**Advances in Applied Mathematics** 应用数学进展, 2019, 8(8), 1410-1417
Published Online August 2019 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2019.88165

Hans汉斯

# Credit Evaluation Based on Improved Naive Bayesian Model

**Gao Wu[1], Xueyuan He[1], Ming Li[2*]**

[1]College of Mathematics, Taiyuan University of Technology, Jinzhong Shanxi
[2]College of Data Science, Taiyuan University of Technology, Jinzhong Shanxi
Email: 2895780769@qq.com, *liming01@tyut.edu.cn

Received: Jul. 29th, 2019; accepted: Aug. 13th, 2019; published: Aug. 20th, 2019

## Abstract

**Along with the rapid development of consumer credit, the demand for personal credit assessment has been aroused. To help financial institutions better understand their personal credit situation, combining the advantages of Fast Independent Component Analysis method (FastICA) and Linear Discriminant Analysis (LDA) to extract data features, a credit evaluation model called FastICA-LDA-NB is proposed, which is based on the improved Naive Bayesian classification algorithm. Applying the model to the UCI German personal credit data set, the proposed model has a good credit evaluation effect on the two evaluation index values of accuracy rate and recall rate.**

## Keywords

**Credit Evaluation, Independent Component Analysis, Linear Discriminant Analysis, Naive Bayesian**

# 基于改进朴素贝叶斯模型的信用评估

**吴　皋[1]，贺雪媛[1]，李　明[2*]**

[1]太原理工大学数学学院，山西　晋中
[2]太原理工大学大数据学院，山西　晋中
Email: 2895780769@qq.com, *liming01@tyut.edu.cn

收稿日期：2019年7月29日；录用日期：2019年8月13日；发布日期：2019年8月20日

## 摘　要

**伴随着消费信贷的快速发展，激起了对个人信用评估的需求。为了帮助金融机构更好的了解个人信用情**

*通讯作者。

文章引用: 吴皋, 贺雪媛, 李明. 基于改进朴素贝叶斯模型的信用评估[J]. 应用数学进展, 2019, 8(8): 1410-1417.
DOI: 10.12677/aam.2019.88165

**况，结合快速独立分量分析方法(FastICA)和线性判别分析(LDA)提取数据特征的优势，提出了一种基于改进朴素贝叶斯分类算法的信用评估模型——FastICA-LDA-NB。将该模型应用于UCI上的德国个人信用数据集，在精确率、召回率两个评价指标值上表明所提模型具有较好的信用评估效果。**

## 关键词

**信用评估，独立分量分析，线性判别分析，朴素贝叶斯**



## 1. 引言与文献综述

随着近年来互联网技术的快速发展，涌现出很多贷款模式，比如 b2C 贷款、P2P 贷款、P2C 贷款等。这些贷款模式的出现给人们带来方便的同时也增加了相关金融机构的风险压力，金融机构只有精确地对个人进行信用评估后才能降低贷款的风险。然而，个人信用由多方面因素组成，其中就包括个人基本信息、信贷记录、个人财产等。因此在复杂的因素环境中设计出一个好的信用评估模型对于金融机构来说具有重要的意义。

常用于构建信用评估模型的机器学习方法有朴素贝叶斯(NB)、多元回归分析、逻辑回归、神经网络等。比如朱毅峰等提出了融合 BP 神经网络和概率神经网络的微型企业信用评估模型，实验结果表明该模型能够降低对差企业的误判率，具有重要的实际价值意义[1]。姜明辉等利用多元回归分析方法建立了是否获得贷款和其他变量的关系，基于该方法能够较为准确的判断是否给个人发放贷款[2]。张国政等使用逻辑回归模型分析了商业银行的信贷数据，研究发现影响个人信用的关键因素主要有贷款金额、婚姻状况等六项指标[3]。李旭升等在信用评估的实验中对比了 David West 的五种神经网络和三种 NB 算法，结果表明 NB 算法具备更好的信用评估能力[4]。

