Statistics and Application 统计学与应用, 2017, 6(3), 292-297
Published Online August 2017 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2017.63033

Hans汉斯

The Comparative Study of Individual Credit Evaluation Model in P2P Lending

Qiuyue Zhang

Yunnan University of Finance and Economics, Kunming Yunnan
Email: 1084695021@qq.com

Received: Jul. 14th, 2017; accepted: Aug. 1st, 2017; published: Aug. 4th, 2017

Abstract

This article compares the statistical analysis methods and the machine learning methods in the credit evaluation field, and according to the applicability of the evaluation index system data, this paper finally uses decision tree and its combination algorithm to evaluate individual credit in P2P lending. Through the comprehensive analysis of the accuracy and stability of the model, the best classification method for the data is random forest classifier.

Keywords

P2P Lending, Credit Evaluation, Model Select, Machine Learning Methods

P2P信贷个人信用评价模型的比较研究

张秋月

云南财经大学，云南 昆明
Email: 1084695021@qq.com

收稿日期：2017年7月14日；录用日期：2017年8月1日；发布日期：2017年8月4日

摘　要

本文通过对信用评价方法中的统计分析方法及机器学习方法进行比较，根据评价指标体系数据的适用性要求，最终选择决策树及其组合方法对P2P信贷个人信用进行评价。通过对模型的准确性及稳定性的综合分析，对本文数据拟合最好的分类方法是随机森林分类方法。

文章引用: 张秋月. P2P 信贷个人信用评价模型的比较研究[J]. 统计学与应用, 2017, 6(3): 292-297.
DOI: 10.12677/sa.2017.63033

关键词

P2P信贷，信用评价，模型选择，机器学习方法

1. 引言

自 2007 年国外网络借贷平台引入中国以来，国内 P2P 网络借贷平台如雨后春笋蓬勃发展，迅速形成一定规模。P2P 网络借贷实质是一种信用债，其方便、快捷、门槛低的特点解决了中小微企业及个人融资难的问题。与传统的担保贷款不同的是，P2P 去担保的运营模式使得能否做好风险控制决定了平台的存亡。

信用评价方法一直在不断演进，从最初的统计分析方法发展到现在的机器学习方法。其中统计分析方法主要有多元判别分析，Logistic 回归等，机器学习方法有分类树及其组合算法、神经网络、k 最近邻方法、支持向量机等多种方法。石庆焱等(2004)就国外个人信用评分领域中常用的线性与非线性 5 种方法利用中国商业银行消费者数据分别建立评分模型并比较，认为不同的模型有各自的优缺点，神经网路等非线性方法的精度高于判别分析等线性评价方法，但线性方法的稳健性高于神经网络等方法，在预测精度范围内线性评价方法还是有较强区分“好”“坏”客户的能力，可用于信贷决策[1]；崔媛媛(2006)对国内外部分商业银行所用信用评价指标体系进行比较后建立了个人信用评价指标体系，兼用经典判别分析、Logistic 回归、神经网络和分类树四种方法对个人汽车贷款数据建模，最后对不同评价模型的准确率及有效性进行比较分析，得出神经网络模型最优的结论[2]；庞素琳、巩吉璋(2009)研究了商业银行个人信用的评级问题，为解决个人信用记录中涉及的非数据型数据的问题，提出使用决策树分类算法，在 C4.5 算法基础上引进 C5.0 算法，对个人信贷数据进行检验，通过比较分析发现基于决策树 C5.0 算法模型的分类结果具有更高的精确度、较低的风险成本与较强的可控性，对信贷决策有一定的指导作用[3]；向晖、杨胜刚(2011)针对个人信用评估单一模型的不足，提出基于多种分类器组合的模型[4]；胡小宁、何晓群(2015)在对个人信用评价构建 Logistic 模型发现，需要建立较多的虚拟变量作为解释变量，然而常用的变量选择方法只能选择某一个虚拟变量而不是将整体相关变量进行选择，故而提出使用 Group Lasso 方法建立 Logistic 模型[5]。

