文章编号:1007-5321(2017)03-0104-06 DOI:10.13190/j.jbupt.2017.03.015

面向配电网故障数据的 BIC 评估后向选择方法

曾兴东[1,2], 林荣恒[1,2], 邹 华[1], 张 勇[3]

(1. 北京邮电大学 网络与交换技术国家重点实验室, 北京 100876;
2. 中国电子科技集团公司第五十四研究所 通信网信息传输与分发技术重点实验室, 石家庄 050081;
3. 国家电网 上海电力公司, 上海 200122)

摘要: 10 kV 配电网所处环境复杂,引发故障的原因很多,在使用数据挖掘方法对配电网故障进行分析时,太多的特征会对挖掘模型造成负面影响. 为了防止挖掘模型考虑过多无用信息,需首先对数据进行特征选择来实现降维,因此提出了基于贝叶斯信息准则(BIC)的模型评估后向选择算法,对故障因素进行降维. BIC 评估准则能够尽可能地简化模型,降低维度,而后向选择算法可以快速得到最优的简化模型,两者的结合提升了降维的速度,并能够得到更加简化的模型. 实验结果表明,采用基于 BIC 评估的后向选择算法有助于后续模型准确性的提升,可提高训练效率.

关 键 词: 配电网故障分析; 降维; BIC 模型评估; 后向选择算法

中图分类号: TP3 **文献标志码**: A

An BIC Selection Method for Distribution Network Fault Data Feature Dimension Reduction

ZENG Xing-dong[1,2], LIN Rong-heng[1,2], ZOU Hua[1], ZHANG Yong[3]

(1. State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Science and Technology on Information Transmission and Dissemination in Communication Networks Laboratory, The 54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang 050081, China; 3. State Grid Shanghai Municipal Electric Power Company, Shanghai 200122, China)

Abstract: Feature selection is important to improve the model accuracy and reduce overfitting. The 10 kV power distribution network is complex and there are too many features for a data mining model to work. Before modeling power fault data, the dimensionality reduction and model selection is necessary. In order to solve this problem, a Bayesian information criterions (BIC) model selection algorithm along with backward selection algorithm was proposed. BIC aims to reduce the complexity of model and the backward selection can reach fast convergence. Experiments show that the algorithm works well. It is proven that the algorithm proposed here is of advantage to improve model accuracy and data training efficiency.

Key words: distribution network fault data; dimensionality reduction; Bayesian information criterions; backward selection algorithm

10 kV 配电网是配电网中规模最大、涉及面积最广的部分,其所处的环境复杂,故障原因呈现出多样化特性[1]. 刘笑园等[2]指出常见的配电网故障原因,包括树木、鸟类、雷击、车辆碰撞、施工开挖等外

收稿日期: 2016-08-04
基金项目: 国家高技术研究发展计划(863 计划)项目(2015AA050203); 北京市自然科学基金项目(4174099)
作者简介: 曾兴东(1992—), 男, 硕士生, E-mail: zengxdbupt@163.com; 邹 华(1969—), 女, 教授, 硕士生导师.

力破坏;设备产品质量不良、工艺差造成的设备故障. 刘等[2-3]还指出电网故障和天气自然因素密切相关,大雾、阴雨天气容易发生单相接地故障,每年雷雨季节造成的短线事故频发[1],风灾、覆冰等现象也会导致频繁跳闸;张等[3]指出了一些季节性因素.

传统的故障分析需要电网工作人员实地考察分析故障原因,实时性不高;在数据挖掘技术高速发展的今天,自动化的辅助决策方案正在快速渗入各行各业. 笔者尝试使用数据挖掘的手段对配电网故障进行分析. 由于配电网所处环境复杂,其故障原因变量维度较高,很大程度上增加了分类模型的复杂度,旨在提出一种特征选择方案,对高维度的特征进行筛选,优化模型的选择.

1　相关工作

特征选择的目的是为了降低维度,降低维度有很多种方案,分别是特征降维、特征选择、模型选择.

传统的特征降维也称为特征抽取,是将高维度的数据映射到低维的一种手段,例如主成分分析(PCA,principal component analysis)通过转化数据,将多列数据合并成一列,此方法有 2 个方面的缺陷,一方面 PCA 算法得到的新主成分没有实际含义,不能表达因变量和自变量之间的关系[4];另一方面 PCA 不方便处理混合类型数据. 本研究要处理的数据就是数值类型变量和离散类型变量的混合,因此 PCA 方法在此并不实用.

