

doi: 10.3969/j.issn.1007-7162.2014.03.001

可调多趟聚类挖掘在电信数据分析中的应用

滕少华, 吴昊, 李日贵, 张巍, 刘冬宁, 梁路

(广东工业大学 计算机学院, 广东 广州 510006)

摘要: 电信业务每天都产生大量数据, 如何从这些数据中提取有用的信息是当今数据挖掘的难题之一. 针对实际应用中存在聚类簇数难以确定、单趟聚类算法有时不能收敛到用户指定的簇数等问题, 提出了可调多趟聚类挖掘方法. 第 1 趟通过引入一个较大的 K 值, 采用 K-means 聚类算法, 获得 K 个簇, 为第 2 趟聚类的簇数及簇中心初始值选择提供参考. 经电信现网业务数据实验, 本文的方法既改善了原聚类方法的局部收敛性, 又能较好地适应用户的不同数据分析需求, 该方法可用于不确定簇数的大数据分析中.

关键词: 电信数据; 多趟聚类; K-means 聚类; 客户细分

中图分类号: TP311.13

文献标志码: A

文章编号: 1007-7162(2014)03-0001-07

The Application of the Adjustable Multi-times Clustering Algorithm in Telecom Data

Teng Shao-hua, Wu Hao, Li Ri-gui, Zhang Wei, Liu Dong-ning, Liang Lu

(School of Computers, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Huge amounts of telecom data are generated every day, so how to extract useful information from the data is one of the data mining problems. Because different applications need different clusters, sometimes a single K-means cluster algorithm cannot generate user-specified K clusters. An adjustable multi-times clustering mining method is proposed. A big value K was used in the K-means clustering algorithm for the first time, and K clusters were obtained. They were used to select the number of the clusters and the initial centers of the clusters for the second time. The experimental results show that our method is effective, and it can be applied to mining different amounts of clusters and big data analysis.

Key words: telecom data; multi-times clustering; K-means clustering; customer subdivision

随着大数据时代到来, 3G 网络建设和商用不断深入, 电信业正从服务经济向体验经济转型, 移动网络运营商之间的竞争也已从市场竞争转向服务竞争. 保障网络通信质量、提升网络服务质量是运营商掌握竞争先机所必须考虑的问题.

Reicheld 的调查报告显示, 每增加 5% 的客户保持率, 行业就会增加 25% ~ 95% 的平均利润率^[1]. 因此, 如何为客户群体提供优质服务, 保持客户, 避免客户流失, 是各大企业的核心工作任务. 为了准确地了解客户、分析客户、服务客户, 企业首先要正确

地划分客户群体. 通过客户细分, 运营商能更好地识别客户群体, 对不同的客户, 采取不一样的客户保持策略, 达到最优化配置与利用资源, 最大化获取利润的目的.

1 相关工作

张巍等^[2]从电信业务出发, 按用户行为、用户类型、用户投诉等多个角度细分用户, 探讨了移动网络语音业务用户体验的建模问题, 给出了建模方案. 徐猛^[3]探讨了面向用户体验的多 QoS 约束多流程调

收稿日期: 2014-07-10

基金项目: 国家自然科学基金资助项目(61104156, 61370229); 国家科技支撑计划项目(2013BAH72B01); 教育部重点实验室基金资助项目(110411); 广东省自然科学基金资助项目(10451009001004804, 9151009001000007); 广东省科技计划项目(2012B091000173); 广东省教育厅项目(粤教高函(2013)113号); 广州市科技计划项目(2012J5100054, 2013J4500028)

作者简介: 滕少华(1962-), 男, 教授, 博士, CCF 会员(E200006870S), 主要研究方向为协同计算、数据挖掘、网络安全和大数据.

度算法,目的是满足尽可能多的用户 QoS 需求,但未涉及基于用户体验的协同建模问题.周天圆^[8]将决策树、模糊神经网络等算法用来分析电信业务的用户体验,研究结果用现网数据进行了验证,但未对具体建模过程进行描述.郑茜茜^[9]将客户数据根据某个给定的阈值划分群体,阈值直接决定了划分效果的好坏.通常阈值的确定需要经过专家仔细研究和分析.但是,在不同地区和不同人文地理环境下,客户群体的划分往往不同,传统的通过某种阈值的统计分析法很难做到准确地细分客户群体.

