引用格式:秦怀强,赵茂先. 基于属性值加权的隐朴素贝叶斯算法[J]. 山东科技大学学报(自然科学版), 2018, 37(3):73-78.
QIN Huaiqiang, ZHAO Maoxian. Hidden Native Bayes algorithm based on attribute values weighting[J]. Journal of Shandong University of Science and Technology (Natural Science), 2018, 37(3):73-78.

基于属性值加权的隐朴素贝叶斯算法

秦怀强,赵茂先

(山东科技大学 数学与系统科学学院,山东 青岛 266590)

摘　要:隐朴素贝叶斯(Hidden Naive Bayes, HNB)算法是一种结构扩展后的朴素贝叶斯分类改进算法,其分类精确率较原算法有了很大的提高,但是在分类过程中,HNB 算法没有考虑测试实例的各个特征属性的不同取值对分类的贡献程度。针对这个问题,构建相应的加权函数计算各个特征属性取不同值时对分类的贡献程度,并利用得到的结果对 HNB 算法中用到的条件概率计算公式加权,得到了一个改进的 HNB 算法,然后利用加利福尼亚大学的埃文斯标准数据集(University of California Irvine,UCI)在 Eclipse 上对其进行数值实验,结果表明,改进的 HNB 算法较原始 HNB 算法的分类精确率有了较大提高。

关键词:隐朴素贝叶斯;朴素贝叶斯;属性值加权;分类

中图分类号:TP181　　**文献标志码**:A　　**文章编号**:1672-3767(2018)03-0073-06

DOI: 10.16452/j.cnki.sdkjzk.2018.03.010

Hidden Naive Bayes Algorithm Based on Attribute Values Weighting

QIN Huaiqiang, ZHAO Maoxian

(College of Mathematics and System Science, Shandong University of Science and Technology, Qingdao, Shandong 266590, China)

Abstract: Hidden Naive Bayes (HNB) algorithm is an improved structure-extended Naive Bayes classification algorithm, whose classification accuracy is much better than the original algorithm. However, the HNB algorithm does not consider the contribution of different values of each characteristic attribute of the test instance to the classification in the classification process. To solve this problem, an improved HNB algorithm was obtained by constructing a corresponding weighting function to calculate the contribution of each characteristic attribute value to the classification and using the results to weight the conditional probability formula used in the HNB algorithm. Then numerical experiments were conducted by using the University of California Irvine (UCI) benchmark datasets with Eclipse. The results show that the improved algorithm exhibits higher accuracy performance than the original HNB algorithm.

Key words: Hidden Naive Bayes; Naive Bayes; attribute values weighting; classification

贝叶斯网络(Bayesian networks, BN)源于概率统计学,作为机器学习的重要方法受到了广泛的关注[1]。在无限制条件下学习最优的 BN 网络结构是一个 NP 难问题,所以 GREGOR[2]建议在一定的限制条件下寻找最优的 BN 网络结构,而朴素贝叶斯(Naive Bayes,NB)分类算法是一个很好的解决思路。朴素贝

收稿日期:2017-09-08

基金项目:国家自然科学基金项目(61370207)

作者简介:秦怀强(1992—),男,山东枣庄人,硕士研究生,主要从事朴素贝叶斯分类算法方面的研究.
E-mail: 1099280742@qq.com
赵茂先(1966—),男,江苏江都人,教授, 主要从事最优化理论、方法及相关问题的研究,本文通信作者.
E-mail: sdzmx66@163.com

叶斯分类算法是一种以概率密度分析为基础,根据已知事件来预测未知事件发生可能性的分类算法[3],具有易于实现、计算速度快和分类精确率高的特点,但是当其特征属性条件独立这一假设在一些数据集上被违背时,其分类精确率会降低。因此学者们纷纷通过放松 NB 算法的假设条件,提出了许多更加优化的改进算法,如树扩展的朴素贝叶斯(tree-augmented Naive Bayes,TAN)算法[4]、平均单一依赖估计(averaged one-dependence estimators, AODE)算法[5]和隐朴素贝叶斯算法[6]等。

其中 HNB 算法具有分类效率高,计算速度快的特点,且因其给训练集中的每一个特征属性虚构了一个隐藏的父属性,这个隐藏的父属性是由其他所有特征属性共同作用产生的,所以 HNB 算法极大的放松了朴素贝叶斯分类算法的假设条件,使 HNB 算法能够在更多的不同种类的数据集上均有较好的分类表现。

