doi:10. 16055/j. issn. 1672-058X. 2022. 0003. 008

结合相似性测度与随机森林的个人信用评估模型

都珂珂, 张 玥, 赵 凯
(安徽工程大学 数理与金融学院,安徽 芜湖 241000)

摘 要:针对客户信用数据款项维度多、数量大、复杂性等问题,提出了一种基于相似性度量的多视角决策融合个人信用评估方法。该方法创新点在于能够细致地考虑不同信用数据的几何形状,多角度划分数据,并进行相似性匹配,此外充分运用随机森林能够进行特征提取的自洽性使得模型的准确性与稳健性同步得到了提高。在 UCI 数据集上的实验结果表明:3 种距离测度在进行特征提取与异常值去除后,性能均得到了大幅提升,且识别率的波动区间相对于数据预处理前显著缩小,展现了优化后的模型具有更强的稳健性;融合 3 种测度的决策可以多角度地综合信用信息,使得识别性能较单一测度显著优化,且与其他经典组合方法比较性能更佳;将随机森林与距离测度相组合应用于个人信用评估领域为个人信用评估方法的多样性增添了新的经验。

关键词:随机森林; Hamming 距离; Chebyshev 距离; Cosine 距离; 多视角决策融合
中图分类号:F830. 5 **文献标志码**:A **文章编号**:1672-058X(2022)03-0054-07

0 引 言

随着大数据时代的到来,大数据技术卓越的数据采集和计算能力,使得数据信息更加完全但同时导致了数据结构变得复杂,数据处理难度也大大增加。客户信用数据项具有维度多、数量大、复杂性等问题[1],处理起来众多且复杂。因此,通过量化和融合多视角信用信息度量客户之间的相似性是合理的,如何选择合适的度量工具是处理信用信息数据的基础。

近年来,由于机器学习和人工智能领域的快速发展,信用评估方法,特别是基于组合模型和集成学习的信用评估方法得到了广泛的应用[2-3]。Breiman[4]引入了基于个人信用评估的随机森林算法,发现所学习的集合模型精度高于任何单个模型;Harris[5]组合多支持向量机解决了非线性支持向量机的局限性,进一步拓展了模型组合的多样性;张棚[6]运用随机森林提取重要特征,并将其用于自适应模糊推理系统的输入数据,最终在 UCI 德国信用数据集上预测借贷人员的信用风险,分类效果良好,验证了随机森林优越的特征提取功能。组合模型方法依据应用场景的不同选择适宜的处理工具,弥补了各单一方法的局限性,是如今个人信用评估领域采用的主流研究方式。在模型组合的多样性方面,个人信用评估方法多基于弱分类器融合角度,很少有学者从相似性匹配角度来研究,这为探究新的组合方式提供了思路。基于测度的信用评估方法创新点在于能够细致地考虑不同信用数据的几何形状,多角度划分数据,并进行相似性匹配,用相对较小的

收稿日期:2021-04-03;**修回日期**:2021-05-28.
基金项目:安徽省教育厅自然科学重点研究项目(KJ2016A064); 安徽工程大学教学科研项目(2018JYXM68).
作者简介:都珂珂(1996—),女,安徽淮北人,硕士研究生,从事数据挖掘研究.

相似距离度量作为评判标准,距离越小则相似性越大,二者划为同一类的偏向就越大。

假设特征空间是欧氏空间,本文针对可进行二值化转码的属性,引入 Hamming 距离量化使用二进制编码产生的固定长度字符串之间的相似性[7]。针对可以数量度量的数据,本文提出使用向量空间中一致范数导出的 Chebyshev 距离度量,用表示向量之间角度的 Cosine 距离量化客户信用差异。此外,考虑到客户所携带的原始特征向量是高维的、稀疏的和冗余的,容易导致相似性匹配的性能退化,本文采用与特征提取具有高度自洽性的随机森林方法提取与信用状态密切相关的特征,从而产生重要的特征向量使得模型的准确性与稳健性同步得到提高。这样,将 3 个基于距离的度量分别进行相似性比较,使得个人信用风险评估结果由基于加权投票的方法融合而成。

