网络出版时间：2018-02-08 15:52:47
网络出版地址：http://kns.cnki.net/kcms/detail/42.1009.C.20180207.1449.005.html

DOI:10.13546/j.cnki.tjyjc.2018.02.005

基于信息值的相关属性约减—加权二分类朴素贝叶斯算法研究

杨立洪，李琼阳，李兴耀
（华南理工大学 数学学院，广州 510640）

摘 要：在经典的朴素贝叶斯分类算法中，往往假设各属性之间相互独立，且对目标变量的影响程度一致，但实际问题几乎不可能满足此假设。实际应用中的二分类问题最多，在二分类问题中考虑到属性相关、样本分布不平衡、各属性影响程度的不一致性对模型性能的影响，文章提出一种基于信息值的相关属性约减—加权二分类朴素贝叶斯模型，同时在判定样本类别归属时，采用自适应学习选择合适的阈值，以此削弱不平衡样本集的影响。实证结果表明，通过引入信息值，进行相关属性的约减—加权，模型结果在准确率上较之传统朴素贝叶斯算法有极大提升。

关键词：信息值；属性约减；加权；二分类；贝叶斯算法；自适应
中图分类号：TP391.7　　**文献标识码**：A　　**文章编号**：1002-6487(2018)02-0023-04

0 引言

朴素贝叶斯分类(Naive Bayes Classification)算法是目前公认的一种简单有效的分类算法，它是一种基于概率的分类方法，被广泛地应用于模式识别、自然语言处理、机器人导航、机器学习等领域。朴素贝叶斯算法是基于特征项间独立、对目标变量影响力一致的假设，但实际应用中极少数问题能满足此假设。为此许多学者致力于改进朴素贝叶斯算法，以期提高算法的普适性和准确率。改进之处主要体现在两个方面：一是在属性选择上预把控；二是衡量各属性对目标变量的影响程度，对属性加权。

在属性选择上，Geenen P Ld等[1]提出一种基于互信息选择特征属性的方法，并利用朴素贝叶斯算法在二分类问题上取得了极好的分类效果。魏浩和丁要军[2]提出用属性关联度表示一个属性和类属性间的相关性，反映这个属性对分类结果影响的程度；用属性冗余度表示一个属性和其他属性之间相关性，反映这个属性和其他属性间的依赖度。王行甫和杜婷[3]提出利用CFS算法选择特征属性。焦鹏等[4]提出将属性先验分布的参数设置加入到属性选择的过程中，并研究当先验分布服从 Dirichlet 分布及广义 Dirichlet 分布情况下的具体实践方案。研究出一种加权朴素贝叶斯算法，通过对不同的特征项提供不同的权值，削弱特征项之间的相关性。

在属性加权上，饶丽丽等[5]提出在传统权重计算基础上，考虑到特征项在类内和类间的分布情况，另外还结合特征项间的相关度，调整权重计算值，加大最能代表所属类的特征项的权重。Jiang L等[6]提出在训练集中深度计算特征

基金项目：国家自然科学基金资助项目(11271140)；广东省产学研协同创新成果转化项目(2016B090918041)
作者简介：杨立洪(1961—)，男，湖南湘潭人，博士，教授，研究方向：大数据、智慧产业、金融工程。
李琼阳(1991—)，女，河南漯河人，硕士研究生，研究方向：数理统计及经济信息管理。
李兴耀(1993—)，男，四川自贡人，博士研究生，研究方向：量化金融。

Improvement of ARMA Model Based on Gevers-Wouters Algorithm

Tang Xinxin[a], Deng Guangming[a,b]

(a.College of Science, b.Institute of Applied Statistics, Guilin University of Technology, Guilin 541006,China)

Abstract: Concerned on the problem that the prediction results are not ideal when the autoregressive moving-average model (ARMA model) is constructed in time series analysis, this paper introduces Gevers-Wouters algorithm to improve the model. Taking the data of Chinese domestic airline passenger turnover as an example, the paper uses BIC criterion to realize orders determination of ARMA model after eliminating the seasonal fluctuations in the data, and then employs Gevers-Wouters algorithm to adjust the parameters of MA part in the model and construct unique ARMA model for every period therefrom. By comparing the predictive value with the real value before and after improving the model, the accuracy of prediction is increased obviously.

