

基于卷积核的港口客户细分方法

王 宇*, 徐 艳

(大连理工大学 管理学院, 辽宁 大连 116024)

摘要: 根据港口客户数据特点,运用信息增益方法对其进行了数据预处理,将其表示为树形结构组织方式,得到 216 棵客户树;引入卷积核,定义了度量客户树之间相似性的卷积树核;随后,将先前提出的核 k -凝聚聚类算法推广到基于卷积核的客户树上,并运用 Matlab 数据处理工具实现对港口客户数据的聚类分析.分析结果表明,卷积核在港口客户细分中得到了良好的应用效果.

关键词: 卷积核;核 k -凝聚聚类算法;港口客户细分
中图分类号: G202 **文献标志码:** A

0 引 言

随着商品经济发展和市场竞争日益激烈,客户关系管理(CRM)越来越受到企业管理者的重视.作为服务型企业的港口的竞争也已从传统能力上的竞争转移到人力资源、客户关系管理上^[1,2].在港口企业中引入客户关系管理理论,实现客户细分,来进一步提高企业的管理水平、拓展业务范围,对于促进企业生产发展具有很重要的现实意义.

客户细分(customer segmentation)概念来源于 20 世纪 50 年代出现的市场细分^[3],是指按照一定的标准将企业的现有客户划分为不同的客户群,并提供有针对性的产品、服务或营销模式的过程.企业实施客户细分的关键是要有恰当的实现方法.国内外学者对此进行了多方面研究,提出了多种方法.Boone 等研究了用 Hopfield 人工神经网络技术进行客户细分的 Hopfield-Kagmar (HK) 聚类方法^[4];Tsai 等提出了基于 k -means 算法的客户行为细分模型^[5];张国政对当前存在的客户细分方法进行了归纳总结,并提出一种基于客户满意度的客户细分方法^[6];文献^[7~9]在电信行业的客户细分中提出了各自的聚类算法;文献^[10]将模糊聚类算法引入客户细分中也取得

了不错的效果;吴春旭等最近提出了一种基于动态模糊聚类算法的客户细分方法^[11];叶强等将“云模型”引入到客户细分之中,较好地处理了动态客户分类问题^[12,13];文献^[14,15]则将核聚类算法应用到客户细分,改进了传统聚类算法的缺点,在聚类分析和客户细分中取得了较好的效果.但在众多客户细分的研究中对港口客户细分的研究却较少.由于港口客户数据结构不同于电信等行业直接以客户记录形式表示,上述客户细分方法无法直接使用.针对这一状况,本文以营口港客户数据为研究对象,在对港口客户数据重新组织成树形结构的基础上,借助卷积核设计一个核 k -凝聚聚类算法,并运用 Matlab 数据处理工具验证其细分效果.

1 港口客户数据组织与表示

营口港 1992 年开始信息化建设,目前已建立起电子港务平台、电子商务平台、电子口岸等 3 个非常稳定的电子平台,并且这 3 个电子平台积累的数据全部存储在信息中心.

港口企业的核心产品是装卸服务,故选取港口最重要的生产相关数据库中存储的数据作为研

究对象. 生产相关数据库包括数据表 300 多张, 如合同货物表(cont_cargo)、船只登记表(ship)、单船合同表(contract)、货物流向表(cont_assign)、

货物种类表(cargo_kind)、合同费用表(cont_fee)等. 图 1 为合同货物表(cont_cargo)中的部分数据.

CONT_NO	CARGO_NO	CARGO_NAM_COD	CARGO_NAM	WEIGHT_WGT	TRADE_ID	SHIPPER_DOC
H06121220	35843	DYM	电熔镁	260	<NULL>	大连港汇国际货运代理有限公司鲅鱼
H06121220	38758	JZD	集装箱	0.6	<NULL>	大连港汇国际货运代理有限公司鲅鱼
H06121179	41168	HXG	H型钢	49.5	1	营口国丰物流有限公司
H06121179	41266	QG	轻轨	45	1	营口国丰物流有限公司
H06121221	30897	DXJ	镀锌卷板	499	<NULL>	鞍钢国贸鲅鱼圈公司
H06121223	36010	LZJB	冷轧卷板	1038.8	1	营口经济技术开发区振岐运输有限公
H06121224	34122	YM	玉米	8700	1	营口港博丰物流有限责任公司

