



文章编号:1001-9081(2019)09-2784-05

DOI:10.11772/j.issn.1001-9081.2019030571

## 基于值差度量和聚类优化的 $K$ 最近邻算法在银行客户行为预测中的应用

李 博<sup>1,2\*</sup>, 张 晓<sup>1,2</sup>, 颜靖艺<sup>3</sup>, 李可威<sup>1</sup>, 李 恒<sup>1,2</sup>, 凌玉龙<sup>1,2</sup>, 张 勇<sup>1,2</sup>

(1. 西北工业大学 计算机学院, 西安 710129; 2. 工信部大数据存储与管理重点实验室(西北工业大学), 西安 710129;

3. 西北工业大学 管理学院, 西安 710129)

(\* 通信作者电子邮箱 2640992944@qq.com)

**摘要:**为提升贷款金融客户行为预测的准确性,针对传统的  $K$ -最近邻( $KNN$ )算法在数据分析中处理非数值因素的不完备问题,提出了一种采用值差度量(VDM)距离的对聚类结果迭代优化的改进  $KNN$  算法。首先对收集到的数据信息进行基于 VDM 距离的  $KNN$  算法的聚类,再对聚类结果进行迭代分析,最后通过联合训练提高了预测精度。基于葡萄牙零售银行 2008—2013 年收集的客户数据比较可知,改进的  $KNN$  算法与传统的  $KNN$  算法、基于属性值相关距离的  $KNN$  改进(FCD-KNN)算法、高斯贝叶斯算法、Gradient Boosting 等现有算法相比具有更好的性能和稳定性,在银行数据预测客户行为中具有很大的应用价值。

**关键词:** $K$ -最近邻算法;值差异度量距离;金融危机;行为预测;数据挖掘

**中图分类号:** TP311.13    **文献标志码:**A

### Application of KNN algorithm based on value difference metric and clustering optimization in bank customer behavior prediction

LI Bo<sup>1,2\*</sup>, ZHANG Xiao<sup>1,2</sup>, YAN Jingyi<sup>3</sup>, LI Kewei<sup>1</sup>, LI Heng<sup>1,2</sup>, LING Yulong<sup>1,2</sup>, ZHANG Yong<sup>1,2</sup>

(1. School of Computer Science, Northwestern Polytechnical University, Xi'an Shaanxi 710129, China;

2. Ministry of Communications Key Laboratory of Big Data Storage and Management (Northwestern Polytechnical University), Xi'an Shaanxi 710129, China;

3. School of Management, Northwestern Polytechnical University, Xi'an Shaanxi 710129, China)

**Abstract:** In order to improve the accuracy of loan financial customer behavior prediction, aiming at the incomplete problem of dealing with non-numerical factors in data analysis of traditional  $K$ -Nearest Neighbors ( $KNN$ ) algorithm, an improved  $KNN$  algorithm based on Value Difference Metric (VDM) distance and iterative optimization of clustering results was proposed. Firstly the collected data were clustered by  $KNN$  algorithm based on VDM distance, then the clustering results were analyzed iteratively, finally the prediction accuracy was improved through joint training. Based on the customer data collected by Portuguese retail banks from 2008 to 2013, it can be seen that compared with traditional  $KNN$  algorithm, FCD-KNN (Feature Correlation Difference  $KNN$ ) algorithm, Gauss Naive Bayes algorithm, Gradient Boosting algorithm, the improved  $KNN$  algorithm has better performance and stability, and has great application value in the customer behavior prediction from bank data.

