



文章编号:1001-9081(2018)09-2507-04

DOI:10.11772/j.issn.1001-9081.2018020460

基于析因设计的大数据相关关系挖掘算法

唐小川*, 罗亮

(电子科技大学 计算机科学与工程学院, 成都 611731)

(*通信作者电子邮箱 xiaochuantang@std.uestc.edu.cn)

摘要:针对高维大数据的降维问题,提出了一种基于统计学析因设计的特征选择算法——FFD。首先,使用析因设计的因子效应作为过滤式特征选择算法中特征与目标变量之间相关关系的度量标准;其次,提出一个分治算法用于搜索适合于输入数据集的最优析因设计;再次,为了解决传统实验设计需要人工执行实验的问题,提出一种数据驱动的方法从输入数据集中自动搜索析因设计的响应值;最后,根据设计矩阵和平均响应值计算因子效应,并使用因子效应对特征和交互作用进行排序,得到显著的特征和交互作用。实验结果表明,FFD 的平均分类错误率比互信息最大化算法(MIM)降低了 2.95 个百分点,比联合互信息最大化算法(JMIM)降低了 3.33 个百分点,比 ReliefF 算法降低了 6.62 个百分点。因此,FFD 在实际数据集中能有效挖掘与目标变量相关的特征和交互作用。

关键词:大数据;相关关系;特征选择;交互作用;析因设计

中图分类号: TP181 文献标志码:A

Big data correlation mining algorithm based on factorial design

TANG Xiaochuan*, LUO Liang

(School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu Sichuan 611731, China)

Abstract: Focused on the issue of dimensionality reduction in high-dimensional big data, a feature selection algorithm based on statistical factorial design was proposed, which was named Full Factorial Design (FFD). Firstly, the factor effect of the factorial design was used to measure the correlation between features and the target variable; secondly, a divide-and-conquer algorithm for finding the optimal factorial design for a given dataset was proposed; thirdly, in order to solve the problem that the traditional experimental design required manual execution of experiments, a data-driven approach was proposed to automatically search the response values for the factorial design from the input dataset; finally, the factor effects were calculated based on the design matrix and the average response values, and the features and interactions were sorted by the factor effects. Then the significant features and interactions could be obtained. The experimental results show that the average classification error rate of FFD over Mutual Information Maximisation (MIM), Joint Mutual Information Maximisation (JMIM) and ReliefF was 2.95, 3.33 and 6.62 percentage points, respectively. Therefore, FFD can effectively identify significant features and interactions that are highly correlated with the target variable in real-world datasets.

Key words: big data; correlation; feature selection; interaction; factorial design

0 引言

随着大数据时代的到来,许多数据集具有大量特征和数据记录^[1],比如社交网络数据和自然语言处理数据。文献[2]指出,这种特征数量多、数据量大的数据集为大数据分析带来了巨大的挑战。对于这类数据,传统的因果关系分析可能变得十分困难,复杂度更低的相关关系分析^[3]迎来了新的机遇。变量之间的相关关系是指目标变量与特征之间的关联性,文献[4]对大数据相关关系分析方法进行了综述。文献[5]指出,对于一些大数据分析问题,相关关系的结果就足以解决问题。在机器学习与数据挖掘领域,特征选择方法广泛应用于挖掘与目标变量相关的重要特征。

特征选择算法通常可分为三类^[6]:嵌入式(Embedding)、封装式(Wrapper)和过滤式(Filter)。嵌入式方法将特征选择作为分类器的一个组成部分。封装式方法枚举所有特征子

集,并计算其分类效果。过滤式方法通过定义一个评分标准对特征进行打分排序,最终选择得分高的特征,文献[6]提出了一个过滤式特征选择算法的框架。相比嵌入式和封装式方法,过滤式方法的效率更高并且独立于具体的分类器,因此,本文研究使用过滤式特征选择方法挖掘大数据相关关系。

文献[7]将过滤式特征选择方法分为单变量算法和多变量算法。单变量方法的效率高但是忽略特征之间的依赖性,比如信息增益(Information Gain, IG)^[8]。多变量算法使用特征之间的依赖性提升了特征选择的效果,比如:文献[9]提出的互信息最大化算法考虑了相关性;文献[10]提出的一种改进的最大相关最小冗余算法考虑了条件冗余性;文献[11]提出的一种两阶段特征选择算法考虑了特征之间的交互作用,但是,多变量方法的复杂度较高,难以直接应用于大数据分析。