图 1 港口数据库合同货物表(cont_cargo)中的部分数据

Fig. 1 Part of data in cont_cargo table in port database

通过对所选数据库进行分析, 发现原始数据表不仅数量大(仅生产相关的数据库中数据表就有 300 多张, 表中记录数量相当大, 如船只登记表(ship)中有 7 800 多条记录), 而且数据表中的数据属性多(有八十维之多). 属性过多, 会降低分类效果, 为此, 需要通过属性相关性分析, 识别出弱相关或不相关的属性, 将它们排除在客户描述之外. 本文使用信息增益分析技术删除对客户分类信息量较少的属性. 计算信息增益的方法如下:

设 S 是样本集合, 其中每个样本的类标号是已知的. 假定有 m 个类, 设 S 包含 s_i 个 C_i 类样本, $i = 1, 2, \dots, m$. 一个任意样本属于类 C_i 的可能性是 s_i/s , 其中 s 是集合 S 中对象的总数. 对一个给定样本分类所需的期望信息是 $I(s_1, s_2, \dots, s_m) = - \sum_{i=1}^m \frac{s_i}{s} \log_2 \frac{s_i}{s}$. 具有值 $\{a_1, a_2, \dots, a_v\}$ 的属性 A 可以用来将 S 划分为子集 $\{S_1, S_2, \dots, S_v\}$, 其中 S_j 包含 A 中值为 a_j 的那些样本. 设 S_j 包含类 C_i 的 s_{ij} 个样本, 则根据 A 的这种划分的期望信息称做 A 的熵, 其定义为加权平均, 即 $E(A) = \sum_{j=1}^v \frac{s_{1j} + \dots + s_{mj}}{s} I(s_{1j}, s_{2j}, \dots, s_{mj})$. A 上该划分获得的信息增益定义为 $G(A) = I(s_1, s_2, \dots, s_m) - E(A)$.

经过计算、选择, 最后选出 27 个与客户细分相关性较大的属性. 具体选择的属性及其信息增益值如表 1 所示.

为了进行客户细分, 必须将经过上述处理后

的数据, 表示成以客户名组织的数据, 将同一个客户对应的所有数据作为一个整体研究. 在港口交易中, 一个客户与港口可能进行多次交易, 相应的在数据表中, 一个客户会对应多个合同号, 一个客户会对应多条数据, 每个合同中可能涉及多种货物或同种货物到岸后的不同处理, 同一合同号也会对应多条数据. 对客户进行聚类分析, 若按照传统的一条记录表示一个合同中一种货物的相关信息, 一个客户与港口的交易将会表示成多条记录, 且记录间有很多属性有相同的属性值. 根据上述客户、合同、货物之间的关系, 本文将一个客户所有交易数据表示为一个树形结构, 如图 2 所示.

表 1 数据属性及其信息增益值

Tab. 1 Data attributes and their information gain values

数据属性	信息增益值	数据属性	信息增益值
货主所处地区	2.526 4	总吨	2.318 3
交易次数	3.823 5	净吨	2.262 2
合同总金额	4.459 7	船舶性质	0.359 2
船主	0.350 2	内外贸	0.349 4
船型	0.069 0	计重方式	0.696 3
代理名称	0.517 0	货名	2.712 0
船国籍	0.476 6	合同方	0.119 5
起运港	0.998 2	完成件数	2.100 7
到达港	1.142 1	费率	1.909 3
进出口	0.110 2	合同金额	4.347 9
船长	2.043 2	包装	0.806 0
船进口重量	2.141 2	堆存地	1.750 5
完成重量	4.058 3	作业地	1.711 5
派遣方式	0.114 8		

