·先进计算与数据处理· 文章编号:1000-3428(2018)11-0027-06 文献标志码:A 中图分类号:TP301

基于粒子群优化的朴素贝叶斯改进算法

邱宁佳,李 娜,胡小娟,王 鹏,孙爽滋

(长春理工大学 计算机科学技术学院,长春 130022)

摘 要: 针对朴素贝叶斯(NB)算法因条件独立性的理想式假设引起分类性能降低的问题,提出一种改进的粒子群优化-朴素贝叶斯(PSO-NB)算法。在文本预处理时,引入权重因子、类内和类间离散因子进行属性约简,基于NB加权模型,将条件属性的词频比率作为其初始权值,利用PSO算法迭代寻找全局最优特征权向量,并以此权向量作为加权模型中各个特征词的权值生成分类器。运用经典数据集对PSO-NB算法进行性能分析,结果表明,改进算法可有效减少冗余属性,降低计算复杂度,具有较高的准确率和召回率。

关键词: 朴素贝叶斯;互信息;属性约简;粒子群优化算法;权值优化

中文引用格式: 邱宁佳,李 娜,胡小娟,等. 基于粒子群优化的朴素贝叶斯改进算法[J]. 计算机工程,2018,44(11):27-32,39.

英文引用格式: QIU Ningjia, LI Na, HU Xiaojuan, et al. Improved native Bayes algorithm based on particle swarm optimization[J]. Computer Engineering, 2018, 44(11):27-32, 39.

Improved Native Bayes Algorithm Based on Particle Swarm Optimization

QIU Ningjia, LI Na, HU Xiaojuan, WANG Peng, SUN Shuangzi

(College of Computer Science and Technology, Changchun University of Science and Technology, Changchun 130022, China)

【Abstract】 Aiming at the problem of classification performance degradation caused by the idealized assumption of conditional independence of Naive Bayes(NB) algorithm, an improved Particle Swarm Optimization-Native Bayes(PSO-NB) algorithm is proposed. In text preprocessing, weight factor, intra-class and inter-class discrete factors are introduced for attribute reduction. Based on NB weighted model, the word-frequency ratio of conditional attribute is used as its initial weight, and PSO algorithm is used to iteratively find global optimal feature weight vector. The vector is used as a weight value to generate a classifier for each feature word in the weighting model. The performance analysis of PSO-NB algorithm is done using classical dataset. Result shows that the improved algorithm can effectively reduce redundant attributes, reduce computational complexity, and has high accuracy and recall rate.

【Key words】 Native Bayes(NB); mutual information; attribute reduction; Particle Swarm Optimization(PSO) algorithm; weight optimization

DOI: 10.19678/j.issn.1000-3428.0048410

0 概述

朴素贝叶斯(Native Bayes, NB)算法是一种简洁而高效的分类算法,在很多情况下达到的分类效果可以与一些复杂分类算法相媲美。但其以假设条件属性变量之间相互独立为前提,而在现实应用中,事务的各属性之间大都有着一定的联系。因此,朴素贝叶斯算法的这种理想式假设是不合理的,这也使得分类性能受到很大的影响。为解决该问题,相关学者提出了不同的方法来弥补朴素贝叶斯分类器的不足之处,提高其分类精度。

文献[1]以Hall所提出的加权方法为目标函数,采用差分进化算法取得属性的最优权值,建立朴素贝叶斯加权模型,使其准确性有所提高。文献[2]提出一种局部加权方法,为经典NB模型中的每个属性分配权重,并使用基于对数的简单假设将其转换成线性形式,利用最小二乘法确定方程中的最优权向量,以该权向量建立加权模型,使算法的复杂度有所简化。文献[3]提出一种局部加权的学习方法来削弱朴素贝叶斯分类器的条件独立性假设,使算