本文的主要内容如下:提出一种快速的多变量过滤特

收稿日期:2018-03-07;修回日期:2018-03-27;录用日期:2018-03-28。 基金项目:国家自然科学基金资助项目(61602094)。

作者简介:唐小川(1986—),男,四川成都人,博士研究生,CCF 会员,主要研究方向:特征选择、机器学习、大数据分析; 罗亮(1980—),男,陕西汉中人,讲师,博士,主要研究方向:云计算可靠性建模、大数据处理。



征选择算法 FFD(Full Factorial Design), 用于挖掘大数据中与目标变量关联性强的特征和交互作用。FFD 将析因设计的因子效应作为一种新的相关关系度量标准, 用于度量特征和交互作用与目标变量的相关性, 从而能对特征和交互作用进行统一排序, 通过交互作用提升多变量过滤式特征选择算法的性能。FFD 使用一种分治算法快速地从输入数据集中获取析因设计的结果, 从而提高计算因子效应的速度。

1 一种大数据特征和交互作用选择方法

记输入数据为 $\mathbf{D} = \{\mathbf{X}, \mathbf{y}\}$, $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n \times p}$ 。 \mathbf{X} 的每一列 $\mathbf{x}_j = (x_{1j}, x_{2j}, \dots, x_{nj})^T$ 表示一个特征, $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ 表示目标变量。 k 阶交互作用记为 $\mathbf{I}_k = (x_{i_1}, x_{i_2}, \dots, x_{i_k})$ 。本文要解决的问题是从 \mathbf{X} 中选择一个具有代表性的特征和交互作用子集 $S = \{\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_s\} \cup \{\mathbf{I}_{j_1}, \mathbf{I}_{j_2}, \dots, \mathbf{I}_{j_t}\}$ 。

1.1 两水平完全析因设计

两水平完全析因设计 FFD 广泛应用于统计学实验设计(Design of Experiments, DOE), 实验设计研究如何制定实验计划和分析实验结果^[12]。传统的实验设计方法是一种模型驱动的方法, 主要包括三个阶段: 设计、测试和分析。在设计阶段, 制定最有效的实验计划; 在测试阶段, 按照实验计划做实验; 在分析阶段, 使用统计学工具分析实验结果。

1.2 基于析因设计的特征和交互作用选择

文献[11]提出一种基于析因设计的特征选择算法。本文提出一个快速的搜索适合输入数据集的最优析因设计的算法, 使得析因设计能够应用于大数据的特征和交互作用选择。式(1)是析因设计的统计模型。该模型是一个包含交互作用项的一般线性模型(General Linear Model, GLM)。

$$f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p) = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3 + \dots + \beta_{12} \mathbf{x}_1 \mathbf{x}_2 + \beta_{13} \mathbf{x}_1 \mathbf{x}_3 + \dots + \beta_{123} \mathbf{x}_1 \mathbf{x}_2 \mathbf{x}_3 + \dots + \varepsilon \quad (1)$$

其中: $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots$ 表示特征; $\mathbf{x}_1 \mathbf{x}_2, \mathbf{x}_1 \mathbf{x}_3, \mathbf{x}_1 \mathbf{x}_2 \mathbf{x}_3, \dots$ 表示交互作用; ε 表示随机误差。特征和交互作用常被称为因子。

式(1)中的系数 β 表示因子与目标变量之间的关联性, 常被称为因子效应。对于给定的特征, 如果该特征的因子效应的绝对值大于其他因子, 那么该特征与目标变量之间的相关关系最强。因此, 本文使用因子效应作为特征与目标变量之间关联性的度量标准。

本文提出一种快速计算因子效应的方法, 该方法是一种数据驱动的方法, 克服了由模型驱动的析因设计方法需要手动执行实验的局限性。该方法也说明了为输入数据集搜索最优析因设计的过程, 具体如下:

1) 用特征排序方法对 \mathbf{X} 中的所有特征进行排序, 比如对称不确定性和最大互信息^[9]。

2) 将 \mathbf{X} 中的所有特征都转化为二值变量, 比如: 对于连续性特征, 使用该特征的均值将其划分为高(+1)和低(-1)两个水平。这个过程常被称为二值化。

3) 计算最大的两水平析因设计, 使得数据集 \mathbf{D} 能够为其提供足够的数据。由于析因设计与因子数量一一对应, 本文提出一个搜索最大析因设计对应的最大因子数量的算法:

输入: 数据集 \mathbf{D} , 其中的特征已经排序和二值化。

输出: 最大因子数量 k 。

$index[0] \leftarrow \{1, 2, \dots, n\}$

$k \leftarrow 1$

```

while true do
    析因设计的行的数量 nrun ← 2k
    k ← k + 1
    for i ← 0 to  $\frac{nrun}{2}$  do
        // 将 index[i] 划分为两部分:
        temp_index[2i] ← 找出所有的  $s_0 \in index[i]$ ,
            满足  $D[s_0][k] == -1$ 
        temp_index[2i + 1] ← 找出所有的  $s_1 \in index[i]$ ,
            满足  $D[s_1][k] == 1$ 
    end for
    if temp_index 中有元素为空 then
        break
    else
        index ← temp_index
    end if
end while

```

该算法使用分治法搜索最大析因设计对应的特征数量 k , 避免了将数据集 \mathbf{D} 与特征数量为 1 到 k 的所有析因设计进行比较, 因此, 该算法能快速地找出适合数据集 \mathbf{D} 的最大析因设计。

4) 为含有 k 个特征的析因设计生成设计矩阵^[11], 记为 $\mathbf{M} = (m_{ij}) \in \{-1, 1\}^{N \times N}$, 其中 $N = 2^k$ 。 \mathbf{M} 的行记为 $\{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N\}$, 其中 \mathbf{r}_i 是一个因子水平组合, \mathbf{r}_i 常被称为一次实验(run)。 \mathbf{M} 的列记为 $\{\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_k, \mathbf{m}_{k+1}, \dots, \mathbf{m}_N\}$, 其中前 k 列 $\{\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_k\}$ 是特征, 接下来的 $\binom{k}{2}$ 列是二阶交互作用, \dots , 最后的一列是 k 阶交互作用。生成 \mathbf{M} 的方法如下:

① 生成单个特征 $\mathbf{m}_i, i \in \{1, 2, \dots, k\}$ 。 \mathbf{m}_i 的值开始于 -1, 然后切换为 1, \dots , 依此类推, 每隔 $N/2^i$ 个元素切换一次正负号。

② 构造二阶交互作用 $\mathbf{m}_i, i \in \{k+1, k+2, \dots, C_k^2\}$ 。令 $\mathbf{m}_i = \mathbf{m}_{i_1} \cdot \mathbf{m}_{i_2}$, 其中 $i_1, i_2 \in \{1, 2, \dots, k\}$ 。

...

③ 构造 k 阶交互作用 \mathbf{m}_{2k} 。令 $\mathbf{m}_{2k} = \mathbf{m}_{i_1} \cdot \mathbf{m}_{i_2} \cdot \dots \cdot \mathbf{m}_{i_k}$ 。

5) 计算因子效应。首先, 计算设计矩阵 \mathbf{M} 的每一次实验(\mathbf{M} 的一行)的目标变量值, 即实验结果, 常被称为响应(response), \mathbf{M} 的平均响应值记为 $\bar{\mathbf{y}} = (\bar{y}_1, \bar{y}_2, \dots, \bar{y}_N)^T$ 。 \mathbf{M} 的第 i 次实验 \mathbf{r}_i 的实验结果已经由 4) 中的算法计算得到, 即 $y[\text{index}[i]]$ 。 \mathbf{r}_i 的实验结果可能不止一个, 取其均值即可。多个实验结果相当于重复做了多次实验, 重复实验有利于减小实验的误差。

因此, 可以通过式(2)计算因子效应, 其中 $w_i (i \in \{1, 2, \dots, k\})$ 是 \mathbf{m}_i 的权重系数, 表示单个特征的因子效应; $w_j (j \in \{k+1, k+2, \dots, N\})$ 是 \mathbf{m}_j 的权重, 表示交互作用的因子效应, 从而可以通过因子效应对特征和交互作用进行统一排序。

$$\mathbf{w} = \mathbf{M}^T \bar{\mathbf{y}} \quad (2)$$

下面举例说明如何使用 FFD 分析两个因子 $\mathbf{x}_1, \mathbf{x}_2$ 和交互作用 $\mathbf{x}_1 \mathbf{x}_2$ 。

在设计阶段, 生成一个设计矩阵, 如表1的左侧3列所示, 得到设计矩阵 \mathbf{M} 如下:

$$\mathbf{M} = \begin{bmatrix} -1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{bmatrix} \quad (3)$$



