计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(9):62−68 [doi: 10.15888/j.cnki.csa.007981]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于谱回归特征降维的客户流失预测①

李国祥[1,2], 蒋怡琳[1,2], 马文斌[1,2], 夏国恩[1]

[1](广西财经学院 教务处, 南宁 530003)
[2](广西师范大学 广西多源信息挖掘与安全重点实验室, 桂林 541004)
通讯作者: 蒋怡琳, E-mail: 290761452@qq.com

摘　要: 针对于大样本数据的客户流失预测, 从特征有效表达的角度, 提出了一种基于谱回归特征约简的预测模型. 模型在原始客户特征基础上, 利用基于谱回归的流形降维, 建立可区分性的低维特征空间, 在此之上采用支持向量机实现客户流失的二分类. 通过在网络客户和传统电信客户两种不同数据集上的大样本实验, 并与不同分类器、不同特征约简或选择方法的对比, 证明了该方法的有效性.

关键词: 谱回归; 客户流失; 特征约简; 分类器

引用格式: 李国祥,蒋怡琳,马文斌,夏国恩.基于谱回归特征降维的客户流失预测.计算机系统应用,2021,30(9):62–68. http://www.c-s-a.org.cn/1003-3254/7981.html

Prediction of Customer Churn Based on Spectral Regression

LI Guo-Xiang[1,2], JIANG Yi-Lin[1,2], MA Wen-Bin[1,2], XIA Guo-En[1]

[1](Department of Academic Affairs, Guangxi University of Finance and Economics, Nanning 530003, China)
[2](Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin 541004, China)

Abstract: In order to predict customer churn with large sample data, a customer churn prediction model based on spectral regression was put forward from the perspective of feature expression, which took advantage of the spectral regression to reduce the dimension of feature. On the basis of the original customer features, a distinguishing feature space of low dimension is established by using the manifold dimension reduction based on spectral regression, and then we used the support vector machine to realize the binary classification of customer churn prediction. The model was evaluated on two different data sets of network customers and traditional telecom customers, and compared with different classifiers, different feature reduction or selection methods, the experiment results verify that the model is effective.

Key words: spectral regression; customer churn; feature reduction; classifier

流失客户通常是指在一定时期内终止使用企业的服务或产品的客户, 其预测水平是衡量客户保持策略有效性和客户关系管理智能化程度的重要标志. 目前对于客户流失的研究对象主要集中在传统的电信客户流失预测和网络客户流失预测两个方面, 研究方法上主要是从特征向量选择和分类器优化两个角度构建客户流失预测模型.

在特征选择方面, 文献 [1] 针对于高维度的样本特征属性, 定义了属性满意度和属性集满意度, 通过满意度函数来开展高维特征属性的选择. 文献 [2] 基于原始

① 基金项目: 广西重点研发计划 (2018AB15003); 广西多源信息挖掘与安全重点实验室开放基金 (MIMS17-02); 广西高校中青年教师基础能力提升项目 (2018KY0520); 2019 年度广西高校中青年教师科研基础能力提升项目 (2019KY0661); 广西财经学院青年教师科研发展基金 (2018QNA02)
Foundation item: Key Research and Development Program of Guangxi (2018AB15003); Open Fund of Guangxi Key Lab of Multi-source Information Mining and Security (MIMS17-02); Basic Ability Promotion project for Young and Middle-aged Teachers in Colleges and Universities in Guangxi (2018KY0520); Year 2019, Basic Research Ability Promotion Project for Young and Middle-aged Teachers in Colleges and Universities in Guangxi (2019KY0661); Young Teachers Research and Development Fund of Guangxi University of Finance and Economics (2018QNA02)
收稿时间: 2020-10-25; 修改时间: 2020-11-23; 采用时间: 2020-12-09; csa 在线出版时间: 2021-09-02

特征引入网络客户价值特征和情感特征, 增加了客户流失预测的新的客户特征属性. 文献 [3] 以网络客户的在线评论信息为依据, 通过技术性的情感分析将其表示为积极与消极情感并作为客户流失预测新属性. 文献 [4] 针对电信数据集中存在的特征维度过高问题, 结合过滤式特征选择和嵌入式特征选择方法的优点, 提出了一种基于 Fisher 比率和预测风险准则的分步特征提取方法.