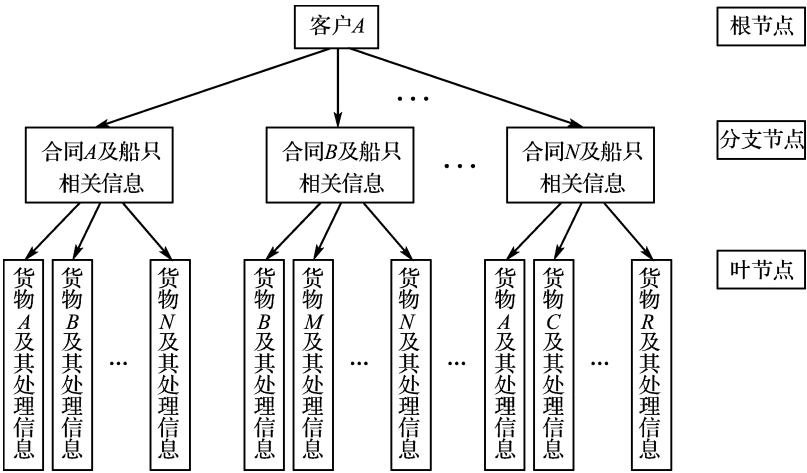


图 2 港口客户 A 树形表示

Fig. 2 Tree structure representation of port customer A

为了实施客户细分,还需再将选出的 27 个属性分别表示为上述的根节点、分支节点、叶节点. 其中根节点是由属性货主所处地区、交易次数、合同总金额组成的向量,描述了客户的总体情况. 分支节点是由属性船主、船型、代理名称、船国籍、起运港、到达港、进出口、船长、船进口重量、总吨、净吨、船舶性质、派遣方式、内外贸组成的向量,描述了客户对应的各合同中交易的相关信息及对于港口交易非常重要的船只信息,节点数量由客户与港口形成的合同数决定,每个合同对应一个节点. 叶节点是由属性合同方、完成件数、完成重量、费率、合同金额、货名、包装、作业地、计重方式、堆存地组成的向量,详细记录和描述了货主与港口交易的具体内容和方式,叶节点总数作为第一层中交易次数的值. 每个叶节点合同金额的加和作为根节点中的合同总金额的值. 这样就清晰地描述了一个客户的交易情况,并减少了数据冗余.

本文从港口相关数据库中选取出的客户数据表示为上述树形结构,得到 216 棵客户树,所有的数据被表示为有对应关系的 3 个矩阵,分别存放根节点、分支节点、叶节点,大小分别为 216×3 维、 $2\,482 \times 14$ 维、 $9\,568 \times 10$ 维,共 131 076 个数据元素. 如果用传统的每次交易一条记录的方式,则需要建立 $9\,568 \times 27$ 维的矩阵,共 258 336 个数据元素. 可以看出树形结构数据表示方式将数据压缩了近一半.

2 卷积核及客户树间相似性定义

经预处理,港口客户数据被表示为树形结构. 对树形结构的数据进行聚类分析,就是将每一个客户树看成一个数据整体,对其进行聚类. 聚类分析的前提是定义数据间的相似性度量. 传统聚类方法中是对数据库中的一条条记录型数据定义距离度量其相似性,这种度量方法无法直接用于树形结构数据间的相似性度量. 为此,引入卷积核概念.

卷积核(convolution kernels)^[16]是一种通过对定义在对象子结构上的核函数的卷积来定义对象的核函数. 这里,核函数 $k(x, x')$ 是指对任意数据点 $x_1, x_2, \dots, x_m \in X$ 和参数 $c_1, c_2, \dots, c_m \in \mathbf{R}$ 均满足 $\sum_{i,j=1}^m c_i c_j k(x_i, x_j) \geq 0$ 的对称函数^[17].

由核函数定义^[17],任意核函数均存在一个映射 $\Phi: X \rightarrow F$,满足 $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$. 将输入空间映射到特征空间 F ,使得利用核函数可以直接计算特征空间中向量之间的内积. 因此,核函数可以看成两个数据点之间的一种相似性度量. 下面就给出卷积核的定义和一个定理^[16].