Key words: ARMA model; seasonal trend; Gevers-Wouters algorithm

加权频率，估计朴素贝叶斯的条件概率。Lungan Zhang等[7]提出了两种自适应特征加权方法：一是基于树的自适应特征加权；二是基于信息增益率的特征加权。

为了解决朴素贝叶斯算法的两个固有缺陷，本文将构建基于信息值的相关属性约减—加权二分类朴素贝叶斯模型，有效解决属性相关、属性加权的问题。在判定样本类别归属时，采用自适应学习选择合适的阈值，以此削弱不平衡样本集的影响，提高模型的准确率。最后在某运营商提供的垃圾短信用户行为消费特征样本数据上进行实证分析，结果表明：基于信息值的相关属性约减—加权二分类朴素贝叶斯较传统朴素贝叶斯模在模型准确率上有显著提升。

1 朴素贝叶斯算法

朴素贝叶斯算法的分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，具有最大后验概率的类则为该对象所属的类。朴素贝叶斯是在贝叶斯分类法的基础上提出的，该算法满足一个简单的假定，即在给定目标值时属性值之间相互条件独立。

朴素贝叶斯分类算法的工作过程如下：

（1）设 A 表示训练样本的属性集，有 m 个属性 $A_1, A_2, \cdots, A_m$；C 表示类集合，有 k 个类 $C_1, C_2, \cdots, C_k$；每个数据样本 X 用一个 m 维特征向量来描述 m 个属性的值，即：$X=(x_1, x_2, \cdots, x_m)$，其中 $x_i \in A_i\,(1 \le i \le m)$。

（2）对训练样本集进行统计，计算得到每个特征属性在各类别的条件概率估计，即：

$$p(x_1|C_1), p(x_2|C_1), \cdots, p(x_m|C_1)$$
$$p(x_1|C_2), p(x_2|C_2), \cdots, p(x_m|C_2)$$
$$\vdots$$
$$p(x_1|C_k), p(x_2|C_k), \cdots, p(x_m|C_k)$$

（3）对每个类别计算后验概率，根据贝叶斯定理及朴素贝叶斯算法的假定可知：

$$p(C_i|X)=\frac{p(C_i)\cdot\prod_{j=1}^{m}p(x_j|C_i)}{p(X)} \tag{1}$$

（4）取最大后验概率项作为样本所属类别：

$$c(X)=\arg\max_{(1\le i\le k)} p(C_i)\cdot\prod_{j=1}^{m}p(x_j|C_i) \tag{2}$$

2 基于信息值的相关属性约减—加权二分类朴素贝叶斯算法

2.1 信息值

信息值简称 IV，衡量自变量对目标变量的影响程度，是建模时筛选变量的一个非常重要的指标。它起源于香农提出的信息理论，与广泛应用的熵有极大的相似性，主要适用于二分类模型。在介绍信息值之前，有必要先引入 WOE (weight of evidence)。为方便表述，将二分类目标变量标识为 0、1，其中 1 表示违约，0 表示正常。WOE 实质上是表示自变量取某个值时对违约比例的影响，例如当自变量取值为 i 时对目标变量违约比例的影响 woe_i 的计算公式如下：

$$woe_i=\ln\left(\frac{B_i/B_T}{G_i/G_T}\right) \tag{3}$$

其中，B_i 指当该自变量取值为 i 时的违约样本数，G_i 指该自变量取值为 i 时的正常样本数。B_T 指建模样本数据中总的违约样本数，G_T 指建模样本数据中总的正常样本数。

信息值衡量一个变量的信息量，例如对于一个有 n 个取值的自变量而言，该自变量的信息值计算公式如下：

$$IV=\sum_{i=1}^{n}IV_i=\sum_{i=1}^{n}\left(\frac{B_i}{B_T}-\frac{G_i}{G_T}\right)\cdot\ln\left(\frac{B_i/B_T}{G_i/G_T}\right) \tag{4}$$

从计算公式上可以看出，信息值是自变量 WOE 值的一个加权组合，其值的大小决定了自变量对目标变量的影响程度。从形式上看来，信息值与信息熵也是极为相似的。

通常认为，$0.1 \le IV \le 0.3$ 时认为该变量对目标变量有中等影响力；$0.3 \le IV \le 0.5$ 时认为该变量对目标变量有较强影响力；$IV \ge 0.5$ 时认为该变量对目标变量有极强的影响力。

2.2 属性约减

朴素贝叶斯算法要求建模的自变量之间相互独立，但在实际应用过程中，参与建模的变量之间往往会存在一定程度的相关性。相关系数是反映两个变量之间相关程度的一个重要度量，计算公式如下：

$$\rho_{A_i,A_j}=\frac{\mathrm{cov}(A_i,A_j)}{\sqrt{D(A_i)}\cdot\sqrt{D(A_j)}} \tag{5}$$

