Hans Journal of Data Mining 数据挖掘, 2016, 6(3), 97-105
Published Online July 2016 in Hans. http://www.hanspub.org/journal/hjdm
http://dx.doi.org/10.12677/hjdm.2016.63012

Hans汉斯

The Application of Credit Approval Based on Machine Learning Classification Method

Yulian Mo, Yu Fei*

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan
Email: 1508447697@qq.com, *feiyukm@aliyun.com

Received: Jul. 23rd, 2016; accepted: Aug. 15th, 2016; published: Aug. 18th, 2016

Abstract

The traditional method of credit card approval is often rely on the experience of credit personnel and is to decide whether the credit card applicants meet the conditions of application. Obviously, this approval method has a lot of randomness and instability. In this paper, we take advantages of R software and introduce the six latest machine learning classification method, decision tree classification, AdaBoost, Bagging classification, random forest classifier, support vector machine (SVM) classification, artificial neural network (Ann) into the credit card application management, then establish the automatic application management system, effectively reducing the randomness and instability of the examination and approval results. Finally we calculate the mean square error of all the classification method through 8-fold cross validation and chose the classification with the best effect. The result shows that the classification error of random forest classification is the smallest.

Keywords

Credit Card Application, Machine Learning Classification, Random Forest

基于机器学习分类方法的信用卡审批应用

莫玉莲，费　宇*

云南财经大学统计与数学学院，云南　昆明
Email: 1508447697@qq.com, *feiyukm@aliyun.com

*通讯作者。

文章引用: 莫玉莲, 费宇. 基于机器学习分类方法的信用卡审批应用[J]. 数据挖掘, 2016, 6(3): 97-105.
http://dx.doi.org/10.12677/hjdm.2016.63012

收稿日期：2016年7月23日；录用日期：2016年8月15日；发布日期：2016年8月18日

摘　要

传统的信用卡审批方法往往是依靠信贷人员的经验进行审批，确定信用卡申请者是否符合申请条件，这种审批方法有很大的随意性和不稳定性。本文利用R软件并将最新的六种机器学习分类方法——决策树分类、Adaboost分类、Bagging分类、随机森林分类、支持向量机分类、人工神经网络引入到信用卡申请管理中，建立了自动化的申请管理体系，有效地降低了审批结果的随意性和不稳定性，并通过八折交叉验证计算出每种方法的分类均方误差并进行对比，筛选出分类效果最好的方法。结果表明：随机森林分类的分类误差是最小的。

关键词

信用卡申请，机器学习分类，随机森林

1. 研究背景

信用卡申请的审批一般是由专门设置的风险控制部门的征信审核岗位员工完成，其目的就是通过拒绝高风险客户的信用卡申请，核准低风险高收益客户的信用卡申请，在保证信用卡部门收益的同时降低持卡人违约风险。随着信用卡业务竞争的加剧，各银行信用卡中心都将规模和市场占有率作为考核指标。在这种大环境下，各家银行都需要特别注意信用卡申请审批环节的效率和质量。而依靠征信审核岗位员工的经验去审核信用卡申请者是否符合申请条件的方法往往具有很大的随意性和不稳定性，其效率和审批的质量也缺乏一定的保障。随着信用卡业务的跨越式发展、同业竞争的不断加剧以及信息技术在银行业广泛而深入的应用，银行业务的经营理念和经营方式发生了很大的转变。例如，由于网络技术和电子商务的发展，银行业开始重视运用数据挖掘和机器学习等审批方法。