但 HNB 算法提出的构建隐藏父属性的方法太过简单,无法详细地描述训练集中各属性间的相互依赖关系,针对这个问题,许多学者又提出了一些改进的 HNB 算法。李晶辉[7]提出了双层隐朴素贝叶斯分类(Double-layer Hidden Naive Bayes Classification, DHNBC)算法,该算法在 HNB 算法的基础上为每个特征属性多引入一个隐藏父属性,表示其他属性与该特征属性相关程度的加权和,其中权值的大小为属性间的条件互信息值。杜婷[8]提出加权隐朴素贝叶斯分类(weighted hidden Naive Bayes classification, WHNBC)算法,该算法利用 KL 距离和分裂信息的属性权值计算公式来构造相应的加权公式,设计了一个改进的 HNB 算法。

上述关于 HNB 的改进算法均是从特征属性出发,而实际上特征属性的不同取值对分类的贡献程度也是不同的[9]。在分类阶段,HNB 算法没有考虑测试实例的特征属性不同的取值对分类的贡献程度,这在一定程度上限制了其表现。针对这个问题,本研究提出利用训练集中的相应特征属性值的统计信息来构建加权函数,在分类阶段计算每个测试实例的特征属性在取不同属性值时对分类的贡献程度,并把计算结果作为权重,对 HNB 算法中用到的条件概率计算公式加权,得到基于属性值加权的隐朴素贝叶斯(attribute value weighting for Hidden Naive Bayes,AVWHNB)算法,然后通过实验验证 AVWHNB 算法较原始的 HNB 算法在分类精确率方面有很大的提高。

1 基于属性值加权的隐朴素贝叶斯算法

构建朴素贝叶斯分类器是一个利用给定类标记的训练集构建分类器的过程,其中训练集定义为 $D=\{X^{(1)},X^{(2)},\cdots,X^{(t)}\}$,包含 t 个训练实例。假设 $A_i(i=1,2,\cdots,n)$ 是训练集中的 n 个特征属性,并且假定训练集中有 m 个类标记,记为 $C=\{c_1,c_2,\cdots,c_m\}$,给定一个具体的测试实例 $X=(a_1,a_2,\cdots,a_n)$,这里 a_i 就是特征属性 A_i 的取值,则可以依据公式(1)来判断测试实例 X 的类标记。

$$C(X)=\underset{c_k\in C}{\operatorname{argmax}}P(c_k)\prod_{i=1}^{n}P(a_i\mid c_k)。\tag{1}$$

HNB 算法是结构扩展后的 NB 改进算法,针对训练集中的每一个特征属性 A_i,给其构建一个隐藏的父属性 A_{hpi},并且 A_{hpi} 是由除了特征属性 A_i 之外的其他所有的特征属性共同作用产生的,a_{hpi} 为 A_{hpi} 的取值。由此得到 HNB 算法的分类公式

$$C(X)=\underset{c_k\in C}{\operatorname{argmax}}P(c_k)\prod_{i=1}^{n}P(a_i\mid a_{hpi},c_k)。\tag{2}$$

本节中将要介绍的 AVWHNB 算法即是在 HNB 算法的基础上得到的。

1.1 AVWHNB 算法介绍

由公式(2)可以看出,在分类阶段,HNB 算法把每个测试实例的特征属性的各个不同取值对分类的贡献看成是一样的,这在一定程度上限制了 HNB 算法的分类精确度。针对这一问题,构建加权函数 w_{ijk} 对公式(2)中的条件概率计算公式进行加权,得到 AVWHNB 算法。其中 w_{ijk} 的计算公式如式(3)所示。

$$w_{ijk}=\frac{Num(A_i)+\dfrac{Count(a_i,a_j,c_k)}{Count(a_j,c_k)}}{Num(A_i)}。\tag{3}$$

其中 $Num(A_i)$ 是特征属性 A_i 在训练集中取值的种类数,$Count(a_i,a_j,c_k)$ 是在训练集中相应特征属性位取

值为 a_i 和 a_j ,并且类标记为 c_k 的训练实例的数目,$\frac{Count(a_i,a_j,c_k)}{Count(a_j,c_k)}$ 越大说明在训练集中特征属性标记 a_i 越倾向于类标记 c_k ,换句话说就是:$\frac{Count(a_i,a_j,c_R)}{Count(a_j,c_R)}$越大,测试实例的特征属性值 a_i 对将其分类到类别标记 c_k 的贡献越大。因此可以用 w_{ijk} 来衡量特征属性值对分类的贡献程度,并且结合公式(2)和(3)可以得到 AVWHNB 算法的计算公式如(4)所示。

$$C(X)=\underset{c_k\in C}{\operatorname{argmax}}P(c_k)\prod_{i=1}^{n}\Big(\sum_{j=1,j\neq i}^{n}W_{ij}P(a_i\mid a_j,c_k)w_{ijk}\Big)。\tag{4}$$

