doi：10.13682/j.issn.2095-6533.2017.03.005

基于主成分分析和分类回归树的客户欠费预测

卢光跃，董静怡，岳 赞，刘 迪

（西安邮电大学 陕西省信息通信网络及安全重点实验室，陕西 西安 710121）

摘 要：针对电信数据维度增加导致的客户欠费预测算法复杂度过高的问题，提出基于主成分分析和分类回归树的电信客户欠费预测算法。该算法将原始电信数据进行数据缺失值处理、数据冗余识别和数据结构化后，进行数据规范化建模，利用主成分分析算法对建模后的电信数据进行降维处理，将降维后的数据作为分类回归树算法的输入数据对客户是否欠费进行分类，预测客户是否将存在欠费行为。利用实际电信数据进行验证，结果表明该算法的预测错误率为4.49%，预测耗时为17.05 s，与分类回归树算法相比，在能够预测客户欠费的同时，还能提高预测效率。

关键词：客户欠费预测；电信数据；主成分分析；分类回归树

中图分类号：TP18 **文献标识码**：A **文章编号**：2095-6533(2017)03-0029-05

Customer owing fee prediction based on the classification and regression tree of principal component analysis

LU Guangyue， DONG Jingyi， YUE Yun， LIU Di

（Shaanxi Key Lab of Information Communication Network and Security，
Xi'an University of Posts and Telecommunications，Xi'an 710121，China）

Abstract：Aiming at the problem of the customer's arrears forecasting algorithm caused by the increase of the dimension of the telecommunication data，an algorithm for estimating the arrears of telecommunication customers based on principal component analysis and classification and regression tree is proposed. The data of the original telecommunication data is processed by data loss，data redundancy identification and data structure are used to normalize the data. The principal component analysis algorithm is used to reduce the dimension data of the modeled telecommunication data. The input data as a classification and regression tree algorithm classifies the customer owes arrears. The results of the classification of the actual telecommunication data show that the prediction error rate is 4.49% and the prediction time is 17.05 seconds. Compared with the prediction efficiency of the classification and regression tree algorithm，the efficiency of forecasting is improved while the customers of arrears are effectively identified.

Keywords：customer owe fee prediction，telecommunication data，principal component analysis，classification and regression tree

客户恶意欠费问题严重危害着运营商的发展[1]。运营商开始借助数据挖掘技术进行电信客户欠费预测[2]，根据已有的电信数据，应用数据挖掘算法[3]进行客户行为分析[4]，预测可能的欠费客户，减少客户恶意欠费对运营商造成的损失。

可用的预测方法主要有人工神经网络算法

收稿日期：2017-02-03

基金项目：陕西省工业科技攻关计划资助项目(2016GY-113，2015GY-013)

作者简介：卢光跃(1971－)，男，博士，教授，从事信号处理研究。E-mail：tonylugy@163.com

董静怡(1992－)，女，硕士研究生，研究方向为宽带无线通信技术。E-mail：1439688341@qq.com

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(artificial neural networks, ANNs)[5]、支持向量机算法(support vector machine, SVM)[6]、贝叶斯算法[7]和分类回归树算法[8-9]等。ANNs 算法对大规模数据具有较好的拟合效果,且无需任何先验知识,但其所依据的经验风险最小化原则,容易导致泛化能力下降且模型结构难以确立[5]。SVM 算法对小样本数据的测试环境适应能力强,分类正确率高,能有效防止过学习,但 SVM 算法求解是一个最优化过程,算法复杂度高[10]。贝叶斯算法适合处理海量数据,但数据集必须满足各属性之间相互独立的前提条件[11]。分类回归树算法不仅可直观对分类规则进行解释,而且可在相对短的时间内对大型数据源做出可行且效果良好的处理[12],但随着电信业务的不断扩张、复杂性的不断增加,使得电信数据具有高维特性,导致运用分类回归树算法进行客户欠费预测耗时较长。