在对信用卡申请审批的监督管理方面，为了提高商业银行的信用卡风险管理能力，降低信用卡的管理风险，国内许多学者进行了大量的研究。刘继海，陈晓剑[1]通过将基于统计学习理论的分类方法 SVM (Support Vector Machine)引入信用卡申请管理，建立了信用卡申请管理的评分模型。同时将 SVM 与信用评分领域常用的 Logistic 回归进行了对比，从而帮助银行挑选优质客户。田晓光，孔德婧[2]建立了一个信用卡监督管理系统，对信用卡申请者进行资信评估，判断是否同意向申请者发行信用卡，从而辅助专家做出决策。并与其它的分类系统作了比较，实验结果表明，本系统在信用卡监督管理领域有很好的应用前景。刘慧[3]针对信用卡信用风险计量模型进行研究分析，将当前数据分析领域最为先进的数据挖掘技术应用于商业银行信用卡信用风险管理中。相对于我国，国外的信用卡业务发展较为靠前，许多学者也对此进行了大量的研究。其中 Sum Sakprasat 和 Mark C. Sinclair [4]对信用卡申请数据集的分类规则进行了 8 种不同的遗传编程的数据挖掘方法的研究。NF Matsatsinis [5]用机器学习的方法对信用卡评估智能决策系统。前人的研究大多只是基于某种机器学习方法或是某种计量模型对信用卡风险管理做了一些应用，几乎没有将科学前沿的几种机器学习方法——决策树分类、Adaboost 分类、Bagging 分类、随机森林分类、支持向量机分类、人工神经网络引入其中并进行对比分析。

机器学习是一门多领域交叉学科，涉及了概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。本文有效地利用机器学习方法的自动学习能力以及其又好又快的学习效率，并将其运用到信用卡的申请审

批上，这样便可有效地减少传统方法存在的随意性和不稳定性，也提升了审批过程的效率和质量。

2. 机器学习分类方法概述

2.1. 决策树分类

决策树分类是多阶分类技术中的一种，它将分类任务分解为多次完成。决策树分类器的分类规则由多个决策结点组成，每个结点仅完成分类任务中的一部分，经过逐次向下分类，最后完成分类任务。

2.2. Adaboost 分类

Adaboost (adaptive boosting 的简写)是 boosting 的一种，是一种组合方法，可以译为“自适应助推法”[6]。其核心思想是针对同一个训练集训练不同的分类器(弱分类器)，然后把这些弱分类器集合起来，构成一个更强的最终分类器(强分类器)。其算法本身是通过改变数据分布来实现的，它根据每次训练集之中每个样本的分类是否正确，以及上次的总体分类的准确率，来确定每个样本的权值。将修改过权值的新数据集送给下层分类器进行训练，最后将每次训练得到的分类器最后融合起来，作为最后的决策分类。

2.3. Bagging 分类

Bagging (bootstrap aggregating 的简写)可以译为“自助整合法”，它利用了自助法(Bootstrap)放回抽样，对训练样本做多次放回抽样，每次抽取的样本量相同的观测值。对每个抽取的样本生成一棵决策树，由这些树的分类结果的“投票”产生 bagging 分类。

2.4. 随机森林分类

随机森林(random forests)分类是一种进行许多次自助放回抽样的分类方法。它较之 bagging 分类的关键区别在于，在生成每棵树的时候，每个节点的变量都仅仅在随机选出的少数变量中产生。故不但其样本是随机的，就连每个节点变量的产生都有相当大的随机性。由随机抽取的样本建立得到的决策树数量要远远多于 bagging 的样本数目，进而产生了随机森林的说法，其最终的分类结果也是由这些决策树的分类“投票”结果而得。

2.5. 支持向量机分类

支持向量机(support vector machine)是一种分类算法，通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。通俗来讲，它是一种二类分类模型，其基本模型定义为特征空间上的间隔最大的线性分类器，即支持向量机的学习策略便是间隔最大化，最终可转化为一个凸二次规划问题的求解。

2.6. 人工神经网络分类

人工神经网络(artificial neural networks)是对自然的神经网络的模仿，其主要是由大量与自然神经细胞类似的人工神经元互联而成的网络。人工神经网络的工作和方法是模仿人脑，即首先要根据输入的信息建立神经元，通过学习规则或自组织等过程建立相应的非线性数学模型，并不断进行修正，使输出结果与实际值之间差距不断缩小，从而求取问题的解。

