文章编号：2095-2163(2019)02-0104-05 中图分类号：TP181 文献标志码：A

面向脏数据的贝叶斯统计建模研究

程炜东，王洪亚，郭开彦
（东华大学 计算机科学与技术学院，上海 201620）

摘 要：为了处理贝叶斯建模中的脏数据，通常会有 2 种解决方法。一种是对整个数据集进行清洗，但这种方法的代价很高，且对中型或大型的数据集可行性较低。另一种是使用点估计，这种点估计的方法虽然能有效减少清洗的代价，但是对训练出来贝叶斯模型的可信程度没有保证。针对上述清洗方法中存在的问题，本文提出了一种基于区间的贝叶斯统计建模方法，简称区间贝叶斯建模。区间贝叶斯建模结合中心极限定理，使用区间估计的方法，保证了真实的后验概率会以一定的概率落在后验概率区间内。实验结果表明，区间贝叶斯建模通过清洗少量的样本，便能够训练出良好的贝叶斯模型，有效改善了清洗成本，并在精度和召回率上比不清洗任何数据的情况有显著的提升。

关键词：贝叶斯分类器；数据清洗；概率区间；区间比较策略

Research on Bayesian statistical modeling with dirty data

CHENG Weidong, WANG Hongya, GUO Kaiyan
(School of Computer Science and Technology, Donghua University, Shanghai 201620, China)

【Abstract】 In order to deal with the dirty data in Bayesian statistical modeling, there are usually two solutions. One is to clean the entire data, but this method is costly and less feasible for medium or large data sets. Another is to use point estimation. Although this method can effectively reduce the cost of cleaning, the authenticity of Bayesian model is not guaranteed. In view of the problems in the above cleaning methods, this paper proposes a novel approach, namely Interval - based Bayesian Statistical Modeling, abbreviated as Interval Bayesian Modeling. Interval Bayesian Modeling combines the Central Limit Theorem, and uses the method of interval estimation to ensure that the posterior probability will fall in the estimated probability interval with constant probability. Experimental results show that Interval Bayesian Modeling can train a good Bayesian model by cleaning a small number of samples, offering significant improvement in cost over cleaning all of the data and significant improvement on precision and recall over cleaning none of the data.

【Key words】 Bayesian classifier; data cleaning; probability interval; interval comparison strategies

0 引 言

分类是一个非常重要的数据挖掘问题，简单来说就是确定对象属于哪个预设目标类的过程。分类通过分析训练样本数据，产生关于类别信息的精确描述，然后去预测类标未知的对象所属的类别[1]。如今，分类技术是数据挖掘中应用价值较高的技术之一，且这项技术已经发展到相对成熟的地步，在许多行业中都起着决策支持的作用。

贝叶斯分类器是一种常见的概率分类器，具有坚实、完善的理论体系，并且在现实生活中被广泛应用[2]。在很多领域中，贝叶斯分类器的分类效果并不逊色于其它分类算法，例如决策树、SVM、以及神经网络，某些时候贝叶斯分类器的性能还会超过这些分类算法。贝叶斯分类器的算法并不复杂，而且当数据集的规模较大或很大时，分类的准确率也有一定的保证。此外，贝叶斯分类器可以用数学公式进行定量描述。因此，贝叶斯分类器一直都是机器学习领域研究人员致力于探讨的热门内容。

贝叶斯分类器可以处理离散的和连续的数据类型。同时，贝叶斯分类器在某种程度上可以看作是一种动态分类模型，换句话说就是当把新数据逐渐加入到原始数据集中，训练过程可以增量进行[3]。贝叶斯分类器的目的就是预测待检测数据所属的类别，为了达到这个目的，贝叶斯分类器会首先计算这些待检测数据属于每一个类别的概率，然后根据最大后验假设，把待检测数据分类到具有最大后验概率的类别中。在上述计算后验概率的过程中，数据集中的所有特征属性都会参与决策，而不是只由某几个特征属性决定着分类结果，所以最后得到的结果是较为准确的[4]。朴素贝叶斯分类器是一种比较简单的贝叶斯分类器，且基于一个朴素的假设，即

