

doi: 10.11830/ISSN.1000-5013.201701020



# 采用机器学习的聚类模型 特征选择方法比较

赵 玮

(北京联合大学 应用科技学院, 北京 100101)

**摘要:** 针对机器学习聚类模型在特征选择时存在的问题,首先,对特征选择在聚类模型中的适用性进行分析并对其进行调整和改进.然后,基于 R 语言中的递归特征消除(RFE)特征选择方法和 Boruta 特征选择方法进行特征选择算法设计.最后,应用聚类内部有效性指标,对在线品牌忠诚度聚类模型优化结果进行分析,进而对特征选择方法进行比较研究.结果表明:Boruta 特征选择方法更具优势.

**关键词:** 特征选择; 聚类模型; 机器学习; 递归特征消除算法; Boruta 方法

**中图分类号:** TP 181      **文献标志码:** A      **文章编号:** 1000-5013(2017)01-0105-04

## Comparison of Feature Selection Method of Clustering Model Using Machine Learning

ZHAO Wei

(College of Applied Science and Technology, Beijing Union University, Beijing 100101, China)

**Abstract:** Targeting at problems during the feature selection process of machine learning clustering model, at first, it makes analysis on the applicability of the feature selection for clustering model and makes adjustment and improvement. Then makes feature selection algorithm design based on R language recursive feature elimination (RFE) feature selection method and Boruta feature selection method. At last, applying cluster interior validity indexes to analyze the optimization result of online brand loyalty clustering model, a further comparative study is made on the feature selection method. The results show that the Boruta feature selection method has more advantages.

**Keywords:** feature selection; clustering model; machine learning; recursive feature elimination algorithm; Boruta method

特征选择对基于机器学习的模型构建和优化有着重要意义. 优质特征的选择和构建可以选择有限合理的特征,剔除不相关的特征,让模型得到数据集中良好的结构,使模型运算速度更快,模型结果更易理解,模型更易维护. 因此,使模型发挥优良效果,需要通过特征选择的方法对模型进行优化. 然而,基于机器学习的聚类模型进行特征选择存在以下 2 个问题. 1) 特征选择方法通常是针对机器学习中的监督学习模型,需要具备目标特征变量,而聚类模型属于无监督学习<sup>[1-2]</sup>,没有目标特征变量. 2) R 语言作为机器学习比较好的实践环境,提供了相应的特征选择包,但需要做比较分析. 本文对特征选择在聚类模型中的适用性进行分析,应用 R 语言进行特征选择算法设计.

**收稿日期:** 2016-11-25

**通信作者:** 赵玮(1981-),女,讲师,博士,主要从事数据挖掘与数据分析、电子商务的研究. E-mail: yykjtzhaowei@buu.edu.cn.

**基金项目:** 北京市教委科研计划项目(KM201511417010)

# 1 理论基础

## 1.1 特征选择适用性分析

特征选择要考虑聚类模型的适用性问题. 聚类属于无监督学习<sup>[3]</sup>, 只有聚类特征变量, 没有目标特征变量. 因此, 从在线消费数据集<sup>[4]</sup>中, 以各品牌作为目标特征变量, 研究聚类特征变量集对目标特征变量的影响, 并根据影响程度的大小, 决定选择哪些特征作为衡量品牌忠诚度的聚类特征.

## 1.2 R 语言特征选择方法分析

以 R 语言作为机器学习特征选择和模型优化的实践环境. 该环境中递归特征消除(RFE)特征选择方法<sup>[5]</sup>和 Boruta 特征选择方法具备特征选择基本方法的功能. 再结合启发式搜索中的序列后向选择对特征进行选择, 即从特征全集开始, 每次减少一个特征用于模型构建, 从而选择最优.