3. 实证分析

3.1. 数据来源与说明

本文实证分析所选用的信用卡申请数据是由 quinlan '@' cs.su.oz.au 向 UCI 机器学习网站捐赠的 Credit

Approval Data Set 数据(数据的网址为：http://archive.ics.uci.edu/ml/machine-learning-databases/credit-screening/)。该数据总共有 690 个信用卡申请客户的相关资料，总共有 16 个变量，因数据涉及一定的商业机密，所以许多变量只是用一些名义型的字母或字符代替(如表 1 所示)。其中第十六个变量是是否同意申请，“+”代表同意申请，“−”代表不同意申请，其中表 2 给出了这两种类型分布统计。数据变量的形式包括有分类、整数和实数，且其有缺失值的存在(如表 3 所示)，对于缺失值的处理，本文采取用随机森林的方式弥补。

3.2. 机器学习分类方法的 R 软件实现

3.2.1. 决策树分类结果

本文根据决策树分类的 R 程序包 rapart 中的函数 rapart( )对信用卡审批数据进行拟合，输出结果如图 1、图 2 以及表 4。其 R 语言处理的主要程序为：library(rpart)；library(rpart.plot)；(f=rpart(V16~.,w))；rpart.plot(f,type=2,extra=4)。从图 1 中，我们可以知道决策树在对信用卡审批进行分类时各变量的分节点在何处、分类偏差和分类结果在该节点的均值等。图 2 是由信用卡审批数据生成的一棵决策树，这棵决策树说明了在对申请者进行审批时，首先要看变量 V9，通过则是不同意申请，不通过则看变量 V10，随着决策树给出的示意图，一步步地分下去，最后得到分类的结果。表 4 给出了决策树的分类效果，其将同意申请的 40 个样本错误的分至不同意申请这一组，不同意申请的 41 个样本分至了同意申请这一组，分类误差为 0.1173913。可见决策树分类方法的正确率高达 88.3%，而其处理的速度也是比较快的。

3.2.2. Adaboost 分类结果

Adaboost 分类的 R 语言主要处理程序为：library(adabag)；set.seed(4410)；a=boosting(V16~.,w)；barplot(a$importance,cex.name=.8)。输出结果如表 5 及图 3 所示。表 5 给出了 Adaboost 分类器对信用卡审批数据的分类效果，我们可以知道其分类误差是为 0，相比决策树的分类误差要低得多。图 3 给出了 Adaboost 在拟合信用卡审批数据时各变量的重要性条形图，其中变量 V2、V3、V6、V14 在 Adaboost 分类器拟合信用卡审批数据时，是否同意申请者得到信用卡起到比较重要的作用。

3.2.3. Bagging 分类结果

Bagging 分类器的 R 语言主要处理程序是：library(adabag)；set.seed(4410)；a=bagging(V16~.,w)；barplot(a$importance,cex.name=.6)。输出结果如表 6 及图 4。表 6 给出了 Bagging 分类器对信用卡审批数据的分类效果，其将同意申请的 32 个样本错误的分至不同意申请这一组，不同意申请的 37 个样本分至了同意申请这一组，分类误差为 0.1，比决策树的分类误差要低一些，但比 Adaboost 分类效果要差得多了。图 4 给出了 Bagging 在拟合信用卡审批数据时各变量的重要性条形图，其中变量 V9 对信用卡审批的通过是否起到至关重要的作用。

3.2.4. 随机森林分类结果

随机森林分类器的 R 语言主要处理程序为：library(randomForest)；set.seed(101010)；(a=randomForest(V16~.,w,importance=T,proximity=T))。输出结果如表 7 及图 5。表 7 给出了随机森林拟合信用卡审批数据的分类效果，得到的分类误差为 0，分类误差同 Adaboost 的一样，比决策树和 Bagging 效果好得多；图 5 显示了在随机森林拟合信用卡时的各变量重要性，其中 V9 对信用卡审批同意与否起到最为重要的作用。