作者简介：程炜东(1994-)，男，硕士研究生，主要研究方向：机器学习；王洪亚(1976-)，男，博士，教授，主要研究方向：数据库管理；郭开彦(1992-)，男，硕士研究生，主要研究方向：数据清洗。

收稿日期：2018-12-16

属性独立假设：当给定一个分类属性中的类别，各个特征属性之间是相互独立的。由于属性独立假设，朴素贝叶斯分类器具有计算高效、精确度高等特点。然而，现实中属性独立假设在大多数时候都无法成立，这样就会对朴素贝叶斯分类器的分类性能造成一定的影响。因此，半朴素贝叶斯分类器以及其它对朴素贝叶斯分类器的改进方法就成为了目前科研的热点问题[5]。下面拟对此展开研究论述如下。

1　脏数据对贝叶斯建模的影响

在贝叶斯建模中使用脏数据会导致很多问题，比如使统计得到的概率不准确。一般来说，不同比例的错误对贝叶斯分类器的分类准确度有不同的影响。下面通过一个例子来具体说明脏数据对贝叶斯分类器分类准确度的影响。

表 1 是通过苹果的重量（Weight）、颜色（Color）、以及形状（Shape），对苹果质量好坏（Good Apple）进行分类的一个二分类例子，其中括号内的值代表数据清洗后的正确值。在特征属性中，Weight 是连续属性，Color 和 Shape 是离散属性。分类属性 Good Apple 是离散属性。

表 1　脏数据的清洗结果

Tab. 1　The cleaning result for dirty data

Weight	Color	Shape	Good Apple
111(106)	green	irregular	No
152	red	irregular	Yes
148	green(red)	circle	Yes
145	red(green)	circle	Yes(No)
147	green	irregular	No
118	red	circle	Yes
135	green	circle(irregular)	No
121	red	circle(irregular)	No
109	green	circle	No
138	red	irregular(circle)	Yes

本文利用表 1 中的脏数据进行一次贝叶斯建模，同时再利用表 1 中括号内数据清洗后的正确值进行一次贝叶斯建模。2 次贝叶斯建模均使用朴素贝叶斯分类器，并假设测试数据为｛126，red，circle｝。当事先不进行数据清洗，直接使用脏数据进行贝叶斯建模时，可以计算出后验概率 $P(\text{No} \mid 126,\text{red},\text{circle}) = 0.004\ 0$，$P(\text{Yes} \mid 126,\text{red},\text{circle}) = 0.001\ 9$，测试数据｛126，red，circle｝分类为 No。当事先进行数据清洗，并使用数据清洗后的正确值进行贝叶斯建模时，可以计算出后验概率 $P(\text{No} \mid 126,\text{red},\text{circle}) = 0.001\ 2$，$P(\text{Yes} \mid 126,\text{red},\text{circle}) = 0.004\ 2$，测试数据｛126，red，circle｝分类为 Yes。通过以上计算可以看出，脏数据会导致每个类别的后验概率发生变化，从而影响到贝叶斯分类器的分类结果，尤其是脏数据所占的比例较大时，很有可能使分类结果出错。

为了处理贝叶斯建模中的脏数据，通常会有 2 种解决方法，这里对其表述如下。

（1）在贝叶斯建模前对整个数据集进行清洗，但这种方法的代价很高，且对中型或大型的数据集可行性较低。

（2）使用点估计，即从包含脏数据的数据集中采样，再对样本进行清洗，而后用清洗过的样本来训练贝叶斯模型。这种点估计方法虽然能有效减少清洗的代价，但是对训练出来贝叶斯模型的可信程度却未能做出基础保证。综合前述脏数据对贝叶斯建模的影响，本文提出了区间贝叶斯建模方法，这种方法只要求用户清洗小部分数据，然后利用清洗过的样本来训练贝叶斯模型，并且可以保证真实的后验概率会以一定的概率落在估计的后验概率区间内。