# 2 R 语言特征选择算法设计

## 2.1 基于 RFE 特征选择方法的算法设计

算法以聚类特征变量集和目标特征变量作为算法输入参数, 具体实现的功能是评价变量的重要性和模型的精确性. 算法核心程序如下所示.

```
f_select_feature_accuracy<-function(x_set,y_set)
{
  x_scale_filter <<- sbf(x_set,y_set,sbfControl=sbfControl(functions=rfSBF,verbose=F,
method='cv'))
  x_set<-x_set[x_scale_filter$optVariables]
  x_scale_profile <<- rfe(x_set,y_set,sizes =c(1,2,3,4,5,6,7,8,9,10,11,12),rfeControl=
rfeControl(functions=rfFuncs,method='cv'))
  plot(x_scale_profile,type=c('o','g'))
}
```

由算法程序可知: 算法核心是针对聚类特征变量集  $x$  和目标特征变量  $y$  实施过滤方法, 并在此基础上确定变量对模型精确性的影响程度. 其中, RFE 方法应用 `rfe()` 函数执行特征选择, 其程序要点如下: `rfe(x,y,sizes=1:MAX,rfeControl=control)`.

特征选择就是根据聚类特征变量集  $x$  对目标特征  $y$  的影响程度进行评价, 从而对  $x$  中的特征重要性进行排序. 其中, Size 是特征个数; `rfeControl` 用来指定特征选择算法的细节.

## 2.2 基于 Boruta 特征选择方法的算法设计

根据特征选择原理和思路, 算法包括两个子算法, 具体实现的功能主要是: 对于子算法 `f_select_feature_Boruta`, 基于随机森林方法思想, 通过循环迭代的方式对各特征的重要性进行评价, 并对结果进行可视化呈现. 对于子算法 `f_select_feature_Boruta_final`, 将特征重要性未被确定的暂定项进行最终判别. 算法核心程序如下所示.

```
f_select_feature_Boruta<-function(x_set,y_set,mr,pv=0.01,mc=TRUE,do=2,hh=TRUE,
gi=getImpRfZ)
{
  ...boruta_total_scale<<- Boruta(x_set,y_set,maxRuns=mr,pValue=pv,mcAdj=mc,do-
Trace=do,holdHistory=hh,getImp=gi)...
  plot(boruta_total_scale,xlab = " ",xaxt = "n",main="Boruta 算法特征重要性")...
  f_select_feature_Boruta_final<-function(boruat_result)
  {
    ...final_boruta_total_scale<- TentativeRoughFix(boruat_result) ...
    plot(final_boruta_total_scale,xlab = " ",xaxt = "n",main="Boruta 算法特征重要性确认")...
    getSelectedAttributes(final_boruta_total_scale,withTentative = F)
    result_df_boruta_total_scale <- attStats(final_boruta_total_scale) ...
  }
}
```

由算法程序可知: 算法核心是对于子算法 `f_select_feature_Boruta` 来说, 主要通过 `Boruta()` 对各特征的重要性进行评价; 对于子算法 `f_select_feature_Boruta_final` 来说, 主要通过 `TentativeRoughFix()` 方法对特征重要性进行最终判别. 其中, Boruta 方法应用 `Boruta()` 函数执行特征选择, 即 `Boruta(x,y,`

pValue=0.01,mcAdj=TRUE,maxRuns=100,doTrace=0,holdHistory=TRUE,getImp=getImpRfZ,...). 特征选择就是根据评价特征集合  $x$  对目标特征  $y$  的影响程度进行评价, $x$  中的特征对  $y$  的影响越显著,则  $x$  中的这个特征越重要.

### 3 实验及结果分析

#### 3.1 实验数据

以在线消费者特征集作为聚类特征变量集  $x$ ,通过特征选择方法选择出有效特征,实现在线品牌忠诚度的度量.其中,聚类特征变量集包含 x\_sum\_count,sum\_count,c\_rate,x\_sum\_price,sum\_price,p\_rate,x\_avg\_score,x\_sum\_score,max\_commentdate,max\_buydate,max\_usrtypenum,max\_usrlocnum.目标特征变量  $y$  为 L4(各品牌名称).

