基于 FCM 算法的电子商务客户分类研究

郑晓薇,马 琳
(辽宁师范大学 计算机与信息技术学院,辽宁 大连 116081)

摘 要:面对电子商务模式下电商对客户竞争的现状,针对传统的客户分类方法的不足,设计了一种基于 FCM 模糊聚类算法客户分类的并行算法。实验结果表明设计的方法能准确地对电商客户分类,在 MATLAB 集群下并行算法的运行取得了明显的并行效果。

关键词:电子商务客户分类;FCM 算法;MATLAB 集群并行

中图分类号:TP393 **文献标识码**:A **文章编号**:1674-7720(2013)15-0090-03

Rescarch of electronic commerce customer classification based on FCM algorithm

Zheng Xiaowei, Ma Lin
(College of Computer and Information Technology, Liaoning Normal University, Dalian 116081, China)

Abstract: Electricity supplier to customer facing E-commerce mode of competition of the status quo, for the shortcomings of traditional customer classification, designed a customer classification based on FCM clustering algorithm for parallel algorithms. This E-commerce website Eslite's historical trading data for experiment, results indicated that this design approach can accurately to customer classification, MATLAB clustering algorithm for parallel run has been made apparent side-effects.

Key words: E-commerce customer classification; FCM algorithm; MATLAB clustering and parallel

市场经济的发展和网络技术的革新促使电子商务迅速普及。在竞争激烈的电子商务经济模式下,客户成为电商竞争的焦点。电商想要对客户进行分析需要将客户分类,找出优质客户、挖掘潜在客户才能制定出针对性的营销策略。电商客户分类是指根据客户的历史交易情况将客户群划分为不同的等级,从中找出共同的要素并对客户的消费需求及消费行为进行研究,制定并实施有效的销售策略。

传统的客户分类方法是基于经验或简单统计方法[1],依据电商客户历史交易数据对客户过去和现在价值进行分析,忽略了客户的潜在价值和未来价值。这两种方法分类主观性强,与分类标准的关联性大,分类效果不理想。FCM 模糊聚类算法是多元统计算法中广泛应用于经济分析的算法,它是在聚类分析算法的基础上,增加"隶属度",用数学的方法定量地确定每一个样本点与各个类别的亲疏关系,分类结果客观。此外,面对电商网站运营产生的海量历史交易数据,本文利用 MATLAB 集群可以发挥其适合执行数据密集型任务的优势,解决"数据大,计算难"的问题,高效地计算出聚类结果。

本文基于 FCM 模糊聚类算法设计了一个针对电商客户分类的方法,以电商网站凡客诚品的历史交易数据为例进行实验测试设计方法的有效性。同时在 MATLAB 集群中针对 3 个规模不同的数据进行并行计算实验,做并行化研究。实验结果表明 FCM 模糊聚类算法能够准确地将电子商务客户分类,利用 MATLAB 集群的多个节点并行计算数据,缩减了计算数据时间。

1 电子商务网站客户分类算法

1.1 电子商务网站客户分类

电子商务客户分类是电商在收集和整理客户交易信息的基础上,按照客户交易记录把某一类的客户分到一个群体的过程,其原理如图 1 所示。

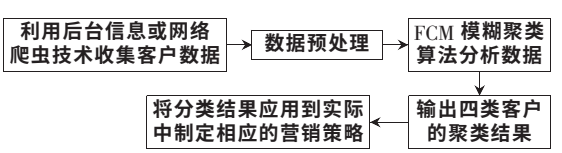

图 1 电子商务客户分类原理

首先收集电子商务客户的原始交易记录数据，利用电子商务后台数据或者爬虫技术爬取。其次是数据预处理环节，要对收集的数据进行规约和清洗，删除其中没有用处的数据。最后通过 FCM 模糊聚类算法对输入数据进行聚类分析，获得聚类分析结果。电商可以针对不同消费群体制定指定的销售策略，实现稳定盈利。

1.2 FCM 模糊聚类算法

K-means 聚类分析算法是依据实验数据本身具备的定性或定量的特征来对数据进行分组归类的方法，方便了解数据集的内在结构，是数据挖掘的主要数据分析方法[2]。算法优势是操作简单、聚类速度快。算法存在的缺陷是容易陷入局部最优值，这样获得的聚类结果是局部最优解而不是全局最优解。由于 K-means 聚类分析算法的缺陷，用于电子商务客户分类的聚类效果不理想。