定义 1 设 $x, y \in X, \vec{x} = x_1, x_2, \dots, x_D \in \mathbf{R}^{-1}(x), \vec{y} = y_1, y_2, \dots, y_D \in \mathbf{R}^{-1}(y)$. 又设 k_i 是 $X_i \times X_i$ 上的核函数, $1 \leq i \leq D, k_i(x_i, y_i)$ 给出 x_i 与 y_i 的相似性度量. x 与 y 的相似性度量由以下

的广义卷积给出：

$$K_{\text{cov}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\substack{\vec{x} \in \boldsymbol{R}^{-1}(\boldsymbol{x}), \\ \vec{y} \in \boldsymbol{R}^{-1}(\boldsymbol{y})}} \prod_{i=1}^D k_i(\boldsymbol{x}_i, \boldsymbol{y}_i) \quad (1)$$

式(1)在 $\boldsymbol{S} \times \boldsymbol{S}$ 上有定义,其中 $\boldsymbol{S} = \{\boldsymbol{x} : \boldsymbol{R}^{-1}(\boldsymbol{x}) \text{ 非空}\} \subset \boldsymbol{X}$. K 在 $\boldsymbol{X} \times \boldsymbol{X}$ 上的零扩张称为核函数 $k_1, k_2, \dots, k_D$ 的 R -卷积,记做 $k_1 * k_2 * \dots * k_D$.

定理 1 设 $k_1, k_2, \dots, k_D$ 分别是 $\boldsymbol{X}_1 \times \boldsymbol{X}_1, \boldsymbol{X}_2 \times \boldsymbol{X}_2, \dots, \boldsymbol{X}_D \times \boldsymbol{X}_D$ 上的核函数, R 是 $\boldsymbol{X}_1 \times \dots \times \boldsymbol{X}_D \times \boldsymbol{X}$ 上的有限关系,则 $k_1 * k_2 * \dots * k_D$ 是 $\boldsymbol{X} \times \boldsymbol{X}$ 上的核函数.

由定义 1 和定理 1 可知,使用卷积核,首先要将数据划分为不同的部分. 针对客户数据树形结构,有两种划分方法将其划分为不同的部分:一是按层划分,分别将根节点、分支节点、叶节点作为不同的 3 个部分;二是按子树划分,将根节点有几个子树就划分为几部分. 由于第 2 种方法中,不同的客户树划分成数量不等的部分,每个部分仍然是一棵树,且树中每个节点都是向量,仍然不适合用传统的核函数比较其相似度. 于是,本文选择第 1 种划分方法,将每棵客户树划分为 3 个部分,构造适合客户树的卷积树核 $K_{\text{cov}}(\boldsymbol{x}, \boldsymbol{y})$,衡量客户树 $\boldsymbol{x}, \boldsymbol{y}$ 之间相似度:

$$K_{\text{cov}}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^3 \lambda_i k_i(\boldsymbol{x}_i, \boldsymbol{y}_i); i = 1, 2, 3 \quad (2)$$

其中 $k_i(\boldsymbol{x}_i, \boldsymbol{y}_i)$ 表示客户树 $\boldsymbol{x}, \boldsymbol{y}$ 中第 i 部分相似度的核函数. 经划分部分后,每个部分都为属性向量,对应数据库中的记录,可以用传统的核函数衡量记录之间的相似度. 客户树中,每个节点都是一个属性向量,由前面的分析,每个属性对分类的信息增益不同,因此对每个部分加入权重因子 λ_i ($0 < \lambda_i < 1$),表示不同部分包含的属性的信息增益值,其值为每个部分包含的属性信息增益加和后经标准化处理的结果.

客户树的第 1 部分(层)为根节点,对应数据库中的一条记录,其核函数 $k_1(\boldsymbol{x}_1, \boldsymbol{y}_1)$ 直接可以用传统的 Gauss 核函数 $k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{\delta^2}\right)$ (δ 为参数) 衡量其相似度.

第 2 部分(层)为分支节点,根据签订的合同数不同,每个节点包含不同维数的向量,其核函数

为 $k_2(\boldsymbol{x}_2, \boldsymbol{y}_2) = \sum_{j=1}^{s_x} \sum_{k=1}^{s_y} k(\boldsymbol{x}^j, \boldsymbol{y}^k)$,其中 s_x, s_y 表示树 $\boldsymbol{x}, \boldsymbol{y}$ 第 2 部分的节点数.

