计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200200031

# 基于 Canopy 和共享最近邻的服务推荐算法

邵欣欣
大连东软信息学院　辽宁 大连 116023

**摘　要**　为辅助银行机构进行精准的服务推荐,提出了基于改进的 Canopy 和共享最近邻相似度的聚类算法。基于该算法对用户进行细分,针对用户群特点进行精准服务推荐。该算法首先采用最大值和最小值对 Canopy 算法进行改进,并应用该算法得出初始聚类结果,然后利用共享最近邻相似度算法对聚类结果中的交叉部分数据进行归类,最终得出用户聚类数据。该算法在某银行对真实客户数据进行应用,选择基于客户的贡献度、忠诚度和活跃度 3 个指标进行聚类,结果表明,该算法提高了客户细分的质量和聚类的效率,聚类结果对于客户的消费数据刻画非常准确,能够为银行的精准服务推荐提供数据支持。

**关键词**:客户聚类;服务推荐;Canopy 算法;共享最近邻相似度;聚类指标
**中图法分类号**　TP391.9

## Service Recommendation Algorithm Based on Canopy and Shared Nearest Neighbor

SHAO Xin-xin
Dalian Neusoft University of Information,Dalian,Liaoning 116023,China

**Abstract**　In order to improve the accuracy of banking institution service recommendation,a clustering algorithm based on the improved Canopy and shared nearest neighbor similarity is proposed. Based on this algorithm,users are subdivided and accurate service recommendation is made according to the characteristics of user groups. First,the improved Canopy algorithm is used to obtain the initial clustering results. Then the shared nearest neighbor similarity algorithm is used to classify the intersecting data in the clustering results. Finally,the user clustering data are obtained. The algorithm is applied to the real customer data of a bank. Three indexes of customer contribution,loyalty and activity are selected for clustering. The results show that the algorithm improves the quality of customer segmentation and the efficiency of clustering. The result of clustering is very accurate in describing the consumption data of customers. Clustering results can provide data support for accurate service recommendation of banks.

**Keywords**　Customer clustering,Service recommendation,Canopy algorithm,Shared nearest neighbor similarity,Clustering index

## 1　引言

银行机构拥有海量的客户,对客户进行细分,根据客户的特点为其推出合适的金融产品,并提供个性化服务推荐,可以使营销更具有针对性,能够有效提高银行的利润和竞争力。

当前,广泛使用的银行客户细分算法是 K-means 算法,该算法需要事先确定聚类个数,并且聚类中心点的选取也具有一定的随机性,因而容易陷入局部最优解[2]。随着数据量的增大,K-means 算法在对数据量激增的情况进行分析时,难以满足速度方面的需求[1-3]。另一个广泛使用的算法是 Canopy 算法,该算法无需指定类别数,运行速度极快,适合处理大规模的数据集,但是在不同类别的边界处,极易出现聚类重叠的现象[4-5]。K 最近邻(k-NearestNeighbor,KNN)分类算法也是一种常用的聚类算法,该算法简单易用、精度高、理论成熟,能够更好地分析相似客户的行为,更好地对客户进行分类,但是该算法本身计算量较大,不适合数据量较大的情况[6-7]。本文把 Canopy 和 KNN 算法结合应用于客户细分,既可以克服 Canopy 重叠数据过多的问题,又可以减少 KNN 的计算量。

本文首先基于 Canopy 算法进行研究,从两个方面对算法进行改进:1)采用最大最小原则对初始中心点的选取进行优化;2)针对聚类重叠的问题,提出采用 KNN 算法的共享最近邻相似度的方式对重叠部分的数据进行归类。基于以上改进,提出 Canopy 和 KNN 结合应用的聚类算法,实验结果表明,改进后的算法能有效处理 Canopy 算法结果中的交叉数据,并减少 KNN 的计算量,对大量数据的处理具有耗时短和处理速度快的优点。

## 2　服务推荐问题描述

为了对银行客户进行精准的服务推荐,首先需要对客户进行聚类,确定聚类指标和客户数据。在应用系统中,基于设计的业务场景,主要采集客户的细分指标,如基本信息、交易信息与外部信息,客户聚类将通过细分模型构建、细分结果分析、细分结果应用 3 个步骤进行建模与分析。针对客户,采用

基金项目:辽宁省自然科学基金(2019-ZD-0354)
This work was supported by the Natural Science Foundation of Liaoning Province,China(2019-ZD-0354).
通信作者:邵欣欣(sxx929@163.com)