2　区间贝叶斯建模方法

本文先研究没有数据错误的情况，并得出一些关于抽样数据区间估计的结论。令有 N 个元组组成的数据集为 D，从 D 中均匀采样（每个元组被采样到的概率是相同的），获得一个样本 S，假设样本 S 中有 K 个元组。下面通过区间估计的相关理论和中心极限定理，给出连续属性均值 $mean(D)$ 的区间估计，以及离散属性某一特定属性值所占比例 m 的区间估计。

首先，本文考虑 D 中的连续属性。令 D 中连续属性的均值为 $mean(D)$，方差为 $var(D)$；令 S 中连续属性的均值为 $mean(S)$，方差为 $var(S)$。根据中心极限定理，样本 S 中连续属性的均值近似服从如下正态分布，即：$N(mean(D), \frac{var(D)}{K})$。因此，通过区间估计的相关理论，可以定义 $mean(D)$ 的置信区间，这个置信区间是关于参数 α 的（95% 置信度表示 $Z_{\frac{\alpha}{2}} = 1.96$）。对此可表示为：

$$mean(S) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{var(S)}{K}}, \tag{1}$$

其次，本文考虑 D 中的离散属性。令 D 中离散属性某一特定属性值所占比例为 m；令 S 中离散属性某一特定属性值所占比例为 m'。由中心极限定

理可知,样本 S 中离散属性某一特定属性值所占比例 m' 近似服从如下正态分布,即:$N(m, \frac{m(1-m)}{K})$。因此,通过区间估计的相关理论,可以定义 m 的置信区间,这个置信区间是关于参数 α 的(95%置信度表示 $Z_{\frac{\alpha}{2}} = 1.96$)。对此可表示为:

$$m' \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{m'(1-m')}{K}}, \tag{2}$$

总之,以上 2 个置信区间(连续属性均值的区间和离散属性某一特定属性值所占比例的区间)可解析为 2 层含义,将其探讨论述如下。

(1)如果重新计算另一个随机样本连续属性的均值或离散属性某一特定属性值所占比例,其结果将以一定的概率落在各自的置信区间内。

(2)真实值 $mean(D)$ 和 m 将以一定的概率落在各自的置信区间内。另外,对 $mean(D)$ 和 m 的估计可以认为是无偏的,因为估计的期望值等于真实的值。

本文假设 D_{clean} 是数据集 D 相对应的干净数据集,并且想要估计 D_{clean} 中属性的分布情况。然而,干净的数据集 D_{clean} 是无法预先得知的,所以直接从 D_{clean} 中采样是不现实的。于是,本文先从含有脏数据的数据集 D 中采样,接着再清洗采集到的样本。鉴于数据错误的类型较多,本文研究中仅涉及了 2 类数据错误:值错误和重复错误。文献[6]中讨论了对这 2 种错误的处理方法。具体研究详见如下。

(1)值错误。是由于数据集中不正确的属性值引起的,这类错误不会影响数据集大小,即 $|D| = |D_{clean}|$。而且,纠正一个值错误只会影响一个元组。因此,如果纠正一个元组中的属性值,仍然可以保留均匀采样的特性。换句话说,某一个元组被采样到的概率不会因为对值错误的纠正而发生改变。

(2)重复错误。在处理上比起值错误的处理就要复杂许多。由于重复错误会影响多个元组,而且清洗了重复错误后的 D_{clean} 大小和 D 的大小不同,因此重复错误会影响均匀采样的特性。另外,重复的数据更有可能被采样到,从而影响对样本中属性分布情况的估计。重复错误的处理要分 2 种情况,即对连续属性重复错误的处理和对离散属性重复错误的处理。假设 D 中只含有重复错误,令 S 是 D 中均匀采样得到的大小为 K 的样本,对于每一个元组 $t_i \in S$,m_i 代表 t_i 在 D 中出现的次数。对于连续属性,其结果受到重复率 d 的影响,重复率为:

$$d = \frac{K}{K'}. \tag{3}$$

其中,$K' = \sum_{i=1}^{n} \frac{1}{m_i}$。

此时,t_i 的正确值就等于 $\frac{d \times t_i}{m_i}$;而对于离散属性,因为其结果不受重复率 d 的影响,所以计算方式相对简单,t_i 的正确值就等于 $\frac{t_i}{m_i}$。