**Key words:**  $K$ -Nearest Neighbors ( $KNN$ ) algorithm; Value Difference Metric (VDM) distance; financial crisis; behavior prediction; data mining

### 0 引言

在贷款金融领域,银行机构营销需要对用户进行分析和分类,以降低营销成本。基于某目标人群,从海量的其他人群中找出和目标人群相似的人群,以拓展目标人群规模。在现实生活中,通过海量数据集,并对数据划分标签,然后对用户行为进行分析和分类,再进行相应的营销手段,可以降低成本,并取得较好的效果<sup>[1-3]</sup>。当前的一些研究指出,银行信息的数据挖掘不应该仅仅局限于会计数据,还需要考虑一些社会因素。

基于数据挖掘和用户行为预测的目的,本文采用数据挖掘方法对葡萄牙银行业金融机构直接营销活动(电话)相关数据进行分析,通过电话营销和电话销售预测银行长期存款的可能性。该数据集由葡萄牙零售银行于 2008—2013 年收集,受到当时金融危机的影响,分类的目的是预测客户是否会订购定期存款。对于该数据集来说,主要的困难在于其特征的选择,数据集中存在无用的或有噪声的特征,这些特征可能会降低预测结果。基于这个目的,本文采用了一种改进的  $K$ -最近邻( $K$ -Nearest Neighbors,  $KNN$ )算法。 $KNN$  算法能够更好地分析相似客户的行为,更好地对客户进行分类。传统的

收稿日期:2019-04-08;修回日期:2019-06-02;录用日期:2019-06-03。    基金项目:国家重点研发计划项目(2018YFB1004401)。

**作者简介:**李博(1994—),男,甘肃陇西人,硕士研究生,CCF 会员,主要研究方向:云存储、数据挖掘; 张晓(1978—),男,河南新乡人,副教授,博士,CCF 会员,主要研究方向:存储系统; 颜靖艺(1993—),女(回族),广西桂林人,硕士,主要研究方向:技术创新管理; 李可威(1993—),男,湖北云梦人,硕士研究生,主要研究方向:数据挖掘; 李恒(1993—),男,河南周口人,硕士研究生,主要研究方向:数据挖掘; 凌玉龙(1995—),男,安徽宿州人,硕士研究生,主要研究方向:数据挖掘; 张勇(1995—),男,安徽六安人,硕士研究生,主要研究方向:数据挖掘。



KNN 算法存在一定的局限性。本文对距离计算和聚类分析方法进行了改进,实验结果表明,改进的 KNN 算法在银行数据挖掘中具有良好的预测效果。

## 1 研究现状

数据挖掘是指通过数据过滤,从大量现有数据中搜索有趣的、有价值的数据点或数据模块的数据处理技术。数据挖掘在商业金融领域有着广泛的应用,根据商业分析的既定目标,可以通过企业内部的财务数据系统进行数据分析,以获得所需的业务运营和市场发展规律,并通过成熟的数据挖掘模型和其他分析工具进行支持,形成了商业化的数据挖掘与分析系统。

2008—2013 年,葡萄牙零售银行业受到金融危机的影响,因此银行需要分析数据挖掘,分析客户是否可以继续存款。根据社会心理学研究,当人们处于压力状态下时,往往有更多的本能表现,数据分析的准确性也会相应提高<sup>[4-5]</sup>。在金融危机期间,人们对金融投资都会持谨慎态度。另一方面,葡萄牙零售银行业有着成熟的数据仓库,对银行客户的个人数据、账户信息、交易历史、业务服务历史、财务管理数据、个人财务风险评估等进行了数据仓储,可以对每个银行客户进行多维度的财务分析。

目前,对银行客户信息挖掘的研究较多,对银行客户信息挖掘的研究需求巨大。一些研究发现:配给大量信息的信贷员并没有比配给少量信息的信贷员预测更准确,现有会计信息可能过量。因此当前的研究应该更多考虑考虑非数值指标,如:职业、学历等。基于属性值相关距离的 KNN(Feature Correlation Difference-KNN, FCD-KNN)改进算法对非数值的因素进行了考虑:比较样本间的距离为属性值的相关距离,从而度量样本间的相似度<sup>[6-7]</sup>。KNN 算法是一种非常常见的算法,简单易用,易懂,精度高,理论成熟;但也存在许多问题,为此人们提出了许多改进的 K 近邻算法。为了解决银行分类问题,本文采用了一种改进的 KNN 算法:用更适合银行业情况的搜索距离函数代替标准欧几里得距离,用更精确的概率估计方法代替简单的投票机制。实验表明,本文提出的改进的 KNN 算法精度得到了很大的提高,是一种有效的算法,具有很好的推广前景。