基金项目: 吉林省科技发展计划重点科技攻关项目(20150204036GX);吉林省省级产业创新专项资金(2017C051)。

作者简介: 邱宁佳(1984—),男,讲师、博士,主研方向为数据挖掘;李 娜,硕士研究生;胡小娟,讲师、博士;王 鹏,副教授、博士;孙爽滋,副教授、硕士。

收稿日期: 2017-08-21 **修回日期:** 2017-10-24 **E-mail:** qiunj@cust.edu.cn

法的分类性能明显提高。文献[4]采用最优带宽选择来估计类条件概率密度函数的方法降低特征间的依赖性,实验结果表明,当特征之间存在依赖时该模型明显优于传统模型。文献[5]采用基于准割线法的局部优化方法确定目标函数中的最优权值,使模型的分类精度明显优于原有 NB 分类模型。文献[6]采用粗糙集对数据集进行属性约简,使用对数条件似然估计法对条件属性求取全局最优权值,明显提高使算法的性能。文献[7]首先采用无标签训练集求得置信度比较高的样本,再结合有标签训练样本不断迭代,使传统的半监督朴素贝叶斯算法的性能明显提高。文献[8]利用主成分分析法提取独立属性的性质,构造新的属性,达到提高分类效果的目的。文献[9]采用最小二乘法确定目标函数,以 FOA 算法优化权值,明显提高分类器的性能。文献[10]利用支持向量机构造一个最优分类超平面降低样本空间规模,并使用朴素贝叶斯算法训练样本集生成分类模型。上述朴素贝叶斯模型均在不同程度上提高了朴素贝叶斯算法的分类性能,然而这些研究存在没有属性约简、权值优化以及设定初始权值的问题。

上述方法在一定程度上降低了算法分类性能,本文针对其存在的问题,提出一种改进的 PSO-NB 算法。通过引入权重因子、类内和类间离散因子对互信息进行改进,以改进的互信息方法进行属性约简,获得彼此相对独立的核心属性,将属性的词频比率作为其初始权值,利用 PSO 算法迭代求得最终的特征权向量生成分类器,以提高算法分类性能。

1 文本预处理

在文本分类中,特征选择方法能从高维的特征空间中选取对分类有效的特征,降低特征的冗余,提高分类准确度。文献[11]对特征选择的各个方法进行了简述。互信息算法是文本分类中常用的特征选择方法之一,但其在理论以及现实应用中分类精确度较低。文献[12]采用一种综合考虑相关度和冗余度的特征选择标准 UmRMR 来评价特征重要性的方法,提出一种基于互信息的无监督特征选择方法使模型的性能有所提高。文献[13]分别从特征项在类内出现的频数、类内分布等方面对传统互信息算法的参数进行了修正,提高了算法的分类精度。文献[14]结合功能特征互信息和特征类互信息,提出了一种基于互信息的贪心特征选择方法以找到最佳的特征子集,提高分类精确度。文献[15]采用归一化互信息代替对称不确定性作为 FCBF 算法的相关性评价标准,并进行相关性分析以获得最优特征子集,提高了平均分类正确率。本文基于上述互信息理论,提出一种改进的特征评价函数,减少冗余属性,提高分类精确度。

1.1 改进的互信息算法

传统的互信息算法按照特征词和类别一起出现的概率来衡量特征词与类别的相关性,特征词 t 和类别 c_i 的互信息公式如下:

$$MI(t,c_i)=\log_a\frac{p(t,c_i)}{p(t)p(c_i)}=\log_a\frac{p(t|c_i)}{p(t)} \tag{1}$$

其中,$p(t,c_i)$ 表示在训练集中类别为 c_i 且包含特征词 t 的文本的概率,$p(t)$ 表示在训练集中包含特征 t 的文本的概率,$p(c_i)$ 表示训练集中属于类别 c_i 的文本的概率。

由于计算各个概率时使用的频数都是包含特征词的文本数量,因此并没有考虑到特征词在各个文本中出现的词频因素。

当类别 c_q 中包含特征词 t_i 和 t_j 的文本数目一致,并且 2 个特征词在其他类别中都甚少出现,那么由 MI 公式计算出的 2 个特征词与类别之间的互信息值基本接近。

然而,当文本中包含的特征词 t_i 的频数远大于 t_j 时(如特征词 t_i 在文本出现的平均频数为 10,而特征词 t_j 在文本中出现的频数为 1,显然,特征词 t_i 更具有代表性,分类能力也越好),利用 MI 公式计算的互信息值仍相近。因此,可以得出特征在某类别各个文本中出现的频数是体现特征词分类能力的重要因素,本文根据这个因素提出了权重因子、类间和类内离散因子 3 个定义,具体的描述如下所示。

定义 1 设特征集 $T=\{t_1,t_2,\cdots,t_m\}$,训练集类别集 $C=\{c_1,c_2,\cdots,c_n\}$,记 f_{ij} 为特征词 t_j 在类别 c_i 出现的总频数,F_j 为特征词 t_j 在训练集中出现的总频数。特征 t_j 在类别 c_i 中的权重因子定义如下:

$$\omega_{ij}=\frac{f_{ij}}{F_j} \tag{2}$$

一个特征词的权重因子就是该特征词在某一类别中出现的频率。特征的权重因子越大,分类性能就越强。在互信息公式中引入权重因子,削弱互信息对低词频的倚重,增强高词频属性的影响力,提高分类准确性。