在测试阶段,为每一个实验随机产生两个响应值,如表1右侧两列所示。

表1 二因子析因设计的设计矩阵和响应值

Tab. 1 Design matrix and responses of a two variable factorial design

x_1	x_2	x_1x_2	y_1	y_2
-1	-1	1	0.4170	0.1468
-1	1	-1	0.7203	0.0923
1	-1	-1	0.0001	0.1863
1	1	1	0.3023	0.3456

在分析阶段,计算各个特征和交互作用的因子效应。由表1右侧2列的每一行分别求平均值可得平均响应值为:

$$\bar{y} = (0.2819, 0.4063, 0.0932, 0.3240)^T \quad (4)$$

因此,根据式(2)可得因子效应为:

$$w = M^T \bar{y} = (-0.2710, 0.3552, 0.1064)^T \quad (5)$$

从而 x_1 , x_2 和 x_1x_2 的因子效应分别为 -0.271, 0.3552 和 0.1063。按因子效应绝对值的大小可将特征和交互作用排序为:

$$x_2 > x_1 > x_1x_2 \quad (6)$$

1.3 计算复杂度分析

输入数据 $D \in \mathbb{R}^{n \times p}$ 由 n 条记录和 p 个特征组成,适合数据集 D 的最大的两水平析因设计由 k 个因子组成, k 的上界为 $O(\ln n)$ 。

搜索最大析因设计的复杂度为 $O(2nk)$,即 $O(2n \ln n)$ 。

为最大析因设计生成设计矩阵的复杂度为 $O(nm)$,计算因子效应的复杂度也为 $O(nm)$,其中 m 是特征和交互作用的总数。在最坏的情况下,需要考虑所有的交互作用,其复杂度为 $O(n^2)$,但是,实验设计领域提出效应稀疏性原则,即:大多数系统只受一部分因子和低阶交互作用的影响,高阶交互作用可以忽略。二阶交互作用的数量为 $\binom{k}{2}$,三阶交互作用的数量为 $\binom{k}{3}$,依此类推。如果只考虑二阶交互作用,那么整个方法的复杂度为 $O(\ln^2 n)$ 。

2 实验与分析

本文通过实验对比了 FFD 与其他特征选择方法。实验采用 3 个 UCI (University of California Irvine Machine Learning Repository) 数据集,包括细颗粒物数据集 (Beijing Particulate Matter 2.5, PM2.5)、垃圾邮件数据集 (Spambase) 和文本分类数据集 (Baseball vs. Hockey, BASEHOCK)。由于本文提供的方法已经显式地考虑了交互作用,所以通过线性回归的分类错误率对比不同的特征选择方法。本文对比了 3 个著名的过滤式特征选择算法:联合互信息最大化 (Joint Mutual Information Maximisation, JMIM) 算法^[14]使用联合互信息度量两个特征和一个目标变量三者之间的交互作用;ReliefF^[15]是一种基于相似度的特征选择算法,利用了特征之间的条件依赖性;互信息最大化 (Mutual Information Maximisation, MIM) 算法^[9]考虑了单个特征与目标变量之间的相关关系。

本文的实验配置^[16]如下。首先,使用十折交叉验证将数据随机划分为训练数据和测试数据。然后对数据进行二值化。其次,用特征选择方法选择 $k = \{2, 4, \dots, 50\}$ 个特征或交互

作用,并更新数据集。交互作用的值可由向量的对应分量的乘积得到,可视为一个新特征。再次,用训练数据训练线性回归模型,然后用得到的模型在测试数据上得到分类错误率。最后,计算十折交叉验证的平均分类错误率。

Spambase 数据集。该数据集由 4601 条记录和 57 个特征组成。图 1 是 Spambase 数据集上的实验结果。FFD 的错误率一直低于其他对比方法。FFD 的最低错误率是 9.06%,其他方法的最低错误率是 11.89%,FFD 将错误率降低了 2.83 个百分点。一个可能的原因是 FFD 选择的交互作用具备单个特征所不具备的信息,比如二阶交互作用 x_7x_{53} 。