Canopy 和 KNN 结合应用的方式，首先利用 Canopy 算法确定初始聚类数 $k$，然后利用共享最近邻相似度算法对重叠数据进行归类，得出用户群，在用户群中按照相似用户的偏好进行智能协同推荐。客户聚类的算法流程如图 1 所示。

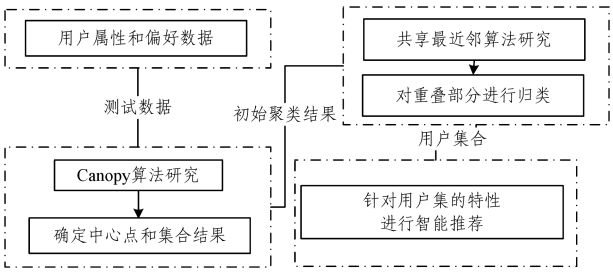



图 1 推荐流程

Fig. 1 Recommendation process

## 3 客户聚类算法

首先使用 Canopy 算法确定初始聚类个数，然后采用最近邻相似度算法优化聚类交叉的问题。

### 3.1 定义

**定义 1** 对于数据集 $D=\{d_i \mid i=1,2,\cdots,n\}$，存在 $\forall x_i \in D$，其中 $x_i$ 如果满足：$\{C_j \mid \exists \parallel x_i - C_j \parallel \leqslant T_1, C_j \subseteq D, i \neq j\}$，则 $x_i$ 属于 Canopy 数据集，$C_j$ 为 $D_{\text{Canopy}}$ 的中心点，$T_1$ 为 Canopy 数据集的半径[8]。

**定义 2**（候选中心点） 对于任意数据集 $D=\{d_i \mid i=1,2,\cdots,n\}$，存在 $\forall x_i \in D$，$x_i$ 如果满足 $\{C_m \mid \exists \parallel x_i - C_m \parallel \leqslant T_2, T_2 < T_1, C_m \subseteq D, i \neq j\}$，则 $C_m$ 为非 Canopy 候选中心点。

**定义 3**（最近邻集） 数据对象 $d_i$ 的最近邻定义为 $N_n(d_i)$，最近邻集为距离 $d_i$ 最近的数据集合[9]。

**定义 4**（共享最近邻集） 数据对象 $d_i$ 和 $d_j$ 的共享最近邻集定义为 $Snn(d_i,d_j)=N_n(d_i) \cap N_n(d_j)$，即 $N_n(d_i)$ 和 $N_n(d_j)$ 的交集。基于“同类相近”的原则，两个数据对象越相近，那么它们的共享最近邻就越相似，因此把数据对象共享最近邻的交集定义为共享最近邻集。

**定义 5**（共享最近邻相似度） 数据对象 $d_i$ 和 $d_j$ 的共享最近邻集的数量定义为共享最近邻相似度，记为 $Sim(d_i,d_j)=Count(Snn(d_i,d_j))$。

### 3.2 算法描述

Canopy 算法利用距离的计算方法把待处理的数据集划分为存在重叠现象的几个子集[10]，需要选取两个阈值 $T_1$ 和 $T_2$ 对需要聚类的数据集进行计算，将其聚类为 $k$ 个相互交叉的集合。首先通过 Canopy 算法进行初始的聚类分析；然后对于交叉部分采用最近邻相似度进行归类。通常对于客户数据集合中的任意两个客户 $d_i$ 和 $d_j$，它们的相似度越高，属于同一个聚簇的可能性就越大[11]，因此对于交叉部分的客户，可以将其归类到和它最相似的最近邻所属的聚簇中。算法如算法 1 所示。

**算法 1**

输入：向量化的银行客户数据集合

输出：聚类集合

Step 1 将客户数据按照贡献度进行排序，得到一个 list。

Step 2 选择距离阈值，分别为 $T_1$ 和 $T_2$，$T_1$ 为 Canopy 集合的半径，$T_2$ 为非 Canopy 候选中心点集合的半径，$T_1 > T_2$。根据定义 1 和定义 2 可知，$T_1$ 和 $T_2$ 值的选取直接影响聚类结果，$T_1$ 范围过大会增加重叠部分数据的数量，增加后续共享最近邻相似度的计算开销；$T_2$ 范围过大会导致客户聚类数量减少，为了合理选取 $T_1$ 和 $T_2$ 的值，文中计划采用最大最小原则优化中心点的选择过程，对于已有的 k 个中心点，第 k+1 个中心点的选择根据最大最小原则，计算候选数据点中与前 k 个中心点之间距离最大的点，选择这 k 个最大距离中的最小者对应的点，作为第 k+1 个中心点。在应用时，首先把客户信息数据划分为两个子集，采用多种阈值进行测试，找出聚类效果最好的阈值 $T_1$ 和 $T_2$[12]。

