改进 k-means 算法在电信 CRM 客户分类中的应用①

左国才, 周荣华, 黎自强

(湖南软件职业学院 软件工程系, 湘潭 411100)

摘 要: 电信市场的竞争在于客户的竞争, 为了在激烈的竞争中保持优势, 必须将客户进行细分, 针对不同的客户, 研究出相应的营销策略. K-means 算法能对大型数据集进行高效分类, 但对 “噪声”敏感, 聚类结果不准确, 本文对该算法进行改进, 使其能够实现更加准确和全面的客户细分.

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

Improved K-Means Algorithm and Its CRM in Telecom Customer Segmentation Application

ZUO Guo-Cai, ZHOU Rong-Hua, LI Zi-Qiang

(Software Engineering, Hunan Vocational Institute of Software, Xiangtan 411100, China)

Abstract: The telecommunication market competition is the competition for customers, in the fierce competition to maintain advantage, must be customer segmentation, for different customers, work out the corresponding marketing strategy. K-means algorithm for large data sets for efficient classification, but the " noise" sensitive, clustering result is not accurate, the algorithm is improved, which can achieve a more accurate and comprehensive customer segmentation.

Key words: telecommunication CRM; customer segmentation; data mining; clustering analysis of K-means algorithm

1 引言

为了在激烈的市场竞争中保持竞争优势, 电信企业引入 CRM, 对客户的认知度有很大的提升, 对于以 “客户为中心”的管理理念有更深层次的认识, 形成了以客户为中心, 业务平台为支撑的企业经营管理系统. 企业的竞争是客户的竞争, 电信企业已经意识到有效的客户细分是企业参与客户竞争的核心竞争力, 企业的服务营销策略也离不开有效客户细分的支持.

数据挖掘利用各种分析工具在海量数据中发现模型和数据间关系[1], 数据挖掘中的聚类分析方法可以进行客户细分. 在聚类分析算法中, K-means 算法是最著名和最常用的划分法之一. K-means 算法能对大型数据集进行高效分类[2], 但却对 “噪声”敏感, 经过改进后的算法能够解决对 “噪声”敏感的问题, 实现更加准确和全面的客户群细分.

2 电信客户关系管理现状分析

客户关系管理(Customer Relationship Management, CRM)最初由 Gartner Group 提出来, CRM 是一种旨在改善企业与客户之间关系的新型管理机制, 它实施于企业的市场营销、销售、服务、技术支持等与客户有关的领域.

电信客户关系管理的特点:

(1) 更注重客户的个性化需求

随着社会的发展, 消费者崇尚一种个性化的消费理念, 定制商品已经成为吸引客户的重要手段, 也是以 “客户为中心”理念的体现. 所以, 在 CRM 中, 客户的细分越来越重要, 只有真正掌握用户的需求, 真正了解客户, 才能够保证客户的忠诚度.

(2) CRM 在电信企业中的独特性更加显著

电信企业要兼顾以小额消费为主的个人用户和以集团消费为主的集团客户. 两类客户消费差距很大,

① 基金项目:湖南省教育厅科学研究项目(11C0724, 11C0723)

收稿时间:2012-03-19;收到修改稿时间:2012-05-01

但带给企业的利益整体却是势均的，因此，在企业实施 CRM 过程中应该从两方面来考虑，针对不同的用户制订相应的营销策略.

(3) 客户定位是客户关系管理中的重要内容

通过客户定位，企业可以更好地识别客户群体，区别对待不同的客户，采取不同的客户保持策略，达到最优化配置客户资源的目的.

通过聚类分析可以实现上述目的，聚类分析是一种非常实用、方便、有效的划分方法[3]，可以发现客户群，并描述具有不同特征的顾客群，帮助管理者制定市场营销策略，达到改善客户关系的目的，并对将来的趋势和行为进行预测、支持企业决策.

