客户行为的有效聚类

刘慧婷[1],倪志伟[2]
LIU Hui-ting[1],NI Zhi-wei[2]

1.安徽大学 计算机科学与技术学院,合肥 230039
2.合肥工业大学 计算机网络所,合肥 230009
1.School of Computer Science and Technology,Anhui University,Hefei 230039,China
2.Institute of Computer Network System,Hefei University of Technology,Hefei 230009,China
E-mail:wangph168@yahoo.com.cn

LIU Hui-ting,NI Zhi-wei.Effective algorithm to cluster customers' actions.Computer Engineering and Applications,2010,46(4):12-14.

Abstract: Clustering customers' transaction data is an important analysis means of customers' behavior.As customers' transaction data have high dimension,the clustering algorithm based on Empirical Mode Decomposition (EMD) and *K*-means is implemented to cluster the customers' actions in supermarkets,that is,employ the EMD and bottom-up algorithms to realize dimension reduction,and further use *K*-means algorithm to support effective clustering on data sequences,which have fewer dimensions. Customers are divided into different categories (or sub-market) by means of clustering customers' transaction data.Each sub-market is then described by the commodities which are purchased with higher rates,so as to make respective promotions and advertisements.The clustering algorithm proposed in this paper can effectively cluster customers' behavior,and as the algorithm has dealt with dimensionality reduction to a sequence of transaction data,this can save a certain amount of storage space.

Key words: Empirical Mode Decomposition(EMD);bottom-up algorithm;*K*-means algorithm;trend extraction;clustering customers' actions

摘 要: 对客户的交易数据进行聚类是客户行为分析的一个重要手段。针对客户交易数据维数高的特点,提出了基于EMD和*K*-means的顾客行为聚类算法。首先利用EMD和自底向上分段算法实现交易数据序列维度的约简,再利用*K*-means算法完成降维后序列的聚类,最后利用每个类别中购买率较高的商品作为该类的描述,为商家提供促销依据。该聚类算法一方面可以有效实现客户行为的聚类,另一方面,由于算法对交易数据序列进行了降维处理,节约了一定的存储空间。

关键词: 经验模态分解方法;自底向上算法;*K*-means算法;趋势提取;客户行为聚类
DOI: 10.3778/j.issn.1002-8331.2010.04.004 **文章编号:** 1002-8331(2010)04-0012-03 **文献标识码:** A **中图分类号:** TP181

1 引言

在目前竞争日益激烈的知识经济环境和电子商务经济模式下,进行客户关系管理已成为所有企业的当务之急。客户行为分析是客户关系管理的研究热点之一,分析的主要目的是提高企业利润、挽留老客户、发现潜在新客户、交叉销售和销售自动化等。目前,客户行为分析广泛应用于金融、保险和投资等众多领域[1]。

客户行为聚类是客户行为分析的一个重要手段,它把大量的客户根据某些属性度量分成不同的类,使得属于相同类的客户拥有相似的属性,而在属于不同类的客户的属性相差较大。

客户行为聚类方法通常依赖于一段时间内顾客的购买行为。将顾客根据购买行为,即交易数据,划分为不同的聚类之后,利用每个聚类中购买率较高的商品作为该聚类的描述,然后有针对性地促销和广告[2]。

因为顾客行为的聚类是通过对顾客的购买行为,即交易数据进行聚类实现的,交易数据的数据量大、维数高,提出了基于EMD和*K*-means的客户行为聚类算法,并把它应用于超市交易数据的分析中。先利用EMD和自底向上分段算法进行维度约简,再利用*K*-means算法进行聚类,该算法节约了存储空间,并且可以有效地完成客户行为的聚类。

基金项目: 国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2007AA04Z116);国家自然科学基金(the National Natural Science Foundation of China under Grant No.70871033);安徽省高校省级自然科学研究项目(No.KJ2007B303ZC)。

作者简介: 刘慧婷(1978-),女,博士,副教授,主要研究方向为算法分析与设计,机器学习;倪志伟(1963-),男,博士生导师,教授,主要研究方向为机器学习、数据挖掘和智能决策支持系统。