在运用数据挖掘方法细分客户群体中,常用的是基于划分的 K-means 聚类^[10]. K-means 聚类简单、易行,时间复杂度是 $O(tKmn)$ ^[11],其中 t 为迭代次数, K 为类别数, m 为样本维数, n 为样本数, K-means 聚类的时间复杂度和样本数成线性关系,因此可用于大数据的客户细分.王扶东等^[12]从客户生命周期价值方面考虑,提出了基于加权的 K-means 客户细分方法.但是, K-means 聚类算法对初始聚类中心的选取依赖性极大,较差的初始聚类中心常导致算法陷入局部最优解^[13].

国内外许多研究集中于降低初始聚类簇中心的选择对算法结果的影响,杜巍等^[14]研究初始聚类中心的选取方法,提出了一个改进的 K-means 聚类方法,并应用于客户细分,获得了一些有益的结果. Md. SohrabMahmud 等^[15]基于平均权重,提出了一个初始化聚类中心的方法.王令剑等^[16]将聚类方法和时间序列分析结合,并用于入侵检测中,通过实验验证了该方法的有效性.曹永春^[17]将改进的人工蜂群算法和 K-means 迭代相结合,降低了初始聚类中心的依赖性和陷入局部最优解的可能性.胡伟^[18]提出了一个层次聚类的思路,通过从上往下地构造一棵 K 均值聚类树,最终在叶子节点达到聚类结果.高灵渲等^[19]在 K-means 算法的欧氏距离及中心点判定阶段改进,提升了 K-means 算法的效率.易倩等^[20]将马氏距离与 K-means 算法相结合,并在目标函数中增加变量权重因子和协方差矩阵调节因子,提高了聚类精度.杜红乐等^[21]针对数据之间距离不均匀的问题,研究了模糊 C 均值聚类问题及其应用.

上述研究在一定程度上对初始聚类中心的确定有所帮助,但尚未有效解决 K-means 算法依赖于初始聚类中心的问题.

除了初始聚类中心的影响外, K-means 算法对噪声点和离群点很敏感.廖松有等^[22]通过引入模糊熵约束,使用了模糊 C 均值聚类算法,在一定程度上能抑制噪声数据对聚类的影响.卢鹏丽等^[23]以密

度敏感距离作为相似性测度,结合近邻传播聚类算法和谱聚类算法,提出了一种密度敏感的层次化聚类算法.潘章明^[24]基于分治策略,提出了一种改进的共享近邻聚类算法.这些算法都是从密度的角度研究噪声点和离群点.

国内外学者针对 K-means 聚类方法及其应用,开展了一系列研究,取得了许多成果,但下列问题尚未得到有效解决.

(1) 初始聚类中心选取: 初始聚类中心选取不合适,会导致局部最优解;

(2) 聚类个数 K : 有时用户需要聚类成不同个数的簇. 同一个数据集,从不同的角度分析将需要不同的聚类个数,产生不同的聚类效果,比如评价可分为好、中、及格、差,也可分为很好、好、中、及格、差、极差等;

(3) 聚类效果: 由于聚类受初始簇中心选取和 K 值影响,聚类结果往往难以保证.

为此,本文提出可调多趟聚类方法,第 1 趟通过引入一个较大的 K 值,采用 K-means 聚类算法,获得 K 个簇,为第 2 趟聚类的簇数及簇中心初始值选择提供参考.由此,用户可以根据实际情况,选择合适的簇数及初始簇中心,通过第 2 趟聚类获取较好的挖掘结果.

2 K-means 聚类

K-means 聚类算法是 1967 年由 MacQueen 首次提出的一种经典划分聚类方法. K-means 算法以 K 为参数,把 n 个数据对象的集合分为 K 个簇,使得簇内对象的相似度高,簇间对象的相似度低.簇的相似度是关于簇中对象的均值度量,如质心、重心、距离等.本文采用欧氏距离计算.

K-means 算法流程如下: 首先,随机地选择 K 个对象,每个对象代表一个簇的初始簇中心.对剩余的每个对象,根据其与其各个簇中心的距离,将它指派到最相似的簇.然后计算每个簇的新均值.这个过程不断重复,直到准则函数收敛.