特征选择经过了数十年的发展,已经衍生出了数十种算法. 钟等[4]从评价准则的角度总结了常用的特征选择方法可以分为嵌入法(embedded)、封装法(wrapper)、过滤法(filter). 嵌入法和封装法依赖分类器的分类性能,容易产生复杂的模型导致过拟合,过滤法的提出同时弥补了封装法和嵌入法的不足. 目前使用最为广泛的特征选择评价准则算法包括 Relief-F[5]、卡方检验[6]、互信息[7]、联合互信息亏损等,它们都属于过滤法. 所有的这些方法都在关注数据 2 个方面的质量:实用性(relevance)和冗余性(redundancy)[8]. 实用性要求筛选出来的特征在分类的时候具有较高的准确率,冗余性要求选出来的特征在满足实用性的前提下数量尽量少.

特征选择算法是一个使用很普遍的方法,但是在确定模型的情况下,使用模型选择算法进行降维,模型将会得到更好的结果. Raftery[9-10]将模型选择算法应用在特征的选择和模型参数选择上,并取得了良好的结果. 模型选择算法是在基于已知模型的基础上通过调整模型参数以及特征维度来得到一个最优的模型. 算法关键在于模型的评价方法,如果以模型的准确率作为模型的评价算法,则模型选择等效于特征选择中的 wrapper 算法,此评价准则对分类器的性能的依赖很高,会产生过拟合. 为了同时满足实用性和冗余性,基于信息熵的信息准则被提出来,最早提出的是赤池信息量准则(AIC,Akaike information criterion)[11],AIC 准则假设存在一个最优模型,通过计算选出的模型和最优模型之间的信息熵来评价模型的好坏,AIC 越小模型越接近最优. 该准则将特征维度作为惩罚量,从而实现冗余性的要求. 贝叶斯信息准则(BIC,Bayesian information criterions)基于 AIC 准则,并在 AIC 基础上加大了惩罚力度[12].

AIC 和 BIC 准则只是提出了模型的评价准则,进行降维还需要结合其他过程. Dash 等[13]指出特征选择的 4 个过程同样适用于模型选择,分别是:子集生成、评估函数、迭代退出、模型验证. 子集生成可以有多种方法,包括完全搜索、启发式搜索,随机搜索. 启发式搜索适合于中等数量集的数据[8],而且能快速收敛,适合在本实验中用于子集生成.

2　问题描述与分析

10 kV 配电网故障分析一直是一个热门话题,表 1 所示为某地配电网络故障数据的一部分,表中各列代表的含义如表 2 所示. 本研究的研究对象(响应变量)是故障原因(之后用 Y 代替),可以看到影响故障发生的原因很多,配电网故障的发生直接或者间接地与多种因素相关,其中气候原因包括温度(T)、湿度(U)、风速(F)、降雨量(R),时间因素包括月份(MON)、早晚时间(H),环境因素包括故障发生所在地区(L1)、故障发生时设备(L2)、线路架设类型(L3),通常分析故障原因还会将故障的表现也当作因变量考虑在内,包括故障持续时间(V1)、故障保护动作(V2)、故障大类(V3)等. 在具体分析原因的时候会发现,特征的维度太高会导致分析的时候没有把握主要原因,使得模型复杂度太高.

表1 配电网故障数据部分内容

Date	T/℃	U/%	F/m·s^{-1}	R/mm	L1	L2	L3	V1	V2	V3	H	MON	Y
2015/1/1	-1.4	18	4	0	XX	跳闸	架空线路	瞬时性	重合成功	运行维护	1	1	鸟害
2015/1/1	-3.4	20	2	0	XX	接地	架空线路	永久性	未跳闸	外力	4	1	异物
2015/1/1	-7.2	34	1	0	XX	跳闸	架空线路	永久性	未跳闸	用户影响	8	1	外力

表2 配电网故障数据特征及其含义

	Date	T	U	F	R	L1	L2	L3	V1	V2	V3	H	MON	Y
含义	时间	温度	湿度	风速	降雨量	地区	归属	线路	性质	动作	大类	小时	月份	原因
类型	时间	数值	数值	数值	数值	类别	类别	类别	类别	类别	类别	数值	数值	类别

图1所示为使用表1所示配电网故障数据,利用贝叶斯分类器作为分类模型,依次将各个特征添加到模型中来增加模型复杂度,得到的分类预测错误率随模型复杂程度增加而变化的曲线. 整个训练的过程中保证训练数据集和测试集不变,从曲线左端开始,第1个数据点为使用温度(T)作为特征训练朴素贝叶斯模型,得到的错误率超过了50%,第2次在温度的基础上增加了湿度(U)特征,得到的模型准确率有小幅的提升.