本文拟提出一种降维的电信客户欠费预测算法。首先,对电信数据进行预处理及规范化建模;其次,应用主成分分析算法[13],进行数据降维处理;再次,将降维后的数据作为分类回归树算法的输入数据,构建出基于主成分分析和分类回归树的电信客户欠费预测算法;最后,通过实际电信数据评估验证所提预测算法的有效性。

1 电信数据的预处理及规范化建模

电信数据中包含客户基本属性、客户价值和客户消费信息这 3 类信息。在对原始电信数据进行客户欠费预测时,原始电信数据中存在数据缺失,如原始电信数据中存在某些无记录的数据样本;数据冗余,如所属城市编码和所属城市名称为同一属性的不同表现形式;数据非结构化,如是否 2G 转 3G 用户、是否公务测试、是否离网、是否主动离网、是否被动离网、是否主动停机的取值为“是”和“否”;数据不规范,如短信费用属性与入网时长属性具有不同量纲。针对这些问题,在应用电信数据之前,首先对原始电信数据进行数据缺失值处理、数据冗余识别和数据结构化[14]后进行数据规范化建模,电信数据预处理流程如图 1 所示。

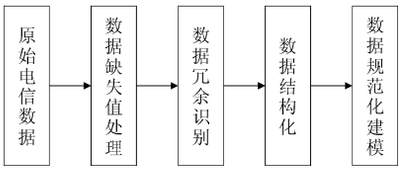

图 1 电信数据预处理流程

1.1 数据缺失值处理

数据缺失值处理的方法可分为 3 类:删除、搁置和插补。若原始电信数据中包含缺失值的数据样本较少,则删除该数据样本是最有效的方法。若要求数据具有较高的完备性,则采用搁置方式对原始电信数据进行处理。若原始电信数据缺失值较多且对数据完备性具有一定要求时,选取数据样本的中值、中位数或固定值等补足缺失处数据,实现数据插补。电信数据具有数据量大、缺失值少的特性,选取删除的方法处理电信数据的缺失值。

1.2 数据冗余识别

数据冗余识别是通过检索原始电信数据中多次出现的同一属性以及不同表现形式的同一属性,剔除重复属性。在维持数据信息完整性的同时,消除数据冗余。例如,当所采用的电信数据中所属城市编码和所属城市名称属于同一属性的不同表现形式时,剔除所属城市名称属性。

1.3 数据结构化

数据结构化是将原始电信数据中的某些文字表现的非结构化数据选取合适规则进行结构化处理,使得数据在保证完备性的前提下更加适应于所采用的算法。针对电信数据,结构化处理规则如下:是否 2G 转 3G 用户、是否公务测试、是否离网、是否主动离网、是否被动离网、是否主动停机、是否欠费属性取值为是(1)、否(0);付费方式取值为预付费(1)、后付费(0);用户属性取值为公众客户(0)、集团客户(1);渠道类型取值为电话营销(0)、电子渠道(1)、集团营销(2)、社会渠道(3)、社区渠道(4)、自由渠道(5)、社区直营(6)、其他(7)。

1.4 数据规范化建模

对原始电信数据进行数据缺失值处理、数据冗余识别、数据结构化后,为了消除因电信数据量纲不同、自身变异或数值相差过大对客户欠费预测算法带来的不良影响,进行数据规范化建模。

数据规范化建模的原理是,将具有 m 个样本,n 个属性的原始电信数据表示为矩阵

$$\boldsymbol{Y}_{m\times n}=\begin{bmatrix} y_{11} & \cdots & y_{1n} \\ \vdots & \ddots & \vdots \\ y_{m1} & \cdots & y_{mm} \end{bmatrix}。 \tag{1}$$

矩阵中任意一行元素代表原始电信数据某一样本的全部数据,任意一列元素代表某一属性的所有数据。对矩阵 $\boldsymbol{Y}_{m\times n}$ 进行中心化处理,即令