定义 1^[25] 平方误差准则函数. 设 x 是多维空间中的点,表示给定对象; m_i 是簇 C_i 的簇中心, $d(\cdot)$ 表示两个对象之间的距离(通常采用欧氏距离),平方误差和 SSE 定义为:

$$SSE = \sum_{i=1}^k \sum_{x \in C_i} d(m_i, x)^2. \quad (1)$$

基于距离的 K-means 聚类算法存在以下不足: 算法需要预先给定簇个数 K ; 算法对初始簇中心值选取依赖性极大,常陷入局部最优; 在数据量小的情

况下,算法对离群点和噪声很敏感;算法不能用于发现非凸形状的簇;算法只能用于数值数据。

K-means 聚类算法需要预先给定簇个数,这一特点限制了对基于客户细分的用户业务分析。有时,用户希望从多个角度对数据进行分析,希望将客户细分为不同的级别,以便进行不同级别上的决策。例如某企业决策层希望把客户细分为潜在流失客户、一般客户和 VIP 客户 3 个类别;而具体业务人员希望把客户细分为潜在流失客户、可挽留客户、一般客户、可升级客户和 VIP 客户 5 个类别。这时,K-means 聚类方法需要开展两次完全无关的挖掘,并且在大数据下,难以保证挖掘的效果。针对 K-means 存在的不足,并兼顾不同用户的需求,本文在进行大量数据挖掘实践的基础上,提出了一个可调的多趟 K-means 聚类方法。

3 可调的多趟 K-means 聚类

本文的多趟 K-means 聚类分为两趟,第 1 趟确定初始簇,为用户选择 K 值和初始簇中心提供依据与参考。第 2 趟为可定制的聚类。其体系结构如图 1 所示。

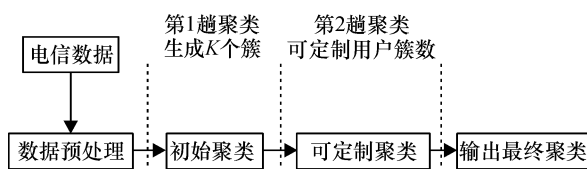


图1 可调多趟聚类过程

Fig. 1 The adjustable multi-times clustering process

3.1 初始聚类

层次聚类分为自顶向下和自底向上两种。本文提出的聚类算法采用自底向上的方式,初始聚类从用户期望分类的簇数出发。聚类过程如下:

- (1) 获取电信业务数据集 D ,对数据进行预处理;
- (2) 根据用户实际需要,获取聚类簇数 K_0 (K_0 为整数);
- (3) 依据专家知识、背景和用户需求将簇数 K_0 放大到聚类簇数 K_1 ($K_1 \gg K_0$, K_1 为大于 K_0 的整数);
- (4) 从 D 中任选 K_1 个数据对象作为初始聚类中心;
- (5) 对数据集 D 做 K_1 -means 聚类;
- (6) 生成大约 K_1 个聚类簇。

初始聚类时,由于选取了多个聚类中心,随机性强,覆盖面广,能够在一定程度上避免选取不合适的聚类中心,导致聚类效果进入局部最优。

初始聚类后,可以用误差平方和、抱团性和分离性等评价指标分析是否有离群簇。

3.2 可定制的 K-means 聚类

可定制的 K-means 聚类,可通过用户需求、反复实验及历史经验,用多个 K 值调整目标聚类个数。例如可设置聚类个数为 $K-2$ 、 $K-1$ 、 K 、 $K+1$ 、 $K+2$,然后比较哪组聚类的效果较优,根据实际需要作出合适的选择。

第 2 趟聚类在第 1 趟聚类基础上,选取 K 个值作为第 2 趟初始簇中心,它们是距离最远的 K 个第 1 趟的簇中心,以避免聚类中心在迭代过程中重合,导致局部最优解。第 2 趟聚类过程如下:

- (1) 获取第 1 趟聚类中心 $C_1, C_2, \dots, C_{K_1}$;
- (2) 计算任意两个簇中心 C_i 与 C_j 之间的距离 DIST_{ij} ;
- (3) 取 K 个不同的 C_i 值 ($i=1, 2, \dots, K$), 计算
$$\max \sum_{i=1}^K \sum_{j=1}^{K_1} [\text{DIST}(C_i, C_j)]$$
;
- (4) 将满足式 (2) 的 K 个 C_i 作为第 2 趟的初始聚类中心;
- (5) 对数据集 D 做 K-means 聚类;
- (6) 最后根据聚类结果合并簇,得到目标划分。