其中 NB 算法因其扎实的概率论基础、简单的模型结构和稳定的分类能力等优点，在很多分类任务中得到了广泛的应用。然而 NB 算法需要满足特征之间独立，这在现实中很难满足，往往是特征之间越不独立，其分类效果就越差。针对此问题，模型融合是一种有效克服特征独立性假设问题的策略。比如叶晓枫等首先通过随机森林算法对信用数据集提取特征，然后用 NB 算法训练所提取的特征，研究表明此两种算法的融合能够提高信用预测精度[5]。徐岫等从三个不同角度提出了三种改进的 NB 信用评估模型，并在其实验中取得较好的评估效果[6]。秦锋等将快速独立分量分析(FastICA)与 NB 算法融合得到 FastICA-NB 模型，该模型将原始样本投影到独立的特征空间，从而获得具有独立性的新样本，进而改善分类效果[7]。李楚进等针对独立性要求采用 PCA 的改进方法得到了融合的 PCA-NB 模型，该模型将原始特征变换为不相关特征，提高了 NB 的分类效果[8]。

独立分量分析(ICA)作为盲源分离的一种重要统计方法，最大特点就是通过其处理后的数据具备独立性，能够满足 NB 算法对特征独立性的要求。随着 ICA 的快速发展，学者们已经提出很多种 ICA 的变体，其中用得最多的当属 HyvarinenAapo 提出的 FastICA [9]。线性判别分析(LDA) [10]作为一种降维方法，优势在于它是一种有监督学习方法，经其降维后不同类别的样本尽可能分开。

综上所述，结合 FastICA 和 LDA 的优势，提出一种基于模型融合改进的 NB 分类算法——FastICA-LDA-NB。将所提模型应用于 UCI 上的德国个人信用数据集，实验结果表明 FastICA-LDA-NB 具有较好的个人信用评估能力。

## 2. 相关理论基础

### 2.1. 朴素贝叶斯算法

假设分类任务的输入特征为 $n$ 维向量 $\left\{x^{(1)},x^{(2)},\cdots,x^{(n)}\right\}$，输出类别 $y=\{c_1,c_2,\cdots,c_K\}$，则朴素贝叶斯公式为：

$$P\left(y\mid x^{(1)},x^{(2)},\cdots,x^{(n)}\right)=\frac{P(y)\prod_{i=1}^{n}P\left(x^{(i)}\mid y\right)}{P\left(x^{(1)},x^{(2)},\cdots,x^{(n)}\right)} \tag{1}$$

考虑到 $P\left(x^{(1)},x^{(2)},\cdots,x^{(n)}\right)$ 对所有类别都相同，因此式(1)可转化为如下 NB 算法的学习模型：

$$\hat{y}=\arg\max_{y}P(y)\prod_{i=1}^{n}P\left(x^{(i)}\mid y\right) \tag{2}$$

由式(2)可得，朴素贝叶斯分类的学习意味着估计 $P(y)$ 和 $P\left(x^{(i)}\mid y\right)$。设 $N$ 表示样本总数，由极大似然估计得到两个概率值为：

$$P\left(y=c_k\right)=\frac{\sum_{j=1}^{N}I\left(y_j=c_k\right)}{N},k=1,2,\cdots,K \tag{3}$$

$$P\left(x^{(i)}=a_{il}\mid y=c_k\right)=\frac{\sum_{j=1}^{N}I\left(x_j^{(i)}=a_{il},y_j=c_k\right)}{\sum_{j=1}^{N}I\left(y_j=c_k\right)} \tag{4}$$

其中 $a_{il}$ 表示第 $i$ 个特征 $x^{(i)}$ 在集合 $\left\{a_{i1},a_{i2},\cdots,a_{iS_i}\right\}$ 中的第 $l$ 个值，$i=1,2,\cdots,n$；$l=1,2,\cdots,S_i$；且 $n$ 表示特征总数，$S_i$ 为特征 $x^{(i)}$ 取值集合的大小。式(3)和(4)中概率估计值的分子可能会出现零，对后验概率计算产生影响。可采用贝叶斯估计解决该问题，即式(3)、(4)分别修改为：

$$P_\lambda\left(y=c_k\right)=\frac{\sum_{j=1}^{N}I\left(y_j=c_k\right)+\lambda}{N+K\lambda} \tag{5}$$

$$P_\lambda\left(x^{(i)}=a_{il}\mid y=c_k\right)=\frac{\sum_{j=1}^{N}I\left(x_j^{(i)}=a_{il},y_j=c_k\right)+\lambda}{\sum_{j=1}^{N}I\left(y_j=c_k\right)+S_i\lambda} \tag{6}$$