在分类器优化方面, 文献 [5] 利用分类回归树算法和自适应 Boosting 算法作为分类算法, 生成通信企业的离网客户的预测模型. 文献 [6] 改进随机森林中生成每棵树时节点划分的方法, 形成新的随机森林分类模型. 文献 [7] 将深度学习引入到客户流失预测中, 构造了基于深度神经网络的流失预测模型. 文献 [8] 通过改进粒子群算法优化支持向量机分类器. 文献 [9] 区分边界样本和非边界样本, 分别采用 K 近邻分类法与支持向量机作为分类器.

上述两类方法在不同数据集上都取得了较好的预测效果, 但随着信息管理技术在客户关系管理中的广泛应用, 客户的属性维度和记录数大规模增长, 原始实验中数据样本体量偏小, 对于预测结果科学性的解释问题日益凸显, 文献 [3] 使用京东运营商手机卡用户的在线评论作为数据源, 将评论星级、会员等级、点赞数作为特征属性, 采集样本共 10 000 余条; 文献 [1] 使用两个数据集, 第 1 个数据集通过在 UCI 中随机抽样, 获得 3333 个训练样本和 1667 个测试样本, 第 2 个数据以国内某电信公司对小灵通客户拆机停号来定义客户流失, 建立 1474 个训练样本, 966 个测试样本; 文献 [8] 选取 UCI 最常用的 8 个数据集, 每个样本集 150~1500 不等; 文献 [6] 以某电信公司 2013 年 9 月至 2014 年 2 月在网和离网的客户样本作为研究对象, 样本数量共计 7913 个; 文献 [5] 选取了 15 个可能影响客户流失的属性, 在 18 万条数据中, 在网数据和离网数据分别随机抽取 3000 条数据, 形成研究样本. 由此可见, 当前客户流失预测研究的数据源大部分为小数据集或者大样本集的抽样, 且特征维度较低. 随着大数据技术的发展, 小样本的抽样数据集已经不能满足对于预测的需要, 大样本的高维度数据计算将成为必然.

大样本的高维度数据计算核心算法包括早期的主成分分析 (Principal Component Analysis, PCA)[10], 线性判别分析 (Linear Discriminant Analysis, LDA) 等, 这类算法理论基础坚实, 且易于执行, 很多学者通过使用核技巧, 将这些线性特征提取算法扩展到核领域, 如核独立主成分分析[11]. 另一类非线性特征提取技术是流行学习方法, 例如, 局部保持投影 (Locality Preserving Projection, LPP)[12]、局部线性嵌入 (Locally Linear Embedding, LLE)[13] 等, 文献 [14] 中 Zhai 等人在 LPP 的基础上提出了一种改进的局部保持投影. 局部保持投影 (LPP) 不但具有简单、快捷等优点, 同时可以考虑到整体数据空间; 此外, LPP 算法最大程度保持了数据的局部结构, 因此在低维空间中表示的最近邻搜索极大可能与高维空间中产生的结果类似. 所以, LPP 算法在数据降维领域有相当高的实用性. 虽然 LPP 算法实用性较强, 但是却有一个不可避免的缺点: 在算法的优化过程中包含一个稠密矩阵分解计算. 这是一个非常消耗时间和计算资源的计算过程, 而谱回归 (Spectral Regression, SR)[12] 将学习嵌入函数的方式转化为一个回归框架, 避免了稠密矩阵分解这一计算过程, 同时提高了优化的效果. 因此本文提出基于谱回归的特征降维更适合大样本高维度数据的计算.

针对以上问题, 本文以网络客户数据集和传统电信客户数据集为研究对象, 从特征向量提取的角度, 提出基于谱回归局部保留投影的客户属性降维算法, 并从特征选择和分类器优化方面与不同的方法做了对比, 实验证明了算法的有效性.

1 基于谱回归的特征降维

基于谱回归的特征降维算法是针对流行结构图嵌入式的典型降维算法, 通过特征提取来构造一个能揭示数据流行的结构图, 其结构图的表示方式为一个投影矩阵, 实现将高维数据特征投影到低维子空间中, 以保持高维空间中数据间的邻近结构, 达到降维的目的. 在该算法模型中, 每个顶点都是一个样本点, 两个样本点之间的边权重采用 K 近邻法计算两个样本点之间的邻接程度, 因此对数据的完整性保持较好.

