K-means算法在电信CRM客户分类中的应用①

左国才 [1,2] 杨金民 [1] (1.湖南大学 软件学院 湖南 长沙 410082;
2.湖南软件职业学院 湖南 长沙 410205)

摘 要: 面对电信市场竞争的日益加剧和信息技术的迅猛发展,电信运营商必须建立以“客户为中心”的管理模式,将客户进行分类,针对不同的客户,研究出相应的营销策略。数据挖掘中的 K-means 聚类算法能对大型数据集进行高效分类。对 K-means 算法进行改进,使其能够应用于复杂的电信客户关系管理,实现更加准确和全面的客户分类。

关键词: 电信 CRM; 数据挖掘; 客户分类; K-means; 聚类分析算法

K-means Algorithm for CRM Customers in the Telecommunications Classification

ZUO Guo-Cai[1,2], YANG Jin-Min[1]
(1. Software School, Hunan University, Changsha, 410082, China; 2. Hunan Vocational Institute of Software, Changsha 410205, China)

Abstract: With the sharp competition in the telecommunications market and the rapid development of information technology, telecommunications operators must establish a “customer-centric” management style to classity customers. For different customers, different marketing strategies are developed. Data Mining in the K-means clustering algorithm for large data sets can be efficiently classified. In this paper, K-means algorithm is used in complex telecom customer relationship management to achieve more accurate and comprehensive customer classification.

Keywords: telecom CRM; data mining; customer classification; K-means cluster analysis algorithm

1 引言

数据挖掘是数据库研究、开发和应用最活跃的分支科学之一,从大量数据中用非平凡的方法发现有用的知识和人们感兴趣的数据模式成了人们的一种自然需求。随着数据挖掘研究的蓬勃发展,出现很多数据挖掘的方法,其中聚簇是最基本的方法,它既可以独立地应用,也可以作为其他数据挖掘方法的前期工作。在聚簇方法中,k-means 算法是最著名和最常用的划分法之一。k-means 算法能有效地处理规模较大和高维的数据集合,但却只能聚簇数值数据,因为数值数据能用欧几里德距离测量不同数据对象之间的相异度。k-means 算法不能处理分类属性型数据。经过改进后的 k-means 算法能够适应复杂的电信 CRM 中非数值数据的处理,实现更加准确和全面的客户分类。

2 电信客户关系管理现状分析

随着中国电信竞争格局的改变和通信技术的飞速发展,中国电信业的市场环境发生了根本性的变化,加入 WTO 后,我国电信市场将逐步对外开放,国内电信运营商正面对一个全新的、更加激烈的、国际国内全方位的市场竞争环境。电信运营商意识到在竞争越来越激烈的商业时代,资源占有成为企业生死成败的关键,客户才是企业生存和发展的根基,而如何改善客户服务,增强客户的满意度和忠诚度,提升客户

① 基金项目:国家自然科学基金(60703097,60703155);国家高技术研究发展计划(863)(2007CB310702)
收稿时间:2009-05-20

价值来扩大自身的收入和利润，如何用信息化管理来替代原有的传统管理手段等问题，是电信企业面临解决的重要难题。

因此，企业必须完成从“产品”导向向“客户”导向的转变，对客户进行有效的管理，从储存大量客户信息的数据仓库中经过深层分析，获得有利于商业运作，提高企业市场竞争力的有效信息。而实现这些有效性的关键技术就是数据挖掘，即从海量数据中挖掘出更有价值的潜在信息。正是有了数据挖掘技术的支持，才使得客户关系管理的理念和目标得以实现，满足现代激烈竞争时代的需求与挑战。

3 电信客户关系管理

3.1 客户关系管理(CRM)

CRM 是一种旨在改善企业与客户之间关系的新型管理方法。它是企业通过富有意义的交流和沟通,理解并影响客户行为,最终实现提高客户获取、客户保留、客户忠诚和客户创利的目的。

3.2 电信客户关系管理的特点

电信运营商作为经营电信业务的服务商有着自己独特的客户特性与市场特性。首先，电信业的客户具有多元性，从政府机关、企业集团和学校宾馆，直到普通个人都是其客户；其次，电信客户的需求特征上具有多样性，表现出极大的行业特征及业务量特征，从个人到团体，从农村到城市，从低收入群体到小康家庭对电信服务有着各种各样的层次需求。同时，电信业的服务比重大，电信产品更新率高。另外，客户使用电信服务的随机性强，用户稳定性差，年均离网率不断上升。最后，中国电信市场中低端用户比重上升。