定义 2 设特征集 $T=\{t_1,t_2,\cdots,t_m\}$,训练集类别集 $C=\{c_1,c_2,\cdots,c_n\}$,记 f_{ij} 为特征 t_j 在属于类别 c_i 的文本中出现的总频数,$f_j=\frac{1}{n-1}\sum_{i=1}^{n}f_{ij}$ 为特征词 t_j 在所有类别文本中出现的平均频数。特征 t_j 的类间离散因子定义如下:

$$\alpha_j=\frac{1}{n}\sum_{i=1}^{n}\frac{2|f_{ij}-f_j|}{f_{ij}+f_j} \tag{3}$$

一个特征词的类间离散因子能够量化该特征词在各个类间的差异分布状况。类间分布差异越大,特征词就越具有类别代表性,其分类性能也就越强。将类间离散因子引入互信息公式,就能剔除在各个类中出现频率相当的、没有分类能力的冗余属性,进而降低计算复杂度,提高分类精确度。

定义 3　设特征集 $T=\{t_1,t_2,\cdots,t_m\}$,训练集类别集 $C=\{c_1,c_2,\cdots,c_n\}$,记 D_i 为类别 c_i 中文档总数,$f_{ik}(t_j)$为特征词 t_j 在属于类别 c_i 的文本 d_{ik}中出现的次数,$f_i=\frac{1}{D_i-1}\sum_{k=1}^{D_i}f_{ik}(t_j)$为特征词 t_j 在类别 c_i 中的文本中出现的平均频数。特征 t_j 的类内离散因子定义如下:

$$\beta_{ij}=\frac{1}{D_i}\sum_{k=1}^{D_i}\frac{2|f_{ik}(t_j)-f_i|}{f_{ik}(t_j)+f} \tag{4}$$

一个特征词的类内离散因子能够量化该特征在某一个类中的差异分布状况。类内差异分布越小,特征词就越有类别代表性,分类性能也就越好。将类内离散因子引入互信息方法中,就能够筛选出在某一类别各个文档中均匀出现的特征词,提高分类性能。

1.2　改进的 CDMI 特征评价函数

本文针对互信息方法中对低频词的倚重,导致冗余属性成为特征词,而有用的条件属性会漏选等不足,引入上文所定义的权重因子、类间离散因子和类内离散因子,提出一种改进的 CDMI 特征选择算法,其公式如下:

$$CDMI(t)=\sum_{i=1}^{n}\frac{\alpha}{\beta_i}\omega_i\log_a\frac{p(t|c_i)}{p(t)} \tag{5}$$

其中,t 为特征词,训练集类别集$C=\{c_1,c_2,\cdots,c_n\}$,α 表示特征 t 的类间离散因子,β_i 表示特征 t 在 c_i 类内的类内离散因子,ω_i 表示特征 t 的权重因子,$p(t)$表示训练集中包含特征 t 的文档数和总文档数的比值,$p(t|c_i)$是训练文本集中含有特征 t 的 c_i 类文档数与 c_i 类文档数的比值。使用 CDMI 算法进行属性约简,获得彼此相对相互独立的核心属性,为朴素贝叶斯模型分类做准备。

2　朴素贝叶斯优化算法

针对朴素贝叶斯分类器的条件独立性假设在众多现实应用中并不成立的缺陷,许多的学者提出可以根据不同特征词对分类的重要程度,给予不同的权值,放大决策属性的影响,从而将朴素贝叶斯模型扩展为朴素贝叶斯加权模型,如式(6)所示。

$$p(c_i|X)=\frac{p(c_i)\prod_{j=1}^{m}p(x_j|c_i)^{\omega_j}}{p(X)} \tag{6}$$

其中,$p(c_i)$表示在现有数据集中 $p(X)$类的先验概率,$p(X)$表示对象 X 出现的先验概率,$p(x_j|c_i)$ 表示特征词 x_j 的条件概率,ω_j 表示为对应于每一个特征值的权重。