## 2 算法分析

### 2.1 传统的 KNN 算法

K-最近邻(KNN)分类算法在模式识别领域得到了广泛的应用。KNN 算法基于类比学习,所有训练基元都存储在 N 维模式空间中。如果特征空间中 k 个最相似的样本中的大多数属于某个类别,那么这些样本就属于这个类别。KNN 算法不仅可以用于分类,还可以用于回归分析。通过寻找样本的 K 最近邻点,并将这些相邻点的属性平均值赋给样本,可以得到样本的预测值。例如,在图 1 中,当一个新的样本值添加到向量空间中时,在样本值附近对其进行分析并进行分类。传统的 KNN 算法得到了广泛的应用,但鉴于银行系统的特殊性,本文对距离选择和判别法进行了改进,使分析预测更加准确,与传统的 KNN 算法相比,其预测精度有了显著的提高。

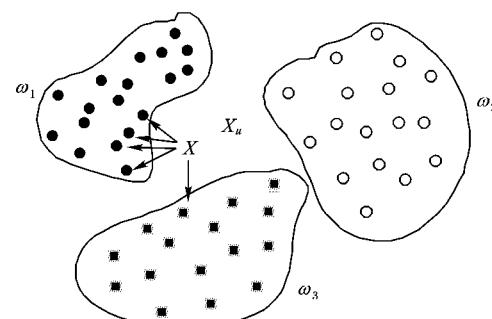


图 1 当样本空间加入新样本时的计算

Fig. 1 Calculation of sample space when new samples are added

### 2.2 本文采用的改进 KNN 算法

针对银行的特殊情况,本文采用了一种改进的 KNN 算法。改进措施包括:用更适合银行业情况的搜索距离函数代替标准欧几里得距离,用更精确的概率估计方法代替简单的投票机制。

#### 1) 采用 VDM 距离修正。

距离计算是数据挖掘聚类的关键步骤。距离计算是计算采样点与采样点之间的距离,并根据计算结果判断采样点之间的关系。传统的 KNN 算法使用欧几里得距离公式计算距离,例如:

$$\rho = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (1)$$

其中  $\rho$  为点  $(x_1, y_1)$  与点  $(x_2, y_2)$  之间的欧氏距离。

欧氏距离通常被用来表示样本的有序属性,在本数据集中只有“年龄”符合这一条件。其他的条件如:婚姻状况、工作类型等这样的无序属性,更适合采用值差度量(Value Difference Metric, VDM) 距离。VDM 距离是指:令  $M_{u,a}$  表示在属性  $u$  上取值为  $a$  的样本数,  $M_{u,a,i}$  表示在第  $i$  个样本簇中在属性  $u$  上取值为  $a$  的样本数,则属性  $u$  上两个离散值  $a$  与  $b$  之间的 VDM 距离为:

$$VDM_p(a, b) = \sum_{i=1}^{n_i} \left| \frac{m_{u,a,i}}{m_{u,a}} - \frac{m_{u,b,i}}{m_{u,b}} \right|^p \quad (2)$$

将欧氏距离和 VDM 结合可处理混合属性。为不失一般性,令有序属性排列在无序属性之前,可得:

$$MinkowDM_p(x_i, x_j) = \left( \sum_{u=1}^{n_c} |x_{iu} - x_{ju}|^p + \sum_{u=n_c+1}^n VDM_p(x_{iu}, x_{ju}) \right)^{1/p} \quad (3)$$