1　研究方法与设计

不同的度量(距离度量或散度)是解决模式识别问题的有效工具,在分类、聚类和检索[8]等领域已被广泛地应用。在分析信用风险时,可以通过不同信用类别客户特征数据之间的差异来进行评估。这样,个人信用评估问题就变成了特征数据的相似性匹配问题。但是在选取 3 种常被用于分类的距离测度进行分类实验时,发现其分类的准确性还有很大的提升空间,分类结果如图 1 所示。为解决这一问题,采用投票分类的方式从多视角出发,将 3 种距离测度所产生的决策进行综合评估来提升模型的性能。除此之外,为了去除原始数据中冗余信息对评估准确性的干扰,从提取重要特征的需求出发,选择与这一需求具有自洽性的随机森林方法,提出了结合相似性测度与随机森林的个人信用评估模型。

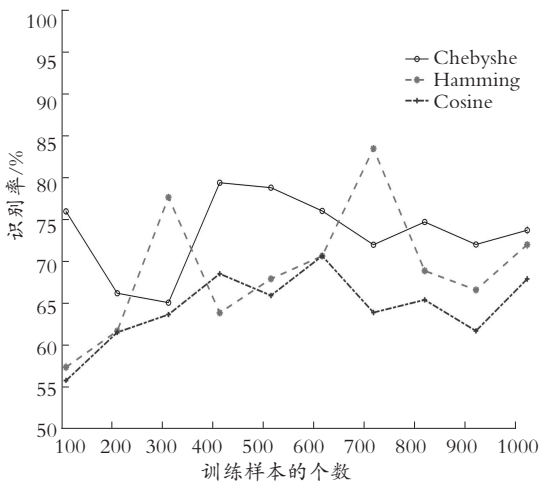

图 1　原始数据集上相似性度量的性能比较

Fig. 1　Performance comparison of similarity measures on primitive data set

如图 2 所示,提出的多视角决策融合包含 3 个关键阶段:重要特征的提取、基于随机森林的相似性匹配和决策融合。第一阶段运用随机森林的方法提取重要特征对原始数据进行降维;第二阶段在重要信用特征之间分别通过 Hamming 距离、Chebyshev 距离和 Cosine 距离进行相似性匹配;最后阶段,通过加权投票进行多视角决策融合。在以下两节中,本文将逐步详细介绍所提出的方法。

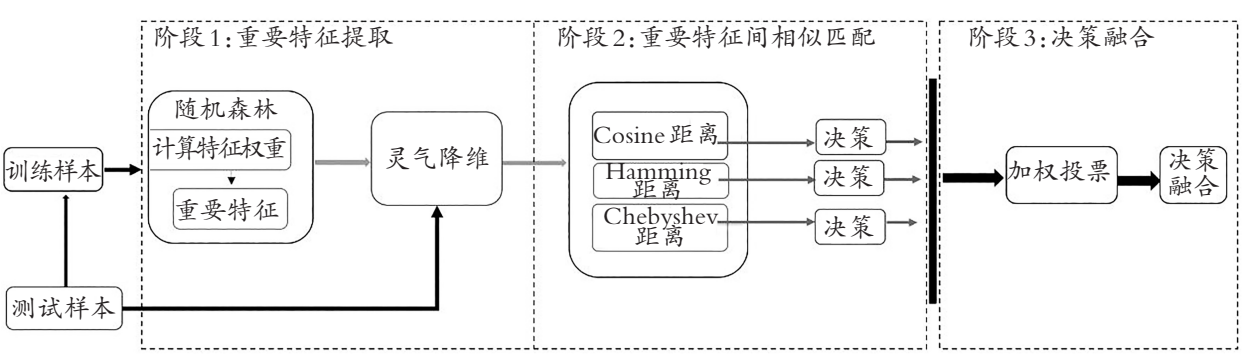

图 2　模型框架

Fig. 2　The framework of our model

本文所涉及的 Hamming 距离、Chebyshev 距离与 Cosine 距离的具体数学表达形式在表 1 中罗列:

表 1 3 种距离度量的公式

Table 1 Formulas of three distance measures

距离度量	公 式	注 释	来 源
Hamming 距离	$d_{\text{Him}}(x,y)=\sum_{i=1}^{n}x_i \oplus y_i$	x,y 是 n 位的编码,这里 x_i 与 y_i 均表示第 i 位编码值,$i\subseteq\mathbf{N}$,$\oplus$表示异或	文献[9]
Chebyshev 距离	$d_{\text{Che}}(\boldsymbol{a},\boldsymbol{b})=\max_{1\leqslant i\leqslant n}(\lvert a_i-b_i\rvert)$	$\boldsymbol{a}$ 和 $\boldsymbol{b}$ 是特征空间中的两向量,a_i 和 b_i 分别是 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 的第 i 个元素	文献[10]
Cosine 距离	$d_{\text{Cos}}(\boldsymbol{a},\boldsymbol{b})=\dfrac{(\boldsymbol{a},\boldsymbol{b})}{\lvert\boldsymbol{a}\rvert\lvert\boldsymbol{b}\rvert}$	$\boldsymbol{a}$ 和 $\boldsymbol{b}$ 表示空间中两向量,$\langle\cdot,\cdot\rangle$是两向量内积,$\lvert\cdot\rvert$表示向量长度	文献[11]

1.1 基于随机森林的重要特征提取

原有的客户信用特征具有高维性、信息冗余性、稀疏性等特点[12],直接使用基于距离的度量会导致风险评估方法的能力退化。为了弥补这些缺陷,采用随机森林方法,对每个特征赋予权重,然后依据权重对特征进行排列,设置权重的阈值,从训练样本中提取与信用具有强影响力的特征。进一步来说,通过重要特征提取实现了数据降维,解决冗余信息对分类的干扰,提升模型的分类效率。

1.2 基于距离的相似度匹配

在提取重要特征后,对测试样本和训练样本进行相似性匹配。为此,由 3 种测度对应的方程式分别计算测试样本与训练样本的重要信用特征之间的 Hamming 距离、Chebyshev 距离和 cosine 距离。

1.3 多视角的决策融合

在完成上述步骤后,本文从 3 个方面获得了基于随机森林的决策,下一个是最终决策。在这一步中,采用加权投票补充基于度量决策的优点,并进一步提高决策的准确性,因此最终决策基于投票矩阵和决策方程产生。

2 实 验

在这一部分中,进行了几个实验来验证所提出方法的有效性。实验数据是德国信用数据集,其中每个数据点包含 20 个属性和类别标签,以表明在 15 万名贷款申请者范围内的良好信用或不良信用风险。评估任务在 Matlab(Version2016a)编程环境下用 Inteli5×8250u 处理器在笔记本电脑上完成。

2.1 模型参数的设定

随机森林分类的准确度很大程度上受树棵数 n 的影响。选择过少的树棵数,会导致预测结果不理想;选择过多的树棵数,会降低分类准确率的提升效果,甚至延长计算速度、降低计算速率。本文选取 100 棵树,分别以 Most Popular 数据与 Excluding In-bag Observation 数据的袋外数据分类误差为评价指标,来探究树棵数的选取对判断准确性的影响,其结果如图 3 所示。

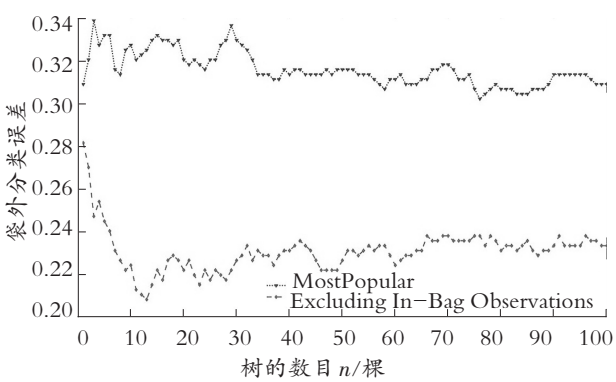

图 3 两数据集的袋外分类误差变化折线图

Fig. 3 Fold line chart of changes in out-of-packet classification error of two data sets