2.1　PSO 算法

加权贝叶斯模型中权值的选取直接影响分类的效果。为了提高分类的准确性,本文引入了 PSO 优化算法对初始权值进行全局寻优,获取最优权值。

在 PSO 优化算法中依照速度与位置公式来调整微粒的速度与位置,求得全局最优解。由于本文设定了合适的初始权值,其大小只需微调,因此在迭代寻优中速度不宜过大,以免得不到精确解。为避免这种情况,设定了最低速度 $v_{\min}$和最高速度 $v_{\max}$,保证其收敛性,改善局部最优的状况。其速度公式和位置公式分别如式(7)、式(8)所示。

$$v_i^{s+1}=\omega v_i^s+\varphi_1 rand()(pbest_i-x_i^s)+\varphi_2 rand()(gbest_i-x_i^s) \tag{7}$$

其中,ω 表示惯性因子,φ_1 和 φ_2 为学习因子,v_i^s 表示第 s 次更新时微粒 i 的速度,x_i^s 表示第 s 次更新时微粒 i 的位置,$rand()$为随机函数。

$$x_i^{s+1}=v_i^{s+1}+x_i^s \tag{8}$$

其中,v_i^{s+1}为第 $s+1$ 次更新时微粒 i 的速度,x_i^s 为第 s 次更新时微粒 i 的位置。根据 PSO 优化算法的思想,可以得出算法 1。

算法 1　PSO 优化算法

输入　微粒群体的规模 N,迭代次数 max,最高速度 $v_{\max}$,最低速度 $v_{\min}$

输出　最优解 $gbest$

初始化位置集合 $x=(x_1,x_2,\cdots,x_i,\cdots,x_N)$和速度集合 $v=(v_1,v_2,\cdots,v_i,\cdots,v_N)$

```
for each x_i ∈ x
初始位置 x_i 作为局部最优解 pbest_i
微粒自适应度计算 fitness(x_i)
end for
gbest = min{pbest_i}
while max > 0
for i = 1 to N
更新 v_i, x_i
if fitness(x_i) < fitness(pbest_i)
当前位置 x_i 设为局部最优解
if fitness(pbest_i) < fitness(gbest)
gbest = pbest_i
```

end for
max = max − 1
end while

2.2 PSO-NB 算法

为了达到提高朴素贝叶斯模型的分类准确性和降低计算复杂度的目的,本文首先使用改进的 CDMI 算法对属性进行约简,然后利用 PSO 优化算法对朴素贝叶斯加权模型中的初始权值进行优化,生成分类器。为能清晰地阐述整个算法流程,下面将该算法划分为 CDMI 特征选择算法和 PSO-NB 分类算法来进行具体描述,完整流程如图 1 所示。

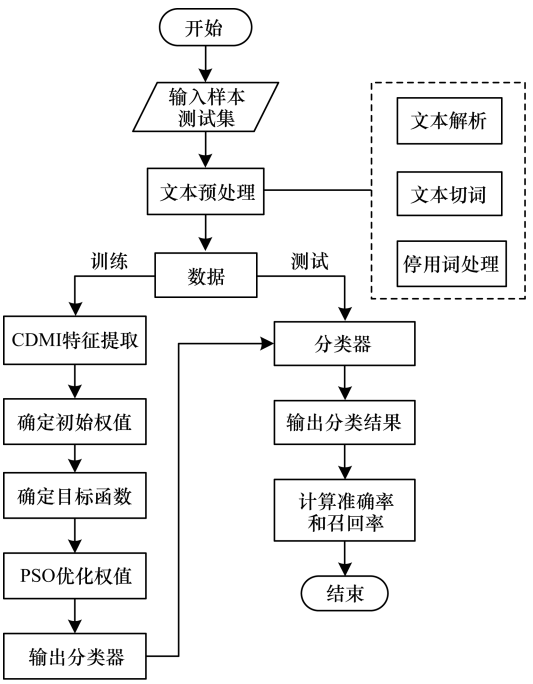

图 1 PSO-NB 算法流程

在特征选择过程中,针对原有互信息计算中忽略词频因素的不足,通过引入权重因子,放大高词频的影响,引入类内离散因子和类间离散因子筛选出具有类别代表性的特征词,具体的算法描述如算法 2 所示。

算法 2 CDMI 算法

输入 数据集,类别集 $C=\{c_1,c_2,\cdots,c_i,\cdots,c_n\}$

输出 特征集 t'

预处理得到初始特征集 $t=\{t_1,t_2,\cdots,t_j,\cdots\}$,$t'=\varnothing$
for each $t_j\in t$
计算 ω_{ij},α_j,β_{ij}
end for
for each $t_j\in t$
计算 CDMI(t_j)
if CDMI(t_j) > ε
$t'=t'\cup t_j$
end for