模糊聚类分析算法 FCM(Fuzzy C-Means algorithm)是在 K-means 聚类分析算法的基础之上，增加“隶属度”，用数学方法定量地确定样本点与其他各个样本的亲疏关系，客观地划分样本集类型。能够客观地计算出每一个客户属于各类样本的概率，分析效果更加精确[3]。FCM 模糊聚类算法步骤如下：

FCM 模糊聚类分析算法的目标函数是：

$$J=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ji}^{m}\|x_j-v_i\|^2 \tag{1}$$

其目标函数的设定为：

$$\sum_{i=1}^{c}u_{ji}=1,1\leqslant j\leqslant n,1\leqslant i\leqslant c,u_{ji}\in[0,1] \tag{2}$$

其中 $d_{ji}=\|x_j-v_i\|$ 是样本 x_j 到聚类中心 v_i 的欧氏距离，c 为聚类个数 $(1<c<n)$，u_{ji} 是第 j 个样本到第 i 个类中心的隶属度，$U=[u_{ji}]$ 是 $n\times c$ 的矩阵，$V=[v_1,v_2,\cdots,v_c]$ 是 1 个 $s\times c$ 的矩阵，s 代表维数。推广到一般化即将 u_{ji}^2 换成 u_{ji}^m，m 是模糊权重因子 $(m>1)$，初始迭代标准为 $\varepsilon>0$，k 为迭代次数，当 $V^{(0)},k=0$。

计算步骤为：

(1)设定聚类的数目 C 和模糊度参数 M(指数，值在 1.5~2.5 之间)；

(2)给出初始隶属度矩阵 $U^{(0)}$；($U^{(0)}$ 中各列元素之和为 1)；

(3)计算 $U^{(k)}$，如果 $\forall j,i,d_{ji}(k)>0$，则有：

$$v_i=\frac{\sum_{j=1}^{n}(u_{ji})^m x_j}{\sum_{j=1}^{n}(u_{ji})^m} \tag{3}$$

(4)计算 $V^{(k+1)}$，则有

$$u_{ji}=\left[\sum_{k=1}^{c}\left(\frac{\|x_j-v_i\|^2}{\|x_j-v_k\|^2}\right)^{1/(m-1)}\right]^{-1} \tag{4}$$

(5)若 $\|V^{(k+1)}-V^{(k)}\|\leqslant\varepsilon$，则停止迭代，否则 $k=k+1$，转向步骤(1)。

FCM 模糊聚类算法流程图如图 2 所示。

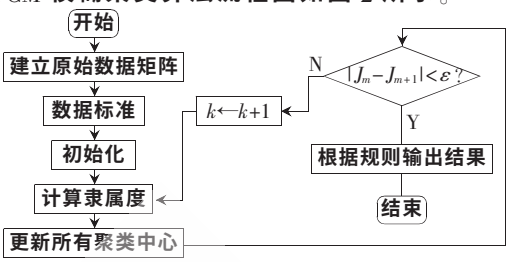

图 2 FCM 模糊聚类算法流程图

2 网站实例测试

2.1 举例实验

本文对电商网站凡客诚品的历史交易数据进行获取和预处理后，选择 100 个客户数据进行分类来举例说明。每 1 组数据包含 4 个指标：交易类别、交易数量、交易单价及总交易金额，数据利用网络爬虫技术获取，均为单笔交易数据。交易类别用数字替代：服装 11、鞋 22、箱包 33、配饰 44、家具 55、断码专区 66。聚类数目为 4 类。各个参数为 $n=100$，$c=4$，U 为 100×4 的矩阵，V 为 4×4 矩阵。模糊度 m 取 2.3。流程为：

(1)每个数据记录除以所在列的最大值来进行数据初始化，使其中的每一个数据大小均在[0,1]之间。目的是不使 FCM 模糊聚类算法在计算的时候倾向较大数值忽略较小数值，致使聚类结果不准确。

(2)设置模糊度 m 参数及初始隶属度矩阵。确定聚类的个数 c，给出原始聚类中心。开始进行聚类分析算法的计算。当前后两次迭代的目标函数值稳定时，也就是两个目标函数值相差值$<\varepsilon$ 时，聚类停止并输出聚类结果。结果如图 3 所示，图中横轴代表数据的 4 个数据点，纵轴代表 x_j 数据标准化后的值。