在聚类过程中,满足式 (2) 的 K 个不同的 C_i 值,确保了 K 个初始簇中心两两之间的距离较大,避免聚类中心在迭代过程中重合。

3.3 可调的多趟聚类效果

可调的多趟聚类效果如图 2 所示。第 1 趟产生 K_1 个簇,第 2 趟用户可定制 K 个簇。在图 2 中, S_i ($i=1, 2, \dots, n$) 为第 1 趟聚类的簇; C_i 与 T_i 分别为用户根据不同需求定制的簇。

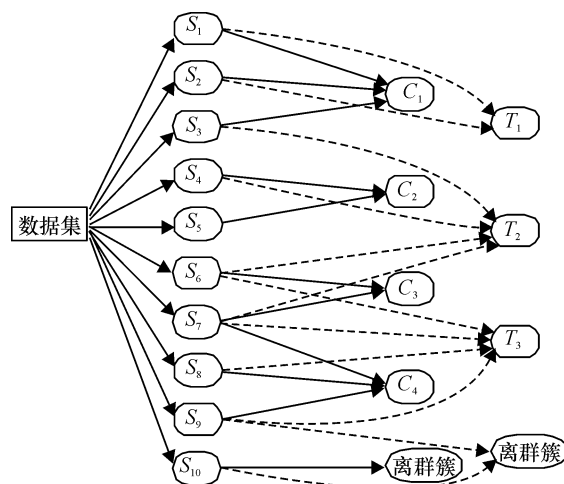


图2 可调多趟聚类结果示意图

Fig. 2 Diagram of the adjustable multi-times clustering results

3.4 可调多趟聚类算法

本文的可调多趟聚类算法如下. 设用户需要划分 K_0 个簇, 数据集为 $X = \{x_1, x_2, x_3, \dots, x_i, \dots, x_n\}$, 其中 x_i 是一条数据记录, 算法的步骤如下:

(1) 设定 K_1 个初始聚类中心, K_1 远大于 K_0 , 可以取 $K_1 = 5 \times K_0$.

(2) 根据 K_1 个聚类中心, 对 X 做 K_1 -means 聚类, 得到 K_1 个簇.

(3) 分析 K_1 个簇, 判断是否有离群簇. 如果有离群簇, 在 X 中分离对应的离群簇数据.

(4) 假设除离群簇外有 K_2 个簇中心 $C = \{c_1, c_2, c_3, \dots, c_{K_2}\}$. 计算两两簇中心之间的距离.

(5) 计算 $\max \sum_{i=1}^K \sum_{j=1}^{K_2} \text{DIST}(C_i, C_j)$.

(6) 选择前 K_0 个不同的 C_i , 满足式(2)的要求, 将 C_i 作为第 2 趟聚类的初始簇中心.

(7) 对数据集用 K_0 个 C_i 做 K_0 -means 聚类. 得到 K_0 个簇 $G = \{g_1, g_2, g_3, \dots, g_{K_0}\}$.

(8) 将步骤 7 的聚类结果输出.

(9) 根据需要修改 K_0 值, 即改变目标聚类个数, 重复步骤 6~8.

步骤 8 对得到的多组聚类结果进行对比和评估, 根据实际情况选择合适的聚类结果.

算法流程如图 3 所示.

使用可调的多趟 K -means 聚类能够排除离群簇, 减少了初始聚类中心对聚类结果的影响, 并能发现非凸形状的簇, 最后还能够根据实际需要调整聚类个数. 本文的方法适合于用户对聚类数及聚类结果都比较模糊, 可为用户提供多种聚类效果的选择.

3.5 算法评估

聚类算法能够使簇内具有较高的相似度, 而簇与簇之间的相似度较低. 为了分析聚类的效果, 需要从簇中对象的“相似性”和不同簇之间对象的“相异性”两方面考虑, 可按定义 1 的评价方法来衡量算法的有效性.

平方误差和, 用于评价簇的集中程度. 小的平方误差和说明簇中的对象集中度高. 平方误差和的计算参见定义 1.