Step 3 从数据集 list 中任意选取一个客户数据 D，计算当前客户数据 D 与 list 中其他数据之间的距离。如果数据 D 与 list 中的某个向量之间的距离在阈值 $T_1$ 的范围内，就把数据 D 聚类到 Canopy 内，根据定义 2 此用户可以作为新的 Canopy 的中心点，使用此数据计算其他的客户数据和此中心点的距离。

Step 4 如果客户数据 D 与某 Canopy 的距离在阈值 $T_2$ 的范围内，则把客户数据 D 从 list 中删除，可以确认客户数据 D 基本属于该 Canopy，数据 D 不能再作为其他 Canopy 的中心。

Step 5 循环 Step2－Step4，list 为空则循环结束。

Step 6 得出的客户数据集 D 的聚类中存在交叉数据，对交叉部分的数据对象 $d_i$ 进行归类。以当前数据对象为中心点（图 2 中的实心三角形），确定距离最近点的阈值，如图 2 虚线部分所示。首先以最小半径查找共享最近邻数量，如果数量不满足阈值，则扩大半径进行查找，直到满足阈值为止，当达到此阈值时，计算与 $d_i$ 距离最近的点所属的集合。然后根据定义 3，计算得出 $d_i$ 的共享最近邻，选取过程如图 2 所示。

Step 7 根据定义 4 和定义 5，确定 $d_i$ 和其他数据元素的共享最近邻相似度，找到与 $d_i$ 相似度最高的客户 $d_j$，把 $d_i$ 划分到 $d_j$ 所属的类别中。

Step 8 重复 Step 7，直到交叉数据处理完毕。

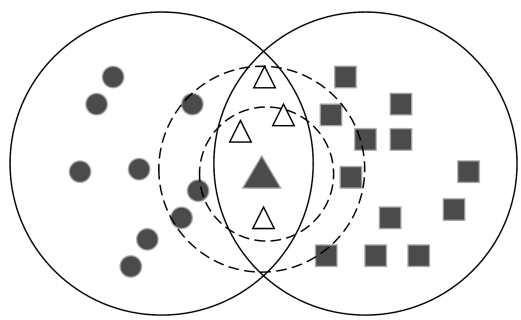


图 2 最近邻选取图

Fig. 2 Nearest neighbor selection

## 4 算法实施

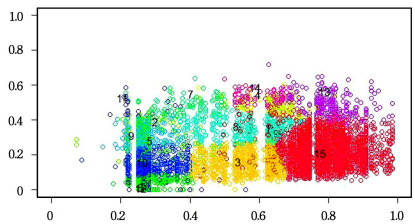
为了测试算法的可行性和性能，将实验结果与 KNN 算法和原始的 Canopy 算法进行了比较。数据集来源于某银行的真实数据，数据覆盖大约 5 万名银行客户，共计 200 万条交易记录，银行数据记录真实有效。

1）算法可行性分析

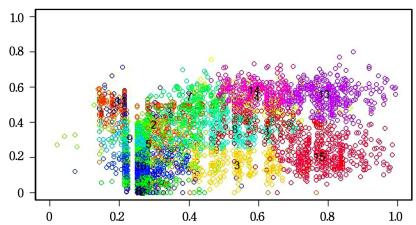
首先对原始数据进行处理，提取客户的贡献度、活跃度和忠诚度的数值。其中贡献度的选取采用客户卡产品 FTP 利润来评价；客户活跃度等于过去 6 个月每月交易活跃度的汇总 * 0.5＋授信额度利用率 * 0.5；忠诚度采用交易总金额、交易笔

数和年均总资产的比值来计算,数值越小代表忠诚度越好。本数据的处理方法采用银行业务中普遍采用的方法。

Canopy 和 KNN 算法在活跃度和贡献度两个维度下进行聚类,首先采用 Canopy 算法进行聚类,对于重叠部分采用共享最近邻相似度算法进行划分,设置共享最近邻的数量分别为 6,8,12,16,20 进行聚类,最后综合考虑计算速度和聚类结果的准确度,确定共享最近邻阈值为 8。样本数据的聚类结果如图 3 所示,两组数据各生成 15 个聚类。