式(4)中的 W_{ij} 可由式(5)求得。

$$W_{ij}=\frac{I_p(A_i;A_j\mid C)}{\sum_{j=1,j\neq i}^{n}I_p(A_i;A_j\mid C)}。\tag{5}$$

式(5)中的 $I_p(A_i;A_j\mid C)$ 可由式(6)求得。

$$I_p=\sum_{a_i,a_j,c_k}P(a_i,a_j,c_k)\log\frac{P(a_i,a_j\mid c_k)}{P(a_i\mid c_k)P(a_j\mid c_k)}。\tag{6}$$

公式(6)表示的是训练集中两个特征属性的条件互信息值。

1.2　AVWHNB 算法步骤

结合 1.1 节中的内容,本节给出 AVWHNB 算法对一个测试实例 $X=(a_1,a_2,\cdots,a_n)$ 的具体分类步骤,如表 1 所示。

表 1　AVWHNB 算法步骤

Tab. 1　Steps of AVWHNB algorithm

	AVWHNB 算法步骤
第一步	利用训练集 D 中的数据计算 $P(c_k)$ 、$P(a_j\mid c_k)$ 和 $P(a_i\mid a_j,c_k)$ 的值,其计算公式如(7)～(9)所示,其中 $c_k\in C$,$a_i,a_j\in D$;
第二步	利用训练集 D 和 X 中的数据计算 w_{ijk} ,其中 w_{ijk} 的计算公式如(3)所示,然后利用第一步中的结果得到 $P(a_j\mid c_k)$ 和 $P(a_i\mid a_j,c_k)$,其中 $a_i,a_j\in X$;
第三步	利用第二步的结果和分类器公式(4)来判断测试实例 X 的类标记。

在实验时需要计算 $P(c_k)$ 、$P(a_j\mid c_k)$ 和 $P(a_i\mid a_j,c_k)$ 的值。为了避免零概率估计对实验的影响,采用拉普拉斯平滑对上述的概率公式进行估计,其具体的公式[10]为:

$$P(c_k)=\frac{Count(c_k)+1}{t+Num(C)},\tag{7}$$

$$P(a_j\mid c_k)=\frac{Count(a_j,c_k)+1}{Count(c_k)+Num(A_j)},\tag{8}$$

$$P(a_i\mid a_j,c_k)=\frac{Count(a_i,a_j,c_k)+1}{Count(a_j,c_k)+Num(A_i)}。\tag{9}$$

在实验前需要对训练集中的数据做如下的预处理:

1) 把训练集中各训练实例的缺失特征属性值补齐,使用的是 weka 中的无监督过滤器 Replace Missing Values;

2) 把训练集中各训练实例的数值型特征属性值离散化,使用的是 weka 中的无监督过滤器 Discretization;

3) 把训练集中无用的特征属性删除,使用的是 weka 中的无监督过滤器 Remove;

4) 把训练集中类标记缺失的训练实例删除,使用的是 weka 中 Instances 类下的方法 delete with Missing Class。

表 1 中的第一步为分类器构建过程的训练阶段,第二步和第三步为分类构建过程的分类阶段。第三步

中主要是利用公式(4)来判断测试实例 X 属于哪个类标记,公式(4)得到的结果可以解释为:在设计的公式中,测试实例属于这个类标记的概率最大。

2 实验分析

本节对 NB 算法、AODE 算法、HNB 算法和 AVWHNB 算法进行分类实验,实验采用的数据是 UCI 标准数据集,数据集的具体描述如表 2 所示[11]。编程使用 Java 语言和 Weka 软件中的 core.jar 算法包,使用的实验平台为 Eclipse,运行程序时的电脑配置为:处理器为 AMD Phenom(tm)II P920,内存大小为 2 GB。

表 2 训练集数据描述

Tab. 2 Training set data description

序号	数据集名称	样本数	属性数	类别值种类	缺失值	数值型数据
1	anneal	898	39	6	Y	Y
2	autos	205	26	7	Y	Y
3	breast-cancer	286	10	2	Y	N
4	balance-w	699	10	2	Y	N
5	colic	368	23	2	Y	Y
6	colic. ORIG	368	28	2	Y	Y
7	credit-a	690	16	2	Y	Y
8	credit-g	1000	21	2	N	Y
9	Glass	214	10	7	N	Y
10	heart-c	303	14	5	Y	Y
11	heart-statlog	270	14	2	N	Y
12	hepatitis	155	20	2	Y	Y
13	hypothyroid	3 772	30	4	Y	Y
14	ionosphere	351	35	2	N	Y
15	kr-vs-kp	3 196	37	2	N	N
16	labor	57	17	2	Y	Y
17	letter	20 000	17	26	N	Y
18	mushroom	8 124	23	2	Y	N
19	segment	2 310	20	7	N	Y
20	sick	3 772	30	2	Y	Y
21	sonar	208	61	2	N	Y
22	soybean	683	36	19	Y	N
23	vehicle	846	19	4	N	Y
24	vote	435	17	2	Y	N
25	vowel	990	14	11	N	Y
26	Waveform-5 000	5 000	41	3	N	Y