因为是在二维分析,可以  $p = 2$ 。无序属性就是通过计算样本簇中在属性  $u$  上样本点的多少来得到该样本簇在该属性上的“距离”。通过修正数据采集的距离,可以使得数据挖掘分析预测结果更为精确。

本文也探讨了马氏距离(Mahalanobis distance)在该问题下的应用,马氏距离是对有序的、数值型的属性,考虑其内在的关联性,从而计算得出结果<sup>[8-9]</sup>。但是本文所提到的数据也有很多无序的属性,使用马氏距离处理会较为复杂,故未采用该处理方法。

#### 2) 对数据处理修正。

传统的 KNN 方法对新增加的样本点进行分类,使其具有更高的相似性。本文同时设置了各采样点的属性,并设置了



划分区域的阈值(比如:70%)。如果超出此阈值,本算法将把采样点添加到一个没有争议的区域。如果点与每个区域之间的距离不明显,本算法将该点标记为疑问点,在初步聚类结束后再考虑它。如图2所示,如果点 $X_a$ 与区域1( $\omega_1$ )和区域2( $\omega_2$ )之间的距离显著不同,则将点 $X_a$ 划分为区域1。然而,在图3中,例如,点 $X_b$ 与区域1和区域2之间的距离没有显著差异。因此,点 $X_b$ 暂时被标记为疑问点。

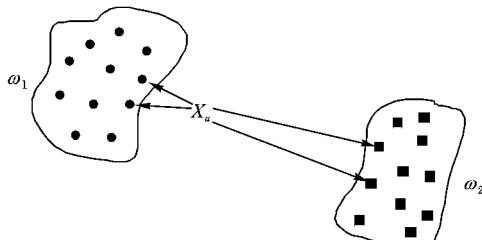


图2 在改进的KNN算法样本空间加入新样本  
(当新样本点特征很明显时)

Fig. 2 New sample points added in the sample space of the improved KNN algorithm (when the characteristics of new sample points are obvious)

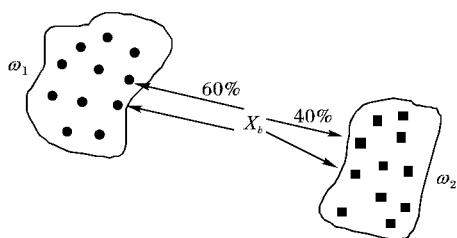


图3 在改进的KNN算法样本空间加入新样本时  
(当新样本点特征不明显时)

Fig. 3 New sample points added in the sample space of the improved KNN algorithm (when the characteristics of new sample points are not obvious)

根据这种方法,最终会发现两种类型的点:区域中心的无争议点和区域边缘的争议点,如图4所示。

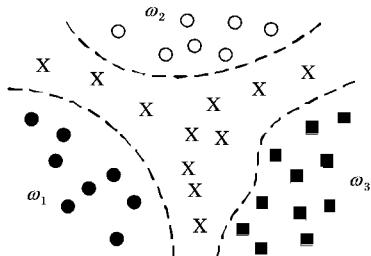


图4 改进KNN可能得到的划分结果  
Fig. 4 Possible classification results of improved KNN

在图4中的情况,需要额外增加判断过程,整体划分,保留整个区域的最小离群值。甚至对于离群值边缘太多,本算法可以将其划分为新的区域或合并原始区域,即对分类结果又进行了一次处理。而对于图5,如果区域外的点内部之间存在更多的相关性,即这一群争议点彼此之间更为相似,如果用距离作标准,即这一群争议点内部彼此之间的距离明显小于它们与现有簇之间的距离(根据本文设置的阈值判断)。首先可以通过在这些争议点中随机找到一个点,计算该点与其他争议点之间的距离。如果发现其内部距离更小,则可以形成一个新的分类;甚至于其内部可能还会进一步的分裂,也可以进一步的处理。在图5,中间的三个点彼此之间的距离