本文首先针对模型的适用性进行分析，然后再对模型的准确性及稳定性进行分析，最后选择合适的模型对 P2P 信贷个人信用进行评估。

2. 数据来源及指标体系的构建

本文使用的数据来自某金融服务公司的现金贷款客户资料库，从个人特征、还款能力及指标数据的可获得性考虑，选取性别、年龄、婚否、学历、住房性质、现地址居住时间、收入、单位性质、职位、在现职工作年限、贷款期数、贷款金额 12 个特征指标。从该公司获得有效样本数据 1119 条，其中贷款审批结果(class)为“通过”的有 794 条记录，“拒绝”的有 325 条记录，class 为被解释变量。特征指标变量的性质、类型及具体的变量取值如表 1 所示。

Table 1. Index system and variables
表 1. 指标体系及变量表

指标名称	变量性质	变量类型	变量取值
性别	定性	离散	男；女
年龄	定量	连续	[19,55]
婚否	定性	离散	离异；未婚；已婚无子女；已婚有子女
学历	定性	离散	小学；初中；高中；中专；大专；本科
住房性质	定性	离散	与父母同住；按揭房；全款房；宿舍；租赁；其他
现地址居住时间	定性	离散	1 年以下；1~2 年；2~6 年；6 年以上
收入	定量	连续	[2000,40000]
单位性质	定性	离散	个体；国有企业；上市公司；事业单位；私营企业
职位	定性	离散	老板；普通员工；熟练员工/公务员；经理/高级职员
在现职工作年限	定性	离散	1 年以下；1~3 年；3~5 年；5 年以上
贷款期数(月)	定性	离散	6；9；12；15；18；24
贷款金额	定量	连续	[3000,22000]

3. 各信用评价模型的适用性分析

3.1. 判别分析

统计学中的判别分析方法较早地在信用评分模型中运用，在分类问题中，假定分类因变量一共有 K 类，即 K 个水平，则可以假定一个对象属于第 k 类的(先验)概率为 π_k，而 $\sum_{k=1}^{K}\pi_k=1$。用 $f_k(x)$ 表示属于第 k 类的分布密度函数，相应于自变量 x 的因变量属于第 k 类的后验概率为：

$$P(G=k|x)=\frac{f_k(x)\pi_k}{\sum_{\ell=1}^{K}f_\ell(x)\pi_\ell}$$

通常经典判别分析假定 X 服从多元正态分布，密度函数为：

$$f_k(x)=\frac{1}{(2\pi)^{p/2}|\sum k|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu_k)^{\mathrm{T}}\sum_k^{-1}(x-\mu_k)\right\}$$

对于线性判别，还假定 $\sum k=\sum(\forall k)$，在这个假定下，不同类的分布仅仅按照正态分布的位置来区别，这时的最优分类为：$\hat{G}(x)=\arg\max_k\left[f_k(x)\pi_k\right]$，由此得：

$$\hat{G}(x)=\arg\max_k\left[x^{\mathrm{T}}\sum{}^{-1}\mu_k-\frac{1}{2}\mu_k^{\mathrm{T}}\sum{}^{-1}\mu_k+\ln(\pi_k)\right]$$

在求解过程中，判别分析对于密度分布函数要求服从正态分布，在数据真实满足这些条件时，判别分析是非常优秀的模型。但是现实情况并不能满足这个要求，且本文的变量指标多数属于定性变量，对于自变量包含有很多水平的定性变量的情况下，经典判别可能根本不能运作。

3.2. Logistic 回归

Logistic 回归用于处理因变量为二分变量的情况，是信用评价模型使用最广泛的模型之一。设 p 是客户“通过”的概率，由于 p 的取值在 0 和 1 之间，对 p 做如下变换：$\log it(p)=\ln\left(\frac{p}{1-p}\right)$，则相应的广义线性模型为：$\ln\left(\frac{p_i}{1-p_i}\right)=x_i^{\mathrm{T}}\beta$。

很显然，对于每一个观测值，我们预测的是 p_i，而不是第 i 个观测值的水平。这时就需要一个阈值 p_t 使得当 $\hat{p}_i > p_t$ 时判定第 i 个观测值属于某一水平。