特征选择属性约简算法的计算复杂度为 $O(|t|)$,$|t|$为初始特征集的大小。相较于计算复杂度为 $O(|t|\times\log_a|t|)$的粗糙集约简算法和计算复杂度为 $O(|t|\times|t|)$的 TSVM-NB 约简算法,本文约简算法计算复杂度明显降低。

在分类算法中,首先将各个属性的词频比率作为其初始权值,然后利用 PSO 优化算法对权值进行优化。而在权值优化之前首先要确定目标函数,下面就针对目标函数确定的问题进行形式化描述。按照朴素贝叶斯算法的思想,假设有类别 $C=\{c_1,c_2,\cdots,c_n\}$,某一样本 $X\in c_1$,那么根据朴素贝叶斯加权式(6)计算出的概率越接近于 1,其他类别的概率越接近于 0,则分类结果就越精确。因此,根据确定目标函数的含义,可将 $p(c_i|X)$与 0 或 1 之间的误差和记为目标函数,记准确值为 γ,测量值为 γ_i,那么具体的公式可描述如下:

$$\gamma=\begin{cases}0, X\notin c_i\\ 1, X\in c_i\end{cases} \tag{9}$$

$$\gamma_i=p(c_i|X)=\frac{p(c_i)\prod_{j=1}^{m}p(x_j|c_i)^{\omega_j}}{p(X)} \tag{10}$$

则目标函数 $f(\omega)$可表示为:

$$f(\omega)=\min\sum_{i=1}^{n}|\gamma_i-\gamma| \tag{11}$$

在目标函数确定之后,就可以利用 PSO 优化算法根据已知的条件对权值迭代优化,每次更新优化都要使目标函数更小,直至目标函数收敛。将最优权值作为朴素贝叶斯加权模型中属性的权值,生成分类器,计算测试文本集的分类结果。

为了在算法 3 中能简单清晰的描述,将算法 2 中提取出的特征集 t'记为特征集 t,具体的算法描述如算法 3 所示。

算法 3 PSO-NB 算法

输入 特征集 t,类别集 C,测试集 X,迭代次数 max

输出 类别结果集 *classify*

初始化权向量 ω = ∅,结果集 classify = ∅
for each $t_j\in t$
计算 $p(c_i)$,$p(t_j|c_i)$,ω_j
$\omega=\omega\cup\omega_j$
end for
ω = PSO(ω,max)
for each $X_k\in X$
best = 0
for each $c_i\in C$
if$p(c_i|X_k)$ > best
当前概率设为最大概率 best

当前类别设为文本所属类别 $classify_k$

end for

end for

3　实验与结果分析

本文将朴素贝叶斯分类模型的改进分为 2 个部分。第 1 部分是对特征选择方法中的互信息方法进行改进,去除冗余特征词,降低维度,减少算法计算的复杂度,同时也改善了算法的分类精度,为了验证改进前后算法的性能,以分类效果作为标准,设计实验对其进行验证。第 2 部分是对加权模型中的权值进行优化,其优化方法采用 PSO 优化算法,并以优化后的权值作为条件属性对分类影响的重要程度。为了验证权值优化前后算法的能力,设计实验将 PSO-NB 算法与 NB 算法以及权值未优化的 WNB 算法的性能进行对比。

本文采用 Newsgroups-18828 中的 10 个类别新闻组作为数据文本集,对算法进行了实验测评,使用五折交叉验证法,将样本集随机地分割成大小相等但互不相交的 5 份,并分别进行 5 次样本训练和验证,计算得出每次分类的召回率与精确率,为了使分类的结果更具科学性,防止实验的随机性和偶然性,本文采取 5 次实验结果的平均值作为最终的衡量标准。

3.1　互信息参数和粒子群参数的选取

本文引入权重因子的 MI 算法为 WMI 算法,引入类间离散因子和类内离散因子的 MI 算法为 CMI 算法,然后将改进的 CDMI 算法与 WMI 算法、CMI 算法以及 MI 算法进行实验对比,确定要筛选的特征词个数。下文进行的对比主要是在不限定总的单词个数情况下,4 种算法能达到的分类结果的最高精确率,以及在相同的单词个数下 4 种算法的精确率和特征词个数。

4 种算法最高精确率对比结果如表 1 所示。

表 1　算法最高精确率对比　　%

实验次数	MI 算法	CMI 算法	WMI 算法	CDMI 算法
1	86.19	87.23	86.29	90.47
2	88.33	88.83	86.93	90.05
3	90.72	90.85	87.63	90.24
4	92.16	92.14	89.04	92.29
5	89.70	89.45	87.63	91.02
平均	89.45	89.70	87.50	90.95