更为接近(超过本文设置的阈值),可以直接增加新的分类,结果如图6所示,这样就有了更合理的集群。

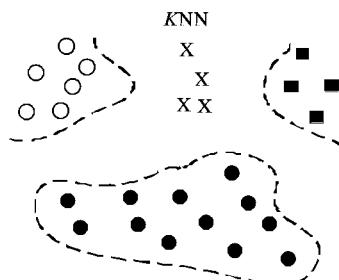


图5 改进的KNN算法划分结果再进行处理  
(可疑点暂不划分分组)

Fig. 5 Process the results of improved KNN algorithm (separate treatment of suspicious points)

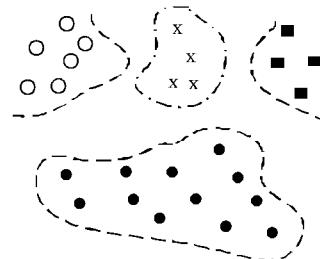


图6 改进的KNN算法划分结果再进行处理  
(将相似的可疑点划分为新分组)

Fig. 6 Process the results of improved KNN algorithm (divide similar suspicious points into new groups)

### 3 实验分析

为了验证改进的K-最近邻算法在银行数据挖掘中的有效性,本文进行了实验分析。选用的数据样本是葡萄牙零售银行在2008—2013年期间收集的数据样本,将数据分为测试集和验证集。数据预处理会有三种情况,分别为:未对原始数据作处理,将原始数据整合为符合正态分布,将原始数据整合到归一化分布。同时为了比较算法的有效性,将传统的KNN算法、FCD-KNN算法,高斯贝叶斯(Gaussian Naive Bayes)算法、Gradient Boosting 4种方法作为对照组实验<sup>[10-11]</sup>。因此共进行了15组实验,然后对实验结果进行分析。

#### 3.1 实验数据处理

为了更为全面地分析数据,本文采用了3种数据预处理的方法,这三种方法各有利弊。本文会通过这5种算法的具体表现,验证其稳定性和有效性。

##### 3.1.1 未对原始数据作处理(只对数据标签数字化)

在这种情况下,只对数据进行了预处理,分析数据本来之间的关系。具体步骤是:将原始数据的标签进行数字化,具体是按序1,2,3的进行转化,“no”是1,“yes”是2,null是3。不进行其他转换,然后进行实验分析。这种情况下,保持了数据的基本特性,但数据中的奇异点可能会对实验精度有较大影响,从而降低一些依赖数值关系算法的精度,如:K-最近邻算法。

##### 3.1.2 将数据标准化成符合正态分布

大部分的数据分析都希望原始数据是满足正态分布的定距变量,这样数据分析更为精确,也会降低数据分析的复杂



度。数据标准化调整是非常有用的。许多机器学习算法在具有不同范围特征的数据中呈现不同的学习效果。例如, Gaussian Naive Bayes 在没有标准化调整过的数据中表现很差,因为可能一个变量的范围是 0 ~ 10 000,而另一个变量的范围是 0 ~ 1。因此,对数据预处理符合正态分布,是一种有效的分析手段。将数据处理为符合正态分布的公式为:

$$z = (x - \mu) / \sigma \quad (4)$$

其中:  $\mu, \sigma$  分别为原始数据集的均值和标准差。该种归一化方式要求原始数据的分布近似为高斯分布,否则归一化的效果会变得很糟糕。本文首先对原始数据进行了分析,发现其大致符合高斯分布,符合将数据正态分布化的先决条件。通过这种方式,可以使数据规范化,同时使数据分析更为简单。