(a)样本 1 的聚类结果



(b)样本 1 的聚类结果

图 3 两个样本的聚类结果

Fig. 3 Clustering results of two samples

图 3 中数字所在的位置即为聚类中心。从聚类结果可以看出,聚类中心选取基本合理,图 3(a)中数据聚类较为集中,图 3(b)客户数据存在明显的交叉重叠现象,经过运算,也都进行了合理的划分,聚类数量较为稳定。全部样本数据的聚类结果如图 4 所示,聚类结果基本与训练集聚类结果一致,由此可见本文提出的聚类算法聚类结果比较稳定。

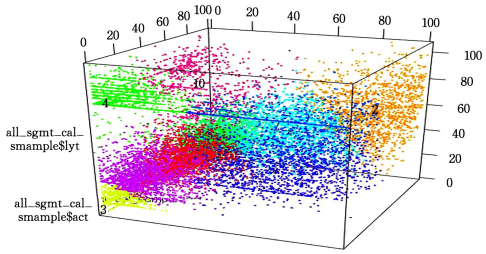



图 4 聚类结果

Fig. 4 Clustering results

2)结果分析

本文算法对于重叠部分的数据聚类采用共享最近邻相似度算法。首先需要设置最近邻数个数,当最近邻个数增多时,会增加算法的运行时间,但是因为只需要计算重叠部分的数据,所以比单纯使用 KNN 算法更节省时间、效率更高。为了测试该方法对传统的 KNN 算法的改进,在共享最近邻个数相同的情况下,对两种算法的运行时间进行对比。图 5 给出了参数设置与聚类质量之间的关系,其中横坐标是参数 $k$ 的取值,纵坐标是运行时间。从图中可以看出,两种聚类算法随着最近邻个数的增加,耗时都相应增加,但是本文算法的耗时明显低于 KNN 算法。

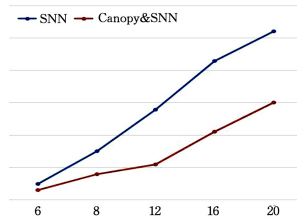



图 5 运行时间比较

Fig. 5 Comparison of running time

由图 4 可见,聚类后生成的数据集合边界较为清晰,能够区分客户类别。根据细分数据把高贡献高活跃、高贡献一般活跃、高活跃一般贡献和高活跃低贡献的客户确定为优质客户,从优质客户的业务特征进行分析,可以明显区分取现偏好与消费偏好的客户群体,高贡献高活跃、高贡献一般活跃、高活跃一般贡献这 3 类客户都属于取现类的优质客户,而高活跃低贡献的客户属于消费类优质客户。优质客户为卡产品业务做出了巨大贡献,统计后发现卡产品某个月份 84.6%的利润都来自于优质客户,因此可以说此类客户的行为特征验证了取现类与消费类的交易偏好,同时刻画了客户的消费偏好。根据以上聚类结果和消费统计结果,可以针对消费类客户,重点推荐分期付款等类别的服务,对取现类的用户重点推荐消费折扣等类别的服务。

**结束语** 本文在海量数据环境下利用聚类方法,对银行卡类客户进行了聚类分析。首先采用 Canopy 算法确定 $k$ 的个数,然后采用最近邻相似度算法对重叠部分数据进行归类。在银行海量客户数据的处理方面,本文算法能有效减少迭代次数,提高处理效率,同时因为聚类数量不需要人为选取,避免了局部最优解,具有较强的可操作性。通过改进的算法分析得出贡献度与活跃度较高的优质客户大约占全行总人数的比例,根据分析结果能够得出优质客户的消费偏好,进而便于银行对优质客户进行各类业务的推荐。

## 参 考 文 献

[1] JI S Q,SHI H B. Optimized K-means clustering algorithm for massive data[J]. Computer Engineering and Applications,2014,50(14):143-147.

[2] HAN L B,WANG Q,JIANG Z F,et al. Improved K-means initial clustering center selection algorithm[J]. Computer Engineering and Applications,2010,46(17):150-152.

[3] CUI X L,ZHU P F,YANG X. Optimized big data K-means clustering using MapReduce[J]. The Journal of Supercomputing,2014(6):1249-1259.

[4] HE H,GUO L,GENG Y. The optimization of CMAC neural network structure based on Canopy-K-Means algorithm[J]. International Journal of Advancements in Computing Technology,2012,4(22):641-647.

[5] WANG Y G,WU C,DAI W. K-means algorithm of random sample based on MapReduce[J]. Computer Engineering and Applications,2016,52(8):74-79.

[6] CAO Y,WANG Y L,HE H M. Intelligent scheduling in preburdening of iron ore:Canopy-Kmeans clustering algorithm and combinatorial optimization[J]. Control Theory & Applications,2017,34(7):947-955.

(下转第 503 页)