第 3 部分(层)为叶节点,其核函数可取为

$$k_3(\boldsymbol{x}_3, \boldsymbol{y}_3) = \sum_{j=1}^{t_x} \sum_{k=1}^{t_y} k(\boldsymbol{x}^j, \boldsymbol{y}^k), \text{ 其中 } t_x, t_y \text{ 表示树 } \boldsymbol{x}, \boldsymbol{y} \text{ 第 3 部分的节点数.}$$

在以上各层核函数中都包含 Gauss 核函数 $k(\boldsymbol{x}^j, \boldsymbol{y}^k)$, Gauss 核函数中的距离 $\|\boldsymbol{x} - \boldsymbol{y}\|^2$ 可定义为^[17]

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^2 = \sum_{i=1}^p \lambda_i (\boldsymbol{x}_i - \boldsymbol{y}_i)^2 + \gamma \sum_{j=p+1}^m \lambda_j \delta(\boldsymbol{x}_j, \boldsymbol{y}_j) \quad (3)$$

这里,设式(3)中每个向量为 m 维,其第 i 维 ($i = 1, 2, \dots, p$) 为数值型属性;第 j 维 ($j = p+1, \dots, m$) 为分类型属性, λ_j 为每个属性的权重,取各个属性的信息增益标准化后的值; γ 为属性平衡因子; $\delta(\boldsymbol{x}_j, \boldsymbol{y}_j)$ 的定义如下:

$$\delta(\boldsymbol{x}_j, \boldsymbol{y}_j) = \begin{cases} 0; & \boldsymbol{x}_j = \boldsymbol{y}_j \\ 1; & \boldsymbol{x}_j \neq \boldsymbol{y}_j \end{cases}$$

3 基于卷积核的核 k -凝聚聚类算法

采用卷积核实现了两客户树之间的相似性度量,为实施客户聚类分析奠定了基础. 考虑到港口客户数据的特点,先将客户树 $\boldsymbol{x}_i, i = 1, 2, \dots, 216$,用非线性映射 $\boldsymbol{\Phi}(\cdot)$ 变换到一个高维特征空间,再在该特征空间上扩展核 k -凝聚聚类算法.

在高维特征空间中,两个客户树之间的距离为

$$\|\boldsymbol{\Phi}(\boldsymbol{x}_i) - \boldsymbol{\Phi}(\boldsymbol{c}_l)\|^2 = K(\boldsymbol{x}_i, \boldsymbol{x}_i) - 2K(\boldsymbol{x}_i, \boldsymbol{c}_l) + K(\boldsymbol{c}_l, \boldsymbol{c}_l)$$

其中核函数 $K(\cdot, \cdot)$ 取卷积核函数 $K_{\text{cov}}(\cdot, \cdot)$. 依据文献[14]中核 k -凝聚聚类算法定义,此时核函数为

$$p(i, l) = p(\boldsymbol{\Phi}(\boldsymbol{x}_i), \boldsymbol{\Phi}(\hat{\boldsymbol{c}}_l)) = \left\{ \exp\left(\left(-\frac{1}{\tau}\right)(K_{\text{cov}}(\boldsymbol{x}_i, \boldsymbol{x}_i) - 2K_{\text{cov}}(\boldsymbol{x}_i, \hat{\boldsymbol{c}}_l) + K_{\text{cov}}(\hat{\boldsymbol{c}}_l, \hat{\boldsymbol{c}}_l))\right) \right\} /$$

$$\left\{ \sum_{j=1}^k \exp \left(\left(-\frac{1}{\tau} \right) (K_{\text{cov}}(\mathbf{x}_i, \mathbf{x}_i) - 2K_{\text{cov}}(\mathbf{x}_i, \hat{\mathbf{c}}_j) + K_{\text{cov}}(\hat{\mathbf{c}}_j, \hat{\mathbf{c}}_j)) \right) \right\} \quad (4)$$