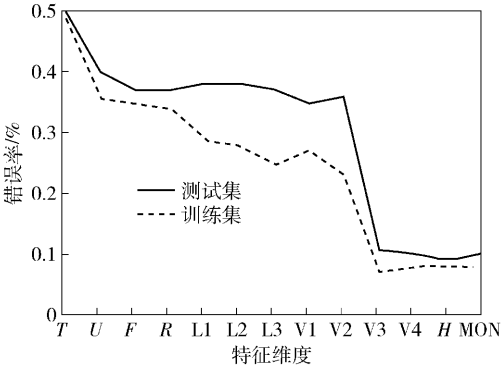

图1 特征维度变化对模型错误率的影响

依此类推. 在整体的错误率变化趋势图中,可以看到部分特征具有决定性因素,例如故障大类 *V3* 显然能够大幅提高对故障原因的预测准确率. 然而其他特征对模型的贡献并不是那么明显,很难人为地筛选出构建模型的特征.

传统的专家系统筛选特征需要结合一些配电网故障分析相关的知识,然而学习电网故障知识以及模型的知识需要时间成本,而且最后得出的模型也不一定是性能优越的. 此时需要使用自动化的方法,结合已有数据进行特征的选择.

3 基于BIC模型评估的后向选择算法

3.1 特征选择策略

如何选择合适的特征,李等[14]提到:在所有可能选择的模型中,能够很好地解释已知数据并且十分简单才是最好的模型. 也即是要选择一个好的模型需要兼顾两方面的要求:

1) 对数据的拟合效果要好;

2) 模型的复杂程度不能太高.

通常来说这两个方面的性能是对立的,拟合程度高的模型会有较为复杂的模型[13],然而较为复杂的模型会产生过拟合的情况,因此要在这两个方面的性能上找一个平衡. 尝试使用基于BIC评价准则,并结合启发式搜索中的后向选择算法来搜索最佳特征组合,从而实现数据降维,期待获得一个能够同时满足实用性和冗余性的结果. 还将使用常用的特征选择方法来解决降维的问题,并和BIC模型选择算法进行对比,分析其中的优势与劣势.

3.2 基于BIC模型评估的特征选择原理

将故障原因 $\boldsymbol{Y}_n$ 抽象为 n 个相互独立的响应变量为

$$\boldsymbol{Y}_n = (y_1, y_2, \cdots, y_i, \cdots y_n)^{\mathrm{T}} \tag{1}$$

其中:n 为样本数据的个数,y_i 表示第 i 个样本的故障原因,本文筛选了4种最常见的故障类型,分别为“雷击”“鸟害”“绝缘子”“施工影响”.

令解释变量为

$$\boldsymbol{X}_n = (x_1, x_2, \cdots, x_i, \cdots x_n)^{\mathrm{T}} \tag{2}$$

$\boldsymbol{X}_n$ 为一个 $n \times p$ 的矩阵,该矩阵的第 i 行 x_i 是和 y_i 对应的 p 维变量,p 代表 p 种特征,这些特征来自表1中的13种特征,n 为样本个数.

使用AIC方式来评价模型好坏,AIC评价指标用 Q_{AIC} 表示,其公式可以表示为

$$Q_{\mathrm{AIC}} = 2p - 2\ln\phi \tag{3}$$

其中:p 为模型中使用的特征维度个数,为惩罚量,$\boldsymbol{\Phi}$ 是模型的似然函数. 不同的模型 M 计算似然函数的方法不一样,后面统一使用式(4)来表示:

$$\phi = \mathrm{ML}(M, \boldsymbol{Y}_n, \boldsymbol{X}_n) \tag{4}$$

AIC越小,模型的质量越好. 于是特征的选择

问题成为求 AIC 最小的最优化问题，可以表达为

$$\min_{0\leqslant p\leqslant P}(Q_{\mathrm{AIC}})=\min_{0\leqslant p\leqslant P}(2p-2\ln\phi) \tag{5}$$