3.2.5. 支持向量机分类结果

支持向量机的 R 语言主要处理程序为：library(e1071)；a=svm(V16~.,w,kernal="sigmoid")；wp=predict(a,w)；z0=table(w[,16],predict(a,w))。输出的结果如表 8。表 8 给出了支持向量机分类器拟合信

Table 1. Type and value of variables for the Credit Approval Data
表 1. Credit Approval Data 各变量类型及取值

变量	类型	取值
V1	分类	a, b
V2	数值	13.75~80.25
V3	数值	0~28
V4	分类	u, y, l, t
V5	分类	g, p, gg
V6	分类	c, d ,cc, i, j, k, m, r, q, w, x, e, aa, ff
V7	分类	v, h, bb, j, n, z, dd, ff, o
V8	数值	0~28.5
V9	分类	t, f
V10	分类	t, f
V11	数值	0~67
V12	分类	t, f
V13	分类	g, p, s
V14	数值	0~2000
V15	数值	0~100,000
V16	分类	+, −

Table 2. Class distribution for the Credit Approval Data
表 2. Credit Approval Data 审批类别分布

类别	样本量
+	307 (44.5%)
−	383 (55.5%)

Table 3. Statistics of missing values in the Credit Approval Data
表 3. Credit Approval Data 缺失数据统计

变量	缺失个数
V1	12
V2	12
V4	6
V5	6
V6	9
V7	9
V14	13

Table 4. Decision tree classification results on the Credit Approval Data
表 4. 信用卡审批数据的决策树分类结果

	+	−
+	267	40
−	41	324

```
n=690

node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 690 307 – (0.55507246 0.44492754)
   2) V9=f 329  23 – (0.93009119 0.06990881) *
   3) V9=t 361  77 + (0.21329640 0.78670360)
     6) V10=f 133  56 + (0.42105263 0.57894737)
      12) V6=aa, c, d, ff, i, j, m 73  31 – (0.57534247 0.42465753)
        24) V14>=111 48  15 – (0.68750000 0.31250000) *
        25) V14< 111 25   9 + (0.36000000 0.64000000) *
      13) V6=cc, e, k, q, r, w, x 60  14 + (0.23333333 0.76666667) *
     7) V10=t 228  21 + (0.09210526 0.90789474) *
```

Figure 1. Decision tree output for the Credit Approval Data
图 1. 信用卡审批数据的决策树输出结果

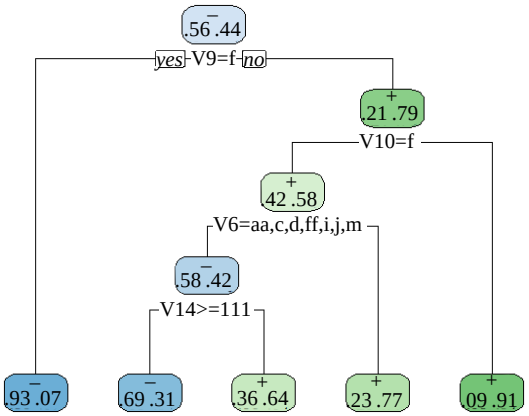

Figure 2. Decision tree for the Credit Approval Data
图 2. 信用卡审批数据的决策树

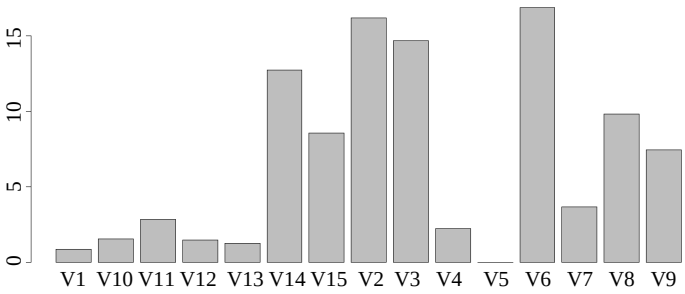

Figure 3. The variables importance of Adaboost fitting the Credit Approval Data
图 3. Adaboost 拟合信用卡审批数据时的变量重要性图