其中 $\lambda>0$，特别地当 $\lambda=1$ 时为拉普拉斯平滑。

### 2.2. 独立分量分析

ICA 是信号处理领域的一种统计方法，其主要任务是把混合信号分解成若干个独立的信号。图 1 为 ICA 模型的简单表示，其中 $S=\left[s_1,s_2,\cdots,s_n\right]$ 为 $n$ 维未知独立成分，一般假设该 $n$ 维分量的均值为 0 方差为 1，$A$ 为 $m\times n$ 维的未知混合矩阵，$X=\left[x_1,x_2,\cdots,x_m\right]^{\mathrm{T}}$ 为 $m$ 维观测变量，$W$ 为解混矩阵，ICA 的主要任务是求解矩阵 $W$，使得输出 $Y=W^{\mathrm{T}}X$ 是 $S$ 的最佳逼近。

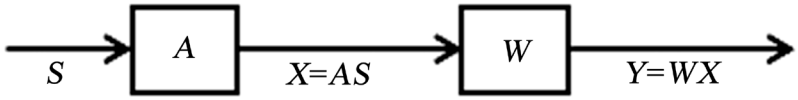


**Figure 1.** Simple block diagram of independent component analysis
**图 1.** 独立分量分析简单框图

评价 $Y$ 中各分量独立性的常见准则有信息最大化、负熵最大化等。基于这些评价标准，衍生出了很多 ICA 的实现方法，比如 JADE、FastICA [11]等。本文采用负熵最大化的 FastICA，负熵定义为：

$$J(Y)=H(Y_{Gauss})-H(Y) \tag{7}$$

其中 $H(Y)=-\int P_Y(\eta)\log P_Y(\eta)\mathrm{d}\eta$ 表示 $Y$ 的熵，$P_Y(\cdot)$ 为 $Y$ 的密度函数，$Y_{Gauss}$ 是与 $Y$ 具有相同协方差阵的高斯随机变量，负熵一般近似为：

$$J(Y)\propto\left\{E\left[G(Y)\right]-E\left[G(Y_{Gauss})\right]\right\}^2 \tag{8}$$

$G(\cdot)$ 是一个非二次函数，常用的形式有如下三个[12]

$$\begin{aligned}G_1(y)&=\frac{1}{a}\log\cosh(ay)\\G_2(y)&=\exp\left(-\frac{1}{2}ay^2\right)\\G_3(y)&=ay^4\end{aligned} \tag{9}$$

其中 $a$ 通常取区间(1,2)的常数。

要使(7)式最大，只需 $E\left[G(Y)\right]=E\left[G\left(W^{\mathrm{T}}X\right)\right]$ 最大，根据 Kuhn-Tucker 条件，在 $E\left[\left(W^{\mathrm{T}}X\right)^2\right]=\|W\|^2=1$ 条件下，当 $E\left[G\left(W^{\mathrm{T}}X\right)\right]$ 最大时有：

$$F(W)=E\left[XG\left(W^{\mathrm{T}}X\right)\right]-\beta W=0 \tag{10}$$

式中 $\beta$ 为恒定值，$\beta=E\left[W^{\mathrm{T}}XG\left(W^{\mathrm{T}}X\right)\right]$，由牛顿法对式(10)求极大值得 $W$ 的迭代式为

$$W^*=W-\frac{E\left[XG\left(W^{\mathrm{T}}X\right)\right]-\beta W}{E\left[G'\left(W^{\mathrm{T}}X\right)\right]-\beta} \tag{11}$$