3.3. 神经网络

人工神经网络是对自然神经网络的模仿，是最早的机器学习方法之一，它可以有效地解决很复杂的有大量互相相关变量的回归和分类问题。神经网络的因变量可以有多个，隐藏层也可以有多个，但一般一个就够了。隐藏层的节点数可多可少，节点数太少可能导致拟合不好，节点数太多可能导致过拟合，可以用交叉验证来选择隐藏层的节点数目。神经网络的原理是把上层节点的值加权平均后送到下层节点，最终到输出层节点，然后根据误差大小反馈到前面的层，再重新加权平均，每个平均值都通过一个激活函数作用，反复训练，直到误差在允许的范围内。

使用 *R* 软件对数据进行神经网络模拟，*R* 软件的 *nnet* 程序包神经网络只有一个隐藏层，为了确定隐藏层中的节点数，用不同的节点数进行试验，最后进行比较后选择合适的节点数。

对训练集本身做预测，当节点数目为 20 时，有 165 个观测值被误判，误判率 14.7%；当隐藏节点数目改为 15 时，有 187 个观测值被误判，误判率 16.7%；再将节点数目减少到 10，就全部分类成“通过”了。随机种子的改变也会大大改变分类结果。所以神经网络不适合本文信用评价数据的分类。

3.4. k 最近邻方法

k 最近邻方法是在 N 个已知样本点中，找出 x 的 k 个近邻，看这 k 个近邻中多数属于哪一类，就将 x 归为哪一类。设 N 个样本中有 c 个类别：$\omega_1,\omega_2,\omega_3,\cdots,\omega_c$，每类有 N_i 个样本，设样本指标有 z 个，则构成一个 z 维特征空间，所有的样本点在这个 z 维特征空间里都有一个唯一的点与之对应。则对任何一个待识别样本 x，把它也放到这个 z 维空间里，通过构造一个距离公式(一般采用欧式距离公式)，可以找到样本 x 的 k 个近邻。定义判别函数为：$g_i(x)=k_i$，$i=1,2,\cdots,c$；若 $g_i(x)=\max k_i$，则分类 x 属于 w_j。

3.5. 支持向量机

两类分类的依据是数据空间中对象的距离准则，即同类对象之间的距离尽可能小，不同类对象之间的距离尽可能大。因为距离运算主要涉及点积运算，所以对于在低维空间中线性不可分问题，可以通过非线性变换将低维空间变换到一个高维空间，使得在高维空间中变得线性可分，这类非线性变换正是通过核函数来实现的。经过核变换的分类函数依靠支持向量形成分类，因此，称为支持向量机。可见，核函数是支持向量机算法中重要的核心，只有通过正确的核函数才能将问题的求解由低维空间转化到高维空间求解。

标准支持向量机有严密的数学理论基础，从数学角度分析，支持向量机是一个求条件极值的问题。支持向量机与 Logistic 回归、K 最近邻方法有一个共同的弱点，其算法都是基于数量变量，当数据的自变量有太多的定性变量或定性变量的水平太多时往往可能无法运作，这和其数学结构有关。

3.6. 决策树及其组合分类算法

分类决策树是一种非参数统计方法，其基本思想是按照一定规则进行分割，产生两个子节点，将这些子节点重复进行划分，直到达到一定的要求停止成为最终的叶节点。这些叶节点所表示的数据子空间的特征决定了它们判属于哪一类别。

组合多个分类器来得到结果的方法称为组合方法，将决策树作为基本学习器的组合方法有 bagging、adaboost 和随机森林。Bagging 分类从训练样本中做多次放回抽样(自助法抽样)，每次建立一棵决策树，假定一共建立 B 棵树，然后对于每一个新的观测值，通过这 B 棵树得到 B 个预测结果，最后按照少数服从多数的原则来投票确定该观测值属于哪一类。随机森林与 bagging 分类非常相似，随机森林的每棵树都不剪枝，让其充分生长，最终所有决策树按照各自分类结果投票，票数最多的类别为预测结果。Adaboost 每次用自助法抽样来构建树时，都根据前一棵树的结果对于误判的观测值增加抽样权重，使得下一棵树能够对误判的观测值有更多的代表性。

4. 模型结果的比较

以上我们对信用评估方法从早期的判别分析及其变形的 logistic 回归，再到后来开发出来的神经网络、k 最近邻方法、支持向量机以及决策树及其分类算法的人工智能方法，它们都互有优缺点。鉴于本文数据的适用性，本文最终选择决策树及其分类算法进行建模并进行比较。