3 电信数据分析

电信企业运营了很多年，积累了海量客户数据. 分析电信业务数据，得到客户信息、客户消费及购买行为等近四个月的历史数据，确定了数据源：客户基本信息、在用产品信息、帐单信息(通话信息、缴费信息、欠费信息)、服务使用信息.

分主题在电信企业数据仓库中选择需要的各类数据，并进行汇总，生成数据基础表.

用户及客户的基本信息：包括客户身份信息、联系方式、产品拥有情况，用户竣工时间、入网时长、服务开通情况、优惠套餐信息、客户服务信息(投诉、咨询、催缴情况)等.

价值信息：包括话音、宽带业务月租费、使用费、优惠费及增值业务、新业务、信息费、缴欠费信息等.

针对电信源数据的特点，改进聚类分析中的 K-means 算法，得到准确的聚类结果，实现更加全面的客户细分.

4 K-means算法在电信客户分类中的应用

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[4]，能对大型数据集进行高效分类，将数据分成几组，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.

4.1 K-means 算法描述

K-means 算法聚类的数量 k 是在算法运行前确定的，先从样本中随机选取 k 个聚类中心，再根据欧氏距离，把每个点分配到最接近其均值的聚类中，然后计算被分配到每个聚类的点的均值向量，并作为新的中心进行递归[5].

具体的算法：假定数据点 $D=\{X1\cdots\cdots Xn\}$，任务是找到 k 个聚类 $\{C1\cdots\cdots Ck\}$：

伪代码如下：

for k=1,…n,令 $R(k)$为从 D 中随机选取的点;

while 在聚类 Ck 中有变化发生

do 形成聚类;

for k=1, …., n do $Ck=\{X$ 属于 $D|D(Rk,x)\leqslant D(Rj,x)$ 对所有 j=1…..k, j≠k}; end;

计算新的聚类中心;

for k=1, …, n do $Rk=Ck$ 内点的均值向量;

end;end;

4.2 K-means 算法的改进思想

k-means 算法寻找质点的过程，是对某类簇中所有的样本点维度求平均值，即获得该类簇质点的维度. 当样本点中有 “噪声”(离群点)时，在计算类簇质点的过程中，受到噪声异常维度的干扰，造成所得质点与实际质点位置偏差过大，从而使类簇发生 “畸变”.

假设：类簇 C_1 中已经包含点 $A(1,1)$、$B(2,2)$、$C(1,2)$、$D(2,1)$，假设 $N(70,70)$为异常点，当它纳入类簇 C_1 时，计算质点 Cen ((1+2+1+2+70)/5, (1+2+2+1+70)/5) =Cen (15,15)，此时可能造成了类簇 C_1 质点的偏移，在下一轮迭代重新划分样本点的时候，将大量不属于类簇 C1 的样本点纳入，因此得到不准确的聚类结果.

为了解决 “噪声”敏感的问题，改进后的 K-means 算法提出了新的质点计算规则，而不是像 K-means 算法采用均值计算法. 在改进后的 K-means 算法中，每次迭代后的质点，都是从聚类的样本点中重新选取，选取的标准是当该样本点成为新的质点后，能提高类簇的聚类质量，使得类簇更紧凑. 该算法使用绝对误差标准来定义一个类簇的紧凑程度.

$E=\sum_1^k\sum_{p\in cj}|p-oj|$ (p 是空间中的样本点, Oj 是类簇 Cj 的质点).

如果某样本点成为质点后，绝对误差能小于原质点所造成的绝对误差，那么该样本点是可以取代原质点的，在一次迭代重计算类簇质点的时候，选择绝对误差最小的样本点成为新的质点.

假设：样本点 A ->$E1$=20　样本点 B ->$E2$=21

样本点 C ->$E3$=22　原质点 O ->$E4$=23,

则选择 A 作为类簇的新质点.