一般情况下，$0.4 \le \rho_{A_i,A_j} \le 0.6$，认为 A_i 与 A_j 之间中等程度相关；$0.6 \le \rho_{A_i,A_j} \le 0.8$，认为 A_i 与 A_j 之间强相关；$0.8 \le \rho_{A_i,A_j} \le 1$，认为 A_i 与 A_j 之间极强相关。

为了尽可能满足朴素贝叶斯算法的假设条件，有必要对变量进行属性约减。约减规则如下：

（1）根据 IV 值筛选一批对目标变量影响程度较大的自变量，一般选择 IV 值大于0.3的变量。

（2）计算（1）中筛选出的自变量之间的相关系数，一般当 $\rho_{A_i,A_j} > 0.5$ 时，即可认为 A_i 与 A_j 之间有较强的相关性，不宜全部进入模型。

（3）若 $\rho_{A_i,A_j} > 0.5$，且变量 A_i 的 IV 值大于变量 A_j 的 IV 值，则只选择变量 A_i 参与建模。

2.3 属性加权

朴素贝叶斯算法选择后验概率最大的类别作为归属，在计算中默认各属性对目标变量的影响程度一致，但由于信息值已知，各自变量对目标变量的影响程度是有差异的，因此考虑利用各属性的 IV 值进行属性加权。经由前文中基于信息值和相关属性约减筛选出对目标变量有强影响力的 m 个变量，并保证了各变量之间几乎独立。假设这 m 个变量蕴含了所有的信息，则属性 j 所占的比重 e_j 即为：

$$e_j=\frac{IV_j}{\sum_{j=1}^{m}IV_j} \tag{6}$$

信息值的比重越大，该变量对目标比变量的影响力就越大，由于条件概率 $p(x_j|C_i)$ 的取值范围是 (0，1)，为此可以按照如下公式修正权重，即：

$$w_j=\frac{1-e_j}{\sum_{j=1}^{m}(1-e_j)}=\frac{1-e_j}{m-1} \tag{7}$$

对后验概率公式进行修正：

$$p(C_i|X)=\frac{p(C_i)\cdot\prod_{j=1}^{m}p(x_j|C_i)^{w_j}}{p(X)} \tag{8}$$

$$c(X)=\arg\max_{(i\in 0,1)}p(C_i)\cdot\prod_{j=1}^{m}p(x_j|C_i)^{w_j} \tag{9}$$

选择后验概率最大的类别作为归属。

2.4 自适应学习选择合适的阈值

在二分类朴素贝叶斯算法中，若 $p(C_1|X)>p(C_0|X)$，则把样本 X 归为 C_1 类。但有时样本集中在二类样本的数量上相差极大，样本分布极不均衡，若仍按上述方式判定类别归属，则极有可能误判。考虑当 $p(C_1|X)>l\cdot p(C_0|X)$ 时，把样本 X 归为 C_1 类，l 值的具体选择可以依赖其在训练数据集上的表现。一般情况下，若 C_1 类的样本错判到 C_0 类的较多，可以考虑在 (0，1) 内选择合适 k 值；若 C_0 类的样本错判到 C_1 类的较多，可以考虑在 (1，50) 内选择合适 k 值，以此来提高模型的准确率。

3 实证

3.1 数据的收集和处理

某运营商提供了用户行为消费特征样本数据共78258条，其中垃圾短信用户样本数据11837条，用1标识；正常用户66421条，用0标识。数据集有当月消费额、品牌、通话时长、发送短信条数、短信回复率、账户余额、是否为垃圾短信用户等共56个属性。下面将利用基于信息值的相关属性约减—加权朴素贝叶斯算法来进行建模。

3.2 基于信息值进行属性约减—加权

在垃圾短信用户识别过程中，是否是垃圾短信用户是目标变量。利用信息值的计算公式，计算除目标变量的其他55个变量的信息值。按照上文中的属性约减规则进行属性约减，最终选定7个变量参与建模，变量的信息值、权重、相关系数如表1和表2所示。

表1 建模变量的信息值、权重

指标	信息值(IV)	权重
当月消费额	0.8	0.13
消费额的波动	0.67	0.14
当月返还金额	0.67	0.14
短信回复率	0.51	0.14
通话时长	0.49	0.15
账户余额	0.45	0.15
品牌	0.36	0.15

表2 相关系数矩阵

	消费额	消费额的波动	当月返还金额	短信回复率	品牌	账户余额	通话时长
消费额	1.00	−0.05	−0.07	−0.24	−0.23	0.49	0.35
消费额的波动	−0.05	1.00	0.29	0.36	0.11	−0.20	−0.14
当月返还金额	−0.07	0.29	1.00	0.32	0.46	−0.05	0.05
短信回复率	−0.24	0.36	0.32	1.00	0.27	−0.33	−0.11
品牌	−0.23	0.11	0.46	0.27	1.00	0.02	0.08
账户余额	0.49	−0.20	−0.05	−0.33	0.02	1.00	0.27
通话时长	0.35	−0.14	0.05	−0.11	0.08	0.27	1.00