1.1 局部保形投影

局部保形投影算法 (LPP) 应该被视为 PCA 的替代方法. PCA 是一种经典的线性技术, 他沿着最大方差的方向投影数据. 当高维数据位于嵌入外围空间的低维流形上时, 通过求流行上 Laplace Beltrami 算子特征函数的最优特征逼近, 得到局部保持投影. 因此, LPP 具有许多非线性技术的数据表示特征.

局部保形投影算法, 属于将图嵌入子空间的学习算法, 其目的是用低维向量表示高维空间中图的节点. 通过求解一个投影矩阵 A 将空间样本节点投影到低维空间从而实现降维. 图中任意两节点之间的关联性用最近邻图模型表示, 因此较好的保留了子空间中节点局部的结构, 实现了局部降维.

假设构建一个无向加权图 Graph 有 m 个节点, 第 i 个节点用 x_i 表示, 任意两个节点之间采用 K 近邻法定义是否关联. 选择与 x_i 邻近的 k 个节点作为 x_i 的邻近点, 若 x_j 在 x_i 的 k 个邻近点中或者 x_i 在 x_j 的 k 个邻近点中, 则 x_i 与 x_j 相连; 反之, 则不相连.

根据上述邻接图计算权值. 矩阵 W 表示权值矩阵, 则两节点 x_i 与 x_j 之间的权值为 W_{ij}, 若 x_i 与 x_j 在相互的邻近域中, 则 W_{ij} 为非 0 值, 反之, W_{ij} 为 0. 用径向基函数计算任意两节点的权值, 则权值矩阵 W_{ij} 可定义为:

$$W_{ij}=\begin{cases} e^{\left(\frac{\|x^i-x^j\|^2}{\sigma}\right)}, 若i与j相连 \\ 0, 其他 \end{cases} \tag{1}$$

最后, 对其做特征分解. 假设总节点数即样本集为 m, 样本集矩阵$X=[x_1,x_2,\cdots,x_m]$, 矩阵 X 通过投影到低维空间的矩阵 $Y=[y_1,y_2,\cdots,y_m]$, 定义线性函数 $y_i=f(x_i)=a^{\mathrm{T}}x_i$, 表示高维空间向量 x_i 通过投影向量 a 投影到低维空间向量 y_i. 为保持图中节点的局部结构, 邻近点 x_i 与 x_j 投影后得到的 y_i 与 y_j 仍需保持邻近, 则需满足下列准则函数值最小:

$$\min\sum_{j}^{i}\left(y_i-y_j\right)^2W_{ij} \tag{2}$$

因线性函数$y_i=f(x_i)=a^{\mathrm{T}}x_i$, 则式 (2) 可变换为:

$$\begin{aligned} &\min\sum_{i,j}\left(a^{\mathrm{T}}x_i-a^{\mathrm{T}}x_j\right)^2W_{ij} \\ &=\sum_{i,j}\left(a^{\mathrm{T}}x_iW_{ij}x_i^{\mathrm{T}}a\right)-\sum_{i,j}\left(a^{\mathrm{T}}x_iW_{ij}x_j^{\mathrm{T}}a\right) \\ &=\sum_{i,j}\left(a^{\mathrm{T}}x_iD_{ii}x_i^{\mathrm{T}}a\right)-\sum_{i,j}\left(a^{\mathrm{T}}x_iW_{ij}x_j^{\mathrm{T}}a\right) \\ &=a^{\mathrm{T}}X(D-W)X^{\mathrm{T}}a \\ &=a^{\mathrm{T}}XLX^{\mathrm{T}}a \end{aligned} \tag{3}$$

其中, D 为$n\times n$的对角阵, $D_{ii}=\sum_j W_{ij}$, 即权重矩阵 W_{ij}每列的和为对角矩阵 D 对角线上的元素. $L=D-W$, L 称为拉普拉斯矩阵. 为了在投影后数据最密集的地方建立坐标轴, 需对 Y 进行一定的约束: $Y^{\mathrm{T}}DY=1$即 $a^{\mathrm{T}}XLX^{\mathrm{T}}a=1$; 则式 (3) 可变换为:

$$\begin{aligned} &\min_{a^{\mathrm{T}}XDX^{\mathrm{T}}a=1}a^{\mathrm{T}}XLX^{\mathrm{T}}a \\ &=\min_{a^{\mathrm{T}}XDX^{\mathrm{T}}a=1}\frac{a^{\mathrm{T}}XLX^{\mathrm{T}}a}{a^{\mathrm{T}}XDX^{\mathrm{T}}a} \\ &=\min_{a^{\mathrm{T}}XDX^{\mathrm{T}}a=1}\frac{a^{\mathrm{T}}X(D-W)X^{\mathrm{T}}a}{a^{\mathrm{T}}XDX^{\mathrm{T}}a} \\ &=\min_{a^{\mathrm{T}}XDX^{\mathrm{T}}a=1}-\frac{a^{\mathrm{T}}XWX^{\mathrm{T}}a}{a^{\mathrm{T}}XDX^{\mathrm{T}}a} \\ &=\max_{a^{\mathrm{T}}XDX^{\mathrm{T}}a=1}\frac{a^{\mathrm{T}}XWX^{\mathrm{T}}a}{a^{\mathrm{T}}XDX^{\mathrm{T}}a} \end{aligned} \tag{4}$$

用拉格朗日乘数法将式 (4) 转化为求解下列方程的最大特征向量a:

$$a^{\mathrm{T}}XWX^{\mathrm{T}}a=\lambda a^{\mathrm{T}}XDX^{\mathrm{T}}a \tag{5}$$

其中, λ为拉格朗日乘数.

LPP 算法实现降维的同时保留了数据节点间的局部空间结构, 具有较好的局部判别能力; 与传统的线性降维方法相比, 该算法能保持数据的流行结构, 克服了非线性方法难以获得新样本低维投影的缺点. 但是 LPP 算法也存在自身的缺陷, 在求解大规模特征值问题时会导致计算量较大, 计算时间较长. 算法只注重数据的局部结构, 而未考虑到数据样本的类别, 另外在噪声影响下算法不能获得较理想的结果, 因此算法的鲁棒性较差.

1.2 基于谱回归的特征降维

为了克服局部保形投影算法计算稠密矩阵的特征值问题, 引入谱回归 (Spectral Regression, SR) 方法用回归模型处理特征函数, 先将特征函数根据图谱理论进行图的谱分析, 再将数据放入回归模型中处理. 其特征降维的优良特性使得在众多领域中得到了广泛应用[15].

在定义线性函数$y_i=f(x_i)=a^{\mathrm{T}}x_i$求解投影向量 a时, 投影向量a可能会无解, 谱回归算法通过最小二乘算法寻找与投影向量a的最佳函数匹配, 使求得的数据与实际向量a之间的误差的平方和为最小, 最大程度逼近投影向量a.

$$a=\min_a\left(\sum_{i=1}^{m}\left(a^{\mathrm{T}}x_i-y_i\right)^2\right) \tag{6}$$

通过对式 (6) 求偏导可得:

$$a=\left(XX^{\mathrm{T}}\right)^{-1}Xy \tag{7}$$

若$\left|XX^{\mathrm{T}}\right|=0$, 即$XX^{\mathrm{T}}$为奇异矩阵, 将会导致$\left(XX^{\mathrm{T}}\right)^{-1}$无解, 因此为了避免因$XX^{\mathrm{T}}$为奇异矩阵导致式 (7) 呈病态, 需对投影向量$a$施加一个正则化项$\alpha\|a\|^2$, α称为正则化参数:

$$a=\min_{a}\left(\sum_{i=1}^{m}\left(a^{\mathrm{T}}x_i-y_i\right)^2+\alpha\|a\|^2\right) \tag{8}$$

当正则化参数α无限趋向于 0 时, 式 (8) 的正则解即为特征问题 (式 (5)) 的最大特征向量解.

2 实验

本文在网络客户和电信客户两个大样本数据集上进行实证研究, 预测流程如图 1 所示, 采用 $F1$ 值、精确率、召回率、准确率等指标评价模型预测结果, 具体参见表 1. 实验所用电脑的内存是 16 GB, 处理器是 Intel(R) Xeon(R) CPU E5-1603 v3, 操作系统为 Win7 64 位, 实验环境为 Matlab 2018a.