真实的后验概率将以大于等于 0.95^n 的概率落在估计的后验概率区间内,这里的 n 表示有多少个概率区间相乘(在计算连续属性和离散属性的置信区间时,置信度都取 95%)。对其证明过程可阐释为:由贝叶斯定理可知,在计算后验概率时,分母对于所有类别都是相同的,所以只需要计算类条件概率和先验概率的乘积。为了表述方便,令 $P(B)$ 表示后验概率,$P(R_1), P(R_2), \cdots, P(R_n)$ 表示类条件概率和先验概率。现在已知 $P(R_1), P(R_2), \cdots, P(R_n)$ 都是概率区间,且真实的类条件概率或先验概率都有 95%概率落在其各自的区间内。根据贝叶斯定理,则有 $P(B) = P(R_1)P(R_2)\cdots P(R_n)$。由于不在概率区间 $P(R_1), P(R_2), \cdots, P(R_n)$ 内的概率相乘之后也有可能落在最后的后验概率区间 $P(B)$ 内,研究推得最差的情况就是真实的后验概率将以 0.95^n 的概率落在估计的后验概率区间内。

3 实验

本文使用了来自 UCI 机器学习库中的 Adult 数据集,这个数据集中包含了人口普查的信息。Adult 数据集大约有 48 000 行,包含 14 列特征属性,以及 1 列分类属性。在 Adult 数据集的 14 列特征属性中,有 6 列特征属性是连续的,还有 8 列特征属性是离散的。另外,对于分类属性,在本次研究中就是把分类属性当成离散属性来处理。

本文还使用了 OPENDATA KC 上的 Business License Holders(简称 Business)数据集。Business 数据集大约有30 000行,包含 4 列特征属性,以及 1 列分类属性。在 Business 数据集中,所有的列都是离散型的。实验中,IBSM(Interval-based Bayesian Statistical Modeling)对应区间贝叶斯建模方法,Dirty Training Set 表示脏的训练集,也就是直接使用脏数据来训练贝叶斯模型。本文拟从 2 个方面来评价区间贝叶斯建模方法的效果。研究设计过程可剖析详述如下。

首先，本文固定采样比例（10%），然后比较区间贝叶斯建模方法的效果和直接使用脏数据来训练贝叶斯模型的效果。对于 Business 数据集，区间贝叶斯建模方法在不同错误比例下的运行效果如图 1 所示。从图 1 中可以看出，随着错误比例的增加，直接使用脏数据进行贝叶斯统计建模的精度和召回率都有所下降。因此，使用脏数据进行贝叶斯建模会影响最终的分类效果。然而，使用本文提出的区间贝叶斯建模方法，可以发现对精度和召回率都有很明显的提升。同时，区间贝叶斯建模方法会使得精度和召回率趋于一个稳定的值，这个稳定的值不会随着错误率的增加而改变。总之，区间贝叶斯建模方法在不同错误比例下都是健壮的，而且可以得到较为准确的结果，即精度和召回率比使用脏数据的情况有很大的提升。