新的类中心为

$$\begin{aligned} \Phi(\hat{\mathbf{c}}_l) &= \frac{\sum_{i=1}^n p(\Phi(\mathbf{x}_i), \Phi(\mathbf{c}_l)) \Phi(\mathbf{x}_i)}{\sum_{i=1}^n p(\Phi(\mathbf{x}_i), \Phi(\mathbf{c}_l))}; \\ l &= 1, 2, \dots, k \end{aligned} \quad (5)$$

并且

$$K_{\text{cov}}(\mathbf{x}_i, \hat{\mathbf{c}}_l) = \frac{\sum_{q=1}^n p(q, l) K_{\text{cov}}(\mathbf{x}_q, \mathbf{x}_i)}{\sum_{q=1}^n p(q, l)} \quad (6)$$

$$K_{\text{cov}}(\hat{\mathbf{c}}_l, \hat{\mathbf{c}}_l) = \frac{\sum_{q=1}^n \sum_{r=1}^n p(q, l) p(r, l) K_{\text{cov}}(\mathbf{x}_q, \mathbf{x}_r)}{\left(\sum_{q=1}^n p(q, l) \right)^2} \quad (7)$$

具体算法描述如下：

(1) 对数据集 $\mathbf{X}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ 进行初始化，确定聚类类数 k 和迭代停止条件 $\epsilon \in (0, 1)$ ，凝聚参数 $\tau > 0$ ；本文数据集为 216 棵客户树，根据港口实际情况，将客户分为四类， τ 分别取 0.10、0.25、0.50、0.80 观察实验结果。

(2) 选择式 (2) 定义的卷积核核函数 $K_{\text{cov}}(\mathbf{x}_i, \mathbf{c}_l)$ ，作为数据集中客户树的相似性度量。

(3) 任选初始类中心 $\mathbf{c}_l (l = 1, 2, \dots, k)$ ；为对比实验，本文选取多组聚类中心。

(4) 按式 (4) 计算每个样本在特征空间中对每个类中心的核函数 $p(i, l)$ 。

(5) 由式 (6) 和 (7) 计算新的核函数值 $K(\mathbf{x}_i, \hat{\mathbf{c}}_l)$ 和 $K(\hat{\mathbf{c}}_l, \hat{\mathbf{c}}_l)$ ，并按式 (4) 更新核函数为 $\hat{p}(i, l)$ ， $i = 1, 2, \dots, 216$ ； $l = 1, 2, 3, 4$ 。同样计算 K_{cov} 时会遇到计算两个属性向量 $\mathbf{x}, \mathbf{y}$ 之间距离 $\|\mathbf{x} - \mathbf{y}\|^2$ 的问题，依式 (3) 计算， γ 取 $0.5 \sim 0.7^{[18]}$ ，本文取 0.6。

(6) 若 $\max_{i,l} |p(i, l) - \hat{p}(i, l)| < \epsilon$ 或迭代直到 $p(i, l)$ 中出现不确定值 $\text{NaN}^{[17]}$ ，则算法停止；否则置 $\tau = \tau/2$ ，更新 $p(i, l) = \hat{p}(i, l)$ ，然后转到

(5)。
(7) 聚类结果由 $\hat{p}(i, l)$ 的值确定。

4 港口客户聚类结果分析

依照上面表述的算法，对港口客户数据进行聚类分析，将 216 个客户分为四类。选取不同的凝聚参数和不同的初始聚类中心，实验结果如表 2 所示。

从表 2 可看出，增大凝聚参数，迭代次数会稍有增加，聚类结果基本不会发生变化。改变初始聚类中心，聚类结果中大部分客户会保持在原来的聚类中心，小部分发生变化。

表 2 基于卷积核的核 k -凝聚聚类算法实验结果
Tab.2 Results of convolution kernels k -aggregate clustering algorithm

τ	初始聚类中心				迭代次数	聚类结果中每类客户数			
0.10	(26	78	125	190)	5	(44	29	33	110)
0.25	(26	78	125	190)	7	(44	29	35	108)
0.50	(26	78	125	190)	8	(44	29	37	106)
0.80	(26	78	125	190)	8	(44	29	37	106)
0.10	(18	62	120	180)	5	(31	29	13	143)
0.25	(18	62	120	180)	8	(31	29	13	143)
0.50	(18	62	120	180)	8	(31	29	13	143)