在相同单词总数情况下,4 种算法的精确率和特征词数对比如图 2、图 3 所示。

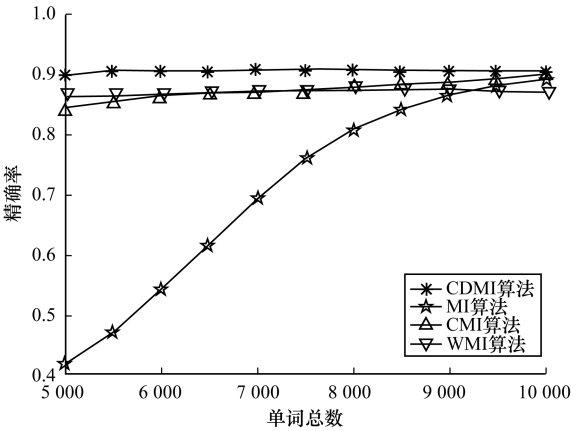

图 2　4 种算法精确率结果对比

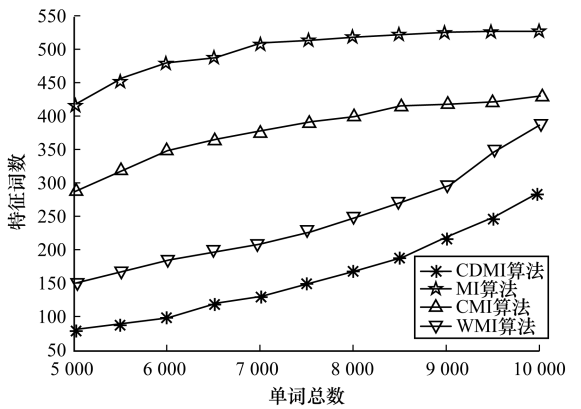

图 3　4 种算法特征词数结果对比

从图 2 可以看出,在数据集的单词由 10 000 下降到 5 000 时,MI 特征选择算法的分类结果呈急速的下降趋势;而改进后的 CDMI 算法的分类结果一直都稳定在 0.9 附近,这就说明了改进后的 CDMI 算法其分类性能比较稳定,不会因为数据集单词总量的变动而发生急剧的变化,并且 CDMI 算法的分类精确度明显优于 MI 算法。结合图 2、图 3 可以看出,当数据集单词数目相同时,CDMI 算法所选取的特征词数量明显少于 MI 算法,而分类精确度却明显优于 MI 算法,这就说明改进后的 CDMI 算法可以降低属性冗余,筛选出具有高分类能力的核心属性,这也在一定程度上降低了算法的计算复杂度。因此,可以得出,CDMI 算法无论是在分类性能上面还是计算精度上面都明显优于 MI 算法。

对于 CDMI 算法而言,在数据集的总单词数为 7 000时,分类结果的精确率最高,为了更加直观地说明这一因素,本文对 5 次实验得到的精确率的平均值进行了描述,如图 4 所示。

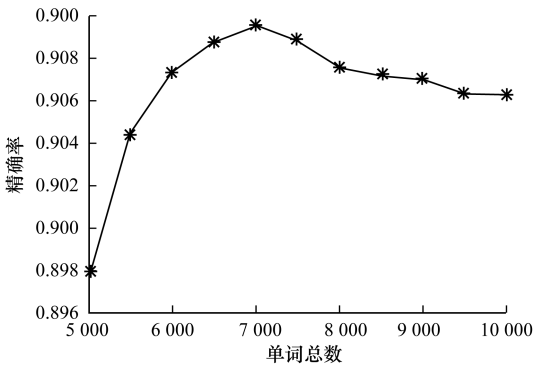

图 4 CDMI 算法准确率对比

对于 CDMI 算法,在数据集的总单词数变化的过程中,特征词的数量变化如表 2 所示。

表 2 特征词数量变化情况

单词总数	特征词数
5 000	80
5 500	90
6 000	100
6 500	120
7 000	130
7 500	150
8 000	170
8 500	190
9 000	220
9 500	250
10 000	285

由表 2 可知,在数据集的单词总数为 7 000 时,特征词的个数为 130,本文将特征词的个数设置为 130。因此,将 PSO-NB 算法中粒子的规模设为 $n=130$,粒子群其他参数的选取分别为 $\varphi_1=2.05$,$\varphi_2=2.05$,$\omega=0.729$,$rand()$ 为(0,1)区间上均匀分布的随机数。