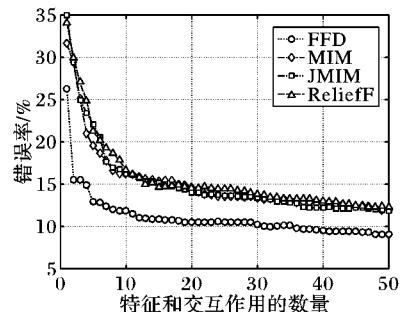


图 1 Spambase 数据集上的分类错误率随特征数量的变化

Fig. 1 Changes of classification error rate with the number of features on the Spambase dataset

PM2.5 数据集。该数据集记录了北京市在 2010 至 2014 年间的 PM2.5 数据,由 43824 条记录和 13 个特征组成。在进行特征选择之前,本文将 PM2.5 数据集的特征和目标变量离散化为二值变量。对于连续型变量,使用其平均值将其离散化为二值变量,将高于平均值的项标记为 1,低于平均值的项标记为 -1。对于一些离散型变量,使用一对多(one-vs-all)分解策略将其离散化为二值变量。对于时间特征,每年 (year) 被离散化为四个季度,每个月 (month) 被离散化为两个半个月,每天 (day) 被离散化为两个半天。对于风向特征 “wind direction”,离散化为东北 (North East, NE)、西北 (North West, NW)、东南 (South East, SE) 和和静风 (Calm and Variable wind, CV)。对于目标变量,设定 PM2.5 的阈值为 75,PM2.5 大于 $75 \mu\text{g}/\text{m}^3$ 表示空气受到污染,记为 1,否则记为 -1。因此,经过离散化以后,得到 21 个二值变量特征。

图 2 是 PM2.5 数据集上的实验结果。可以看到,随着已选特征和交互作用数量的增加,FFD 的错误率逐渐降低,当特征和交互作用大于 12 时,FFD 的错误率持续低于其他对比方法。FFD 的最低错误率为 32.72%,低于其他对比方法的最低错误率 (33.61%)。一个可能的原因是 FFD 选择的交互作用具有其他特征所不具备的关键信息,比如:交互作用 x_6x_8 (风速和东南风之间的交互作用) 的重要性排第三,可能对北京 PM2.5 有显著的影响。

BASEHOCK 数据集。该数据集由 1993 条记录和 4862 个特征组成。图 3 是 BASEHOCK 数据集上的实验结果。FFD 的错误率一直低于对比方法,FFD 将最低错误率降低了 5.07 个百分点。一个可能的原因是 FFD 选择的交互作用(比如二阶交互作用 $x_{2965}x_{3302}$)具备其他方法所忽略的关键信息。

表 2 对比了 FFD 与其他算法的分类错误率。FFD 分别将 MIM、JMIM 和 ReliefF 的平均错误率降低了 2.95、3.33 和 6.62 个百分点。FFD 在 Spambase、PM2.5 和 BASEHOCK 数



据集上的最低错误率都低于对比方法。

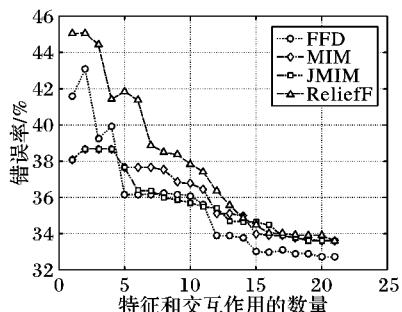


图2 PM2.5数据集上的分类错误率随特征数量的变化
Fig. 2 Changes of classification error rate with the number of features on the PM2.5 dataset

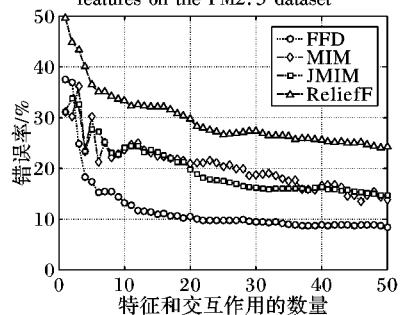


图3 BASEHOCK 数据集上的分类错误率随特征数量的变化
Fig. 3 Changes of classification error rate with the number of features on the BASEHOCK dataset

表2 各特征选择算法的分类错误率对比
Tab. 2 Comparison of lowest classification error rate for each feature selection algorithm

算法	各数据集分类错误率			平均分类错误率
	Spambase	PM2.5	BASAHOCK	
MIM	11.93	33.62	13.45	19.67
JMIM	11.89	33.61	14.65	20.05
ReliefF	12.30	33.62	24.09	23.34
FFD	9.06	32.72	8.38	16.72

