

# 改进的 K-Means 聚类算法在保险客户信用分析中的算法实现

宋加升, 陈 琰

(哈尔滨理工大学 经济管理学院, 黑龙江 哈尔滨 150040)

**摘 要:** 针对保险业对客户信息的分析中缺乏考虑客户信用分析的问题, 根据聚类分析算法理论和保险公司客户数据库特点, 进一步对 K-means 聚类算法在大样本环境下初始聚类中心的选取提出有效改进, 同时选取一家财产保险公司的客户信用数据, 来探讨聚类算法在保险客户信用分析中的应用.

**关键词:** 聚类分析; K-means 聚类算法; 保险客户

**中图分类号:** TP181 **文献标识码:** A **文章编号:** 1007-2683(2009)01-0116-04

## Algorithm Realization of Improved K-Means Clustering Algorithm in Credit Analysis of Policyholders

SONG Jia-sheng, CHEN Yan

(School of Economy Management, Harbin University of Science and Technology, Harbin 150040, China)

**Abstract:** In order to solve the problem that the analysis of policyholders' information lack of considering to analyse policyholders' credit, this paper gives a kind of K-means clustering algorithm that further improved initial center of cluster on environment condition for large sample in terms of cluster analysis theory and trait of policyholders' database, investigates that the algorithm excavates the practice application of policyholders' credit analysis through clients' datum of a property insurance company.

**Key words:** cluster analysis; K-means clustering algorithm; policyholder

## 1 引 言

随着信息技术的飞速发展, 围绕着人类各种日常活动和经济活动的数据, 以多种形式被储存在数据库中. 尤其对企业而言, 日益扩张的历史数据是研究商业模式和客户行为模式的丰富资源. 如何对大量的数据进行准确分析, 从而找出对企业决策有帮助的信息并使之更好地发展, 被越来越多的企业所关注, 数据挖掘技术为其实现提供了可能性.

保险业的经营活 动是以特定的风险为前提, 以集合多数单位和个人为条件, 以 大数法则为基础, 以经济补偿和给付为基本功能的<sup>[1]</sup>. 可以看出, 保险产品的实现依赖于保险人和投保人基于对未来风险的预期所制定的契约, 是建立在双方信用基础之上的. 但随着我国保险业的不断发展壮大, 由于信息不对称、道德风险和 心理风险以及逆选择的存在<sup>[2]</sup>, 保险客户信用风险控制已为保险公司与保险监管部门所重视. 如何对客户信用进行分类, 进而分析其价值是客户信用风险控制所要解决的首要问题. 聚类

作为数据挖掘的一种方法被广泛应用于客户信息的挖掘.通过聚类算法,可以识别密集和稀疏的区域,从而发现全局的分布模式以及数据属性间的关系<sup>[3]</sup>.

本文在分析 Kmeans算法的实现过程及其优缺点的基础上,针对提高大型数据库聚类分析的准确性及高维数据聚类的有效性,对 Kmeans算法进行了改进,同时选取了一家财产保险公司的个人客户信用信息,探讨了聚类分析在客户信用分析中的应用.

## 2 聚类分析算法

聚类分析源于多个研究领域,包括数据挖掘、统计学、生物学,以及机器学习.数据挖掘中的聚类 (clustering)是根据数据对象的属性值计算相异度并将其分组成为多个类或簇 (cluster),同一个簇中的对象相似度较高,不同簇间的对象差别较大,它是一种无监督的学习过程.聚类与分类不同.在分类中数据对象有哪些类是预先知道的,而聚类所要研究的是如何在一定约束下对目标数据库中的未知类进行划分,不依赖预先定义的类和带类标号的训练实例.聚类算法包括如下几类:划分方法 (partitioning method)、层次方法 (hierarchical method)、基于密度的方法 (density-based method)、基于网格的方法 (grid-based method),及基于模型的方法 (model-based method)<sup>[4]</sup>.

划分聚类方法主要有 Kmeans算法、Kmedoids算法和基于大型数据库的 CLARANS算法.主要思想是:给定一个包含  $n$  个元组或对象的数据库,构建  $K$  个分组,每个组代表一个簇,且  $K \leq n$ .这些分组满足如下两个条件:① 每个组至少包含一个对象;② 每个对象必须属于且只属于一个组 (这个要求在某些聚类算法中可以放宽).具体实现时是给定划分数目  $K$  之后,创建一个初始划分,然后利用迭代重定位技术不断优化分组. Kmeans算法中每个簇用该簇对象的平均值来表示, Kmedoids算法中每个簇用接近聚类中心的一个对象来表示.