实验采用的是十折交叉验证的方法。十折交叉验证指的是将一个原始训练数据集平分成 10 份,进行 10 次实验,每一次都是将这 10 份数据中的 1 份作测试集、9 份做训练集,10 次实验结果的平均值为最终的结果[12]。在上面的准备工作后,通过数值实验得到了 NB、AODE、HNB 和 AVWHNB 算法的分类精确率,如表 3 所示。

表 3　各算法分类精确率对比

Tab. 3　Classification accuracy comparison of different algorithms　　%

数据集名称	NB	AODE	HNB	AVWHNB
anneal	94.10	97.22	98.66	98.33
autos	66.34	75.12	79.51	79.02
breast-cancer	72.73	73.08	73.43	73.78
balance-w	97.42	97.42	95.71	95.99
colic	78.80	80.43	81.25	81.52
colic. ORIG	76.36	77.17	73.91	75.54
credit-a	84.93	86.09	84.49	84.93
credit-g	75.90	76.30	77.30	77.50
Glass	58.41	61.68	57.48	58.41
heart-c	83.83	82.51	82.51	80.86
heart-statlog	84.81	84.07	81.85	81.85
hepatitis	83.87	83.23	81.29	83.23
hypothyroid	92.90	93.50	93.32	93.27
ionosphere	91.17	91.45	92.59	92.59
kr-vs-kp	87.80	91.11	92.43	92.08
labor	96.49	94.74	89.47	91.23
letter	70.03	85.24	86.19	86.17
mushroom	95.43	99.95	99.96	99.96
segment	89.05	92.94	94.68	94.63
sick	96.77	97.53	97.80	97.72
sonar	75.48	80.77	82.69	81.73
soybean	92.24	92.83	94.73	94.88
vehicle	60.99	71.63	73.52	73.64
vote	90.11	94.25	94.25	94.48
vowel	66.16	88.59	92.83	92.53
Waveform-5000	79.90	84.36	83.42	83.42
平均分类精确度	82.39	85.89	85.97	86.13

对比这 4 个算法在每一个训练集上的表现得到表 4。

表 4　各算法在每个数据集上的分类精确率对比

Tab. 4　Classification accuracy comparison of different algorithms at each dataset

赢/平/输	NB	AODE	HNB
AODE	21/1/4	—	—
HNB	18/0/8	15/2/9	
AVWHNB	18/2/6	16/1/9	12/4/10

对比上述 4 个算法的时间复杂度得到表 5。

表 5　各算法时间复杂度对比

Tab. 5　Time complexity comparison of different algorithms

算法	NB	AODE	HNB	AVWHNB
训练时间	$O(nt)$	$O(tn^2)$	$O(tn^2+mn^2v^2)$	$O(tn^2+mn^2v^2)$
分类时间	$O(nm)$	$O(mn^2)$	$O(mn^2)$	$O(mn^2)$

在表 5 中，m 是类标记的种类数，n 是特征属性的数目，v 是一个特征属性的各个属性值的平均数目，t 是训练集中训练实例的数目[13]。

由表 3 可知 AVWHNB 算法的平均分类精确率大于 NB 算法、AODE 算法和 HNB 算法。由表 4 看出 AVWHNB 算法分类效果好的数据集数目多于 NB、AODE 和 HNB 算法。由表 5 可以看出 AVWHNB 算法的训练时间、分类时间和 HNB 算法相同，即 AVWHNB 算法的时间复杂度和 HNB 算法相同。综合上面的分析可知，AVWHNB 算法在提高分类精确率的同时并未增加算法的时间复杂度，这充分说明了 AVWHNB 算法的分类效果比 HNB 算法好。

从表 3 和表 4 的数据中可以看出 AVWHNB 算法也存在着一些不足。首先，当数据集中各特征属性间的关联程度较弱[14]时，其在某些数据集上的表现不如 NB 算法。其次，在某些数据集上的表现不如原始