图 3 中,折线图的横坐标表示树棵数,纵坐标表示模型的袋外数据分类误差率。可以看出,两条折线的转折点都在 $n=32$ 左右。当 $n<32$ 时,随着树数目的增多,Excluding In-bag Observations 数据的错误率由 28% 下降到 22% 左右,Most Popular 数据的判断错误率略有波动但差距极小;当 $n>32$ 时,两组

数据的错误率趋于平稳。故选择 $n=32$ 作为随机森林模型树的数目。

2.2　重要特征的提取

计算单个特征变量的重要性是随机森林方法拥有的一个显著特性[13],这使得随机森林方法与本模型所提取出的重要特征提取要求具有高度自洽性。随机森林方法比较特征变量重要性的评判标准通常采用袋外数据分类误差。这是因为生成随机决策树时采用随机有放回的方法采样,不会将所有的样本引入用于生成一棵树,这个过程使得袋外数据(OOB)得以产生。通过对比加入噪声后特征的袋外数据分类误差变化幅度来判断重要性,变化幅度越大则特征越重要。这种方法可以在模型生成过程中取得真实误差的无偏估计,且不损失训练数据量。具体的选择过程如下:

步骤 1　计算每个特征的重要性;

步骤 2　确定要剔除的比例,在此基础上依据特征重要性降序剔除冗余特征,得到一个新的特征集;

步骤 3　用新的特征集重复上述过程,直到剩下事先设定的特征值的数目;

步骤 4　对比上述过程提取的特征集的袋外误差率,选择袋外误差率最低的特征集。

使用随机森林方法,按照上述方式对特征进行权重赋值,并依据权重值对其降序排列,选取出 23 条评分较高的款项,其结果如图 4 所示。并以此图为评判依据选出重要的 23 个款项,对其进行属性划分后,最终提取 13 个重要特征,并实现数据的降维,具体属性如表 2 所示。

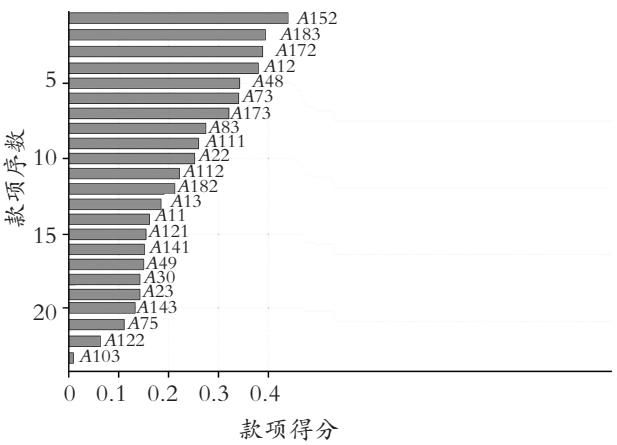

图 4　重要款项得分

Fig. 4　Scores of important attribute items

表 2　德国个人信贷数据的重要特征

Table 2　Important features of individual credit data in Germany

属　性	款　项	属　性	款　项
现有支票账户存额	*A*11: …<0DM *A*12:0<=…<200DM *A*13 : … >= 200 DM	贷款年限	*A*22 : 1 <= … < 4 years *A*23 : 4 <= … < 7 years
信用历史	*A*30 :无信用历史/所有信用均还清	财产	*A*121 : real estate *A*122: bank savings/ life insurance
贷款目的	*A*48 : retaining *A*49 : business	分期付款率占可支配收入的百分比	*A*83 : 50% <= … < 75%
目前自住宅	*A*111 : YES *A*112 : NO	工作年限	*A*73 : 1 <= … < 74years *A*73 : .. >= 7 years
担保人	*A*103 : guarantor	住宅所有权	*A*152:own
供养人数	*A*182:2 *A*183 : 3	职业	*A*173 : 有技能的员工/公务员 *A*274::管理者/个体
其他投资计划	*A*141 : 银行 *A*143 : 无		