收稿日期: 2009-10-12 **修回日期:** 2009-11-27

2　交易数据维度的约简

该文聚类算法分为两部分：交易数据序列维度的约简,以及降维后的交易数据序列的聚类。下面将重点阐述交易数据维度约简方法的主要思想。

2.1　EMD 理论和自底向上算法

2.1.1　EMD 理论

EMD 方法是 Huang 等[3]在 1998 年提出的一种信号处理方法。它依据数据自身的时间尺度特征来进行信号分解,无须预先设定任何基函数。这一点与建立在先验性的谐波基函数和小波基函数上的傅里叶分解与小波分解方法具有本质性差别。正是由于这样的特点,EMD 方法在处理非平稳及非线性数据上,具有明显的优势。所以,EMD 方法一经提出就在不同的工程领域得到了有效的应用[4]。

从本质上讲,该方法是对一个信号进行平稳化处理,对序列进行平稳化处理的过程称为"筛"过程。如果序列 $x(t)$极值点的数量与过零点个数的差别超过一个,或者上包络线和下包络线的均值不是处处为零,序列将由"筛"过程分解成若干个本征模函数(Intrinsic Mode Function,IMF)$c(t)$及序列的趋势 $r(t)$。分解结果表示为:

$$x(t)=\sum_{i=1}^{n}c_i(t)+r_n(t)$$

EMD 方法对原始序列进行分解，提取出的趋势和原始序列的关系如图 1 所示。

从图 1 可以看出,EMD 方法对原始序列中的噪声进行了过滤,产生的趋势序列准确反映了原序列的趋势走向,序列变得更加清晰,而信息量丢失相对较少。

2.1.2　自底向上算法

自底向上(Bottom-Up)算法首先得到最精细的线性分段表示,即数据序列上相邻两点组成最小分段。然后计算合并两个相邻分段所产生的拟合误差,合并拟合误差最小的两个邻接分段,直到该拟合误差超过某个指定阈值[5-6]。

2.2　交易数据序列维度约简算法

交易数据序列维度的约简包含以下几步:(1)利用 EMD 方法对交易数据序列进行趋势提取;(2)利用自底向上算法对趋势序列进行分段,生成分段序列;(3)把分段序列转换成齐序列;(4)把齐序列转换成{-1,0,1}构成的序列,实现交易数据序列的降维。下面对上述几个步骤进行详细说明。

(1)趋势序列的提取

该文客户行为聚类算法中,是通过 EMD 方法的"筛"过程对交易数据序列进行平稳化处理来获取趋势序列的。"筛"过程停止的条件是分解得到的剩余部分的极值点个数小于预定值[7],在实际应用中,如果提取出的趋势序列不能反映出原序列的走势,则需要适当调节预定值的大小。若交易数据序列表示为 $x(t)$,IMF 分量用 $c_i(t)$表示,趋势序列用 $r_n(t)$表示,交易数据序列的趋势可以表示为:

$$r_n(t)=x(t)-\sum_{i=1}^{n}c_i(t)$$

从图 1 可以看出,EMD 方法对原始序列中的噪声进行了过滤,产生的趋势序列准确反映了原序列的趋势走向。

(2)利用 Keogh[5]提出的自底向上分段算法对趋势序列进行分段。

(3)把分段后的序列转换成齐序列

首先给出齐序列的定义,如下:

定义 1　假设用$[(lx,ly),(rx,ry)]$表示分得的每一段的最左端数据点的横、纵坐标,最右端数据点的横、纵坐标,分段序列 S_1 和 S_2 是齐序列当且仅当分段数相等,且满足 $lx_{1i}=lx_{2i}$,$rx_{1i}=rx_{2i}$,$1\leqslant i\leqslant N$,其中 N 是序列分段后分得的段数。

$S_1=\{[(lx_{11},ly_{11}),(rx_{11},ry_{11})],\cdots,[(lx_{1i},ly_{1i}),(rx_{1i},ry_{1i})],\cdots,[(lx_{1N},ly_{1N}),(rx_{1N},ry_{1N})]\}$

$S_2=\{[(lx_{21},ly_{21}),(rx_{21},ry_{21})],\cdots,[(lx_{2i},ly_{2i}),(rx_{2i},ry_{2i})],\cdots,[(lx_{2N},ly_{2N}),(rx_{2N},ry_{2N})]\}$

实际的趋势序列分段后,基本不存在齐序列,如图 2 所示。

图 2 中，每个趋势序列长度都是 166，而且均被分成 13 段。如果用$[lx,rx]$表示分得的每一段的最左端数据点的横坐标,最右端数据点的横坐标，第一个趋势序列被分成以下 13 段:[1,8],[9,16],[17,24],[25,48],[49,56],[57,74],[75,96],[97,104],[105,118],[119,138],[139,144],[145,150],[151,166];第二个趋势序列被分成以下 13 段:[1,8],[9,16],[17,24],[25,46],[47,60],[61,70],[71,96],[97,102],[103,118],[119,128],[129,142],[143,152],[153,166]。