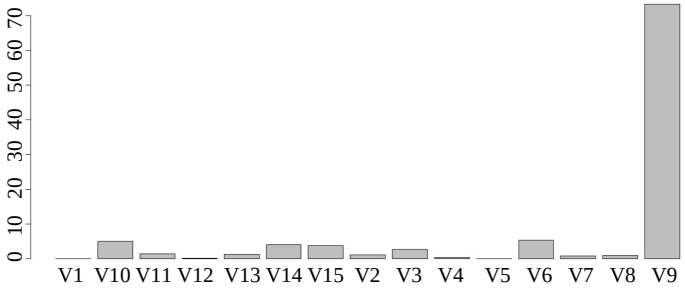

Figure 4. The variables importance of Bagging fitting the Credit Approval Data
图 4. Bagging 拟合信用卡审批数据时的变量重要性图

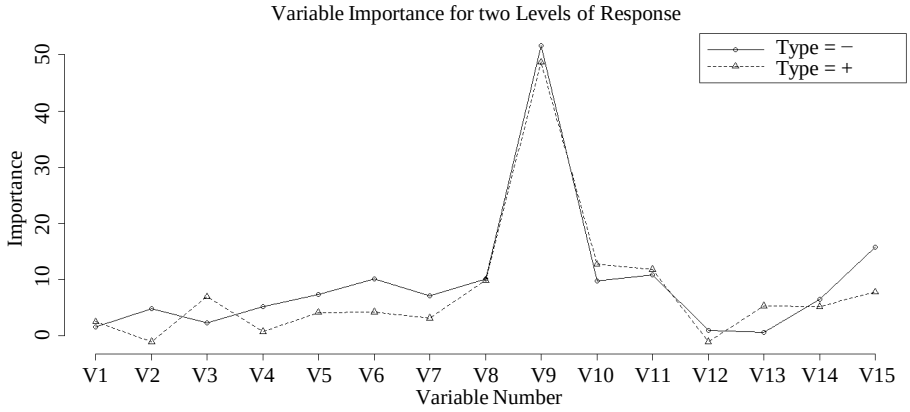

Figure 5. The variables importance of Random forest fitting the Credit Approval Data
图 5. 随机森林拟合信用卡审批数据时的变量重要性图

Table 5. Adaboost classification results on the Credit Approval Data
表 5. 信用卡审批数据的 Adaboost 分类结果

	+	−
+	307	0
−	0	383

Table 6. Bagging classification results on the Credit Approval Data
表 6. 信用卡审批数据的 Bagging 分类结果

	+	−
+	275	32
−	37	346

Table 7. Random forest classification results on the Credit Approval Data
表 7. 信用卡审批数据的随机森林分类结果

	+	−
+	307	0
−	0	383

Table 8. SVM classification results on the Credit Approval Data
表 8. 信用卡审批数据的支持向量机分类结果

	+	−
+	285	22
−	75	308

用卡审批数据的效果，其将同意申请的 22 个样本错误的分至不同意申请这一组，不同意申请的 75 个样本分至了同意申请这一组，分类误差为 0.1405797，即正确分类的比例为 85.9%。其分类误差是这几个方法中最大的一个。

3.2.6. 人工神经网络分类结果

人工神经网络的 R 语言主要处理程序为：library(nnet)；set.seed(1010)；a=nnet(w$V16~., data=w,

subset=1:n,size = 5,rang = 0.1,decay=5e−4,maxit = 200)；wp=predict(a, w, type = "class")。得到的误判率为0.07536232，正确分类的比例高达 92.4%。其分类效果次于 Adaboost 和随机森林，比其他三种的分类误差要低一些。