结合图 4 与表 2,在德国信用数据集上,对信用影响最深的款项是贷款人当前住宅所有权情况。权重排名前两项的款项与借款人的个人背景信息相关,如住宅所有权、供养人数、工作年限等,而大部分

的重要属性皆与个人征信信息相关,如分期付款率占可支配收入的百分比、信用历史、贷款目的等。表 2 依据实验结果选出与德国客户信用强相关性的重要特征,以此建立本文模型的信用评价指标以进行最终分类。此外,提取这些重要特征相关的数据进行分类实验,有效解决了信息的复杂性带来的分类困难。

2.3 异常值的处理

在重要特征提取后,考虑到原始数据中由于操作失误或者机械设备故障等原因导致的数据缺失情况。这种情况会造成数据出现不完整性、不适用性以及缺乏一致性等问题,因此需要对原始数据进行清洗,将无意义、缺失值较多的字段删除。本文在完成此步骤的基础上,为进一步优化模型的性能,通过计算个体数据间的相关性找到更具普遍代表意义的样本数据以避免个体差异较大的离群数据对实验结果的影响。使用 matlab 计算并绘制出显示数据间相关性程度的数量分布直方图,由图 5 表示。

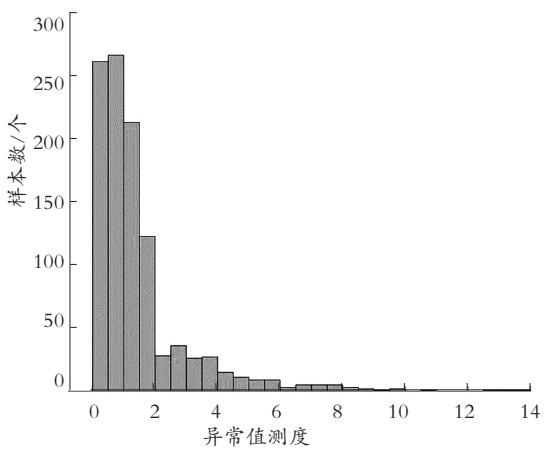

图 5 数据相关性分布直方图

Fig. 5 Histogram of data correlation distribution

图 5 中,X 轴代表数据的相关性程度,越靠近原点相关性越大;Y 轴代表样本数据数量。由图 5 可以得知,当相关性值达到 10 之后,数据的个体差异性出现断层,离群性体现突出,因此在进行接下来的分类准确性实验比较之前,要去除这些离群数据,以保证模型性能的准确性。

2.4 特征提取后模型性能的比较

针对 3 种基于距离的相似度匹配分类性能较差的问题,利用 inranked 算法实现随机森林来提取重要特征,对原始信用特征进行降维和信息浓缩。本次实验中,在叶节点数为 1, 决策树数为 32 的情况下运行 inranked 算法,然后在相同的实验环境下,对基于关键项和项集的样本风险性进行评估比较。实验结果如图 6 所示。

从图 6 所示的结果可以看出:3 种距离测度在基于重要特征的相似性匹配性能均有明显提高。与图 1 相比, Chebyshev 距离的识别率最高,其次是 Hamming 距离。当训练样本数为 800 时,Chebyshev 距离、Hamming 距离和 Cosine 距离的识别率分别达到 88.25%,87.65%和 86.75%。即使当样本量为 700 时,3 种距离度量的谷值为 76.77%,也表现出了优良的分类准确性。实验结果表明:在经过重要特征提取后的数据上,使用 3 个距离测度进行相似度匹配,大大提升了模型的分类性能。