这两个趋势序列的分段序列虽然都是 13 段，但并不是每一段的起始点和终止点都相同,所以它们两个不是齐序列。

可以通过以下步骤把分段序列转换成齐序列:

①找出所有分段序列中每一段的最右端数据点的横坐标,即 rx,形成集合 P,对集合 P 进行排序,并删除其中重复的数据点。

②对于分段序列的每一个分段,如果 P 中小于或者等于该段的最右端数据点的横坐标 rx,大于该段的最左端数据点的横坐标 lx 的元素的个数为 m,该段就被分为 m 个子段。

按照上述转换算法，图 2 中的两个分段序列转换成了 20 段,它们的$[lx\ \ rx]$分别为[1 8],[9 16],[17 24],[25 46],[47 48],[49 56],[57 60],[61 70],[71 74],[75 96],[97 102],[103 104],

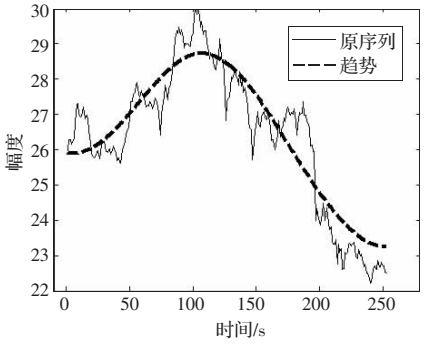

图 1　原始序列及其趋势

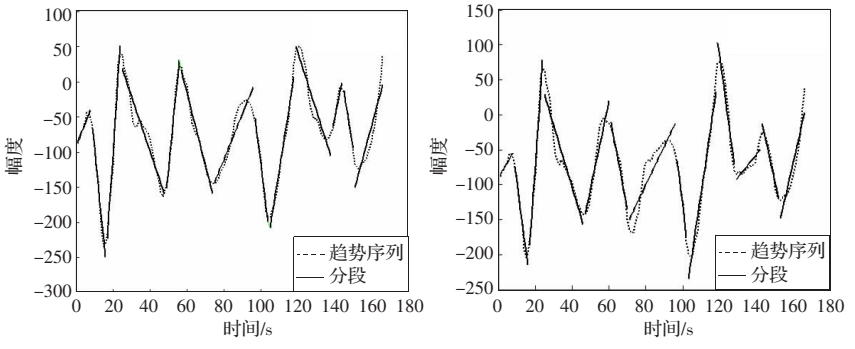

图 2　两个趋势序列及它们的分段序列

[105 118],[119 128],[129 138],[139 142],[143 144],[145 150],[151 152],[153 166]。分段数增加了,每一段的横坐标的区间相同了,成为齐序列。可以对齐序列进行各种操作。

(4)把齐序列转换成{-1,0,1}构成的序列

假设齐序列包含 m 段,转换后的序列为 $STR=(h_1,h_2,\cdots,h_m)$,h_i 可以表示如下:

$$h_i=\begin{cases}-1, & \text{若 } ry_i-ly_i<0\\ 0, & \text{若 } ry_i-ly_i=0\\ 1, & \text{若 } ry_i-ly_i>0\end{cases}\qquad i=1,2,\cdots,m$$

其中,ry_i 表示线段 i 最右端采样点的纵坐标值,ly_i 表示线段 i 最左端采样点的纵坐标值。

如图 3 所示的序列,通过变换后,转变成(1,-1,-1,1,1)。

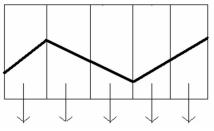

图 3 分段序列的符号转换

图 2 中两个趋势序列的长度是 166,通过第 2 步的分段后均被分成 13 段,再通过齐序列的转换成为 20 段,最后变成长度为 20 的{-1,0,1}构成的序列,实现了数据维度的约简。

3 基于 K-means 算法的客户行为聚类

提出的基于 EMD 和 K-means 的客户行为聚类算法先对交易数据序列进行 EMD 分解,提取它们的趋势,再对趋势序列进行分段、转换。从图 1 可以看出,EMD 方法提取出的趋势序列准确反映了原序列的趋势走向,序列变得更加清晰,所以在趋势序列的基础上进行分段、转换,一方面有效地实现了降维,另一方面,信息量丢失相对较少。经过上文得到的由{-1,0,1}构成的齐序列可以代替原序列用于交易数据序列聚类中。