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$y'_{ij}=\frac{y_{ij}-\frac{1}{m}\sum_{i=1}^{m}y_{ij}}{\sqrt{\sum_{j=1}^{n}\left(y_{ij}-\frac{1}{m}\sum_{i=1}^{m}y_{ij}\right)^2}}。\tag{2}$$

处理后生成规范化矩阵

$$\boldsymbol{Y}'_{m\times n}=\begin{bmatrix}y'_{11}&\cdots&y'_{1n}\\ \vdots&\ddots&\vdots\\ y'_{m1}&\cdots&y'_{mn}\end{bmatrix}。\tag{3}$$

2　基于主成分分析和分类回归树的电信客户欠费预测算法

利用主成分分析算法对电信数据 $\boldsymbol{Y}'_{m\times n}$ 进行降维处理，将降维后的数据利用分类回归树算法划分为欠费类和不欠费类。

2.1　主成分分析

主成分分析算法是在保证数据信息较小损失的前提下，对数据进行降维处理，而数据信息主要表现在数据的方差上，方差越大，包含的信息量越多，故将累积方差贡献率的大小即衡量所包含信息量的多少作为指标，选取主成分个数，确立降维矩阵，将数据转化为低维数据，使得这个低维数据尽可能多的反映原数据的信息[15]。主成分分析算法处理步骤如下：

（1）计算电信数据 $\boldsymbol{Y}'_{m\times n}$ 的协方差矩阵

$$\boldsymbol{P}_{n\times n}=(\boldsymbol{Y}'_{m\times n})^{\mathrm{T}}\boldsymbol{Y}'_{m\times n}。\tag{4}$$

（2）确定主成分个数。对电信数据的协方差矩阵 $\boldsymbol{P}_{n\times n}$ 进行特征值分解，并将其 n 个特征值进行降序排列，得到 $\lambda_1\geqslant\lambda_2\geqslant\lambda_n$，与此对应的特征向量记为 $\boldsymbol{\varepsilon}_1$，$\boldsymbol{\varepsilon}_2$，…，$\boldsymbol{\varepsilon}_n$。计算 p 个主成分数据包含的信息占全部电信数据信息的比重，记为累计方差贡献率

$$\mu(p)=\sum_{k=1}^{p}\frac{\lambda_k}{\sum_{k=1}^{n}\lambda_k}。\tag{5}$$

其中，λ_k 表示降序排列的特征值中排在第 k 位的特征值的大小；$\mu(p)$ 用来衡量主成分信息的完备程度，可以依据其大小选取主成分个数 p，使得所选取的 p 个主成分能够包含电信数据矩阵 $\boldsymbol{Y}'_{m\times n}$ 中绝大部分信息。实际应用中，根据所设计系统的精度要求选取 p，当精度要求一般时根据 $\mu(p)>80\%$ 选取 p，而当精度要求较高时则根据 $\mu(p)>90\%$ 选取 p。

（3）数据降维处理。根据确定的主成分个数 p，选取 $\boldsymbol{\varepsilon}_1$，$\boldsymbol{\varepsilon}_2$，…，$\boldsymbol{\varepsilon}_n$ 的前 p 个特征向量组成的降维矩阵 $\boldsymbol{Q}_{n\times p}$，用以实现对 $\boldsymbol{Y}'_{m\times n}$ 的降维处理，即

$$\boldsymbol{X}_{m\times p}=\boldsymbol{Y}'_{m\times n}\times\boldsymbol{Q}_{n\times p}。\tag{6}$$

利用以上步骤处理后，n 个属性的电信数据通过降维成为具有 p 个属性的数据。

2.2　分类回归树算法处理

将电信数据经降维处理后的数据矩阵 $\boldsymbol{X}_{m\times p}$ 作为分类回归树算法的输入数据集进行分类。分类回归树算法采用最小 GINI 系数[16]选择内部节点的分类属性。根据类别属性的取值是连续的还是离散的，分类回归树算法生成的决策树可分为分类树和回归树[17]。客户是否欠费属于离散属性，因此采用分类树。其方法如下：