## 3 基本思想

通常基于划分的 Kmeans算法要求在给定的包含  $n$  个数据对象的数据库中,由使用者制定所要生成的簇数目  $K$  ( $K < n$ ),也就是分类数目,然后采用

一定的相似度函数对目标数据进行划分.具体实现过程如下<sup>[4]</sup>:

输入:包含  $n$  个数据对象的数据库,簇的个数为  $K$

输出:满足方差最小标准的  $K$  个聚类.

1) 在目标数据库中任意选择  $K$  个对象作为初始的各簇中心;

2) 重复以下步骤,直到各簇不再发生变化;

3) 计算余下对象与簇中心的距离,并将其赋给最近的簇;

4) 重新计算每个簇中对象的平均值.

Kmeans聚类算法是一种不依赖于顺序的,基于划分的算法,其时间复杂度与数据集的大小呈线性关系,缺点是必须指定  $K$  值,且对初始中心点的选择比较敏感,如果初始值选择不当,将会收敛成为一个局部最小的准则函数<sup>[5]</sup>.并且极端的情况是在大型数据集的聚类中,从海量数据中随机产生的大样本也不能涵盖全部类型的样本数据.这些情况的出现会影响聚类结果的准确性.为了提高对样本集和初始聚类中心的选取对聚类结果造成不同准确率,文 [6] 提出了对初始聚类中心选取的改进,并使用 UC 数据库对此进行验证.对样本集  $U$  中的  $n$  个数据对象,计算其两两之间的相异度,找出相异度最小的两个点,形成样本集  $v_1$ ,并把  $v_1$  从样本集  $U$  中除去;再计算  $v_1$  中的每一个对象与  $U$  中的每一对象的相异度,并找出在  $U$  中与  $v_1$  最接近的点,将其并入  $v_1$ ,同时从  $U$  中除去,直到达到一定阈值 (设立参数  $\alpha = n/k$ ).重复以上过程,得到集合  $v_2$  直到取得  $K$  个小样本集,再对各小样本集进行平均,形成  $K$  个初始聚类中心,之后对原样本集  $U$  按照 Kmeans算法的 2)~4) 步骤进行聚类<sup>[6]</sup>.

但是保险客户的信用信息数据库可能储存了一个公司数十年的客户基础资料和相关业务的历史数据,对这样的巨型数据库,只生成一个样本集,很可能不能反映所有客户类型的信息.在数据仓库中选取样本时,为了提高稳定性,需要进行多样本选取.在整个数据仓库中随机产生  $P$  个不同的样本集,这里  $P$  的大小可根据具体分析要求制定.对每个样本集按照上述步骤进行 Kmeans聚类,产生  $P \times K$  个样本集内符合方差最小标准的簇.将各样本集合并为大样本,计算各样本集间的簇间距离,合并距离小于  $\tau$  ( $\tau$  的制定根据实验情况进行调整) 的簇.合并簇的个数应小于或等于  $P$ ,若合并个数大于  $P$ ,则说明  $\tau$  参数设置过大,聚类过于密集,适当加以调整.

然后再以合并簇中各子簇的平均值作为该合并簇的平均值.最后对各簇的平均值进行调整.

对初始聚类中心选取进行改进,可减小避免由于对初始聚类中心的选择随机性带来的聚类结果准确率的不稳定性,使初始聚类中心与实际聚类中心的分布更加吻合,得到好的聚类效果.并且小样本中的聚类结果准确性的提升对大样本的分析也更为有利.多样本选取,同时对子簇进行合并类似于层次聚类的凝聚方法.这种针对大型数据库的处理使最终聚类结果与实际分布更为接近,提高了运算效率,适用于非球状聚类的发现.