分离性, 即任意两个簇中心的距离中的最小值^[13]. 分离性值越大, 簇与簇之间相差越大. 分离性的定义如下:

定义 2^[13] 分离性. 设 C_i 为簇 i ($i = 1, 2, \dots, k$) 的簇中心, 分离性定义为

$$F = \min_{\substack{i=1,2,\dots,k \\ j=1,2,\dots,k}} (c_i - c_j). \quad (3)$$

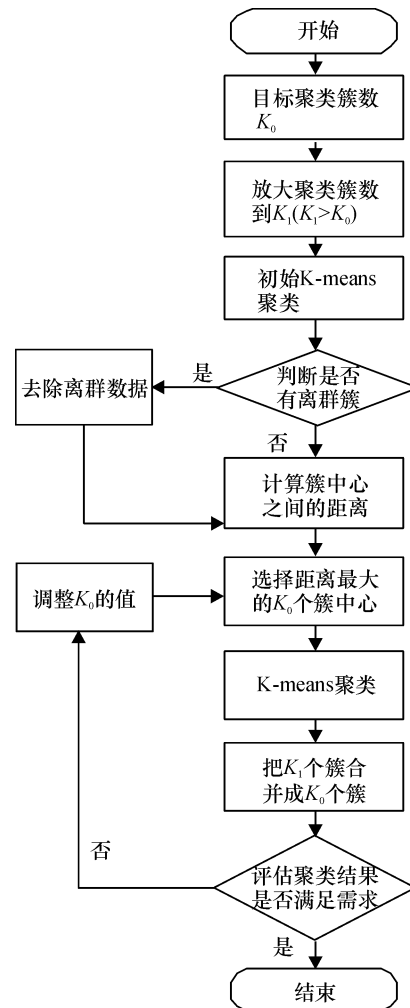


图 3 算法流程图

Fig. 3 The flowchart of the adjustable multi-times clustering algorithm

4 结果与讨论

4.1 实验数据

为了验证可调多趟 K -means 聚类方法在客户细分中的可行性和有效性, 本文用某电信企业的移动用户数据进行了实验, 并对实验结果做了分析和总结.

本实验数据采用电信客户的通话数据, 根据客户的通话质量情况细分客户. 实验中把客户分为 5 个等级, 通信质量好、通信质量较好、通信质量中、通信质量较差、通信质量差, 分别代表 5 个等级的服务情况. 对于通信质量较差与差的客户, 移动运营商应该着重改善该客户群体的服务质量, 防止客户流失; 对于通信质量好及较好的客户, 移动运营商可以有针对性地为他们提供额外的服务, 引导该客户群体提升消费水平.

客户的通话数据来源于测量报告, 如表 1 所示.

数据中包括客户标识、上下行接收电平、接收质量、时间提前量等反映网络质量水平的数据。在 GSM 网络^[20-21]中误码率(RxQual)和上、下行电平(Rxlev)是影响通话质量的重要因素。误码率决定了通话时的语音质量。误码率也称作质量,其值越小,语音质

量越好。上、下行电平反应手机的信号好坏。当上、下行电平太低时,可能会导致手机信号弱,无法进行电话的正常拨打和接听。当上、下行电平不平衡时,也可能会导致频繁的掉话。