确定数据集 $\boldsymbol{X}_{m\times p}$ 的某一属性 Z，将其 m 个取值进行排序，取两相邻值的平均值点作为分隔点，记为 $\eta_s(s=1,2,\cdots,m-1)$。将数据集 $\boldsymbol{X}_{m\times p}$ 按照在属性 Z 上的取值划分为大于 η_s 和小于等于 η_s 的两个数据子集 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$，记这种分类方法的 GINI 系数为

$$G_Z^{\eta_s}(\boldsymbol{X})=\frac{|\boldsymbol{X}_1|}{p}I(\boldsymbol{X}_1)+\frac{|\boldsymbol{X}_2|}{p}I(\boldsymbol{X}_2)。\tag{7}$$

其中

$$I(\boldsymbol{X}_j)=1-\sum_{i=1}^{2}\left(\frac{|C_i|}{|\boldsymbol{X}_j|}\right)^2(j=1,2)\tag{8}$$

代表数据集 $\boldsymbol{X}_j$ 的纯度，而 $|\boldsymbol{X}_j|$ 表示数据集 $\boldsymbol{X}_j$ 的总样本个数，$|C_i|$ 为数据集 $\boldsymbol{X}_j$ 中属于 C_i 类的样本个数。遍历 p 个属性以及 $m-1$ 个分隔点的 GINI 系数 $G_Z^{\eta_s}(\boldsymbol{X})$，选取 GINI 系数最小的分隔点划分数据集 $\boldsymbol{X}_{m\times p}$。

依照此方法递归建立树的子节点，如此循环直到所有子节点的样本属性属于同一类别或无可选的分裂属性为止。

利用分类回归树算法将数据集 $\boldsymbol{X}_{m\times p}$ 按照选择 GINI 系数最小的原则，可将电信客户划分为欠费类客户 C_1 和不欠费类客户 C_2。

3　实证结果及分析

采用陕西某电信运营商 2013 年 5 月份的客户消费数据，通过数据建模，得到具有 6 999 条样本，92 个属性的电信数据。在 Matlab 2012R 平台环境下，从电信数据建模后的数据中随机选取 2 400 条不欠费数据和 200 条欠费数据构成训练数据集，对基于主成分分析和分类回归树的电信客户欠费预

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

测算法进行训练，当主成分个数 p 为 2 时电信数据的累积方差贡献率可以达到 98%，由此确定算法中的主成分个数为 2。构建 5 组测试数据集，每组共包含 650 条数据，其中包括 600 条不欠费数据和 50 条欠费数据，对算法的预测性能进行评估验证。

对基于主成分分析和分类回归树的电信客户欠费预测算法进行评估验证时，需通过一些指标对算法预测效率即准确性及运行效率进行量化分析。

各指标的计算表达式分别为

$$e_l = \frac{w_l}{650}, \tag{9}$$

$$a = \frac{1}{5}\sum_{l=1}^{5} e_l, \tag{10}$$

$$h = \frac{1}{5}\sum_{l=1}^{5} t_l。 \tag{11}$$

其中，e_l 为第 l 组（$l = 1,2,3,4,5$）的预测错误率；w_l 为第 l 组测试数据集中预测错误的样本个数；a 为预测平均错误率；h 为预测平均耗时；t_l 为第 l 组的预测耗时，即测试数据集完成客户欠费预测所需的时间。根据这些指标对基于主成分分析和分类回归树的电信客户欠费预测算法进行评价。