K-means 算法是目前应用最为广泛的聚类算法之一,它的基本思想是:对于给定的聚类数目 K,首先随机创建一个初始划分,即随机选择某些数据代表点作为初始聚类中心,根据其余数据点到各聚类中心的距离将其分到各个类中,然后重新确定新的聚类中心,以此类推采用迭代方法将聚类中心不断移动来尝试进一步改进划分,直到聚类中心不再发生变化[8]。

下面给出 K-means 算法的聚类过程:

给定数据集合,$D=\{d_1,d_2,\cdots,d_n\}$。

(1)随机从集合 D 中选择 K 个对象作为初始簇中心;

(2)将 D 中所有的点分配到最近的簇;

(3)重新计算每个簇的中心;

(4)重复(2)、(3)直至簇中心不再发生变化或迭代次数超过设定的最大迭代次数。

4 实验

使用 FoodMart2000 数据库中的交易数据对提出的算法进行测试。从销售表 sales fact 1998 中取出 34 047 条销售记录,这部分数据涉及编号为 1~1 580 共 1 580 个顾客和 1 559 种商品。

首先把顾客号、顾客购买的商品号从销售表中提取出来,然后把顾客的交易数据生成一个 1 580×1 559 的矩阵,每一行代表一个顾客,每一列代表一种商品。如果第 i 个顾客买了第 j 种商品,则该矩阵中的第 i 行第 j 列元素为 1;如果第 i 个顾客没有买第 j 种商品,则该矩阵中的第 i 行第 j 列元素为 0。

数据预处理以后,就可以利用 EMD 和 K-means 算法对客户行为进行聚类。

(1)对矩阵的每一行,即每一个顾客的交易数据进行 EMD 分解,提取交易序列的趋势;

(2)利用自底向上(Bottom-Up)分段算法对趋势序列进行分段,把序列分成 20 段,形成 1 580×20 的矩阵;

(3)实际的趋势序列分段后,基本不存在齐序列,因此要把分段后的序列转换成齐序列;

(4)把分段序列转换成由{-1,0,1}构成的 1 580×779 齐序列矩阵,实现交易数据序列的降维;

(5)利用 K-means 聚类算法对降维后的序列进行聚类;

(6)计算出每一个聚类中,每种商品被购买的次数,找出购买次数频繁的商品。

令结果的簇数为 3,表 1~表 3 给出了不同顾客分段的描述,其中"商品被购买的次数"字段是聚类中顾客在某种商品上的消费次数,它刻画了聚类中的顾客相对于其他顾客更偏爱某种商品的程度。

表 1、表 2 和表 3 分别是聚类 C_1、聚类 C_2 和聚类 C_3 中顾客消费量最高的前 8 种商品,不同聚类的顾客其口味和生活习惯有所不同。

C_1 中的顾客对零食和肉类食品比较感兴趣;C_2 中的顾客比较喜欢烹饪;相比之下,C_3 中的顾客好像特别钟爱奶制品中的奶酪和罐装食品,而且她们在卫生用品和止疼药上消费量也偏高。

知道了每类顾客的偏好,然后有针对性地促销和广告,能够有效地指导商业行为。

5 结论

现代的市场营销由过去的产品中心论演变为客户中心论。客户中心论要求对客户进行全方位的认识,进行一对一的准确

表 1 聚类 1 喜爱购买的商品表

商品被购买的次数	商品名称
18	Dips(Snack Foods)
17	Ice Cream
16	Chocolate Candy
15	Dried Fruit(Snack Foods)
15	Deli Meats
14	Muffins(Bread)
14	Waffles
14	Cookies

表 2 聚类 2 喜爱购买的商品表

商品被购买的次数	商品名称
19	Shrimp(Canned Shrimp)
17	Tofu(Packaged Vegetables)
17	Bagels(Bread)
17	Preserves(Jams and Jellies)
16	Spices(Baking Goods)
16	Fresh Fruit
16	Pot Cleaners
16	Frozen Vegetables

表 3 聚类 3 喜爱购买的商品表

商品被购买的次数	商品名称
16	Tuna(Canned Tuna)
16	Soup(Canned Soup)
16	Acetaminophen(Pain Relievers)
15	Canned Vegetables
15	Cheese(Dairy)
14	Aspirin(Pain Relievers)
14	Personal Hygiene
14	Cheese(Dairy)

(下转 28 页)