### 3.1.3 将数据进行归一化到[0,1]

对原始数据进行标签数字化后,再对数据进行线性函数归一化。利用线性函数将原始数据线性化的方法转换到[0,1]的范围,归一化公式如下:

$$X_{\text{norm}} = (X - X_{\min}) / (X_{\max} - X_{\min}) \quad (5)$$

该方法实现对原始数据的等比例缩放,其中  $X_{\text{norm}}$  为归一化后的数据,  $X$  为原始数据,  $X_{\max}, X_{\min}$  分别为原始数据集的最大值和最小值。通过这种方法可以避免奇异点对数据分析造成的影响,但是会对数据的完整性和对比度造成影响。

## 3.2 实验流程

本文使用 Eclipse3 + Python3 + pydev 的开发环境,也可以使用 Java 开发环境 (JDK1.8 以上),进行仿真模拟实验。一共做 12 组实验,随机选取样本集的 70% 为训练集,30% 为测试集,先对处理后训练数据进行训练,然后再在测试集上进行训练,最后根据预测的精度来验证实验。

## 3.3 实验结果

### 1) 未对数据进行预处理的精度情况。

当未对数据进行预处理时(仅对标签进行数字化),Gaussian Naive Bayes 和 Gradient Boosting 算法表现的并不是特别理想,相比之下 3 种 KNN 算法的准确性更好,FCD-KNN 算法作为一种较新颖的算法在这种情况下表现略优于本文提出的改进 KNN 算法。未对数据进行预处理时,实验结果如表 1 所示。

表 1 未对数据进行预处理的结果

Tab. 1 Results of no pre-processed data

选用方法	精度
Improved KNN	0.949 613 252 437 653 3
Traditional KNN	0.930 808 448 652 585 6
FCD-KNN	0.961 258 147 309 458 0
Gaussian Naive Bayes	0.846 807 477 543 092 9
Gradient Boosting	0.912 600 145 666 423 9

### 2) 对数据预处理标准化成正态分布的精度情况。

根据 KNN 算法的特性,KNN 算法一般会很好地处理奇异点(比如:不归类),而本文改进的 KNN 算法会尽可能得将数据进行合理的分类;相比于 FCD-KNN 算法,对数据分类进行了进一步的处理,从而在银行数据分析预测中有更好的表现。对数据预处理标准化成正态分布时,实验结果如表 2 所示。

表 2 对数据预处理标准化成正态分布的结果

Tab. 2 Results of data standardized into normal distribution

选用方法	精度
Improved KNN	0.945 567 434 278 632 0
Traditional KNN	0.932 993 445 010 925 0
FCD-KNN	0.940 950 556 877 904 0
Gaussian Naive Bayes	0.848 992 473 901 432 4
Gradient Boosting	0.924 496 236 950 716 1

### 3) 对数据预处理归一化到[0,1]的精度情况。

相比于对数据进行正态化分布预处理的情形,对数据进行归一化处理得到的结果很相似。归一化后加快了梯度下降求最优解的速度。同时,如果一个特征值域范围非常大,那么距离计算就主要取决于这个特征,从而与实际情况相悖(比如这时实际情况是值域范围小的特征更重要)。这种方法非常适用于采用距离判断的 K-最近邻算法,通过这种方法,虽然此时 5 种预测算法的精度都有所下降,但是 3 种 KNN 算法还是明显优于其他 2 种算法,同时改进的 KNN 算法略优于其他两种的 KNN 算法。对数据预处理归一化到[0,1]时,实验结果如表 3 所示。

表 3 对数据预处理归一化到[0,1]的结果

Tab. 3 Results of data normalized to [0,1]

选用方法	精度
Improved KNN	0.946 080 055 467 216 3
Traditional KNN	0.919 397 912 114 590 9
FCD-KNN	0.941 346 587 430 683 0
Gaussian Naive Bayes	0.836 125 273 124 544 8
Gradient Boosting	0.910 172 371 934 935 6