为评价分类效果，本文将分类后每个客户所有交易的数值属性值与其相应的权值相乘、加和，再除以交易次数，求出平均值 M ，即客户分类信息加权值，从中得出对一个客户的普遍一般的描述。选取分类结果最均匀、稳定的第 3 组数据，将其绘制成图，具体如图 3 和 4 所示。

图 3 表示客户第一、二、三类的分类情况，横轴表示客户数 N ，纵轴为 M 值。从图中可以看到三类客户的分类情况，个别客户在两类之间产生交叉，其余基本被清晰地分离。由于四类客户 M 值差异较大，无法绘制在同一幅图中，将第一类客户与第四类客户绘制在图 4 中。由图 4 可清晰看到类一和类四的差别。综合图 3 和 4 可见，基于卷积核的核 k -凝聚聚类算法能较好地将港口客户分为四类。

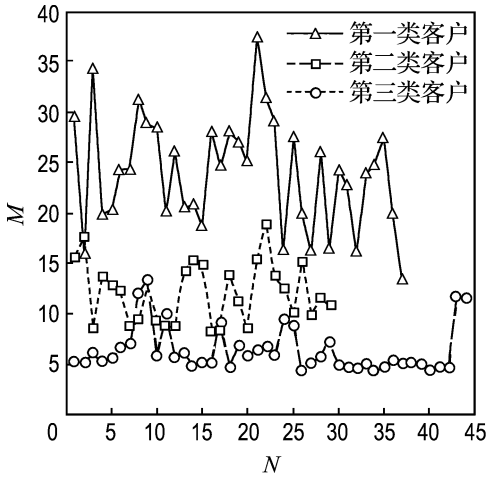


图 3 第一、二、三类客户的分类图

Fig. 3 The classification chart of the first, second and third class customers

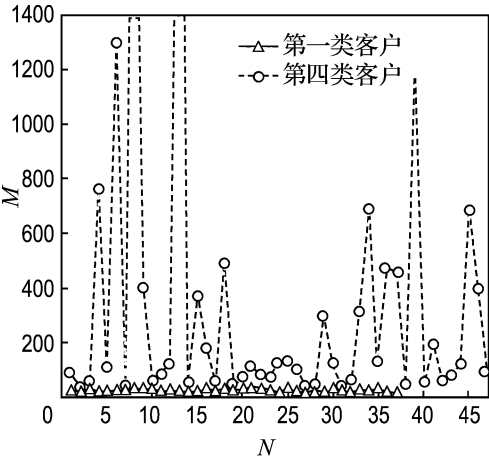


图 4 第一、四类客户的分类图

Fig. 4 The classification chart of the first and fourth class customers

依照文献[17]中给出的 k -prototypes 算法,对初始的 9 568 条客户记录进行聚类.对数据只进行数据完整性的预处理,而不重新组织,因为使用该算法,不能处理树形结构的数据,只能依照传统的一条条记录进行聚类.实验中,第一次取第 1 280、3 250、6 000、8 500 条记录作为聚类中心,迭代到第 7 次时,新的聚类中心数值属性变为无穷大,而无法继续迭代.更换聚类中心,取第 650、3 200、7 600、9 400 条记录为初始聚类中心,算法迭代到第 4 次,新的聚类中心数值属性变为无穷大,而无法继续迭代.查看此时客户的分类情况,

同一个客户的不同交易记录被分在不同的类中,意味着同一客户被分入不同的类,显然是不合理的.由此表明, k -prototypes 算法不适合用于港口客户数据分类.

5 结 语

本文基于卷积核的核 k -凝聚聚类算法,在营口港的客户数据分析中取得了较好的应用效果,为港口客户细分提供了一条可靠途径,同时将卷积核用于客户细分中,拓广了卷积核的应用领域.将该方法推广到所有港口企业,得出港口客户细分的一般方法,以及推广到其他行业,还需要进一步的实验和研究,也是下一步要做的工作.