4.1. 各模型的拟合

用 R 软件对所有样本数据进行决策树分类、bagging 分类、adaboost 分类及随机森林分类，各模型的错误分类比率如表 2 所示。其中 A 类错误率表示将拒绝的错分到通过一类的比率， B 类错误率表示将通过的错分到拒绝一类的比率。显然，犯 A 类错误的损失要大于犯 B 类错误的损失，在两种损失已知时以总损失最小为标准来评价模型的优劣是最合适的方法。但在实际中损失难以衡量，对个人信用进行评价时，以错误分类率为评价准则之一也是一种标准做法。

各模型中错误分类率最低的是 adaboost 分类方法，但是其分类结果与其他几个方法相差很大，存在过拟合问题。而 A 类错误率普遍高于 B 类错误率，可能是由于样本数据中“拒绝”的客户远远小于“通过”的客户。一般而言，决策树及其组合算法对训练样本的分类效果较好，错误分类率都会在一个相对较低的水平，而对训练样本分类时，其错误分类率可能会有所提升，模型的稳定性也是需要重点考察的一个标准。

4.2. 各模型的交叉验证及比较结果

仅仅从全部数据的拟合结果错误率来比较模型还不够，需要用交叉验证方法来比较模型的稳定性。本文选用 10 折交叉验证，结果如表 3 所示。从模型的稳健性考虑，随机森林分类模型最优，预测准确性达到 81.93%。

Table 2. Error rate of each classification model
表 2. 各分类模型的错误率

	总错误率(%)	A 类错误率(%)	B 类错误率(%)
决策树	15.19	37.54	6.05
bagging	14.38	43.69	2.39
adaboost	0.08	0	0.13
随机森林	17.07	45.85	5.29

Table 3. The 10-fold cross-validation misjudgment rate of 4 classification models
表 3. 4 种模型的 10 折交叉验证误判率

折次	决策树	Bagging	adaboost	随机森林
1	0.1592	0.1592	0.1681	0.1327
2	0.1415	0.1769	0.1946	0.1858
3	0.1592	0.1238	0.1415	0.1150
4	0.1681	0.1946	0.1769	0.1769
5	0.1785	0.1428	0.1696	0.1428
6	0.2972	0.2612	0.2342	0.2792
7	0.1981	0.1981	0.2252	0.1981
8	0.1981	0.1621	0.1801	0.1531
9	0.1711	0.1711	0.1441	0.1621
10	0.2342	0.2342	0.2252	0.2612
NMSE	0.1905	0.1824	0.1860	0.1807

5. 结论

本文对某公司 P2P 个人信贷数据进行个人信用评价，将信用评价方法演变过程中的各种方法都对本文数据的适用性做了分析，数据本身的不同，决定了模型的适用性。本文数据中大量的定性变量及定性变量水平较多的特性，使得以数量变量为基础的分析方法的适用性受到限制。决策树及其组合分类算法对本文同类型数据的分类有较高的准确率，至于哪种方法最优，与数据本身的特性有关，需具体分析。本文数据经分析后表明，随机森林分类效果最优。

尽管 P2P 信贷市场的个人征信体系不够完善，本文在建立个人信用评分模型所用到的指标并不能完全反映个人信用行为，但是利用已有信息对未来客户的信用进行预测及区分，对信贷决策具有一定的参考作用，能有效减小信贷损失。

参考文献 (References)

[1] 石庆焱, 靳云汇. 多种个人信用评分模型在中国应用的比较研究[J]. 统计研究, 2004(6): 43-47.

[2] 崔媛媛. 个人信用评价模型比较研究[D]: [硕士学位论文]. 北京: 北方工业大学, 2006.

[3] 庞素琳, 巩吉璋. C5.0 分类算法及在银行个人信用评级中的应用[J]. 系统工程理论与实践, 2009, 29(12): 94-104.

[4] 向晖, 杨胜刚. 基于多分类器组合的个人信用评估模型[J]. 湖南大学学报, 2011(3): 30-33.

[5] 胡小宁, 何晓群. 基于 Group Lasso 的个人信用评价分析[J]. 数学的实践与认识, 2015(6): 89-94.