3.2 评价指标

为有效地评估 PSO-NB 模型的分类效果,实验采用以下 3 个评价指标:

1)召回率(R)。指所有类别为正的样本集有多少被分类器判别为正类别样本,即召回。将由分类器得到的类别为正的样本集合记为 A,真正的类别为正的样本集合记为 B,则有:

$$R=\frac{|A\cap B|}{|B|} \tag{12}$$

2)精确率(P)。指分类器判断其类别为正的样本集中,真正类别为正的样本数有多少。将由分类器得到的类别为正的样本集合记为 A,真正的类别为正的样本集合记为 B,则有:

$$P=\frac{|A\cap B|}{|A|} \tag{13}$$

3)F-Measure。一个综合考虑指标,其综合考虑了召回率与精确率 2 个因素。

$$F=\frac{2\times R\times P}{R+P} \tag{14}$$

3.3 PSO-NB 算法验证

为验证本文所提出 PSO-NB 算法的效果,设计实验分别测试使用改进互信息的 NB、WNB、PSO-NB、文献[6]提出的 NWRNB、文献[9]提出的 FOA-NB 以及文献[10]提出的 TSVM-NB 这 6 种不同的算法,为避免实验的随机性和偶然性,选取互不相交的 5 个测试集进行 5 次实验,取 5 次结果的平均值为最终结果,得到 3 种分类模型的召回率、精确率以及 F-Measure 的值,进而分析分类器的分类性能,其结果对比如表 3 所示。

表 3 分类器的分类性能结果对比

算法	R	P	F-Measure
NB 算法	91.20	90.95	91.07
WNB 算法	91.09	93.66	92.00
PSO-NB 算法	94.98	94.67	94.82
NWRNB 算法	93.42	94.45	93.93
FOA-NB 算法	92.10	92.66	92.38
TSVM-NB 算法	92.32	91.88	92.10

由表 3 可以看出,PSO-NB 算法的召回率和精确率均高于其他 5 个算法。其中,NWRNB 算法和 TSVM-NB 算法分别使用粗糙集技术和支持向量机进行了属性约简,WNB 算法和 FOA-NB 算法使用不同的加权方法来评估特征词的重要程度,以提高分类性能,PSO-NB 算法首先使用改进的 CDMI 算法进行了属性约简,然后将特征词的词频比率作为初始权值,利用 PSO 优化算法对权值更新,每次更新都会使目标函数更小,一方面使得权值更加贴近特征词的重要程度,因此精确率更高,大大降低了文本类别误判的概率;另一方面所有特征词的合适权值使得文本属于某一类别的概率更加精确,因此召回率更高。

4 结束语

为提高朴素贝叶斯算法文本分类准确率并降低计算复杂度,本文提出一种改进的 PSO-NB 算法。首先利用改进的 CDMI 方法进行属性约简,然后以特征词的词频比率作为初始权值,使用绝对误差方法确定目标函数,设定速度更新中的最低和最高速度,通过 PSO 优化算法对初始权值进行优化,直至目标函数收敛,生成分类器。通过在 Newsgroups 语料集上的分析结果表明,该算法具有更高的分类精度以及更低的计算复杂度。

参考文献

[1] WU J,CAI Z. Attribute weighting via differential evolution algorithm for attribute weighted Naive Bayes[J]. Journal of Computational Information Systems,2011,7(5):1672-1679.

[2] ORHAN U, ADEM K, COMERT O. Least squares approach to locally weighted naive Bayes method[J]. Journal of New Results in Science,2012(1):71-80.

(下转第 39 页)

参考文献

[1] HMEDM A, MAHMOOD A N, HU J. A survey of network anomaly detection techniques[J]. Journal of Network and Computer Applications,2016,60:19-31.

[2] KIM G, LEE S, KIM S. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection[J]. Expert Systems with Applications, 2014,41(4):1690-1700.

[3] VASAN K K,SURENDIRAN B. Dimensionality reduction using principal component analysis for network intrusion detection[J]. Perspectives in Science, 2016, 8(C): 510-512.

[4] 李　强,严承华,朱　瑶. 基于决策树的网络流量异常分析与检测[J]. 计算机工程,2012,38(5):92-95.

[5] AMMAR A. A decision tree classifier for intrusion detection priority tagging[J]. Journal of Computer and Communications,2015,3(4):52-58.

[6] 许　倩,程东年. 基于层次聚类的网络流量异常分类算法[J]. 计算机工程,2012,38(23):131-136.