表1 测量报告的主要数据及说明

Tab.1 Main data and instructions in measurement reports

属性名称(英)	属性名称(汉)	说明
TIMESTAMP	时间戳	采集该条测量报告数据的时间。
IMSI	国际移动客户识别码	区别移动客户的标志,储存在 SIM 卡中,可用于区别移动客户的有效信息。
TMSI	临时识别码	每次通话时生成,标识一次通话服务。
SEQUENCENUMBER	序号	本条 MR 的序号
LAC	位置区编码	LAC + CI 构成同一移动网络内的小区唯一编码。
CI	小区编码	
TEI (terminal endpoint identifier)	终端机末端点识别符	标识小区间传输的连接情况。一般设置为 0~63。
TA (timing advance)	时间提前量	信号在空间传输是有延迟的。为了避免时隙重叠,基站在下行信道上以 480ms 一次的频率向移动台发送指令,指示移动台提前发送的时间,这个时间就是 TA。其值为 0~63,单位为 bit,1bit 对应传输距离是 554m。
BSPOWER	基站功率	
RXLEVFULLUP	主服务小区上行电平(full)	指 TCH 和 SACCH TDMA 帧全集的 RXLEV 和 RXQUAL,单位为 db。
RXQUALFULLUP	主服务小区上行质量(full)	
RXLEVSUBUP	主服务小区上行电平(sub)	指在开通间歇发射条件下(DTX)下的 RXLEV 和 RX-QUAL 值,比 full 准确。单位为 db。
RXQUALSUBUP	主服务小区上行质量(sub)	
MSPOWER	移动终端功率	
RXLEVFULLSERVINGCELL	主服务小区下行电平(full)	指 TCH 和 SACCH TDMA 帧全集的 RXLEV 和 RXQUAL,单位为 db。
RXQUALFULLSERVINGCELL	主服务小区下行质量(full)	
RXLEVSUBSERVINGCELL	主服务小区下行电平(sub)	指在开通间歇发射条件下(DTX)下的 RXLEV 和 RX-QUAL 值,比 full 准确。单位为 db。
RXQUALSUBSERVINGCELL	主服务小区下行质量(sub)	
NEIGHBOUR1RXLEV	相邻基站 1 的场强	
NEIGHBOUR2RXLEV	相邻基站 2 的场强	
NEIGHBOUR3RXLEV	相邻基站 3 的场强	
NEIGHBOUR4RXLEV	相邻基站 4 的场强	
NEIGHBOUR5RXLEV	相邻基站 5 的场强	
NEIGHBOUR6RXLEV	相邻基站 6 的场强	

实验环境采用数学软件 Matlab,实验选取了测量报告中与网络质量相关的 11 条属性做分析,构成了 8 组数据集,数据集的数据量从 11322 条到 82665 条。

以最后一组数据为例,实验使用可调多趟 K-means 聚类方法,对 82665 条数据进行聚类。由于需要把客户分为 5 个等级,所以设定第 1 次聚类个数为 20 倍,即 100 个。聚类后分析,得到 1 个离群簇,离群簇中包含数据 609 条。排除这 609 条离群数据

后进行第 2 次聚类,对 99 个簇中心施行可定制的聚类。经过实验对比和用户实际需求,决定第 2 次聚类个数为 5 个。这样定制后聚类簇数为 5 个簇,这 5 个簇分别对应通信质量好、通信质量较好、通信质量中、通信质量较差和通信质量差,合并后每个簇的数据量分别是 40 869、26 300、8 879、2 930、3 078。

实验中分别记录了 11 组可调多趟 K-means 聚类算法和基于密度的 DBSCAN^[4]算法的运算时间,如图 4 所示。

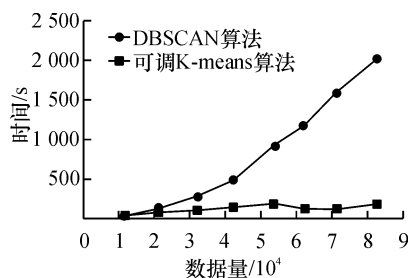


图4 算法效率示意图

Fig. 4 The diagram of the algorithm efficiency

从图4可以看出,由于初始值的选取有一定随机性,本文提出的可调多趟K-means聚类算法达到收敛时的迭代次数存在一定的不确定性,算法的执行时间有一定的波动.当数据量达到一定规模时,多趟K-means聚类算法的效率远远高于传统的DBSCAN算法.

由实验结果可知,通信质量好的客户数有40 869个,通信质量差的客户数有3 078个,说明当前环境下,通信质量总体情况令人满意.但是移动运营商仍需要分析通信质量差的3 078位客户与通信质量较差的2 930位客户,找出通信质量差的原因并及时进行处理,以尽量降低客户流失.

4.2 聚类效果的对比

实验将传统K-means算法和本文提出的可调多趟K-means的聚类效果进行比较,从误差平方和与分离性两个评价指标进行对比,结果如图5和图6所示.

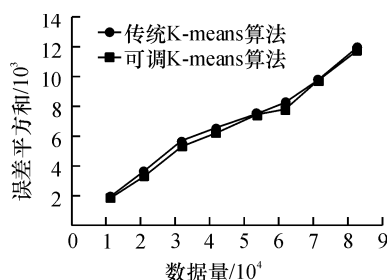


图5 算法误差平方和示意图

Fig. 5 The diagram of the error squares of the algorithm

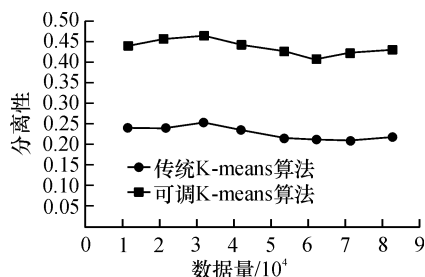


图6 算法分离性示意图

Fig. 6 The diagram of the algorithm separation

从图5中可以看出,本文提出的可调多趟K-means聚类算法效果在“相似性”方面略优于传统的K-means算法.由图6可知,在“相异性”方面,可调多趟K-means聚类算法远远优于传统K-means算法.