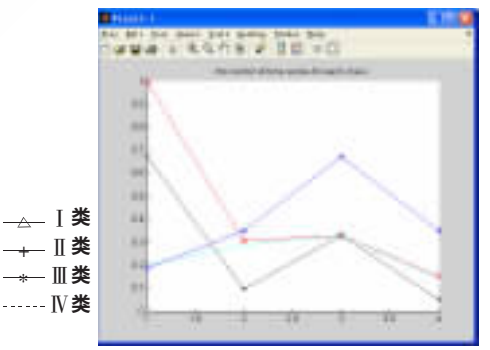

图 3 客户分类实验结果

表 1 是实验取得的隶属度矩阵表，结尾保留 4 位有效小数。列代表客户编号，行代表 4 个类别。对应的数值就是每一个客户属于每一类的概率。每一列概率数值相加之和为 1，代表概率越大，属于那一类的可能性越大。

2.2 举例实验分析

本文先后分别对这 100 个客户数据进行聚类，分为 3 类、4 类和 5 类。结果如图 4 和图 5 所示。

表 1 隶属度矩阵

	1	2	3	4	…	99	100
I	0.016 7	0.062 5	0.453 2	0.047 1	…	0.102 8	0.058 7
II	0.007 9	0.036 6	0.215 7	0.021 1	…	0.840 0	0.031 3
III	0.024 3	0.753 9	0.158 7	0.058 0	…	0.025 6	0.110 1
IV	0.951 0	0.147 0	0.172 4	0.873 7	…	0.031 5	0.799 9

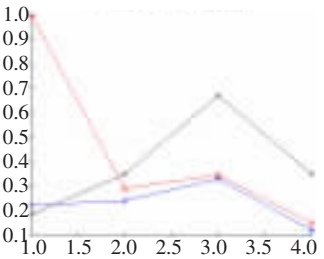

图 4 分为 3 类的聚类结果

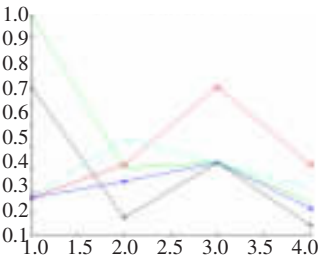

图 5 分为 5 类的聚类结果

由图 4 可以看出 3 类数据的特征性差别大，分为 3 类只是达到了粗略分类，数据特征不明显。图 5 显示分类繁杂，其中有的类别走势基本相同，不是聚类数目最优值。而图 3 的 4 类聚类结果图显示直观，数据特征性明显，所以选择 4 类进行用户分类。

2.3 举例实验结果

上述过程中，FCM 模糊聚类算法根据客户消费模式把客户分为 4 类，数据对应的归属类别如表 2 所示。

表 2 客户数据归类统计

	数据编号
I	4 9 14 18 34 37 51 67 68 69 71 77 81 82 89 90
II	7 16 20 21 22 23 24 25 27 28 45 62 63 70 73 74 75 78 80 83 84 88 98 99
III	2 8 10 12 13 91 92 93 17 19 30 31 35 36 38 52 54 57 59 97
IV	1 3 5 6 11 15 26 29 32 33 39 40 41 42 43 44 46 47 48 49 50 53 55 56 58 60 61 64 65 66 72 76 79 85 86 87 94 95 96 100

根据聚类结果图中的 4 个类别的聚类中心点坐标可以明显地看到 4 类不同消费群体的特征，分析如表 3 所示。

表 3 聚类中心点统计

	类别	单价	数量	金额
I	66	120	1	120
II	44	35	1	35
III	11	140	2	280
IV	11	110	1	110

第一类潜在客户：该类消费群体关心价格，喜欢打折促销。流失倾向偏大，对网站信任度低。虽然具有一定的价值，但给企业带来的利润小。

第二类小客户：该群体主要购买饰品，企业从这类消费群体可以获得的利润较小。流失倾向偏小，应该通过营销方法使其成为一般客户。维持该类客户对电子商务的发展仍具有一定的意义。

第三类优质客户：群体主要购买服装，这类群体是企业可以从中获得利润最大的群体。该类群体购买优质产品，且购买的数量多，是企业的高端顾客。该网站的客户忠诚度高，在一定时间内购买的商品种类和交易数量多，是企业需要重点维护的对象。

第四类一般客户群体：主要购买服装，该类客户偏向于购买普通服装，电商的该类客户数量最多。对网站的产品持肯定态度，虽然没有为电商提供高利润，但是交易会稳定持续地进行，是企业稳定生存的基础。

3 MATLAB 集群并行化

MATLAB 是一套高性能的数值计算和可视化软件，集数值分析、矩阵运算、图形处理和信号处理于一体。MATLAB 最大的优势在于它的强大的科学计算能力，专用工具箱具备全面的数学函数，能够执行数据复杂型任务和数据密集型任务[4]。