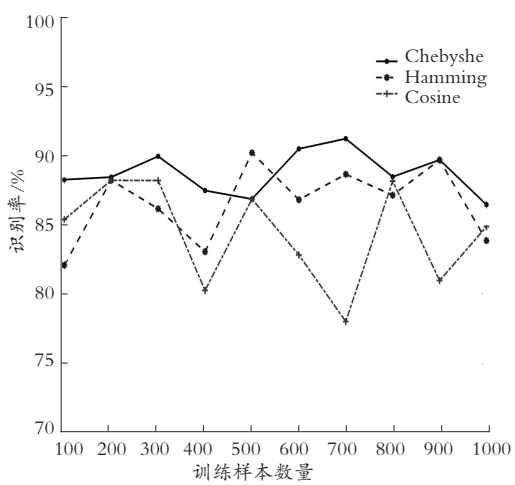

图 6 3 种测度在重要特征间的识别性能比较

Fig. 6 Comparison of recognition performance of three measures among important features

2.5 与其他经典模型的比较

为了验证结合相似性测度与随机森林的多视角决策融合个人信用评估模型的改进效果,本文在拥有 15 条信用数据的德国信用数据集上随机选取 1 000 个训练样本,分别对 8 种方法进行了 10 次交叉验证,其平均识别率与标准差如表 3 所示。结果表明:本文算法平均识别率在 93%以上,标准差为 1.140 7,而 3 种基于差异的算法平均识别率均在 90%以下,标准差均高于 1.4,表明多视角决策融合方法具有可行性和有效性,且模型稳定性更佳。

表 3　与经典的方法比较

Table 3　Comparison of classical methods

方　法	平均识别率/%	标准差
Proposed	93.48	1.140 7
Chebychev	88.53	1.463 0
Hamming	89.87	2.597 5
Cosine	84.53	3.428 7
HAR-AWDF	92.82	1.276 8
WCBA	70.91	2.314 4
CMAR	66.53	2.906 7
CBA	62.46	3.727 8

表 3 还表明:本文所提出的方法优于包含 HAR-AWDF,CBA,WCBA 和 CMAR 这 4 种具有代表性的基于重要特征提取的方法。在平均识别率方面,该算法的识别率最高为 93.48%,其次是 HAR-AWDF 算法,识别率为 92.82%,CBA 算法的识别率最低,为 62.46%。除了本文的模型算法和 HAR-AWDF 之外,基于 3 种不同距离度量的平均识别率都高于其他基于关联规则的分类算法。这进一步证实了准确测量客户信用调查数据之间的信用调查相似性能力对于评估个人信用调查至关重要。

3　结束语

结合随机森林方法建立的信用评价体系将 3 种测度与信用数据相似性匹配后得到信用风险预测的准确性。在 UCI 数据集中的德国信用数据集上的实验结果表明:随机森林方法自有的特征提取特性能够有效提取与分类结果有强关联性的信用特征。3 种距离测度在进行特征提取与异常值去除后性能均得到了大幅提升,且识别率波动区间相对于数据预处理前显著缩小,表明了优化后的模型具有更强的稳健性。通过融合 3 种测度的决策可以多角度地综合信用信息,使得识别性能较单一测度显著优化,且与其他经典组合方法比较性能更佳。将随机森林与距离测度相组合应用于个人信用评估领域为个人信用评估方法的多样性增添了新的经验。

参考文献(References):

[1] YU L Q, CAO F Y, ZHAO X W, et al. Combining attribute content and label information for categorical data ensemble clustering[J]. Applied Mathematics and Computation, 2020(381):2—15.

[2] 张晨,万相昱. 大数据背景下个人信用评估体系建设和评估模型构建[J]. 征信,2019,37(10):66—71.
ZHANG Chen, WAN Xiang-yu. Construction of individual credit evaluation system and evaluation model under the background of big data[J]. Credit Investigation, 2019, 37(10):66—71.

[3] 雒腾. 基于 Stacking 选择性集成算法的个人信用风险评估研究[D]. 昆明:云南财经大学,2020.
LUO Teng. Research on personal credit risk assessment based on stacking selective integration algorithm[D]. Kunming: Yunnan University of Finance and Economics, 2020.

[4] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1):5—32.

[5] HARRIS T. Credit scoring using the clustered support vector machine[J]. Expert Systems with Applications, 2015,42(2):741—750.