4 数据挖掘(DM)

CRM 中的数据挖掘是指从大量的有关客户的数据中挖掘出隐含的、先前未知的、对企业决策有潜在价值的知识和规则。

常用的数据挖掘方法有：关联分析；序列模式分析；分类分析；聚类分析；孤立点分析等。

聚类分析是从给定的数据集中搜索数据对象之间所存在的有价值联系。在商业上，聚类可以通过顾客数据将顾客信息分组，并对顾客的购买模式进行描述，找出他们的特征，制定针对性的营销方案。

5 K-means聚类分析算法在电信CRM中进行客户分类的应用

5.1 聚类分析中的数据结构

许多基于内存的聚类算法选择两种有代表性的数据结构：数据矩阵和相异度矩阵。

数据矩阵是一个对象-属性结构。它是由 n 个对象组成，如：人；这些对象是利用 p 个属性来进行描述的，如：年龄、高度、重量等。数据矩阵采用关系表形式或 n×p 矩阵来表示，如图 1 所示。

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

图 1　数据矩

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \cdots & \cdots & \cdots & \cdots \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

图 2 相异度矩

相异度矩阵是一个对象-对象结构。它存放所有 n 个对象彼此之间所形成的差异。它一般采用 n×n 矩阵来表示，如图 2 所示。

其中 d(i,j)表示对象 i 和对象 j 之间的差异(或不相似程度)。通常 d(i,j)为一个非负数；当对象 i 和对象 j 非常相似或彼此“接近”时，该数值接近 0；该数值越大，就表示对象 i 和对象 j 越不相似。由于有 d(i,j)=d(j,i)且 d(i,i)=0，因此就有下面所示三角矩阵。

5.1.1 数据点的数据结构

数据点的数据结构：

```
class Point{
string pk;
float x,y;
public float distance(class Point){
float d;
d=abs(x-Point.x)+abs(y-Point.y);
return d;}
public float distance(class Reference){
float d;
d=abs(x-Reference.x)+abs(y-Reference.y);
return d;
}}
```

其中，pk 代表该数据在数据库中的主键，用于唯一地标识数据。x、y 分别表示该数据点第一维和第二维的

坐标。其中重载的方法 distance 一个用于计算该数据点与其它数据点之间的距离，另一个用于计算数据点与代表点之间的距离，以便确定代表点代表区域内的点集。另外需要说明，表示数据点的类 Point 存放在一个一维数组 PointSet 中,通过访问数组中的每个元素，就可以得到每个数据点的信息。在扫描数据点时，就是从数组中提取信息。

5.1.2 代表点的数据结构

代表点的数据结构：

```
class Reference{
float xm, ym, xs, ys;
int ns, clflag;
ArrayList ps;
public Reference(){
clflag=0;
}
public float distance(class Reference){
float d;
d=abs(xm-Reference.xm)+abs(ym-Reference.y
m);
return d;
}}
```

其中：x_m 表示代表点的 x 坐标，y_m 表示代表点的 y 坐标；x_s 表示代表点代表区域内的所有点的 x 坐标之和，y_s 表示代表点代表区域内的所有点的 y 坐标之和；n_s 表示代表点代表区域内点的总数，在算法中将要用这个参数来表示代表点的密度；p_s 是一个集合的实例，表示代表点代表区域内的点的集合；clflag 是在对代表点进行簇的划分时，用于标记该代表点被划分在第几个簇中。其中方法 distance 用于对代表点进行簇的划分时计算各个代表点之间的距离，如果两个代表点的距离小于或等于 2 倍的 Radius 则他们为邻接代表点，即它们处于同一簇中。

5.1.3 改进后的 k-means 算法的 java 实现

(1) 什么是 k-means 聚类算法?

K-means 是最常用的聚类算法之一，能有效地处理规模较大和高维的数据集合，能对大型数据集进行高效分类，把数据分成几组，按照定义的测量标准，同组内数据与其他组数据相比具有较强的相似性，这就叫聚簇。

k-means 算法的效率比较高；缺点是只能处理数值型数据，不能处理分类数据，对例外数据非常敏感，不能处理非凸面形状的聚簇。

k-means 算法接受输入量 k ；然后将 n 个数据对象划分为 k 个聚类以便使得所获得的聚类满足：同一聚类中的对象相似度较高；而不同聚类中的对象相似度较小。聚类相似度是利用各聚类中对象的均值所获得一个“中心对象”(引力中心)来进行计算的。