从以上实验结果可知,本文提出的可调多趟K-means聚类算法能够有效地处理海量数据的挖掘问题,该聚类算法应用于电信领域的客户细分有效、可行.本文为电信运营商有效地细分客户提供了参考.

5 结束语

K-means是目前数据挖掘中常用的聚类方法,本文针对传统K-means方法的缺点,提出了一个可调的多趟K-means聚类方法,该方法能够在保留传统K-means聚类方法优点的基础上有效地改进K-means存在的不足,并能根据实际需要分析并调整聚类结果的个数.实验表明,该算法应用于电信数据分析及其客户细分有效、可行,下一步工作将探讨本文的方法在大数据挖掘中的应用.

参考文献:

- [1] Reicheld F F, Sasser W E. Zero defections: Quality comes to services [J]. Harvard Business Review, 1990, 42(9): 105-110.
- [2] 张巍,滕少华,彭重嘉,等. 基于业务的移动网络用户体验协同建模方法 [J]. 兰州大学学报: 自然科学版, 2012, 48(4): 80-84.
- [3] Zhang W, Teng S H, Peng C J, et al. Business-based cooperative modeling method for QoE of the mobile network [J]. Journal of Lanzhou University: Natural Sciences Edition, 2012, 48(4): 80-84.
- [4] 徐猛. 面向用户体验的多QoS约束多流程调度方法研究 [D]. 济南: 山东大学计算机科学与技术学院, 2010.
- [5] 周天圆. 智能算法在电信业务用户体验感知分析中的应用 [D]. 长春: 吉林大学计算机科学与技术学院, 2011.
- [6] 郑茜茜. 基于数据挖掘的客户细分研究 [D]. 重庆: 重庆交通大学管理学院, 2013.
- [7] Han J W, Kamber M, Pei J. 数据挖掘概念与技术 [M]. 3版. 北京: 机械工业出版社, 2012.
- [8] 白亮. 聚类学习的理论分析与高效算法研究 [D]. 山西: 山西大学计算机与信息技术学院, 2012.
- [9] 王扶东, 马玉芳. 基于数据挖掘的客户细分方法的研究 [J]. 计算机工程与应用, 2011, 47(4): 215-218.
- [10] Wang F D, Ma Y F. Research of method for customer segment based on data mining [J]. Computer Engineering and Applications, 2011, 47(4): 215-218.