图 1 运动目标误判效果

2.1 某电子商务网站网络客户数据

该数据集来源于某电子商务网站. 采用过抽样和随机抽样形成训练数据集和测试数据集, 以自然年度为周期共得到训练样本 20 006 个, 测试样本 8574 个. 其中训练集中流失客户 10 002 个, 非流失客户 10 004 个. 测试集中非流失客户 856 个, 流失客户 7718 个. 非流失客户与流失客户的比例基本为 1:9, 主要包括客户首次购买时间、客户关系长度、客户消费新鲜度、客户消费频度、客户消费金额、客户对商品的评分、客户评论情感共 7 个属性特征[2]. 这里我们从不同特征约简算法和分类器两个层面进行对比. 特征约简算法则包括 KPCA、PCA, 分类器包括原始线性核 SVM、优化 SVM 算法 (网格算法, 遗传算法, 种群优化算法) 和 DBN (深度置信网络), 其中 KPCA, PCA, SR-LPP 的约简维度统一设置为 3, DBN 设置为 3 层隐藏层, 每层 30 节点. 鉴于企业获取新客户的成本是保留老客户成本的数倍, 将流失客户判别为非流失客户称为导致严重后果的第一类错误 (*FN*)[1], 将非流失客户判别为流失客户称为第二类错误 (*FP*). 对于企业而言, 模型导致的第二类错误会增加客户保持成本, 而犯第一类错误则将面临着客户流失的巨大风险, 因此在该实验中添加导致严重后果的第一类错误发生率作为辅助评价指标.

实验结果混淆矩阵如图 2 所示 (其中 0 代表了非流失类, 1 代表了流失类). 基于谱回归的预测方法在精确率、召回率、准确率等方面都优于其他方法. 且第一类错误的发生概率仅为 1.7%. 在分类器优化的方法中, 基于遗传算法 (GA) 和种群算法 (PSO) 优化的 SVM, 并不能显著提高客户流失预测效果, 相比与非优化 SVM 各项指标基本持平, 但第一类错误发生率在 35% 左右, 略高于非寻优 SVM 的 32%, SVM+Grid 预测效果则更不理想. 而 DBN 分类器在非流失客户与流失客户明显不平衡的测试集中, 全部将测试集判断为非流失客户. 在特征约简方法中, 除了 KPCA 外, PCA和本文的 SR-LPP 都在不同程度上提高了客户流失预测效果, 其中 SVM+SR-LPP 综合 *Precision*、*Recall*、*Accuracy* 和第一类错误率 4 个指标较其他方法最优, 也在一定程度上说明特征层面的选择优化更为重要.

表 1 混淆矩阵

客户状态	非流失(负例)	流失(正例)	评价指标
预测非流失(负例)	*TN*	*FN* (第一类错误)	—
预测流失(正例)	*FP*	*TP*	*Precision*=*TP*/(*TP*+*FP*)
评价指标	—	*Recall*=*TP*/(*TP*+*FN*)	*Accuracy*=(*TP*+*TN*)/(*TP*+*FN*+*FP*+*TN*)

2.2 电信客户数据

电信客户数据采用高维度、大样本的美国 DUKE 大学电信客户行为数据. 数据样本共计 151 306 个, 其中训练集共 100 000 个样本, 包含流失客户 49 562 个, 非流失客户 50 438 个, 两类客户的比例基本为 1:1; 测试集共 51 306 个样本, 包含流失客户 924 个, 非流失客户 49 514 个, 客户流失率为 1.8%, 数据类别严重不平衡. 其属性值包含产品特征、客户方案、客户信息 3 大类, 共计 87 个初始属性指标. 鉴于数据样本大、维度高, SVM 分类器的参数寻优已无法在实验计算机有效时间内的求出结果, 这里重点进行特征选择和约简算法的对比, 采用 PCA、KPCA、MCFS[16]、SRLPP 算法分别在 1–87 维度之间做了比对, 分类器统一使用线性 SVM.