(2) k-means 算法的处理流程

① 从 c 个数据对象任意选择 k 个对象作为初始聚类中心；

② 循环③到④直到每个聚类不再发生变化为止；

③ 根据每个聚类对象的均值(中心对象)，计算每个对象与这些中心对象的距离；并根据最小距离重新对相应对象进行划分；

④ 重新计算每个(有变化)聚类的均值(中心对象)；

(3) k-means 算法的改进

k-means 算法是在数据挖掘领域中普遍应用的聚类算法，它只能处理数值型数据，而不能处理分类属性型数据。例如表示人的属性有：姓名、性别、年龄、家庭住址等属性。而改进后的 k-means 算法就能够处理分类属性型数据。采用相异度来代替 k-means 算法中的距离，相异度越小，则表示距离越小。一个样本和一个聚类中心的相异度就是它们各个属性不相同的个数，不相同则记为一，最后计算一的总和。这个和就是某个样本到某个聚类中心的相异度。该样本属于相异度最小的聚类中心。

相异度测量：设 X、Y 是分类数据集中的两个对象，该对象是 m(x1, x2, …, xm) 维的，则这两个对象之间的相异度为：

$d_i(X,Y)=\sum_{j=1}^{m}\delta(X_j,Y_j)$ ；其中 $\delta(x_j,y_j)$

$=\begin{matrix}0(x_j=y_j)\\1(x_j\neq y_j)\end{matrix}$ ，

该过程可以被描述为如下数学问题：

最小化 $P(W,Q)=\sum_{i=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m}w_{i,l}\delta(x_{i,j},q_{i,j})$

其中 $w_{i,l}\in W$，并且 $Q_l=\left[q_{i,1},q_{i,2},\cdots,q_{i,m}\right]\in Q$。

算法实现的思想：

1) 确定聚簇数 k 以及阈值 t；

为了确定聚簇数 k，用一个基于距离计算相异度的聚簇方法聚簇样本数据，下面是具体操作过程。

① 从原始数据中随机地取 m 个点作为样本数据放到集合 S 中，定义一个初始阈值 t = t0。

② 给所有的样本数据加上未被聚簇的标记，并定义一个 k = 0。

③ 从样本数据集未被聚簇的点中选择一个初始点 P。{标记 P 属于聚簇 Ck；从 P 开始递归地按照深度优先方式遍历各点，P' = NEAR(P, t)，如果 P'不是空就属于 Ck，否则退回到前一点接着进行搜索，更新 d，d 是聚簇 Ck 中各对顶点之间的平均距离，更新 t = t0×d，t 是阈值。}

④ 如果还有未聚簇的点，使 k=k+1，重复(3)进行聚簇。

其中，函数 NEAR(P，t)是：

{寻找 P 的最近邻居 P'，即 dist (P'，P)≤t；如果没有 P'被发现就返回 NULL，否则返回 P'。}

2) 用样本数据得到的聚簇结果求聚簇模式 Q，用基于频率的方法求模式向量 Q 的过程如下：

求每一簇中在某一属性上对象数的百分数 fr，fr 称为相对频率，找出每个簇中相对频率最大的一个的属性作为 Q 向量的属性值，即让 n(k, j)是第 k 组属性 j 上的对象数，fr(Aj =?ck, j/X) = n(k, j)/n 最大者就是 qj 属性。每个属性都这样求，可以得到 Q 向量。

(4) 算法的实现

```
//找出最接近这个数据项的簇类
public Cluster findMostNearestNeighbor(){
int mindex=0; //最接近簇类在簇类集中的下标
ArrayList<Cluster>
cList=this.neighborClusters.clusterList; //簇类集
int length=cList.size();//目标簇集
Cluster rCls=cList.get(0);
PointND clsCenter=rCls.center;
double mindistance=
this.rowdata.squareDistance(clsCenter);
if(mindistance==-1.0)
 System.out.println("Error!");
//遍历簇集,对每个簇计算与该数据项的距离
if(length>1){
for(int i=1;i<length;i++){
  Cluster cls=cList.get(i);
  PointND center=cls.center;
  //查找距离最小的
double newDistance=
this.rowdata.squareDistance(center);
if(newDistance<mindistance){
mindex=i;
mindistance=newDistance;}}
 rCls=cList.get(mindex);}
 this.mostNearestCluster=rCls;
return rCls;}
//根据聚类结果对数据集分类
private void classify(Matrix theDataSet){
ArrayList<DataItem>
rows=theDataSet.rows;
for(int i=0;i<rows.size();i++){
DataItem rowdata=rows.get(i);
ClusterSet
neighbors=rowdata.neighborClusters;
Cluster
cluster=rowdata.mostNearestCluster;
int  labelIndex=
neighbors.clusterList.indexOf(cluster);
labelIndex++;
rowdata.setLabel(""+labelIndex);}}
//随机选择 N 个中心点进行聚类,直到分类精度超过界限
public void randomSelectTest2(int num) throws
IOException{ //进行聚类          Matrix
ranMatrix=new Matrix();
ranMatrix.rows=(ArrayList<DataItem>)dat
Set.rows.clone();
ranMatrix.height=dataSet.height;
ranMatrix.width=dataSet.width;
ClusterSet
ranClusters=this.initClusterSet(ranMatrix, num);
this.kMeansClustering(ranClusters,ranMatrix);
this.classify(ranMatrix);
this.saveFile();}
```