最优化问题的求解有多种方法，常规的方法为牛顿法、线性规划法等依赖求导或者数学公式的方法. 但是由于故障数据有大量分类变量，只能使用相关工作中提到的 4 种子集生成方法. 在 4 种生成方法里面，最合适的是启发式搜索，而启发式搜索分为前向和后向 2 种搜索方法. 特征搜索的过程是依次减少特征数量，因此该方法属于后向选择. 后向选择类似最速下降法，它的基本思路是每一轮找出使得 AIC 减少最多的那个特征，将其从模型中剔除，搜索的停止条件为 AIC 不再减少. 当 AIC 最小时，此时的特征组成的模型为最优模型.

使用 AIC 进行降维时，对高维度数据的惩罚不够大，为了达到减少特征的目的，加大对高纬模型的惩罚力度，使用 AIC 的优化公式 BIC 为

$$Q_{\mathrm{BIC}}=2p\ln N-2\ln\phi \tag{6}$$

其中 $2p\ln N$ 是惩罚量. 同理得到使用 BIC 求解最优特征组合问题可以简化为

$$\operatorname*{Min}_{0\leqslant p\leqslant P}(Q_{\mathrm{BIC}})=\min_{0\leqslant p\leqslant P}(2p\ln N-2\ln\phi) \tag{7}$$

其中 N 为样本个数.

3.3　基于启发式后向搜索算法的特征选择

3.2 定义了模型的评估准则，在将评估准则应用到降维的过程中，需要在全集中使用特征选择过程，这里使用的是后向选择算法.

算法 1　BIC_后向选择降维算法

```
1   Φ = ML(M, X_n, Y_n) //最大似然
2   len = dim(X_n) //len 为特征维度个数
3   Q_BIC = 2ln(N)len - 2ln(Φ) //求解初始 BIC
4   for i = len to 1 //外循环，每次循环降低一个维度
5       pos = -1; //记录满足要求的特征位置
6       Q_BIC_min = Q_BIC //记录最小 BIC 值
7       p = dim(X_n) //当前数据维度
8       for j = p to 1 //内循环寻找需要去掉的特征
9           Φ = ML(M, X_n[, -j], Y_n)
//X_n[, -j]表示剔除 X_n第 j 维列向量
10          Q_BIC_curr = 2ln(N)len - 2ln(Φ)
11          if Q_BIC_curr < Q_BIC_min
//求最小 BIC 并记录对应的维度
12              Q_BIC_min = Q_BIC_curr
13              pos = j
14      if Q_BIC_min = = Q_BIC //BIC 不再减小
15          break; //退出启发搜索
16      else //得到新的 BIC 并降维
17          Q_BIC = Q_BIC_min
18          X_n = X_n[, -pos]
19  Return X_n //返回经过特征选择的数据
```

如伪代码所示，此处的模型选择算法以 BIC 为模型评价指标，借鉴了启发式搜索中的后向选择(Backwards)算法的思想，逐步筛选特征. 后向选择算法的初始状态是将所有的特征加入集合中，如第 1 ~ 5 行所示为使用所有特征来初始化 BIC 值. 在后续的操作中，依次从集合里排除掉多余的特征，第 10 ~ 15 行为筛选出当前集合中的多余特征，第 16 ~ 20 行为判断当前集合是否最优，并删除多余特征. 后向选择算法相比完全搜索能够更加快速地收敛；相比前向搜索和随机搜索，后向搜索方向与降维的过程一致，因此这里选择后向搜索算法更加有效.

将 BIC 评估标准作为后向选择的对比标准，相比其他准则如 AIC，评估得到的模型将更加精简，这是 BIC 的定义式(6)所导致的. BIC 评估准则的使用是为了平衡模型准确率和特征集合的冗余性. 如果单纯使用准确率会产生过于复杂的模型，相比之下 AIC、BIC 准则符合降维的要求[12].

4　实验分析与验证

4.1　配电网故障数据特征选择结果

使用上述基于 BIC 准则的后向特征选择算法，得到如表 3 所示的结果. 表格中第 $i(12>i>1)$ 行表示第 i 次特征筛选，第 j 列表示第 j 个特征. 每次迭代都会从特征集合中剔除一个特征，剔除特征后的新的 BIC 值都会减少，表明筛选后的特征集合更加接近理想的模型. 当减少集合中的特征，BIC 值不再减少，则特征筛选终止，得到的最终特征为 L1、V2、V3、MON.