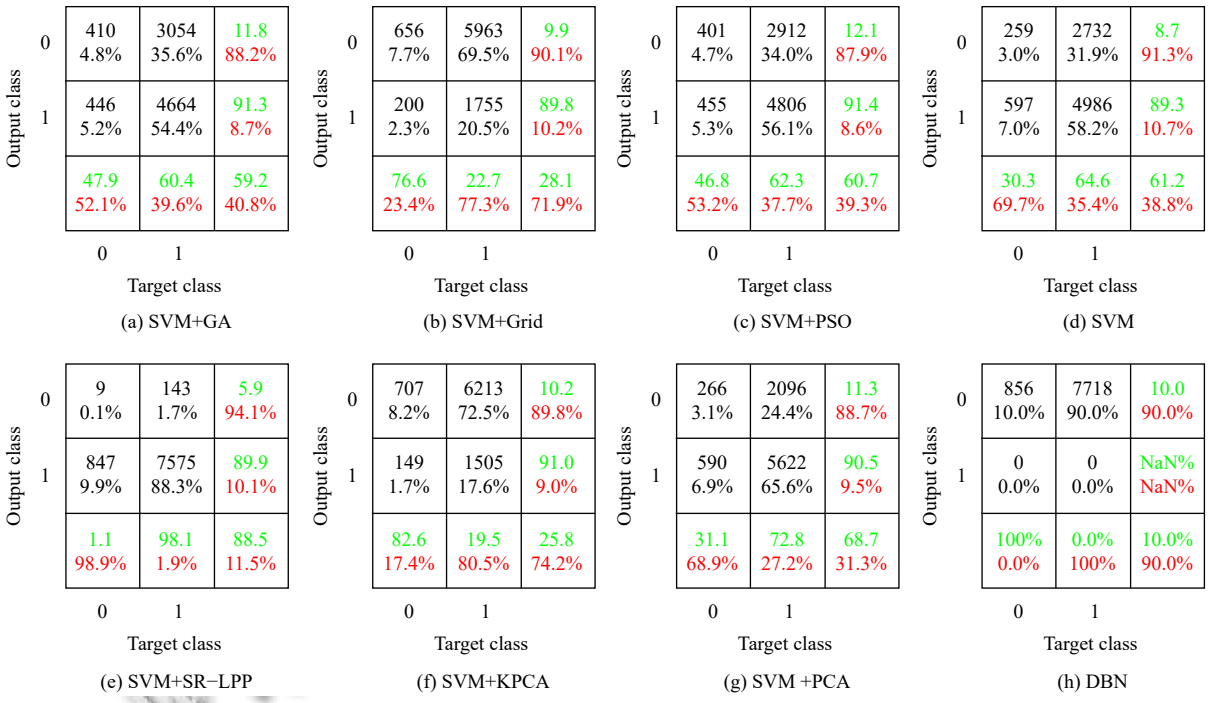

图 2 不同算法的结果混淆矩阵对比

通过图 3–图 5 可见, 降维或选择后的特征在一定程度上优于全部特征作为 SVM 输入的预测方法, 全部特征虽然包含的信息量大, 但不同属性数据间交叉、重合所产生的信息冗余也容易引起不同类别的误判, 这种特征本质的混淆在上述不同分类器并没有得到良好的解决. 在特征降维和选择的算法中, KPCA 对于不同维度稳定性较差, 在不平衡数据中容易将测试集全部预测为流失或非流失, 从而造成大部分实验召回率非 0 即 1, 使得 $F1$ 和精确度指标失去意义. 同时核函数方法需要对核矩阵计算和特征分解来完成高维空间的映射, 对于大样本数据时间复杂度高. 传统 PCA 降维, 虽然没有优异的预测效果, 但计算简单, 结果稳定, 不失为一种有效的特征降维方法. 作为特征选择方法代表的 MCFS 三项指标都略低于其他方法, 说明每一维度的特征都具有一定隐含的语义, 对于单纯维度的剔除难以满足分类的需要. SRLPP 方法则 3 项指标较为稳定, 能够对不同维度特征进行有效的融合, 在 87 个维度的约简中, 大概率的高于其他方法.