为了验证所提算法的有效性，对 5 组测试数据集分别应用该算法与分类回归树算法进行电信客户欠费预测，预测效率如图 2 和图 3 所示。

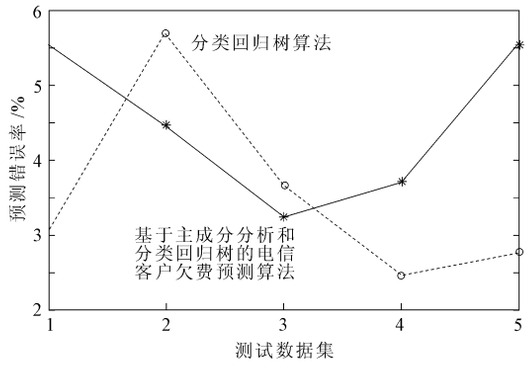

图 2 预测错误率对比

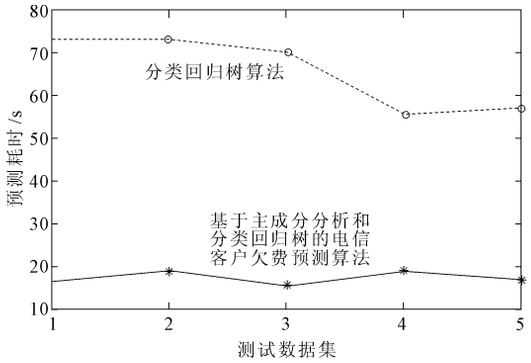

图 3 预测耗时对比

图 2 显示，分类回归树算法预测的平均错误率为 3.54%，基于主成分分析和分类回归树预测算法的预测平均错误率为 4.49%，2 个算法的预测平均错误率差别较小。

图 3 显示，分类回归树算法预测的平均耗时为 65.66 s，基于主成分分析和分类回归树算法的预测平均耗时为 17.05 s。

对比图 2 和图 3 可知，利用基于主成分分析和分类回归树的电信客户欠费预测算法进行客户欠费预测时，可在基本保证预测精准度的前提下，有效降低客户欠费预测耗时。

4 结语

针对电信数据数据量大、维度高的特点，提出了一种基于主成分分析和分类回归树算法的电信客户欠费预测算法。通过主成分分析将具有 92 个属性的电信数据，降维为 2 个属性的数据，将此数据作为分类回归树算法的输入数据进行客户欠费预测，预测平均错误率为 4.49%，预测平均耗时为 65.66 s，与分类回归树算法的预测平均错误率为 3.54%，预测平均耗时为 17.05 s 相比，预测精准度相差 0.95%，预测耗时降低 74%。

参 考 文 献

[1] 王昕. 数据挖掘技术在提升电信业欠费预测及控制能力中的应用[J/OL]. 电信技术，2014(7)：36-39[2016-09-27]. http://dx.chinadoi.cn/10.3969/j.issn.1000-1247.2014.07.008.

[2] 包志强，崔妍. 电信客户欠费模型评估[J/OL]. 西安邮电大学学报，2015，20(4)：97-101[2016-08-17]. http://dx.chinadoi.cn/10.13682/j.issn.2095-6533.2015.04.020.

[3] KRZYWICKI A，WOBCKE W，BAIN M，et al. Data mining for building knowledge bases：techniques，architectures and applications[J/OL]. The Knowledge Engineering Review，2016，1(2)：1-27[2016-12-27]. https://www.researchgate.net/publication/299552854_Data_mining_for_building_knowledge_bases_Techniques_architectures_and_applications. DOI：10.1017/S0269888916000047.

[4] 谷红勋，杨珂. 基于大数据的移动用户行为分析系统与应用案例[J/OL]. 电信科学，2016，32(3)：139-146

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[2016-12-27]. http://d.wanfangdata.com.cn/Periodical/dxkx201603024. DOI: 10.11959/j.issn.1000-0801.2016039.

[5] **张明光,张钰.基于 ANN 伪量测建模的配电网状态估计**[J/OL].**计算机工程与应用**,2016,52(17):253-256[2016-10-29]. http://d.wanfangdata.com.cn/Periodical/jsjgcyyy201617044. DOI: 10.3778/j.issn.1002-8331.1410-0282.

[6] **李海燕. 基于支持向量机算法的股市拐点预测分析**[J/OL]. **郑州大学学报(哲学社会科学版)**,2015(1):96-99[2016-12-27]. http://www.cnki.com.cn/Article/CJFDTotal-ZZDX201501024.htm.