求出 $W$ 之后，根据 $Y=W^{\mathrm{T}}X$ 便可得出独立成分 $S$ 的估计。

## 2.3. LDA 降维

LDA 是一种监督降维方法，其核心思想是通过广义特征值分解把高维特征映射到低维特征，使得在低维特征空间中的异类样本尽可能分离。设 $N$ 为样本总量，$K$ 为类别总数，$N_k$ 为第 $k$ 类样本量，$\overline{X_k}$ 为第 $k$ 类样本均值向量，$\bar{X}$ 为全体样本均值向量，$X_{(k)j}$ 为类别 $k$ 中第 $j$ 个样本。LDA 的计算步骤如下：

a) 计算第 $k$ 类样本的协方差阵

$$S_k=\sum_{j=1}^{N_k}\left(X_{(k)j}-\overline{X_k}\right)\left(X_{(k)j}-\overline{X_k}\right)^{\mathrm{T}} \tag{12}$$

b) 计算类内散度矩阵

$$S_w = \sum_{k=1}^{K} \frac{N_k}{N} S_k \tag{13}$$

c) 计算类间散度矩阵

$$S_b = \sum_{k=1}^{K} \frac{N_k}{N} \left(\overline{X_k} - \overline{X}\right)\left(\overline{X_k} - \overline{X}\right)^{\mathrm{T}} \tag{14}$$

d) 计算 $S_w^{-1} S_b$ 的前 $d$ 个特征值对应的特征向量，组成投影矩阵 $W$

e) 计算训练样本 $X$ 降维后的特征数据 $Y = W^{\mathrm{T}} X$

### 2.4. FastICA-LDA-NB 信用评估模型

结合 FastICA 能够分离出独立分量以及 LDA 算法作为有监督降维的优势，将 FastICA 与 LDA 的结合作为 NB 算法的预处理系统，提高了 NB 算法在信用评估中的性能。FastICA-LDA-NB 信用评估的设计流程如图 2：

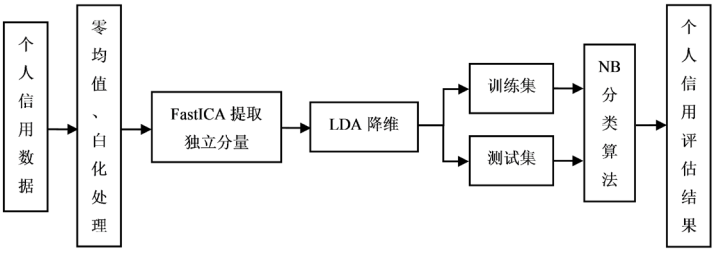


**Figure 2.** FastICA-LDA-NB credit evaluation process
**图 2.** FastICA-LDA-NB 信用评估流程

FastICA-LDA-NB 模型的信用评估过程如下：

a) 对信用数据进行零均值、白化处理，使得模型在提取独立分量阶段取得更好的收敛性。

b) 通过 FastICA 方法提取独立分量。

c) 利用 LDA 对所提取的独立分量进行降维处理。

d) 对降维后的数据随机划分为训练集和测试集。

e) 利用训练集训练 NB 算法模型。

f) 使用训练后的 NB 算法对测试集测试，从而获得个人信用评估结果。

## 3. 实验与结果分析

### 3.1. 实验数据与评价指标

本文选取 UCI 上的德国某银行信用卡个人信用数据集 german.data-numeric，该数据集由 700 个信用好的用户和 300 个信用差的用户组成。每个样本由客户的信用账户额度、年龄、资产状况等 24 个特征组成。该数据集质量好，不存在缺失情况。

为了比较各模型的信用评估效果，选取准确率 $P$、召回率 $R$ 作为评价指标。各指标的具体计算公式为

$$\begin{aligned} P &= \frac{TP + TN}{TP + FP + TN + FN} \\ R &= \frac{TP}{TP + FN} \end{aligned} \tag{15}$$

其中 $TP$、$TN$、$FP$、$FN$ 分别表示真正例、真反例、伪正例、伪反例。

### 3.2. 实验过程

对于模型选取：为了验证提出模型的可行性，采用 python 编程语言实现各个模型，以 NB、PCA-NB、FastICA-NB 和 RF-NB 四个模型作为参照模型，比较 FastICA-LDA-NB 模型在信用评估中的效果，每个模型重复 50 次实验。

对于各模型参数：FastICA 中的 $G(\cdot)$ 函数形式选取式(9)中的 $G_1(y)$ 形式、$a=1.5$，解混矩阵 $W$ 随机初始化；PCA 中主成分的选取原则为特征值贡献率大于 90%；RF 中决策数为 500 个、决策树最大深度为 5。