改进后的算法采用欧几里得距离来衡量样本点属于哪个类簇. 欧几里得距离定义: 欧几里得距离(Euclidean distance)也称欧式距离, 在 m 维空间中两个点之间的真实距离. 欧式距离的公式:

$$d(x_i,x_j)=\sqrt{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2}$$

令 $d_{ij}=d(x_i,x_j),D=(d_{ij}),\cdots$, 形成一个距离矩:

$$\begin{bmatrix} 0 & d_{12} & \cdots & d_{1p} \\ d_{21} & 0 & \cdots & d_{2p} \\ \vdots & \vdots & & \vdots \\ d_{n_1 1} & d_{n_1 2} & \cdots & 0 \end{bmatrix}$$, 其中 $d_{ij}=d_{ji}$

4.3 改进 k-means 算法的基本步骤

进后的 K-means 算法不采用簇中样本点的平均值作为质点, 而是选取簇中的样本点作为质点, 该样本点在簇中的位置是最中心的. 算法的具体步骤如下:

(1) 随机选择 k 个样本点作为初始的聚类质点集.

(2) 将剩余样本点分配给离它最近的质点代表的簇中, 重新计算簇集合.

(3) 重新计算各类簇的质点, 随机选择一个非质点的样本点, 计算用该样本点替代质点的紧凑效果.

(4) 如果该样本点使得类簇更紧凑, 则用该样本点替代原质点, 成为新的质点.

(5) 重复(2),(3),(4), 直到 k 个质点不再发生变化.

5 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 186200. 实验环境: PC 计算机, CPU 为 PⅣ2.2G, 内存为 2G, 操作系统为 WindowsXP, 编程环境: Eclipse3.5/ myEclipse8.5GA, Tomcat6.0.

5.1 聚类结果分析

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, 得到更加准确的聚类结果, 实现对客户进行更全面的分类, 为制定相应的营销政策提供决策依据.

数据集的聚类结果如图 1:

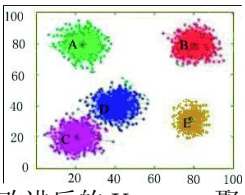

图 1 改进后的 K-means 聚类结果

聚类结果可以看出, 改进后的 K-means 算法的聚类结果更有效, 更准确, 将客户分为五类, 第一、第二类客户群体差异较大, 应保持与稳定客户. 第三、第四类客户群, 价值差异较少, 是企业最忠诚的客户群及企业价值的主要来源. 第五类客户群价值偏低, 具有开发价值和挖掘利润潜力的客户群.

5.2 性能分析

与典型的 K-means 算法一样, 改进后的 K-means 算法也是采用欧几里得距离, 来衡量某个样本点属于哪个类簇. 终止条件是当所有的类簇的质点都不再发生变化时, 则聚类结束.

该算法有效地改善了 K-means 算法对 “噪声”敏感的问题, 但是, 由于采用新的质点计算规则, 每次迭代后, 重新选取质点, 也使得算法的时间复杂度略有上升. K-medoids 算法对初始中心选择敏感, 大数据集聚类应用中性能低下, 该算法在进行中心轮换时需遍历所有非中心点, 执行代价高[6]. 改进后的 K-means 算法性能比较如图 2.

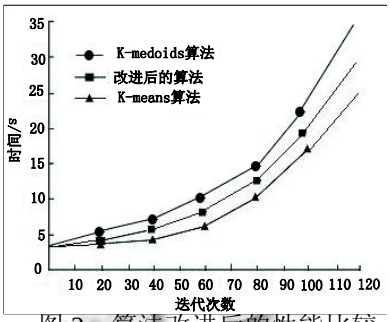

图 2 算法改进后的性能比较

从图 2 可知, 当迭代次数增大时, 改进后的 K-means 算法所需要的时间, 比改进前的算法所需要的时间略多, 算法的时间复杂度略有上升.