[7] **朱志勇,徐长梅,刘志兵,等.基于贝叶斯网络的客户流失分析研究**[J/OL].**计算机工程与科学**,2013,35(3):155-158[2016-10-27]. http://d.wanfangdata.com.cn/Periodical/jsjgcykx201303026. DOI:10.3969/j.issn.1007-130X.2013.03.026.

[8] RUTKOWSKI L, JAWORSKI M, PIETRUCZUK L, et al. The CART decision tree for mining data streams[J/OL]. Information Sciences, 2014, 266(5): 1-15[2016-09-30]. https://www.researchgate.net/publication/260212465_The_CART_decision_tree_for_mining_data_streams. DOI: 10.1016/j.ins.2013.12.060.

[9] DAI Q Y, ZHANG C P, WU H. Research of Decision Tree Classification Algorithm in Data Mining[J/OL]. Journal of East China Institute of Technology, 2016,9(5):1-8[2016-12-29]. http://www.earticle.net/Article.aspx?sn=275551. DOI: 10.14257/ijdta.2016.9.5.01.

[10] **初光磊. SVM 在数据挖掘中的应用**[D/OL]. **北京:北京邮电大学**,2015:1-15[2016-10-13]. http://d.wanfangdata.com.cn/Thesis/Y2848386.

[11] **房丙午,黄志球,李勇,等. 基于贝叶斯网络的复杂系统动态故障树定量分析方法**[J/OL]. **电子学报**,2016,44(5):1234-1239[2016-10-12]. http://d.wanfangdata.com.cn/Periodical/dianzixb201605032. DOI: 10.3969/j.issn.0372-2112.2016.05.032.

[12] **黄文思,郝悍勇,李金湖,等. 基于决策树算法的电力客户欠费风险预测**[J/OL]. **电力信息与通信技术**,2016(1):19-22[2016-12-29]. http://www.cnki.com.cn/Article/CJFDTotal-DXXH201601007.htm.

[13] JAGADISH, RAY A. Optimization of process parameters of green electrical discharge machining using principal component analysis (PCA)[J/OL]. International Journal of Advanced Manufacturing Technology, 2015,87(5): 1299-1311[2016-06-26]. https://link.springer.com/article/10.1007%2Fs00170-014-6372-8. DOI:10.1007/s00170-014-6372-8.

[14] HAYETE B,BIENKOWKSA J R. Gotrees: predicting go associations from protein domain composition using decision trees[C/OL]//Proceedings of the Pacific Symposium on Biocomputing. Hawaii: PMC, 2005: 127-138[2016-06-27]. http://dx.doi.org/10.1142/9789812702456_0013.

[15] BROWN N J, LI G P, KOSZYKOWSKI M L. Mechanism reduction via principal component analysis[J/OL]. International Journal of Chemical Kinetics, 2015, 29(6):393-414[2016-06-27]. https://www.mendeley.com/research-papers/mechanism-reduction-via-principal-component-analysis/. DOI: 10.1002/(SICI)1097-4601(1997)29:6<393::AID-KIN1>3.0.CO;2-P.

[16] BALDWIN J F, LAWRY J, MARTIN T P. A mass assignment based ID3 algorithm for decision tree induction[J/OL]. International Journal of Intelligent Systems, 2015,12(7): 523-552[2016-06-27]. http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1098-111X(199707)12:7<523::AID-INT3>3.0.CO;2-N/citedby.

[17] ZHANG N, WU Y S, ZHANG Q H. Detection of sea ice in sediment laden water using MODIS in the Bohai Sea: a CART decision tree method[J/OL]. International Journal of Remote Sensing,2015,36(6):1661-1674[2016-11-27]. http://dl.acm.org/citation.cfm?id=2836950. DOI: 10.1080/01431161.2015.1015658.

[责任编辑:杨洵]

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net