对于数据集划分：每次实验随机划分训练集和测试集，其中测试集占 20%。

### 3.3. 实验结果与分析

为了更直观地了解各模型的分类效果，画出了 50 次实验中准确率 $P$、召回率 $R$ 两个指标的箱线图，如图 3、图 4 所示：

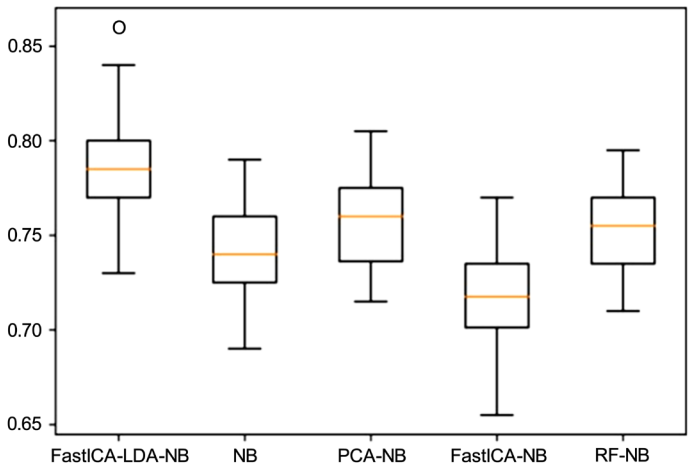


**Figure 3.** Boxplot with accuracy $P$
**图 3.** 准确率 $P$ 的箱线图

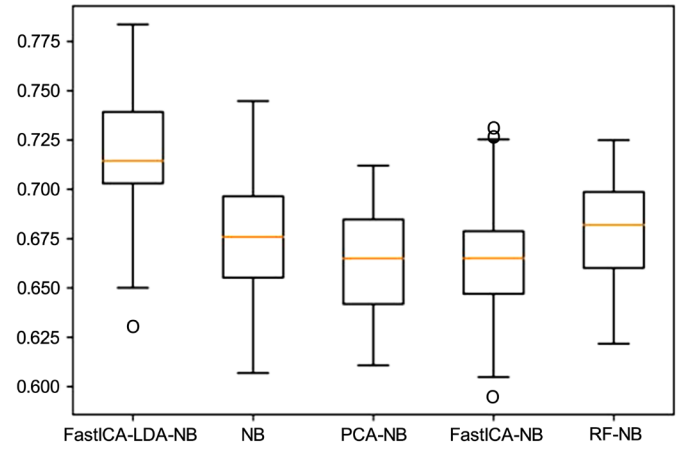


**Figure 4.** Boxplot of recall rate $R$
**图 4.** 召回率 $R$ 的箱线图

对 50 次实验后的 $P$、$R$ 取平均得表 1：

**Table 1.** Comparison of *P* and *R* indicators of five models (%)
**表 1.** 五个模型的 *P*、*R* 指标比较(%)

| | FastICA-LDA-NB | NB | PCA-NB | FastICA-NB | RF-NB |
|---|---|---|---|---|---|
| *P* | 78.24 | 74.21 | 75.72 | 71.85 | 75.29 |
| *R* | 71.56 | 67.78 | 66.27 | 66.13 | 67.81 |

从上面图 3、图 4 的箱线图可以看出，在 50 次实验中所提模型 FastICA-LDA-NB 的两个评价指标 *P*、*R* 整体上优于 NB、PCA-NB、FastICA-NB 和 RF-NB 四个模型。由表 1 得出提出的模型较其他四个模型在准确率 *P* 上提高了大约 2.5%~6.5%，在召回率 *R* 上提高了大约 3.5%~4.5%，表明 FastICA-LDA-NB 模型在个人信用评估中有较好的实用性。

由 FastICA-LDA-NB 与 FastICA-NB 对比可以看出引入 LDA 进行降维的恰当性，充分利用了 LDA 属于有监督降维的优势。FastICA-NB 相对于 NB 来说没有提升精度，可能是提取独立分量后的特征存在信息冗余造成的，因为经过特征降维或特征选择处理后的 PCA-NB 和 RF-NB 两种模型均较 NB 来说有所提高。