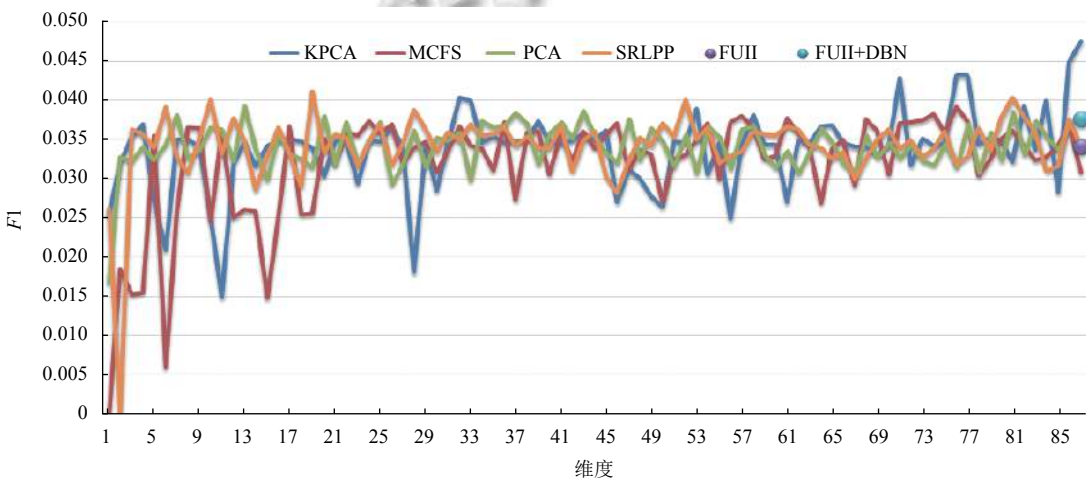

图 3 $F1$ 指标值

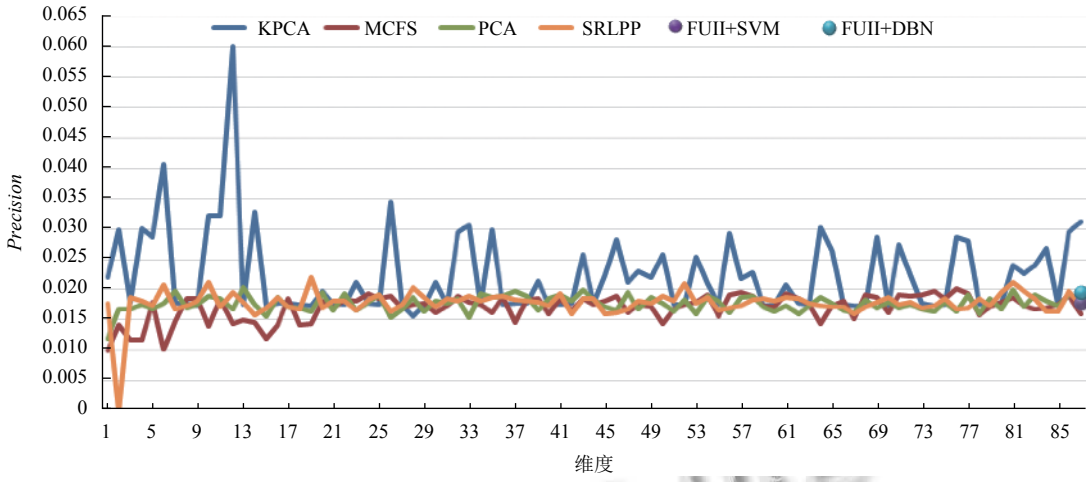

图 4 *Precision* 指标值

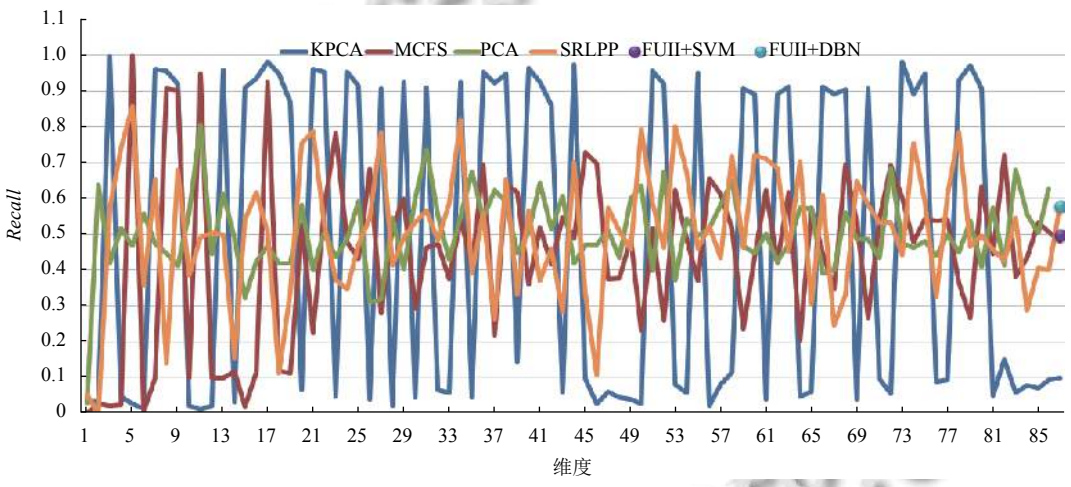

图 5 *Recall* 指标值

我们求取不同维度下的各方法的均值和标准差, 如表 2 所示, 其中 KPCA 平均精度最高, 但是其各指标值偏离程度较大, 尤其是召回率标准差达 0.44 是 PCA 方法的 4 倍, 因此表现出图 4 中连续的大波峰和波谷. MCFS 则综合表现一般, 不如全部特征输入 SVM 的预测效果. SRLPP 平均精度仅次于 KPCA, 其他指标在 4 种特征降维方法中相对最优, 整体表现稳定.