- [9] 杜巍,赵春荣,黄伟建. 改进的 K-means 聚类算法在客户细分中的应用研究 [J]. 河北经贸大学学报,2014,35(1): 118-121.
Du W, Zhao C R, Huang W J. Application of improved K-means cluster algorithm to customer segmentation [J]. Journal of Hebei University of Economics and Business, 2014,35(1): 118-121.
- [10] Mahmud M S, Rahman M M, Akhtar M N. Improvement of K-means Clustering algorithm with better initial centroids based on weighted average [C] // 2012 7th International Conference on Electrical and Computer Engineering. Dhaka: IEEE, 2012: 647-650
- [11] 王令剑,滕少华. 聚类和时间序列分析在入侵检测中的应用 [J]. 计算机应用,2010,30(3): 699-701,714.
Wang L J, Teng S H. Application of clustering and time-based sequence analysis in intrusion detection [J]. Journal of Computer Applications, 2010,30(3): 699-701, 714.
- [12] 曹永春,蔡正琦,邵亚斌. 基于 K-means 的改进人工蜂群聚类算法 [J]. 计算机应用,2014,34(1): 204-207, 217.
Cao Y C, Cai Z Q, Shao Y B. Improved artificial bee colony clustering algorithm based on K-means [J]. Journal of Computer Applications, 2014,34(1): 204-207,217.
- [13] 胡伟. 改进的层次 K 均值聚类算法 [J]. 计算机工程与应用,2013,49(2): 157-159.
Hu W. Improved hierarchical K-means clustering algorithm [J]. Computer Engineering and Applications, 2013,49(2): 157-159.
- [14] 高灵渲,张巍,霍颖翔,等. 改进的聚类模式过滤推荐算法 [J]. 江西师范大学学报,2012,36(1): 106-110.
Gao L X, Zhang W, Huo Y X, et al. Improved clustering filtering recommendation algorithm [J]. Journal of Jiangxi Normal University: Natural Science Edition, 2012,36(1): 106-110.
- [15] 易倩,滕少华,张巍. 基于马氏距离的 K 均值聚类算法的入侵检测 [J]. 江西师范大学学报,2012,36(5): 284-287.
Yi Q, Teng S H, Zhang W. Mahalanobis distance-based K-means clustering algorithm for intrusion detection [J]. Journal of Jiangxi Normal University: Natural Science Edition, 2012,36(5): 284-287.
- [16] Teng S H, Du H L, Zhang W, et al. A cooperative network intrusion detection based on heterogeneous distance function clustering [C] // [s. l.]: CSCWD, 2010: 140-145.
- [17] 廖松有,张继福,刘爱琴. 利用模糊熵约束的模糊 C 均值聚类算法 [J]. 小型微型计算机系统,2014,35(2): 379-383.
Liao S Y, Zhang J F, Liu A Q. Fuzzy C means clustering algorithm by using fuzzy entropy constraint [J]. Journal of Chinese Computer Systems, 2014,35(2): 379-383.
- [18] 卢鹏丽,王祖东. 密度敏感的层次化聚类算法研究 [J]. 计算机工程与应用,2014,50(4): 190-195.
Lu P L, Wang Z D. Density-sensitive hierarchical clustering algorithm [J]. Computer Engineering and Applications, 2014, 50(4): 190-195.
- [19] 潘章明,陈尹立. 面向大数据集的共享近邻聚类研究 [J]. 小型微型计算机系统,2014, 35(1): 50-54.
Pan Z M, Chen Y L. Research on shared nearest neighbor clustering for large dataset [J]. Journal of Chinese Computer Systems, 2014,35(1): 50-54.
- [20] 曾昭才. 移动统计——移动网络分析的另一宝藏 [J]. 移动通信,2005(11): 93-95.
- [21] 苗守野,肖瑞. 基于 MR 的 3G 无线网络质量评估方法研究 [J]. 移动通信,2011(18): 22-26.



广东工业大学滕少华教授

滕少华博士，为广东工业大学计算机学院教授，广东工业大学协同计算与知识工程团队负责人，任中国计算机学会协同计算专委会常务委员、中国计算机学会Petri网专委会常务委员、广东省计算机学会理事、广东省协同计算专委会副主任兼秘书长。自2000年以来，滕少华及其团队致力于“协同计算”、“网络安全”、“大数据与数据挖掘”、“云计算”和“复杂网络群组计算与协同建模”等方面的研究，提出了系列协同检测模型与方法，提出了具有创新性的基于对象监控的协同入侵检测模型，研究成果在国内知名安全企业中获得应用。

目前，滕少华教授累计在IEEE Transactions on Cybernetics、Journal of Networks、Information Science、计算机学报等国内外SCI、EI收录的专业期刊上发表论文150多篇，主持或参与完成了国家自然科学基金资助项目、教育部重点实验室项目、广东省自然科学基金资助项目、省部产学研重点项目、省市科技计划项目多项，主讲省、校级精品资源共享课程与研究生示范课程，多次获得广东省科学技术二等奖、广州市科学技术二、三等奖、广东省教学成果二等奖等奖励，其指导的研究生杜红乐获2011年度广东省百篇优秀硕士毕业论文等奖励。

在滕少华等教授带领下，协同计算与知识工程团队，围绕协同计算与知识工程，面向大数据、数据挖掘、网络安全、高性能计算、复杂网络环境的计算与服务、传感网等高新技术领域，“协同求索，求是求知”。

“有弗学，学之弗能弗措也；有弗问，问之弗知弗措也；有弗思，思之弗得弗措也；有弗辨，辨之弗明弗措也；有弗行，行之弗笃弗措也。”在“协同”环境下，协同团队齐心协力、互助互勉、以求知为原点，求索以径，求是以恒。