用以上数据进行特征选择和模型优化后,应用内部有效性指标针对模型优化结果对特征选择效果进行检验和分析.内部有效性指标包括 SSE,SSB,IntraDPS 和 InterDPS<sup>[6-7]</sup>.衡量的含义是 SSE 越小,SSB 越大,IntraDPS 越小和 InterDPS 越大,模型有效性越好,特征选择的效果也越好<sup>[8-10]</sup>.未进行特征选择的初始模型内部有效性指标 SSE,SSB,IntraDPS,InterDPS 分别为 225 429.700 0,101 762.300 0,0.689 0,0.311 0.

#### 3.2 应用 RFE 特征选择算法的实验分析

应用 RFE 特征选择算法之前,将聚类特征变量集  $x$  和目标特征变量  $y$  作为参数,执行函数 sbf,实施过滤方法.其结果作为 rfe 函数的参数,实施封装方法,实现变量重要性的评价及对模型精确性的判断,其结果如图 1 所示.由图 1 可知:通过递归特征消除算法,在数据集的 9 个特征中选取 4 个特征作为重要特征,分别是 x\_sum\_price,sum\_price,max\_buydate,x\_sum\_count.当重要特征数量达到 4 个时,模型的精确性最高,如图 2 所示.

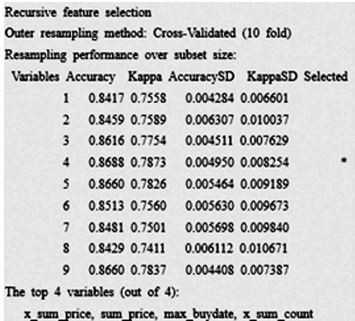


图 1 模型精确性结果

Fig. 1 Model accuracy result

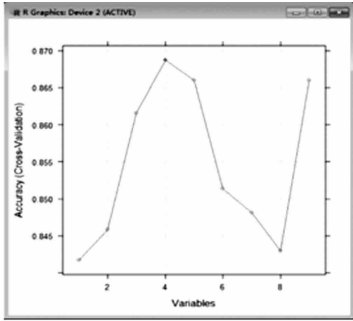


图 2 模型精确性可视化结果

Fig. 2 Model accuracy visualized result

基于 4 个重要特征的模型优化的结果,需要结合初始模型和 RFE 特征选择优化模型内部有效性检验的结果,进行比较分析,如表 1 所示.由表 1 可知:在 4 个指标中,RFE 特征选择优化模型在 SSE,InterDPS,InterDPS 这 3 个指标上优于初始模型,表明经过 RFE 特征选择的模型,其有效性得到提升.

表 1 初始模型和 RFE 特征选择优化模型内部有效性指标

Tab. 1 Internal validity index of initial model and RFE feature selection optimization model

模型	SSE	SSB	IntraDPS	InterDPS
初始模型	225 429.70	101 762.30	0.689 0	0.311 0
RFE 特征选择优化模型	58 155.99	50 908.01	0.533 2	0.466 8

#### 3.3 应用 Boruta 特征选择算法的实验分析

将聚类特征变量集  $x$ 、目标特征变量  $y$  和 Boruta 方法参数要素作为参数,执行函数 f\_select\_feature\_Boruta,没有特征被确定为重要特征和非重要特征,12 个特征均被确定为暂定项.

由于存在暂定项,需要以 f\_select\_feature\_Boruta 算法的执行结果集作为参数,执行 f\_select\_feature\_Boruta\_final 算法中的 TentativeRoughFix 函数以判别暂定项的重要性.经过进一步的判定,12 个暂定项最终被归为重要特征.对于特征重要性确认结果,还可以通过 getSelectedAttributes 和 attStats

函数确认特征列表和创建一个数据框架进行描述.

对数据框架信息中的各项重要性指标进行分析. 按照 medianImp 指标, 可得指标结果及特征重要性的排序: 38.404 31(x\_sum\_count)>26.929 42(sum\_count)>26.783 07(x\_avg\_score)>25.623 44(c\_rate)>25.094 46(x\_sum\_price)>24.051 46(x\_sum\_score)>23.583 09(p\_rate)>19.840 42(sum\_price)>15.650 59(max\_buydate)>14.739 77(max\_commentdate)>12.200 94(max\_usrlocnum)>11.395 70(max\_usrtypenum).