表 2 不同维度下各方法指标均值与标准差

方法	*Precision*		*Recall*		*F*1	
	Mean	SD	Mean	SD	Mean	SD
KPCA	**0.0224**	0.0069	0.4789	0.4377	0.0337	0.0051
MCFS	0.0171	0.0022	0.4461	0.2331	0.0324	0.0059
PCA	0.0177	0.0014	0.5034	**0.1080**	0.0341	0.0031
SRLPP	0.0181	**0.0013**	**0.5091**	0.1761	**0.0346**	**0.0027**
Full	0.0177	—	0.4978	—	0.0342	—

3 结论与展望

随着互联网+的广泛应用, 无论是客户数量还是属性的数据体量都在指数式增长, 且呈现出数据类型严重不平衡的特点, 传统抽样已经不能满足预测结果的解释性要求, 本文针对于高维度多属性的大规模客户流失预测, 利用基于谱回归的流形降维建立可区分性的低维特征空间, 使用线性支持向量机分类, 相比于参数优化的分类器和不同的特征降维方法, 预测效果有了不同程度的提高.

参考文献

1 夏国恩. 基于满意控制的客户流失两类错误. 系统工程, 2016, 34(3): 136–141.

2 夏国恩, 马文斌, 唐婵娟, 等. 融入客户价值特征和情感特

征的网络客户流失预测研究. 管理学报, 2018, 15(3): 442–449. [doi: 10.3969/j.issn.1672-884x.2018.03.016]
3 冯鑫, 王晨, 刘苑, 等. 基于评论情感倾向和神经网络的客户流失预测研究. 中国电子科学研究院学报, 2018, 13(3): 340–345. [doi: 10.3969/j.issn.1673-5692.2018.03.019]
4 徐子伟, 王鹏, 陈宗海. 一种基于 Fisher 比率和预测风险准则的电信客户流失预测分步特征选择方法. 中国科学技术大学学报, 2017, 47(8): 686–694. [doi: 10.3969/j.issn.0253-2778.2017.08.008]
5 张玮, 杨善林, 刘婷婷. 基于 CART 和自适应 Boosting 算法的移动通信企业客户流失预测模型. 中国管理科学, 2014, 22(10): 90–96.
6 丁君美, 刘贵全, 李慧. 改进随机森林算法在电信业客户流失预测中的应用. 模式识别与人工智能, 2015, 28(11): 1041–1049.
7 马文斌, 夏国恩. 基于深度神经网络的客户流失预测模型. 计算机技术与发展, 2019, 29(9): 76–80. [doi: 10.3969/j.issn.1673-629X.2019.09.015]
8 于小兵, 卢逸群. 电子商务客户流失预警与预测. 系统工程, 2016, 34(9): 37–43.
9 卢光跃, 王航龙, 李创创, 等. 基于改进的 K 近邻和支持向量机客户流失预测. 西安邮电大学学报, 2018, 23(2): 1–6.
10 Fowler JE. Compressive-projection principal component analysis. IEEE Transactions on Image Processing, 2009, 18(10): 2230–2242. [doi: 10.1109/TIP.2009.2025089]
11 Zhao CH, Wang YL, Mei F. Kernel ICA feature extraction for anomaly detection in hyperspectral imagery. Chinese Journal of Electronics, 2012, 21(2): 265–269.
12 Zhang LM, Qiao LS, Chen SC. Graph-optimized locality preserving projections. Pattern Recognition, 2010, 43(6): 1993–2002. [doi: 10.1016/j.patcog.2009.12.022]
13 Roweis ST, Saul LK, *et al*. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000, 290(5500): 2323–2326. [doi: 10.1126/science.290.5500.2323]
14 Zhai YG, Zhang LF, Wang N, *et al*. A modified locality-preserving projection approach for hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters, 2016, 13(8): 1059–1063. [doi: 10.1109/LGRS.2016.2564993]
15 陈朋, 张建华, 文再治, 等. 基于核谱回归与随机森林的脑电情感识别. 华东理工大学学报 (自然科学版), 2018, 44(5): 744–751.
16 Cai D, Zhang CY, He XF. Unsupervised feature selection for Multi-Cluster data. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA. 2010. 333–342.