HNB 算法说明 AVWHNB 算法的稳定性有待提高。针对上述问题,在分类中常用的多分类器思想是一个很好的解决办法,而针对于多个分类器的输出,则可以用投票机制来进行综合以给出最终的分类结果[15-16]。

4 结束语

本研究提出的 AVWHNB 算法为一种改进的 HNB 算法,其核心思想是利用构建的加权函数计算各个特征属性值对分类的贡献程度,并将得到的结果对 HNB 算法中用到的条件概率计算公式加权来改进 HNB 算法,然后通过实验对比了 AVWHNB、HNB、NB 和 AODE 算法的平均分类精确率、在每个数据集上的分类精确率和时间复杂度,结果显示 AVWHNB 算法的整体分类效果要优于原始的 HNB 算法。

虽然 AVWHNB 算法的整体分类效果要优于 HNB 算法,但在对比每个数据集上的分类效果时,AVWHNB 算法分类效果好的数据集的数目只是略高于 HNB 算法,这说明改进的算法还是不够稳定,所以在以后的研究中,可以将特征属性值加权和特征属性加权相结合,并借鉴 AODE 算法聚合分类器的思想。具体的思路是:先找到一个合适的方法来判断数据集中各个特征属性的关联程度。然后设置一个阈值,当关联程度低于这个阈值时可以使用 NB 算法来对数据集进行分类,而当关联程度高于这个阈值时可以采用 AVWHNB 算法对数据集进行分类。对于这两类分类器,在每一类上均可以设置多个分类器,在具体分类时可采用某种方法将原始数据集分成若干份,每一份数据都由一个分类器来处理。最后用投票机制综合多个分类器的分类结果来确定测试实例的类标记。经过上述处理,理论上可以得到分类效果好且稳定的 HNB 改进算法。

参考文献:

[1]秦锋,任诗流,程泽凯,等.基于属性加权的朴素贝叶斯分类算法[J].计算机工程与应用,2008,44(6):107-109.
QIN Feng,REN Shiliu,CHENG Zekai,et al. Attribute weighted Naive Bayes classification[J]. Computer Engineering and Applications,2008,44(6):107-109.

[2]GREGORY F C. The computational complexity of probabilistic inference using Bayesian belief networks[J]. Artificial Intelligence,1990,42(2/3):393-405.

[3]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.
WANG Hui,HUANG Ziwei,LIU Shufen. Novel weighted rough naive Bayes algorithm and its application[J]. Application Research of Computers,2015,32(12):3668-3672.

[4]FRIEDMAN N,GEIGER D,GOLDSZMIDT M. Bayesian network classifiers[J]. Machine Learning,1997,29:131-163.

[5]GEOFFREY I W,JANICE R B,WANG Z H. Not so Naive Bayes:Aggregating one-dependence estimators[J]. Machine Learning,2005,58(1):5-24.

[6]JIANG L X,ZHANG H,CAI Z H. A novel Bayes Model:Hidden Naive Bayes[J]. IEEE Transactions on Knowledge and Data Engineering,2009,21(10):1361-1371.

[7]李晶辉.基于互信息的多层隐朴素贝叶斯算法研究[D].长沙:湖南大学,2012.

[8]杜婷.基于属性选择的朴素贝叶斯分类[D].合肥:中国科学技术大学,2016.

[9]CHANG H L. A gradient approach for value weighted classification learning in Naive Bayes[J]. Knowledge -Based Systems,2015,85:71-79.

[10]ZHONG L X,XIANG R Y,DAE K K. Experimental analysis of Naive Bayes classifier based on an attribute weighting framework with smooth kernel density estimations[J]. Applied Intelligence,2016,44(3):611-620.

[11]MERZ C,MURPHY P,AHA D. UCI repository of machine learning database[DB/OL]. [2017-09-08],http://www.ics.uci.edu/mlearn/MLRpository.html.

[12]袁梅宇.数据挖掘与机器学习 WEKA 应用技术与实践[M].北京:清华大学出版社,2014:330-333.

[13]ZHONG L X,DAE K K. Attribute weighting for averaged one-dependence estimators[J]. Applied Intelligence,2017,46(3):616-629.

[14]JUN Y. Correlation coefficient between dynamic single valued neutrosophic multisets and its multiple attribute decision-making method[J]. Information,2017,8(2):41.

[15]CAGATAY C,MEHMET N. A sentiment classification model based on multiple classifiers[J]. Applied Soft Computing,2017,50:135-141.

[16]ANDRONIKI T,GEORGE E T,ANASTASIOS R,et al. A methodology to carry out voting classification tasks using a particle swarm optimization-based neuro-fuzzy competitive learning network[J]. Evolving Systems,2017,8(1):49-69.

(责任编辑:傅 游)