根据 Boruta 特征选择算法的特征重要性排序结果, 结合序列后向选择, 从聚类特征变量集的全集开始, 每次减少一个进行模型构建. 针对这些聚类特征变量集依次进行模型构建, 分别命名为 Boruta 优化模型 1~10, 并从这些模型中选取最优. 应用内部有效性检验指标分别对 Boruta 优化模型 1~10 进行检验. 由检验结果可知: Boruta 优化模型 10 各项指标明显优于初始模型及 Boruta 优化模型 1~9.

3.4 RFE 和 Boruta 特征选择方法的比较

RFE 方法属于最小优化方法, 依赖于特征的子集, 其优势是最大限度地减少了随机森林模型的误差, 而劣势是会丢失一些相关的特征. Boruta 方法依赖全体特征, 其特点是找到所有的特征, 无论其与决策变量的相关性强弱与否. 与传统的 RFE 特征选择算法相比, Boruta 方法的优势是能够返回变量重要性的更好结果, 解释性更强. 因此, 从方法上看, Boruta 方法的特征选择更加有效. 应用 RFE 和 Boruta 特征选择方法进行模型优化的内部有效性评价指标, 如表 2 所示. 由表 2 可知: 采用 Boruta 方法选择的特征用于模型优化时, 模型的聚类效果和有效性最好.

表 2 RFE 和 Boruta 特征选择方法进行模型优化的内部有效性评价指标  
Tab. 2 Internal validity index of RFE and Boruta feature selection optimization model

模型	SSE	SSB	IntraDPS	InterDPS
初始模型	225 429.70	101 762.30	0.689 0	0.311 0
RFE 特征选择优化模型	58 155.99	50 908.01	0.533 2	0.466 8
Boruta 优化模型 10	14 160.71	40 371.29	0.259 7	0.740 3

4 结束语

解决了聚类模型特征选择存在的问题, 设计了基于 R 语言中的 RFE 特征选择方法和 Boruta 特征选择方法实现特征选择的算法. 在此基础上, 应用聚类内部有效性指标, 对在线品牌忠诚度聚类模型优化结果进行分析, 并对 RFE 特征选择方法和 Boruta 特征选择方法进行比较. 由此可知: Boruta 特征选择方法对在线品牌忠诚度聚类模型优化更加适用和有效.

参考文献:

[1] 钱彦江. 大规模数据聚类技术研究 with 实现[D]. 成都: 电子科技大学, 2009: 4-5.  
[2] 汪永旗, 王惠娇. 旅游大数据的 MapReduce 客户细分应用[J]. 华侨大学学报(自然科学版), 2015, 36(3): 292-296.  
[3] 方匡南. 基于数据挖掘的分类和聚类算法研究及 R 语言实现[D]. 广州: 暨南大学, 2007: 78-84.  
[4] 刘蓉, 陈晓红. 基于数据挖掘的移动通信客户消费行为分析[J]. 计算机应用与软件, 2006, 23(2): 60-62.  
[5] 卢运梅. SVM-RFE 算法在数据分析中的应用[D]. 长春: 吉林大学, 2009: 16-28.  
[6] 王曰芬, 章成志, 张蓓蓓, 等. 数据清洗研究综述[J]. 情报分析与研究, 2007(12): 50-56.  
[7] 周开乐, 杨善林, 丁帅, 等. 聚类有效性研究综述[J]. 系统工程理论与实践, 2014, 34(9): 2417-2431.  
[8] 胡勇. 聚类分析结果评价方法研究[D]. 包头: 内蒙古科技大学, 2014: 62-63.  
[9] KANUNGO T, MOUNT D M. A local search approximation algorithm for *k*-means clustering[J]. Computational Geometry, 2004, 28(2/3): 89-112.  
[10] ELKAN C. Using the triangle inequality to accelerate *k*-means[C]// Proceedings of the Twentieth International Conference on Machine Learning. Menlo Park: AAAI Press, 2003: 147-153.

(责任编辑: 钱筠 英文审校: 吴逢铁)