## 4 算法实现及实验结果分析

### 4.1 算法描述

实际数据库中的数据类型复杂,一个数据对象往往同时包含多个不同类型的变量.在进行计算之前需要对数据进行处理.假设数据集包含  $M$  个不同类型变量,数据矩阵为

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1M} \\ x_{21} & x_{22} & \cdots & x_{2M} \\ \cdots & \cdots & \ddots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nM} \end{bmatrix}$$

为了简化计算过程,将不同类型的变量转换到共同的值域空间  $[0, 0, 1, 0]$  上,定义对象  $i$  和  $j$  之间的相异度<sup>[4]</sup>为

$$d(i, j) = \frac{\sum_{f=1}^M \delta_{ij}^{(f)} \cdot d_{ij}^{(f)}}{\sum_{f=1}^M \delta_{ij}^{(f)}}$$

其中:  $f$  代表变量;  $x_{if}$  或  $x_{jf}$  代表对象  $i$  或对象  $j$  的变量  $f$  的度量值.

1) 当  $f$  为二元变量或标称变量时,如果  $x_{if} = x_{jf}$ ,  $d_{ij}^{(f)} = 0$ , 否则  $d_{ij}^{(f)} = 1$ . 若  $x_{if}$  或  $x_{jf}$  缺失,或者  $x_{if} = x_{jf} = 0$ , 且均为不对称二元变量,指示项  $\delta_{ij}^{(f)} = 0$ , 否则  $\delta_{ij}^{(f)} = 1$ .

2) 当  $f$  为序数型变量时,假设变量  $f$  有  $N_f$  个状态,对应于序列  $\{1, \cdots, N_f\}$ ,  $r_{ij}$  为与之相对应  $x_{if}$  的秩,且  $r_{ij} \in \{1, \cdots, N_f\}$ , 此时可用权重  $Z_{if}$  代替  $x_{if}$

$$Z_{if} = \frac{r_{if} - 1}{N_f - 1}$$

3) 当  $f$  为区间标度变量时,将  $f$  的度量值进行标准化,计算平均绝对偏差  $S_f$  为

$$S_f = \frac{1}{n} \sum_{i=1}^n |x_{if} - m_f|$$

其中,  $m_f$  为  $f$  度量值的平均值,即

$$m_f = \frac{1}{n} \sum_{i=1}^n x_{if}$$

则标准化度量值  $Z_{if}$  为

$$Z_{if} = \frac{x_{if} - m_f}{S_f}$$

计算相异度时:

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$$

这里  $h$  遍取变量  $f$  的所有非空缺对象.

该文中涉及三种距离:点与点之间的距离;点与簇之间的距离;簇与簇之间的距离.

1) 点与点之间的距离采用最常用的欧几里德距离,即

$$d(i, j) = \sqrt{\sum_{k=1}^M (x_{ik} - x_{jk})^2}$$

2) 点与簇之间的距离定义为

$$d(i, c) = \min(d(i, j), j \in c)$$

3) 簇与簇之间的距离定义为两个簇的平均值,有

$$d_{\text{mean}}(c_i, c_j) = |m_i - m_j|$$

假设在目标数据库中随机选取  $P$  个不重复的小样本集  $u_1, u_2, \cdots, u_p$ , 每个小样本集  $u_i$  ( $1 \leq i \leq P$ ) 包含  $n$  个对象,输出类数目为  $K$ ,  $c_m$  表示小样本的簇,且  $i = j = r = 1$ .

算法描述如下:

1) 对  $u_i$  进行聚类:计算  $u_i$  中任意两个对象  $x, y$  间的相异度  $d(x, y)$ , 当  $d(x, y) = \min d(x, y)$  时,由  $x, y$  组成  $v_j$  ( $1 \leq j \leq K$ ), 且  $u_i = u_i - v_j$ .

2) 计算  $u_i$  中点与  $v_j$  的距离,将距离最近的点加入  $v_j$ , 并从  $u_i$  中删除此点,重复此步直到  $v_j$  中点数大于等于  $(n/k, 0 < k < 1)$ .

3) 若  $j < K$ , 则返回 1) 步,生成  $v_{j+1}$ , 直到  $j = K$ .

4) 分别计算  $v_1, v_2, \cdots, v_K$  的平均值,形成  $K$  个初始聚类中心,并使用  $K$ -means 算法 2) ~ 4) 步骤对  $u_i$  完成聚类.

5) 计算大样本  $U: \{u_1, u_2, \cdots, u_p\}$  中任意两个簇之间的距离  $d(c_{m1}, c_{m2})$ , 当  $d(c_1, c_2) = \min d(c_{m1}, c_{m2}) < \tau$  ( $0 < \tau < 1$ ) 时,合并  $c_1, c_2$  形成合并簇  $C_r$  ( $1 \leq r \leq K$ ), 更新  $C_r$  的平均值.

