML Workshop on Learning for Text Categorization. Menlo Park, CA: AAAI Press, 1998: 41-48

[5] Jiang Liangxiao, Wang Shasha, Li Chaoqun, et al. Structure extended multinomial naive Bayes [J]. Information Sciences, 2016, 329: 346-356

[6] Jiang Liangxiao, Cai Zhihua, Zhang Harry, et al. Naive Bayes text classifiers: A locally weighted learning approach [J]. Journal of Experimental & Theoretical Artificial Intelligence, 2013, 25(2): 273-286

[7] Wang Shasha, Jiang Liangxiao, Li Chaoqun. Adapting naive Bayes tree for text classification [J]. Knowledge and Information Systems, 2015, 44(1): 77-89

[8] Jiang Liangxiao, Wang Dianhong, Cai Zhihua. Discriminatively weighted naive Bayes and its application in text classification [J]. International Journal on Artificial Intelligence Tools, 2012, 21(1): No.1250007

[9] Yang Yiming, Pedersen J O. A comparative study on feature selection in text categorization [C] //Proc of the 14th Int Conf on Machine Learning. San Francisco, CA: Morgan Kaufmann, 1997: 412-420

[10] Zhang Lungan, Jiang Liangxiao, Li Chaoqun. A new feature selection approach to naive Bayes text classifiers [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2016, 30(2): No.1650003

[11] Jiang Liangxiao, Li Chaoqun, Wang Shasha, et al. Deep feature weighting for naive Bayes and its application to text classification [J]. Engineering Applications of Artificial Intelligence, 2016, 52: 26-39

[12] Zhang Lungan, Jiang Liangxiao, Li Chaoqun, et al. Two feature weighting approaches for naive Bayes text classifiers [J]. Knowledge Based Systems, 2016, 100: 137-144

[13] Ruan Shufen, Li Hongwei, Li Chaoqun, et al. Class-specific deep feature weighting for naïve Bayes text classifiers [J]. IEEE Access, 2020, 8: 20151-20159

[14] Kim S B, Han K S, Rim H C, et al. Some effective techniques for naive Bayes text classification [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(11): 1457-1466

[15] Li Yanjun, Luo Congnan, Chung Soon M. Weighted naive Bayes for text classification using positive term-class dependency [J]. International Journal on Artificial Intelligence Tools, 2012, 21(1): No.1250008

[16] Jiang Liangxiao, Li Chaoqun. Bayesian Network Classifier: Algorithmand Application [M]. Wuhan: China University of Geosciences Press, 2015 (in Chinese)
(蒋良孝, 李超群. 贝叶斯网络分类器: 算法与应用[M]. 武汉: 中国地质大学出版社, 2015)

[17] Leo B. Random forest [J]. Machine Learning, 2001, 45(1): 5-23

[18] Leo B, Friedman J H, Olshen R A, et al. Classification and Regression Trees [M]. London: Chapman & Hall, 1984

[19] Cortes C, Vapnik V N. Support vector networks [J]. Machine Learning, 1995, 20(3): 273-297

[20] Witten I H, Frank E, Hall M A, et al. Data Mining: Practical Machine Learning Tools and Techniques [M]. 4th ed. San Francisco, CA: Morgan Kaufmann, 2016

[21] Triguero I, Gonzalez S, Moyano J M, et al. KEEL 3.0: An open source software for multi-stage analysis in data mining [J]. International Journal of Computational Intelligence Systems, 2017, 10(1): 1238-1249

[22] Singh P K, Sarkar R, Nasipuri M. Significance of non-parametric statistical tests for comparison of classifiers over multiple datasets [J]. International Journal of Computing Science and Mathematics, 2016, 7(5): 410-442

Zhang Wenjun, born in 1998. Master candidate. His main research interests include machine learning and data mining.
张文钧,1998 年生.硕士研究生.主要研究方向为机器学习和数据挖掘.

Jiang Liangxiao, born in 1977. PhD, professor, PhD supervisor. His main research interests include machine learning and data mining.
蒋良孝,1977 年生.博士,教授,博士生导师.主要研究方向为机器学习和数据挖掘.

Zhang Huan, born in 1994. PhD candidate. His main research interests include machine learning and data mining.
张　欢,1994 年生.博士研究生.主要研究方向为机器学习和数据挖掘.

Chen Long, born in 1998. Master candidate. His main research interests include machine learning and data mining.
陈　龙,1998 年生.硕士研究生.主要研究方向为机器学习和数据挖掘.