## 3.4 整体实验结论分析

在整体结果中,本文提出的改进的 KNN 方法和 FCD-KNN 算法表现更好,说明本文提出的改进的 KNN 算法有一定的研究价值。分析原因,银行用户数据集不适合进行标准化,其噪声可以通过 SVM 的 RBF 核函数的处理,RBF 将数据集映射到高维上进行分类,从而有效减少了噪声的影响,在低维上进行计算。进一步的展望是先进行聚类算法,假设噪声都是一些离群点,将识别出来的很小的集合划为噪声,从而将噪声识别出来并剔除,进一步提高精度。改进的 KNN 方法采用了 VDM 距离法,而样本集中很多无法数字化比较的标签(如婚姻状态、工作状态等)很难作为数字因素考虑。FCD-KNN 算法也是对非数值的指标进行了考虑,但是本文提出的改进的 KNN 算法在数据分类过程中有更多的考虑,对实验结果产生了一些有利的结果。

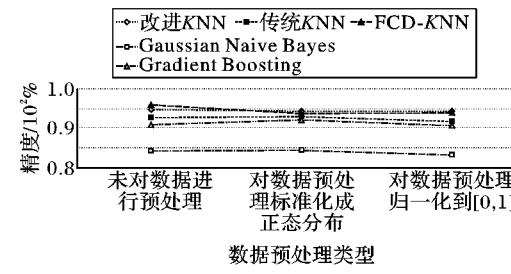


Fig. 7 Comparison of experimental results by different algorithms

而 Naive Bayes 方法相比于其他方法精度较低,原因可能



是:1)朴素贝叶斯方法需要先知道先验分布和数据来决定后验的概率从而决定分类,所以分类决策存在一定的错误率;2)理论上,朴素贝叶斯模型与其他分类方法相比具有最小的误差率。但实际上,因为朴素贝叶斯模型假设属性之间相互独立,这个假设在实际应用中往往是不成立的,在属性个数比较多或者属性之间相关性较大时,分类效果不好。

分析原因,可能是数据集中样本的属性之间有联系,分析银行客户资料,“工作类型”“教育”“住房”“贷款”等属性之间都可能不会有联系,所以这也是 Naive Bayes 方法精度比其他三种方法更低的原因。

3 种 KNN 方法在三组实验中均有优秀的实验结果,精度均在 0.92 左右或以上,预测精度都非常稳定。整体实验结果为:在不同预处理方式之间,不标准化(仅对标签数字化)>对数据预处理正态分布化>对数据预处理线性函数归一化。因为在本次数据集,标签并没有太多的数值关系,因此使用欧氏距离传统的 KNN 方法精度会下降,而采用 VDM 距离的改进的 KNN 方法和 FCD-KNN 方法均有突出的表现。而综合三种情况分析,本文提出的改进的 KNN 方法无疑是在银行数据挖掘分析预测中表现作为优秀和稳定的算法,其对于距离计算和聚类方式的改变,非常适用于银行情况,因此具有很大的潜力。

#### 4 结语

在大数据的背景下,对数据进行充分分析,可以减少实际工作中的成本。在金融行业对客户的分析预测显得尤为重要,数据分析聚类,可以给客户提供相应的个性化服务。本文所提出的改进的 K-最近邻算法,对传统的 K-最近邻算法进行距离计算和聚类方式的改变,通过实验分析与数据验证,以 2008—2013 葡萄牙银行数据作为样本集和测试集,对该算法进行验证,取得了非常理想的计算结果。与目前主流的其他算法相比,具有更好的稳定性和精确性,该算法在金融数据分析方面有良好的效果,有乐观的应用前景。

本文未来还会做以下工作:

1)本文研究的是处于金融危机下的人群,从社会学角度,这一时期的人群处于敏感时期,理财行为更为谨慎,因此要考虑本文研究的价值。

2)对数据的预处理是通常的数据挖掘中采用的手段,本文所提到数据预处理手段都较为简单,本文会未来尝试更多的预处理手段,使预测度更为精确。

#### 参考文献 (References)

- [1] GUO J Y, WANG X, LI Y. *k*NN based on probability density for fault detection in multimodal processes [J]. *Journal of Chemometrics*, 2018, 32(7): e3021.
- [2] FEKI-SAHNOUN W, NJAH H, HAMZA A, et al. Using general linear model, Bayesian networks and Naive Bayes classifier for prediction of Karenia selliformis occurrences and blooms [J]. *Ecological Informatics*, 2018, 43: 12–23.
- [3] SAINI I, SINGH D, KHOSLA A. QRS detection using *K*-Nearest Neighbor algorithm (*KNN*) and evaluation on standard ECG databases [J]. *Journal of Advanced Research*, 2013, 4(4): 331–344.
- [4] 职为梅, 张婷, 范明. 基于影响函数的 *k*-近邻分类 [J]. *电子与信息学报*, 2015, 37(7): 1626–1632. (ZHI W M, ZHANG T, FAN M. *k*-nearest neighbor classification based on influence function [J]. *Journal of Electronics and Information Technology*, 2015, 37(7): 1626–1632.)
- [5] 宓文斌. 数据挖掘在银行信贷业务中的应用 [D]. 上海: 上海交通大学, 2012. (MI W B. Application of data mining in the bank credit [D]. Shanghai: Shanghai Jiao Tong University, 2012.)
- [6] JIANG L, CAI Z, WANG D, et al. Survey of improving *k*-nearest-neighbor for classification [C]// Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery. Piscataway, NJ: IEEE, 2007: 679–683.
- [7] 肖辉辉, 段艳明. 基于属性值相关距离的 KNN 算法的改进研究 [J]. *计算机科学*, 2013, 40(S2): 157–159. (XIAO H H, DUAN Y M. Improved the KNN algorithm based on related to the distance of attribute value [J]. *Computer Science*, 2013, 40(S2): 157–159.)
- [8] 周治平, 苗敏敏. 改进的马氏距离动态时间规整手势认证方法 [J]. *计算机应用*, 2015, 35(5): 1467–1470. (ZHOU Z P, MIAO M M. Dynamic time warping gesture authentication algorithm based on improved Mahalanobis distance [J]. *Journal of Computer Applications*, 2015, 35(5): 1467–1470.)
- [9] de MAESSCHALCK R, JOUAN-RIMBAUD D, MASSART D L. The Mahalanobis distance [J]. *Chemometrics and Intelligent Laboratory Systems*, 2000, 50(1): 1–18.
- [10] TAHERI S, MAMMADOV M. Learning the naive Bayes classifier with optimization models [J]. *International Journal of Applied Mathematics and Computer Science*, 2013, 23(4): 787–795.
- [11] BIAU G, CADRE B, ROUVIÈRE L. Accelerated gradient boosting [J]. *Machine Learning*, 2019, 108(6): 971–992.
- [12] 杨朔, 陈丽芳, 石娟, 等. 基于深度生成式对抗网络的蓝藻语义分割 [J]. *计算机应用*, 2018, 38(6): 1554–1561. (YANG S, CHEN L F, SHI Y, et al. Semantic segmentation of blue-green algae based on deep generative adversarial net [J]. *Journal of Computer Applications*, 2018, 38(6): 1554–1561.)

This work is partially supported by the National Key Research and Development Program of China (2018YFB1004401).

**LI Bo**, born in 1994, M. S. candidate. His research interests include cloud storage, data mining.

**ZHANG Xiao**, born in 1978, Ph. D., associate professor. His research interests include storage system.

**YAN Jingyi**, born in 1993, M. S. Her research interests include technology innovation management.

**LI Kewei**, born in 1993, M. S. candidate. His research interests include data mining.

**LI Heng**, born in 1993, M. S. candidate. His research interests include data mining.

**LING Yulong**, born in 1995, M. S. candidate. His research interests include data mining.

**ZHANG Yong**, born in 1995, M. S. candidate. His research interests include data mining.