3.2.7. 六种机器学习分类方法的八折交叉验证结果

八折交叉验证(8-fold cross-validation)是用来测试算法的准确性，其将数据集分成八份，其中七份作为训练数据集，一份作为测试数据集，进行测试试验。其每次试验都会得到相应的正判率(或错判率)。8次的结果的正判率(或错判率)的平均值作为对算法精度的估计。表 9 给出了六种方法对信用卡审批数据集分类的八折交叉验证的平均误判率。对于训练集来说，分类效果最好的是 Adaboost 分类、随机森林分类和人工神经网络分类这三种方法，其平均误判率都为 0，其次是 Bagging 分类和决策树分类，最差的是支持向量机。但相对于测试集来说，随机森林的分类效果依然是最好的，且平均分类误差都为 0；Adaboost 和人工神经网络的误判率就比训练集的相差得比较大，故它们分类的稳定性还是较差的；决策树和 Bagging 的测试集平均分类误差比训练集的平均分类误差高 5%左右。故综合来看，随机森林的分类效果是这几个方法中最好的。

4. 结论

本文从实际问题出发，回顾前人对信用卡申请的一些研究，结合科技前沿的六种机器学习分类方法(决策树分类、Adaboost 分类、Bagging 分类、随机森林分类、支持向量机分类、人工神经网络分类)分别对信用卡审批数据进行分类拟合，建立了自动化的审请管理方式，并构造了八折交叉验证，计算出每种机器学习方法八折交叉验证的分类平均误差率，最后选出了对此信用卡数据拟合效果最好的机器学习方法——随机森林分类。

1) 本文将机器学习方法引入信用卡风险管理中有效地解决传统审批方法的不稳定性和随意性，大大提升了信用审批的效率和自动化过程。六种分类方法对信用卡数据集(Credit Approval Data Set)的拟合效果都比较不错，它们的分类精度都≥83%。

2) 随机森林分类方法对信用卡数据集(Credit Approval Data Set)在这六种方法中具有最好的拟合效果。不管是对训练集或者是测试集，随机森林分类的平均分类误差都为 0，这也显示出了随机森林分类器强大的稳定性。

3) 本文仅仅只是根据这组数据得到结果，没有对其他的信用卡申请审批数据进行验证，并且未对其他的信用卡申请审批方法进行对比。

Table 9. 8-fold cross-validation average false positive rate of six classification methods on the Credit Approval Data
表 9. 六种方法对信用卡审批数据分类的八折交叉验证平均误判率

分类方法	训练集的平均误判率	测试集的平均误判率
决策树分类	0.1080755	0.1552789
Adaboost 分类	0	0.1393912
Bagging 分类	0.09461941	0.1394584
随机森林分类	0	0
支持向量机分类	0.1389271	0.1451383
人工神经网络分类	0	0.1696285

基金项目

国家自然科学基金项目“广义估计方程(GEE)框架下的回归诊断：基于均值和协方差结构同时拟合的研究”(11561071)，云南省哲学社会科学研究基地 2015 重点项目“云南省社会经济可持续发展竞争力指标体系研究”(JD2015ZD20)。

参考文献 (References)

[1] 刘继海, 陈晓剑. SVM 模型在信用卡申请管理中的创新应用[J]. 哈尔滨工业大学学报: 社会科学版, 2007, 9(4):133-136.

[2] 田晓光, 孔德婧. 数据挖掘在信用卡发行中的应用[J]. 科技信息, 2008(5): 64-66.

[3] 刘慧. 基于数据挖掘技术的信用卡申请评分模型研究[D]: [硕士学位论文]. 大连: 东北财经大学, 2010.

[4] Sakprasat, S. and Sinclair, M.C. (2007) Classification Rule Mining for Automatic Credit Approval Using Genetic Programming. *IEEE Congress on Evolutionary Computation*, Singapore, 25-28 September 2007, 548-555.

[5] Matsatsinis, N.F. (2002) An Intelligent Decision Support System for Credit Card Assessment Based on a Machine Learning Technique. *Operational Research*, **2**, 243-260. http://dx.doi.org/10.1007/bf02936329

[6] 吴喜之. 复杂数据统计方法——基于 R 的应用(第二版) [M]. 北京: 中国人民大学出版社, 2013.