6) 计算余下各簇与  $C_r$  的距离,将距离最近的簇与  $C_r$  合并,同时更新  $C_r$  的平均值,重复此步,直到

被合并的簇数大于等于  $P$ ,此时将  $C_r$  从大样本  $U$  中删除.

7) 若  $r < K$ , 返回 4) 步, 重新生成  $C_{r+1}$ , 直到  $r = k$

至此聚类过程结束.

4.2 实验结果

该文将经过进一步改进的 Kmeans聚类算法应用于保险个人客户的信用信息划分. 在保险专业人士的帮助下,在某财产保险公司的客户信息数据库中抽取了一些个人客户的属性和业务指标来描述个人客户信用,如年龄、性别、学历、婚姻情况、工作情况、续保率、赔付率、保费金额等. 实验中从保险客户信息数据库截取了近两年的客户数据作为目标数据库,随机选择 5 个包含 400 条客户信息的小样本集来构成大样本集. 数据对象包含了多种类型变量,聚类之前需要对其进行处理,以年龄属性为例:分为  $[20, 25]$ 、 $(25, 30]$ ... $(55, 60]$  等 8 个区间,以  $\{1, 2 \cdots 8\}$  为对应的状态值,计算相应的权重  $z_{ii}$  作为变量值. 设置输出类别数为 4. 由于篇幅限制,下表为部分经处理的数据.

表 1 处理后的保险客户信息数据

| 年龄    | 性别 | 婚否 | 工作情况  | 学历    | 续保率   | 赔付率   | 保费金额  |
|-------|----|----|-------|-------|-------|-------|-------|
| 0.286 | 0  | 1  | 0.500 | 0.667 | 0.800 | 0.144 | 0.768 |
| 0.143 | 0  | 0  | 0     | 1     | 0.833 | 0.091 | 0.814 |
| 0.143 | 1  | 1  | 0.500 | 0.667 | 1     | 0.015 | 0.297 |
| 0.429 | 1  | 1  | 1     | 0.333 | 1     | 0     | 0.175 |
| 0.143 | 0  | 0  | 0     | 0.667 | 0.667 | 0.412 | 0.205 |
| 0.714 | 0  | 1  | 1     | 0.333 | 0.714 | 0     | 0.713 |

实验采用了原 Kmeans聚类算法和经改进的 Kmeans算法,从处理时效来看,改进的算法提高了对大样本集的运算效率;从聚类效果来看,参照专业知识,改进的算法选取的初始聚类中心更具有代表

性,更符合实际分布,对输出的 4 类客户信息聚类结果的类别间距离进行方差分析,分析结果表明,类别间差异的概率值均小于 0.001,聚类效果好. 经多次对样本数据的聚类,改进算法的稳定性为 0.795,高于原算法.

5 结 语

为了降低 Kmeans聚类算法的对初始聚类中心的敏感性,提高算法在大数据库中的伸缩性,该文提出了一种在大型数据库中解决高维多类型数据聚类的改进型 Kmeans算法,并结合保险客户数据进行了实验,证明了该算法具有更高的时效性和稳定性. 今后可进一步对该算法进行测试,扩展其应用领域,并加强对降低大型数据库中噪声数据和孤立点对聚类结果敏感性的研究,或者将该算法与其他数据挖掘技术结合,进一步对客户的信用级别进行分析,如模糊推理理论等.

参 考 文 献:

[1] 王绪瑾. 保险学 [M]. 北京: 经济管理出版社, 2001: 178 - 179.  
[2] 赵晓菊,柳永明. 金融机构信用管理 [M]. 北京:中国方正出版社, 2004: 12 - 14.  
[3] PENG J M, XIA Y. A New Theoretical Frame Work for Kmeans Clustering[C] //To Appear in Foundation and Recent Advances in Data Mining, Eds Chu and Lin, Springer Verlag, 2005.  
[4] HAN Jiawei, MICHEL NE Kamber 数据挖掘概念与技术 [M]. 北京:机械工业出版社, 2006: 231 - 233.  
[5] 张光建,黄贤英. 基于最小聚类单元的聚类算法研究及其在 CRM 中的应用 [J]. 计算机科学, 2006, 33: 188 - 189.  
[6] 袁 方,孟增辉,于 戈. 对 KMeans聚类算法的改进 [J]. 计算机工程及应用, 2004, 36: 177 - 178.

(编辑:付长缨)