筛选出来的特征 L1、V2、V3、MON 在实际分析中具有实际的意义. 本实验的响应变量 y_i 有 4 种类别，分别为“雷击”“鸟害”“绝缘子”“施工影响”. 经过统计，雷击和鸟害的季节性因素 MON 的关系非常明显，鸟害的时间大部分发生在 3 ~ 5 月，发生的季节多为春季；雷击引发的故障大部分集中在 5、6、4、8、9 月，其中 7 ~ 8 月雷击引发故障的尤其多. 不同的故障其保护动作(V2)不一样. 原因大类 V3 与因变量的关系更是密切，例如“雷击”和“施工影响”这两种故障小类就分别对应不同的原因大类.

BIC虽然可以综合评价一个模型的好坏,但是大多数时候,人们更加关注实用性,用模型预测的准确性来评价模型的实用性. 为了验证模型选择算法的实用性,本实验使用5折交叉验证法对特征筛选前后模型的准确率进行验证,将数据随机抽样均分成5份,其中4份数据用于分类器模型训练,剩下的一份用于预测,得到如表4所示的数据.

可以看到通过模型选择得到的新特征构成的模型,平均准确率有明显的提高,此结果说明提供的特征筛选方案是可行的.

4.2 冗余性与实用性对比

仍然从实用性和冗余性2个角度分析降维的效果. 本节使用常用的特征选择算法以及模型选择算法进行对比,包括:Relief-F、卡方过滤器(Chi-squared filter)、信息增益(information gain)、一致性过滤器(consistency-based filter)、Boruta、基于贝叶斯分类器准确率的评估方法、AIC、BIC. 其中前4个评估方法属于特征选择算法中的过滤法,Boruta[15]和基于贝叶斯分类器准确率的评估方法属于特征选择算法中的封装法,而AIC和BIC属于模型选择算法. 最后加上不做任何降维处理时的情况.

4.2.1 冗余性

表5所示为故障数据分别使用上述各个算法进行降维之后的维度数,从中可以得到如下结论.

表3 基于BIC评估的后向选择算法过程

序号	T	U	F	R	L1	L2	L3	V1	V2	V3	H	MON
1	104.17	108.41	105.82	106.86	107.86	106.97	109.58	105.53	206.60	104.85	110.30	117.88
2	—	100.12	97.67	97.29	99.28	102.96	100.96	95.40	232.19	97.81	98.44	113.88
3	—	90.40	89.50	89.27	90.22	93.20	93.39	—	227.63	90.92	89.24	110.95
4	—	86.10	83.64	81.96	85.18	87.74	87.88	—	239.80	85.55	—	103.18
5	—	78.39	76.00	—	78.18	78.86	82.42	—	240.80	78.79	—	96.04
6	—	72.01	—	—	73.62	72.32	77.11	—	237.01	74.94	—	89.58
7	—	—	—	—	70.82	67.00	71.14	—	261.06	71.89	—	86.30
8	—	—	—	—	69.79	—	64.83	—	277.50	72.01	—	91.97
9	—	—	—	—	72.48	—	—	—	305.61	69.03	—	93.74

表4 基于BIC评估的降维算法对准确率的作用 %

算法	训练集1	训练集2	训练集3	训练集4	平均值
基于BIC	87.34	91.34	92.41	93.67	91.14
使用所有特征	88.61	91.14	88.61	82.28	87.66

1)基于封装法的特征选择方法(Boruta和基于贝叶斯分类器准确率)降维得到的维度高,因为该方法基于分类器准确率,而复杂的模型在数据集上往往具有较高的准确率,复杂的模型维度也比较高.

2)基于过滤法的特征选择方法都取得了较好的冗余性优化.

3)模型选择算法AIC和BIC选出的模型冗余性也比较低,而且BIC优于AIC以及其他特征选择方法.

4.2.2 实用性

为了比较各种降维算法在电网故障数据降维中的优劣,通过5折交叉验证来对结果进行对比,并多次重复实验用平均值. 表6为使用不同的降维方法得到的特征,然后分别使用3种不同的分类器进行分类,使用5折交叉验证得到预测准确率的平均值. 从表中可以得到如下结论:

1)使用BIC方法得到的特征的模型准确率较高,而且在多个算法的预测中都取得了较为平稳的值.

2)基于过滤器的特征选择算法得到的结果也有较高的准确率和稳定性.