3 结语

本文提出了一种新的过滤式特征选择方法 FFD 用于大数据相关关系分析。FFD 使用析因设计的因子效应作为特征与目标变量之间的关联性的度量标准。提出一种为输入数据集快速搜索最佳析因设计的分治算法,理论分析表明,这个分治算法的复杂度为 $O(lb^2n)$,有效降低了计算因子效应的复杂度。为了对特征和交互作用进行统一排序,FFD 将因子效应作为排序标准。

实验结果表明,FFD 在数据集 BASEHOCK、Spambase 和 PM2.5 数据集上的最低错误率分别比其他对比方法低 5.07、2.83 和 0.89 个百分点,也就是将实验中的所有数据集的最低错误率降低了 2.93 个百分点。FFD 成功地发现了影响 PM2.5 数据集的一个关键因素是风速与风向的交互作用,即东南方向的风速可能对北京 PM2.5 有重要影响。

参考文献 (References)

- [1] TAN M, TSANG I W, WANG L. Towards ultrahigh dimensional feature selection for big data [J]. Journal of Machine Learning Research, 2014, 15(4): 1371–1429.
- [2] FAN J, HAN F, LIU H. Challenges of big data analysis [J]. National Science Review, 2014, 1(2): 293–314.
- [3] 李国杰, 程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J]. 中国科学院院刊, 2012, 27(6): 647–657. (LI G J, CHENG X Q. Research status and scientific thinking of big data [J]. Bulletin of the Chinese Academy of Sciences, 2012, 27(6): 647–657.)
- [4] 梁吉业, 冯晨娇, 宋鹏. 大数据相关分析综述[J]. 计算机学报, 2016, 39(1): 1–18. (LIANG J Y, FENG C J, SONG P. A survey on correlation analysis of big data [J]. Chinese Journal of Computers, 2016, 39(1): 1–18.)
- [5] MAYER-SCHNBERGER V, CUKIER K. Big Data: A Revolution That Will Transform How We Live, Work, and Think [M]. New York: Houghton Mifflin Harcourt, 2013: 50–72.
- [6] BROWN G, POCOCK A, ZHAO M J, et al. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection [J]. Journal of Machine Learning Research, 2012, 13 (1): 27–66.
- [7] SAEYS Y, INZA I, LARRAÑAGA P. WLD: review of feature selection techniques in bioinformatics [J]. Bioinformatics, 2007, 23 (19): 2507–2517.
- [8] YANG Y, PEDERSEN J O. A comparative study on feature selection in text categorization [C]// ICML '97: Proceedings of the 14th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann, 1997: 412–420.
- [9] LEWIS D D. Feature selection and feature extraction for text categorization [C]// HLT '91: Proceedings of the Workshop on Speech and Natural Language. Stroudsburg, PA: Association for Computational Linguistics, 1992: 212–217.
- [10] VINH N X, ZHOU S, CHAN J, et al. Can high-order dependencies improve mutual information based feature selection? [J]. Pattern Recognition, 2016, 53(C): 46–58.
- [11] TANG X, DAI Y, SUN P, et al. Interaction-based feature selection using factorial design [J]. Neurocomputing, 2018, 281: 47–54.
- [12] MONTGOMERY D C. Design and Analysis of Experiments [M]. 9th ed. Hoboken: John Wiley and Sons, 2017: 179–220.
- [13] ZHAO Z, LIU H. Searching for interacting features [EB/OL]. [2018-01-04]. <http://www.ijcai.org/Proceedings/07/Papers/187.pdf>.
- [14] BENNASAR M, HICKS Y, SETCHI R. Feature selection using joint mutual information maximisation [J]. Expert Systems with Applications, 2015, 42(22): 8520–8532.
- [15] ROBNIK-ŠIKONJA M, KONONENKO I. Theoretical and empirical analysis of ReliefF and RReliefF [J]. Machine Learning, 2003, 53(1/2): 23–69.
- [16] SONG Q, NI J, WANG G. A fast clustering-based feature subset selection algorithm for high-dimensional data [J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 25(1): 1–14.

This work is partially supported by the National Natural Science Foundation of China (61602094).

TANG Xiaochuan, born in 1986, Ph. D. candidate. His research interests include feature selection, machine learning, big data analysis.

LUO Liang, born in 1980, Ph. D., lecturer. His research interests include cloud computing reliability modeling, big data processing.