由参与建模的7个变量的相关系数矩阵可以看出，变量之间基本满足条件独立的假设。

3.3 模型构建及对比检验

在数据集中按照7:3的比例进行分层随机抽样，划分训练样本与检验样本。在利用模型求出每一个样本的后验概率之后采用自适应学习选择合适的 l 值，判别每一个样本的归属。经实验发现，当 $l=0.25$ 时模型的准确率最高。至此，当 $p(C_1|X)>0.25\cdot p(C_0|X)$ 时，判断样本 X 为垃圾短信用户。

分别利用改进的朴素贝叶斯算法和传统朴素贝叶斯算法建立模型，二者在训练集和测试集的建模结果如表3所示：

表3 模型改进前后效果对比

	改进的朴素贝叶斯		朴素贝叶斯	
	训练集	测试集	训练集	测试集
(0,0)	44618	19157	29123	12596
(0,1)	1877	769	17372	7330
(1,0)	3274	1362	637	237
(1,1)	5012	2189	7649	3314
准确率(%)	90.6	90.92	67.12	67.76

注：(1,0)表示垃圾短信样本被判定为正常用户。

由表3可以看出基于信息值的相关属性约减—加权朴素贝叶斯算法较传统的朴素贝叶斯算法在建模效果的准确率上有显著提升。

4 结论

朴素贝叶斯算法是目前比较高效经济的分类算法之一，也是常用的十大算法之一。本文在朴素贝叶斯算法的基础上，针对朴素贝叶斯算法的固有缺陷，提出基于信息值的属性约减—加权改进方法，该方法能有效处理相关属性，使之尽可能满足朴素贝叶斯的理论假设。同时利用属性的信息值为属性赋予的权重，采用自学习选择合适的判定阈值，以降低不平衡样本对模型的影响，从而最大程度上提高模型的准确率。

在垃圾短信客户识别的实际应用过程中也发现，基于信息值的属性约减—加权改进朴素贝叶斯算法较传统的朴素贝叶斯算法在准确率上有显著提升。

参考文献：

[1]Geenen P L, Gaag L C V D, Loeffen W L A, et al. Constructing Naive Bayesian Classifiers for Veterinary Medicine: A Case Study in the

Clinical Diagnosis of Classical Swine Fever[J]. Research in Veterinary Science, 2011, 91(1).
[2]魏浩, 丁要军. 一种基于相关的属性选择改进算法[J]. 计算机应用与软件, 2014, 31(8).
[3]王行甫, 杜婷. 基于属性选择的改进加权朴素贝叶斯分类算法[J]. 计算机系统应用, 2015, 24(8).
[4]焦鹏, 王新政, 谢鹏远. 基于属性选择法的朴素贝叶斯分类器性能改进[J]. 电讯技术, 2013,(3).
[5]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[J]. 厦门大学学报:自然科学版, 2012, 51(4).
[6]Jiang L, Li C, Wang S, et al. Deep Feature Weighting for Naive Bayes and Its Application to Text Classification[J]. Engineering Applications of Artificial Intelligence, 2016,(52).
[7]Wang S, Jiang L, Li C. A CFS-Based Feature Weighting Approach to Naive Bayes Text Classifiers[J]. Knowledge-Based Systems, 2016, (100).

（责任编辑/亦　民）

Research on Relevant Attribute Reduction—Weighted BinaryClassification Naive Bayesian Algorithm Based on Value of Information

Yang Lihong, Li Qiongyang, Li Xingyao

(School of Mathematics, South China University of Technology, Guangzhou 510640, China)

Abstract: In the classic naive Bayesian classification algorithm, it often assumes that each attribute is independent and consistent with the degree of influence on the target variable, but in practice, it is impossible to satisfy this assumption. Most of the practical problems are associated with binary classifications, in which must be considered the influence of the relevance of attributes, the imbalance distribution of samples and the inconsistence of each property's influence level on the model performance. This paper puts forward a relevant attribute reduction—weighted binary classification naive Bayesian model based on the value of information, and that at the same time in determining the home category of sample, self-adaptive learning is used to select appropriate threshold so as to weaken the influence of unbalanced sample set. The empirical result shows that compared with the traditional Naïve Bayesian algorithm, the accuracy of the proposed model's result has been greatly elevated through introducing the information value and performing relevant attribute reduction—weighting.

Key words: information value; attribute reduction; weighting; binary classification; naive Bayesian algorithm; self-adaption