3)基于封装器的特征选择算法的准确率只在特定的分类器中效果较好.

4)BIC算法的准确率和稳定性优于AIC.

5)所有的降维方法都比不做任何降维处理的情况要好.

4.3 小结

BIC模型评估准则的后向选择算法在兼顾了模

型准确率(实用性)的同时,加入了冗余度的惩罚因子,避免了出现过拟合的情况,同时在已有模型的基础上降低了特征的维度,减少了模型的复杂度. 使用启发式的后向选择算法有效地缩短了搜索时的收敛时间. 而且提出的方法也具有普适性,适用于混合型数据的降维. 相比于主流的降维算法,BIC 模型评估准则的后向选择算法在保证准确率接近的情况下,具有更加优良的冗余性优化.

表 5　各评估方法冗余性

Relief-F	卡方过滤器	信息增益	一致性过滤器	Boruta	贝叶斯	AIC	BIC
5	5	5	7	9	10	6	4

表 6　不同算法的模型准确率　%

算法	Relief-F	卡方过滤器	信息增益	一致性过滤器	Boruta	贝叶斯	AIC	BIC	全特征
贝叶斯	88.51	89.72	91.06	89.91	88.95	89.51	88.76	90.79	88.3
SVM	90.42	90.29	91.38	91.47	86.97	86.36	90.59	91.45	84.4
决策树	89.01	88.92	89.78	89.30	88.92	88.85	88.90	89.26	88.2

5　结束语

根据上述实验,可以得到如下结论:在分析配电网故障数据时,降维能够降低模型的复杂度,并提高模型的准确率;基于 *BIC* 的模型选择算法在数据降维的实用性和冗余性上和目前主流的特征选择算法接近,甚至在降低冗余性上优于特征选择算法. 使用 *BIC* 模型选择算法降维的优势还在于,在已知模型的前提下,通过使用 *BIC* 算法不仅降低了特征维度,还能确定了模型以及模型参数,具有一定的优势.

参考文献:

[1] 彭向阳, 詹清华, 周华敏. 广东电网同塔多回线路雷击跳闸影响因素及故障分析 [J]. 电网技术, 2012, 36(3): 81-87.

[2] 刘笑园, 岑梁, 王震宇, 等. 配电网故障分析及故障指示器的现场应用 [J]. 上海电力学院学报, 2015, 31(B05): 12-14.

[3] 张勇, 陈凤. 长沙地区配电运行事故统计分析[J]. 中国电力, 2007, 40(5): 28-30.

[4] 钟意伟, 赵杰煜, 朱绍军. 基于贝叶斯和谐度的特征选择算法 [J]. 计算机工程与应用, 2015, 51(23): 125-130.

[5] Robnik M, Kononenko I. Theoretical and empirical analysis of ReliefF and Relief-F [J]. Machine Learning, 2003, 53 (1/2): 23-69.

[6] Elham C, Taheri M, Katebi S, et al. An improved fuzzy feature clustering and selection based on Chi-Squared-Test [J]. Proceedings of the International MultiConference of Engineers and Computer Scientists 2009. Hong Kong: IEEE, 2009: 18-20.

[7] Kumar G. Mutual information based feature selection techniques for intrusion detection [J]. International Advanced Research Journal in Science, Engineering and Technology, 2014, 1(2): 70-75.

[8] Qu Guangzhi, Hariri S, Yousif M. A new dependency and correlation analysis for features [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(9): 1199-1207.

[9] Raftery A E, Dean N. Variable selection for model-based clustering [J]. Journal of the American Statistical Association, 2006, 101(473): 168-178.

[10] Dean N, Raftery A E. Latent class analysis variable selection [J]. Annal Institute of Statistical Mathematics, 2010, 62: 11-35.

[11] Akaike H. A new look at the statistical model identification [J]. IEEE Transactions on Automatic Control, 1974, 19(6): 716-723.

[12] Burnham K P, Anderson D R. Multimodel inference understanding AIC and BIC in model selection [J]. Sociological Methods & Research, 2004, 33(2): 261-304.

[13] Dash M, Liu H. Feature selection for classification [J]. Intelligent Data Analysis, 1997, 1(1-4): 131-156.

[14] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012: 11-14.

[15] Kursa M B, Rudnicki W. Feature selection with the Boruta package [J]. Journal of Statistical Software, 2010, 36(11): 1-13.