[7] HORNG S J, SU M Y, CHEN Y H, et al. A novel intrusion detection system based on hierarchical clustering and support vector machines[J]. Expert Systems with Applications,2011,38(1):306-313.

[8] MALIK A J, SHAHZAD W, KHAN F A. Network intrusion detection using hybrid binary PSO and random forests algorithm[J]. Security and Communication Networks,2015,8(16):2646-2660.

[9] WANG H, GU J, WANG S. An effective intrusion detection framework based on SVM with feature augmentation[J]. Knowledge-Based Systems, 2017, 136:130-139.

[10] RAMAN M R G, SOMU N, KIRTHIVASAN K, et al. An efficient intrusion detection system based on hypergraph-genetic algorithm for parameter optimization and feature selection in support vector machine[J]. Knowledge-Based Systems,2017,134:1-12.

[11] 马林进,万　良,马绍菊,等. 基于词袋模型聚类的异常流量识别方法[J]. 计算机工程,2017,43(5):204-209.

[12] ABUROMMAN A A, REAZ M B I. A survey of intrusion detection systems based on ensemble and hybrid classifiers[J]. Computers and Security,2017,65:135-152.

[13] JI S Y, JEONG B K, CHOI S, et al. A multi-level intrusion detection method for anomaly network behaviors[J]. Journal of Network and Computer Applications,2016,62:9-17.

[14] 杨宏宇,徐　晋. 基于改进随机森林算法的 Android 恶意软件检测[J]. 通信学报,2017,38(4):8-16.

[15] GEURTS P, ERNST D, WEHENKEL L. Extremely randomized trees[J]. Machine Learning,2006,63(1): 3-42.

[16] TAVALLAEE M,BAGHERI E,LU W,et al. A detailed analysis of the KDD CUP 99 data set[C]//Proceedings of IEEE International Conference on Computational Intelligence for Security and Defense Applications. Washington D. C.,USA:IEEE Press,2009:53-58.

[17] PEDREGOSA F, GRAMFORT A, MICHEL V, et al. Scikit-learn:machine learning in Python[J]. Journal of Machine Learning Research,2013,12(10):2825-2830.

编辑　金胡考

(上接第 32 页)

[3] JIANG Liangxiao,CAI Zhihua,ZHANG H,et al. Naive Bayes text classifiers: a locally weighted learning approach[J]. Journal of Experimental & Theoretical Artificial Intelligence,2013,25(2):273-286.

[4] HE Y L, WANG R, KWONG S, et al. Bayesian classifiers based on probability density estimation and their applications to simultaneous fault diagnosis[J]. Information Sciences,2014,259:252-268.

[5] TAHERI S, YEARWOOD J, MAMMABOV M, et al. Attribute weighted naive Bayes classifier using a local optimization[J]. Neural Computing and Applications, 2014,24(5):995-1002.

[6] 王　辉,黄自威,刘淑芬. 新型加权粗糙朴素贝叶斯算法及其应用研究[J]. 计算机应用研究,2015,32(12):3668-3672.

[7] 董立岩,隋　鹏,孙　鹏,等. 基于半监督学习的朴素贝叶斯分类新算法[J]. 吉林大学学报(工学版),2016,46(3):884-889.

[8] 李楚进,付泽正. 对朴素贝叶斯分类器的改进[J]. 统计与决策,2016(21):9-11.

[9] 刘月峰,苑江浩,张晓琳. 改进 NB 算法在垃圾邮件过滤技术中的研究[J]. 微电子学与计算机,2017,34(4):115-120.

[10] 杨　雷,曹翠玲,孙建国,等. 改进的朴素贝叶斯算法在垃圾邮件过滤中的研究[J]. 通信学报,2017,38(4):140-148.

[11] 段宏湘,张秋余,张墨逸. 基于归一化互信息的 FCBF 特征选择算法[J]. 华中科技大学学报(自然科学版),2017,45(1):52-56.

[12] 徐峻岭,周毓明,陈　林,等. 基于互信息的无监督特征选择[J]. 计算机研究与发展,2012,49(2):372-382.

[13] 刘海峰,姚泽清,苏　展. 基于词频的优化互信息文本特征选择方法[J]. 计算机工程,2014,40(7):179-182.

[14] HODUE N,BHATTACHARYYA D K,KALITA J K. MIFS-ND:a mutual information-based feature selection method[J]. Expert Systems with Applications, 2014, 41(14):6371-6385.

[15] CHANDRASHEKAR G,SAHIN F. A survey on feature selection methods[J]. Computers and Electrical Engineering,2014,40(1):16-28.

编辑　索书志