(1)实验环境：由 3 台 PC 机搭建的 MATLAB 集群。硬件配置：Intel (R)Core (TM)、i3CPU530@2.93 GHz (2CPUs)，2 GB 内存。软件配置：系统环境 Windows XP、MATLAB(R2011b)。文件大小：规模大小为 1 GB、2.2 GB、3 GB 的 3 个数据表。

(2)实验结果及分析：本文采用数据分割的方式对 FCM 模糊聚类算法进行集群并行计算。实验分别在单节点与多节点环境下执行，首先在双节点环境下的运行时间小于单节点下运行的时间，并行效果明显。其次又分别在 4 个节点与 6 个节点下分别执行聚类计算，实验结果表明时间缩短的增量与集群节点数目成正比，随着集群节点的增加而增大。说明用 MATLAB 集群来处理本文的数据是有效的，发挥了 MATLAB 集群处理数据密集型任务的优势，体现了 MATLAB 集群的高性能。实验结果如表 4 所示。

表 4 FCM 模糊聚类算法多节点并行执行效果分析表

数据规模/GB	计算时间/min				加速比		
	p=1	p=2	p=4	p=6	p=2	p=4	p=6
1	55	45	25	18	1.22	2.2	3.05
2.2	121	96	50	38	1.26	2.42	3.18
3	165	127	65	51	1.29	2.53	3.23

将记录在表中的数据作运行时间和加速比对比分析，如图 6、图 7 所示。

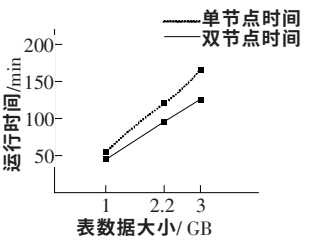

图 6 单节点与双节点运行时间对比

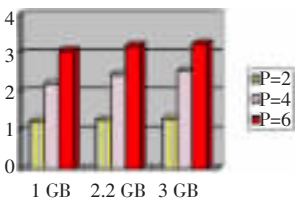

图 7 加速比变化图

通过对图 6 单节点与双节点环境下运行时间的对比，可以看出并行计算时间短于串行计算的时间，且随着数据规模的加大，时间缩短增量逐渐提高。图 7 显现

了加速比的变化，不同规模大小数据的加速比均随着集群节点数目的增大而增大。由此可以证实，FCM 模糊聚类算法在并行集群中应用于电子商务客户分类适用，能够取得良好的并行效果，输出结果时间缩短。充分说明了 FCM 模糊聚类算法并行化的可行性和 MATLAB 集群的高性能性。

电子商务处于蓬勃发展阶段，如何准确有效地对消费客户进行分类并制定针对性的营销策略是其盈利的关键。本文针对这一现实问题，选定多元统计分析中的 FCM 模糊聚类算法进行客户分类并做了并行化研究。实验结果表明，在 MATLAB 集群中运行并行后的 FCM 模糊聚类算法能够取得良好的并行效率，同时也验证了 MATLAB 集群在处理数据密集型任务的高效性。本文设计的方法可以应用于电子商务中，对电子商务客户分析方面有一定的实际意义。

参考文献

[1] 朱晶晶.电子商务网站分类体系理解的用户心智模型研究[D].南京：南京理工大学，2010.

[2] SELIM S Z. K-Means-type algorithms: A generalized convergence theorem and characterization of local optimality[J]. IEEE Transactions on Pattern Analysis and Machine Intelligce, 1984,6(1): 81-87.

[3] DUNN J C. A fuzzy relative of the IOSDATA process and its use in detecting compact well separated clusters [J]. Cybemet.3,197:32-57.

[4] MathWorks. MATLAB Distributed Computing Server 5 System Administrator's Guide [EB/OL]. http://www.mathworks.com/access/helpdesk/help/pdf_doc/mdce/mdce.pdf, 2010.

[5] 徐瑞，黄兆东，阎凤玉.MATLAB2007 科学计算与工程分析[M].北京：科学出版社，2008.

[6] 瞿小宁.K 均值聚类算法在商业银行客户分类中的应用[J].计算机仿真，2011，28(6)：357-360.

[7] 李容. 基于 K 均值聚类算法的图书商品推荐仿真系统[J].计算机仿真，2010，27(6)：346-349.

（收稿日期：2013-04-14）

作者简介：

郑晓薇，女，1957 年生，教授，主要研究方向：并行计算，数据库与计算机决策支持。

马琳，女，1987 年生，硕士，主要研究方向：并行计算。