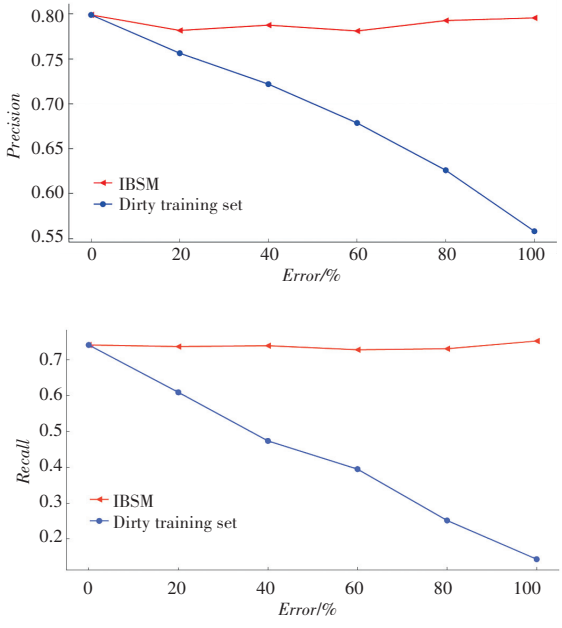

图 1　对于 Business 数据集，区间贝叶斯建模方法在不同错误比例下的效果

Fig. 1　The effect of IBSM with different error ratio for Business dataset

其次，本文固定错误比例（30% 错误比例），然后将区间贝叶斯建模方法、AllDirty、以及 AllClean 在不同采样比例下进行研究对比。

对于 Adult 数据集，区间贝叶斯建模方法在不同采样比例下的运行效果如图 2 所示。从图 2 中可以看出，随着采样比例的增加，区间贝叶斯建模方法、AllDirty、以及 AllClean 的变化趋势都非常小，导致这种现象的原因很有可能是当样本的数量达到一定规模时，统计的概率就会趋于一个稳定的值。另外，区间贝叶斯建模方法在精度和召回率上和 AllClean 的效果非常接近，而且比 AllDirty 的效果要好很多。本文通过对 Adult 数据集的多次实验发现，在 Adult 数据集中只需要清洗大约 600 个元组（600 个元组占整个数据集总元组数的 1.3%），就能在精度和召回率上比 AllDirty 的效果好很多。总之，区间贝叶斯建模方法只需要清洗一小部分的样本，就能在精度和召回率上获得很好的效果。因此，无论数据集中脏数据的比例是很大、还是很小，区间贝叶斯建模方法都具有很强的可行性。

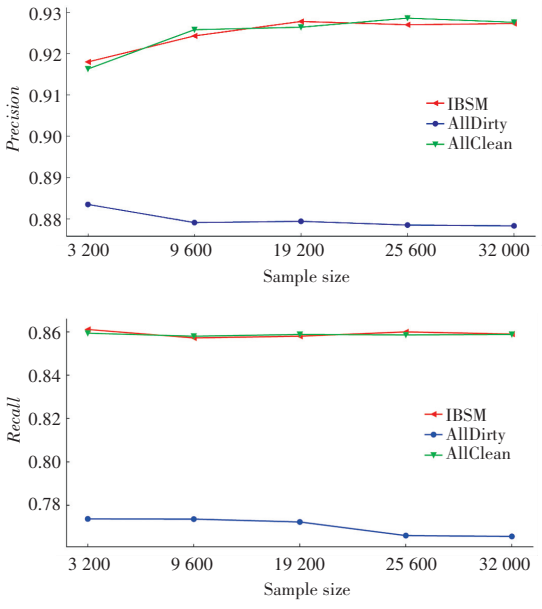

图 2　对于 Adult 数据集，区间贝叶斯建模方法在不同采样比例下的效果

Fig. 2　The effect of IBSM with different sample size for Adult dataset

4　结束语

本文研究了 2 种可能影响贝叶斯统计建模准确度的数据错误，即值错误和重复错误，并提出了区间贝叶斯建模方法来处理这 2 种数据错误。另外，在数据清洗后，本文给出了连续属性均值和离散属性某一特定属性值所占比例的置信区间，通过对置信区间的处理，将置信区间转化为类条件概率区间或先验概率区间。

参考文献

[1] SAHAMI M. Learning limited dependence Bayesian classifiers[C]// KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Portland, Oregon: ACM, 1996: 335-338.

[2] 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.

（下转第 111 页）