[6] 张棚. 基于 RF-ANFIS 的供应链金融个人信用风险评估模型研究[D]. 杭州:浙江工业大学,2017.
ZHANG Peng. Research on individual credit risk assessment model of supply chain finance based on RF-ANFIS[D]. Hangzhou:Zhejiang University of Technology,2017.

[7] 马敏耀,徐艺,刘卓. 隐私保护 DNA 序列汉明距离计算问题[J]. 计算机应用,2019,39(9):2636—2640.
MA Min-yao, XU Yi, LIU Zhuo. Hamming distance calculation of DNA sequence with privacy protection[J]. Computer Applications,2019,39(09):2636—2640.

[8] HOANG NGUYEN. Anovel similarity/dissimilarity measure for intuitionistic fuzzy sets and its application in pattern recognition[J]. Expert Systems with Applica-tions, 2016(45):97—107.

[9] 汤建明,寇小强. 海量网络文本去重系统的设计与实现[J]. 计算机应用与软件,2018,35(12):33—37.
TANG Jian-ming, KOU Xiao-qiang. Design and implementation of massive network text removal system[J]. Computer Applications and Software,2018,35(12):33—37.

[10] RODRIGUES E O. Combining minkowski and cheyshev:

new distance proposal and survey of distance metrics using k-nearest neighbours classifier[J]. Pattern Recognition Letters, 2018(110):66—71.

[11] XU J, MU J P, CHEN G R. A multi-view similarity measure framework for trouble ticket mining[J]. Data & Knowledge Engineering, 2020(127):101800-101811.

[12] 钟金宏,邵晶晶,李兴国. 基于组合分类策略的个人信用风险评估研究[J]. 合肥工业大学学报(自然科学版),2020,43(7):996—1002.
ZHONG Jin-hong, SHAO Jin-jin, LI Xing-guo. Research on individual credit risk assessment based on portfolio classification strategy[J]. Journal of Hefei University of Technology (Natural Science Edition), 2020,43(7):996—1002.

[13] 马晶,蔡文杰,杨利. 基于机器学习的心音识别分类研究[J]. 中国医学物理学杂志,2021,38(1):75—79.
MA Jin, CAI Wen-jie, YANG Li. Heart sound recognition and classification based on machine learning[J]. Chinese Journal of Medical Physics, 2021,38(1):75—79.

Personal Credit Assessment Model Based on the Combination of Similarity Measurement and Random Forest

DU Ke-ke, ZHANG Yue, ZHAO Kai

(School of Mathematics, Physics and Finance, Anhui Polytechnic University, Anhui Wuhu 241000, China)

Abstract: Aiming at the problems of the attribute items of customer credit data, such as many dimensions, large number and complexity, this paper proposes a multi-perspective decision-making fusion personal credit evaluation method based on similarity measurement. The innovation of this method lies in that it can carefully consider the geometric shapes of different credit data, divide the data from multiple angles, and carry out similarity matching. In addition, it makes full use of the self-consistency that random forest can carry out feature extraction to improve the accuracy and robustness synchronization of the model. The experimental results on UCI dataset show that the performance of the three distance measures is greatly improved after feature extraction and outlier removal, and the fluctuation range of the recognition rate is significantly reduced compared with that before data preprocessing, which shows that the optimized model has stronger robustness. By combining the three measures, credit information can be integrated from multiple angles, which makes the identification performance significantly better than that of a single measure, and the performance is better compared with other classical combination methods. The combination of random forest and distance measure in the field of personal credit evaluation adds new experience to the diversity of personal credit evaluation methods.

Key words: random forest; Hamming distance; Chebyshev distance; Cosine distance; multi-view decision fusion

责任编辑:李翠薇

引用本文/Cite this paper:

都珂珂,张玥,赵凯. 结合相似性测度与随机森林的个人信用评估模型[J]. 重庆工商大学学报(自然科学版),2022,39(3):54—60.

DU Ke-ke, ZHANG Yue, ZHAO Kai. Personal credit assessment model based on the combination of similarity measurement and random forest[J]. Journal of Chongqing Technology and Business University (Natural Science Edition),2022,39(3):54—60.