参考文献:

[1] 付昌辉. 国内集装箱港口的现状与发展趋势[J]. 港口科技动态, 2005(6):3-5

[2] 彭志宁. 港口企业实施 CRM 初探[J]. 中国港口, 2006 (8):44-45

[3] SMITH W R. Product differentiation and market segmentation as alternative marketing strategies [J]. **Journal of Marketing**, 1956, **21**(1):3-8

[4] BOONE D S, ROEHM M. Retail segmentation using artificial neural networks [J]. **International Journal of Research in Marketing**, 2002, **19**(3):287-301

[5] TSAI C Y, CHIU C C. A purchase-based market segmentation methodology [J]. **Expert System with Applications**, 2004, **27**:265-276

[6] 张国政. 客户关系管理中基于数据挖掘的客户细分研究[J]. 商业研究, 2006(13):153-155

[7] 周 颖,吕 巍,井 森. 基于数据挖掘技术的移动通信行业客户细分[J]. 上海交通大学学报, 2007, **41**(7):1142-1145

[8] 许昌加,高 阳. 数据挖掘在电信客户细分中的应用研究[J]. 成组技术与生产现代化, 2004, **21**(1):43-46

[9] 陈治平,胡宇舟,顾学道. 聚类算法在电信客户细分中的应用研究[J]. 计算机应用, 2007, **27**(10):2566-2569

- [10] WONG K W. Data mining using fuzzy theory for customer relationship management [C] // **4th Western Australian Workshop on Information Systems Research (WAWISR 2001)**. Perth:The University of Western Australia, 2001
- [11] 吴春旭,吴 铺,蒋 宁. 一种基于动态模糊聚类算法的客户细分方法[J]. 计算机系统应用, 2008, **17**(2):21-24
- [12] 叶 强,卢 涛,闫相斌,等. 客户关系管理中的动态客户细分方法研究[J]. 管理科学学报, 2006(2): 44-52
- [13] 阎长顺,李一军. 基于云模型的动态客户细分分类模型研究[J]. 哈尔滨工业大学学报, 2007, **39**(2): 299-302
- [14] 王 宇,李晓利. 核 k -凝聚聚类算法[J]. 大连理工大学学报, 2007, **47**(5):763-766
(WANG Yu, LI Xiao-li. Kernel k -aggregate clustering algorithm [J]. **Journal of Dalian University of Technology**, 2007, **47**(5):763-766)
- [15] WANG Yu, GUO Qiang, LI Xiao-li. A kernel aggregate clustering approach for mixed data set and its application in customer segmentation [C] // **Proceedings of 2006 International Conference on Management Science & Engineering**. Harbin:Harbin Institute of Technology Press, 2006:121-124
- [16] HAUSSLER D. Convolution kernels on discrete structures [R/OL]. UC Santa Cruz Technical Report UCS-CRL-99-10, 1999. [2008-03-11]. <http://citeseerx.ist.psu.edu/haussler99convolution.html>
- [17] 邓乃扬,田英杰. 数据挖掘中的新方法——支持向量机 [M]. 北京:科学出版社, 2004
- [18] HUANG Zhe-xue. Clustering large data sets with mixed numeric and categorical value [C] // **Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining**. Singapore: World Scientific Publishing Company, 1997:21-34

Method of port customer segmentation

based on convolution kernels

WANG Yu* , XU Yan

(School of Management, Dalian University of Technology, Dalian 116024, China)

Abstract: According to the characteristics of port customer data, a data preprocessing was done applying the method of information gain, and 216 customer trees were got by representing it into tree structure. A convolution tree kernel by measuring the similarity between customer trees was defined by introducing the concept of convolution kernels. Kernel k -aggregate clustering algorithm presented before was extended to customer trees based on convolution kernels and the clustering analysis of port customer data was realized by using Matlab tools. The results of analyses show that application of convolution kernels in port customer segmentation gets good effect.

Key words: convolution kernels; kernel k -aggregate clustering algorithm; port customer segmentation