FastICA 能够分离出相互独立的特征，刚好满足 NB 算法的独立性假设。但是不确定分离出来的特征对分类任务的重要性，如果存在特征冗余问题，那么反而会降低 NB 算法的分类准确率。所以本文有机结合了 FastICA、LDA 和 NB 的优势，构造了 FastICA-LDA-NB 模型，并取得较好的实验效果，因此在个人信用评估领域有一定的价值意义。

## 4. 结语

构建科学的信用评估模型，可以帮助银行等金融机构更加准确地判断客户信用状况，进而作出合理的决策，有效避免了风险。目前 NB 算法被广泛应用于信用评估领域，但其对属性独立性假设的要求很难满足。因此本文应用了盲源分离领域的 FastICA 算法，该算法能够分离出具有独立性分量，恰好满足 NB 算法的要求。考虑到数据集存在特征冗余的情况，在 FastICA 的基础上引入 LDA，该方法属于有监督降维，降维后的特征能够使不同类别之间的样本尽可能分离。综上所述，本文综合利用了 FastICA、LDA 和 NB 三者的优势，提出了 FastICA-LDA-NB 的信用评估模型。在德国个人信用数据集上的结果表明，提出模型不仅提高了 NB 算法的分类精度，而且与文献中几种 NB 的融合模型相比也取得较好的分类效果，表明了 FastICA-LDA-NB 模型对于提高金融机构的风险管理具有一定的意义。

此外，本文提出模型虽然有所改善，但准确率 *P* 和召回率 *R* 的提升幅度不够大，仍有提升的空间。下一步研究将会考虑深度学习中的表征学习思想，把 FastICA-LDA-NB 集成具有深层结构的模型，使其能够充分挖掘信用数据中的信息，进而构建更加科学的信用评估模型。

## 基金项目

国家自然科学基金项目(11771321)；山西省社会发展科技攻关计划项目(201703D321032)。

## 参考文献

[1] 朱毅峰, 孙亚南. 基于神经网络的微型企业信用评估特征选择及其效果评价[J]. 统计与信息论坛, 2008, 23(4): 48-51.

[2] 姜明辉, 姜磊, 王雅林. 线性判别式分析在个人信用评估中的应用[J]. 管理科学, 2003, 16(1): 53-55.

[3] 张国政, 陈维煌, 刘呈辉. 基于 Logistic 模型的商业银行个人消费信贷风险评估研究[J]. 金融理论与实践, 2015(3): 53-57.

[4] 李旭升, 郭耀煌. 基于朴素贝叶斯分类器的个人信用评估模型[J]. 计算机工程与应用, 2006, 42(30): 197-201.

[5] 叶晓枫, 鲁亚会. 基于随机森林融合朴素贝叶斯的信用评估模型[J]. 数学的实践与认识, 2017, 47(2): 68-73.

[6] 徐岫. 基于改进朴素贝叶斯方法的个人信用评估研究[D]: [硕士学位论文]. 武汉: 华中科技大学, 2015.

[7] 秦锋, 任诗流, 程泽凯, 等. 基于 ICA 方法的朴素贝叶斯分类器[J]. 计算机工程与设计, 2007, 28(20): 4873-4874.

[8] 李楚进, 付泽正. 对朴素贝叶斯分类器的改进[J]. 统计与决策, 2016(21): 9-11.

[9] Hyvarinen, A. (1999) Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. *IEEE Transactions on Neural Networks*, **10**, 626-634. https://doi.org/10.1109/72.761722

[10] 董虎胜. 主成分分析与线性判别分析两种数据降维算法的对比研究[J]. 现代计算机, 2016(29): 36-40.

[11] 万坚, 涂世龙, 廖灿辉, 等. 通信混合信号盲分离理论与技术[M]. 北京: 国防工业出版社, 2012: 36-60.

[12] He, X.S., He, F. and He, A.L. (2017) Super-Gaussian BSS Using Fast-ICA with Chebyshev-Pade Approximant. *Circuits, Systems and Signal Processing*, **37**, 305-341. https://doi.org/10.1007/s00034-017-0554-1