① 共有 336 组数据。KMEANS 用于存放进行标

准化处理过的数据，KMCLASS 用于存放进行聚类(K-Means)处理过的结果。

② jfreechart，相应在的 jar 包。jcommon-1.0.12.jar，jfreechart-1.0.9.jar

③ 实现环境，是在 myeclipse6.0GA，tomcat6.0.10

6 测试数据及运行结果分析

测试数据集是某电信公司的客户信息数据库，数据量为 158200。实验的硬件环境：PC 计算机，CPU 为 PⅣ2.0G，内存为 2G；软件环境：操作系统为 Windows Professional 2000，编程环境 Eclipse 3.3/ myeclipse6.0GA，tomcat6.0.10。

运行结果：

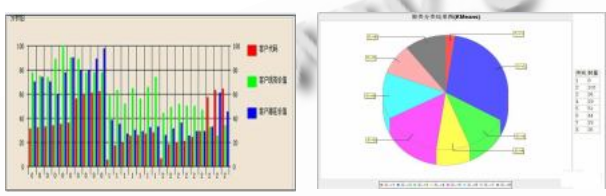

图 3 改进后的 K-means 算法运行结果

改进后的 K－means 算法在本文中主要是对客户的现有价值和潜在价值进行聚类分析，从而对客户进行分类，制定相应的营销政策。实验结果表明，改进后的 K-means 算法能够高效地实现客户分类，为电信部门解决客户分类的难题。

通过以上分析，我们最终把客户分为 8 类，分别为：高端商务型客户、普通商务型客户、时尚追求型客户、费用节省型客户、普通消费型客户、流动型客户、稳定型客户、潜在客户。分类的同时，我们可以得到每个顾客被分到哪个组，以及每组的客户选择不同产品的概率，对于电信运营商提供个性化服务、确定客户价值、进行套餐设计和实行深度营销等都有很强的现实意义。

当然，本研究也有一定的局限性：

① 样本量相对不大。由于数据来源问题，样本量的选取相对不是很多，如果有更多的样本供研究，研究的推广价值应该更大。

② 如果能对连续很多期的数据进行研究，应该可以对客户未来的消费情况进行预测，也是客户流失管理所关心的内容之一。

③ 如果能结合客户的背景资料综合分析，对现实的指导意义会更大。

综上所述，改进后的 k-means 算法在电信客户聚类分组中的应用是相对有效的，对于各类行业中关于客户数据的聚类分组同样具有很强的指导意义。同时希望在今后的研究中能够通过对更加完备的数据的分析，来完善该研究方法。

参考文献

1 Huang ZX. Extensions to the K-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 1998,2:283－304.

2 Barbara D. Using self-similarity to cluster large data sets. Data Mining and Knowledge Discovery, 2003,7:123－152.

3 Modha DS, Spangler WS. Feature Weighting. K-Means Clustering. Machine Learning, 2003.52:217－237.

4 James D. Better together marketing research and CRM. Marketing News, 2002,36:1015－16.

5 Hand D, Mannila H, Smyth P.张银奎,廖丽,宋俊,等译.数据挖掘原理.北京:机械工业出版社, 2003.

6 Michael J, Berry A, Gordon S.Linoff.袁卫等译.数据挖掘—客户关系治理的科学与艺术.北京:中国财政经济出版社, 2004.

7 朱爱群.客户关系治理与数据挖掘.北京:中国财政经济出版社, 2001.

8 李雄飞,李军.数据挖掘与知识发现.北京:高等教育出版社, 2005.