实验结果表明, 改进后的K-means算法采用新的质点计算规则, 有效地解决了 “噪声”敏感的问题, 能够全面、准确地实现客户细分, 为电信运营商解决客户分类的问题. 对于电信运营商进行套餐设计和实行深度营销等都有很强的现实意义.

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, 希望在今后的研究中, 通过更全面的数据分析, 来完善该研究方法.

参考文献

1 Barbara D. Using Self-Similarity to Cluster Large Data Sets.

(下转第 186 页)

表中的数字表示检测的准确率.

表 1 算法检测结果(%)

	S-Tools	Hide4GP	误报率	
100%	100	100	0	0
80%	96.91	94.81	1.21	1.34
50%	90.22	89.16	2.17	2.01
20%	85.25	83.12	5.34	6. 14
5%	75.41	70.68	8.23	8.35

表 2 对比检测率(%)

	S 检测	RS	RQP	SPA
100%	100	100	100	100
80%	96.91	96.67	96.72	96.85
50%	90.22	95.62	95.81	95.98
20%	85.25	86.57	87.54	87.74
5%	75.41	78.54	77.21	79.51

由表中的信息可以看出，在嵌入比率较大的情况下，本文提出的检测算法具有较高的效率.

4 结语

本文提出的一种利用相邻像素相关性的隐写分析算法，通过实验得到了预期的结果，从而验证了算法的正确性与可行性. 理论分析和实践表明，该方法只需统计横向和纵向像素的差值情况，统计量小、方法实现简单，在嵌入率比较高的情况下能够达到非常好的检测效果，嵌入量较低的情况下检测效率就不如其他几种方法. 但是该算法只对 LSB 隐写有效，一旦图像采用其他的隐写方法，检测算法就会失效. 因此下一步研究方向为根据自然图像固有的自然特性，找一个通用且有效的隐写分析方法.

参考文献

1 Li B, Fang YM, Huang JW. Steganalysis of Multiple-Base Notational System SteganograPhy. IEEE Signal Processing Letters. 2008, 15: 493–496.

2 Yang CF, Liu FL, Luo XY, Liu B. Steganalysis Frameworks of Embedding in Multiple Least-Signifieant Bits. IEEE Trans. on Information Forensies and Secunty. 2008, 3(4):662–672.

3 Westfeld A, Pfitzmann A. Attacks on steganographic systems. Proceedings of Information Hiding, Third International Workshop. Berlin: Springer-Verlag, 2000:61–67.

4 Fridrich J, Goljanm DR. Detecting LSB stegano graphy in color and grayscale images. IEEE Multmiedia, 2001, 8(4): 22–28.

5 Fridrich J, Du R, Meng L. Steganalysis of LSB encoding in color images. Proc. of IEEE International Conference on Multimedia and Expo. New York, 2000: 1279–1282.

6 Dumitreseu S, Wu XL, Wang Z. Deteetion of LSB SteganograPhy via Sample Pair Analysis. IEEE Trans. on Signal Proeessing, 2003,51(7): 1995–2007.

7 张涛,平西建.基于差分直方图实现 LSB 信息伪装的可靠检测.软件学报,2004,15(1):151 –158.

8 Zhang XP, Wang SZ, Zhang KW. Steganography with leasthistogram abnormality. Computer Network Security, Lecture Notes in Computer Science. Springer-Verlag, 2003: 395–406.

9 张新鹏,王朔中,张开文.数字密写和密写分析.北京:清华大学出版社,2005.

10 贾玉珍,靳冰,刘琮,等.BMP 文件结构的信息隐藏方法与实现.江西理工大学学报,2009,30(1):42–44.

11 周文锦,范明钰,王光卫.一种针对 BMP 格式图像的 LSB 数字隐藏方法.信息安全与通信保密,2005:253–255.

12 CBIR Image Database. University of Washington. [2012-03-22]. http://www.cs.washington.edu/ research/ imagedata-base/groundtruth/

13 USC-SIPI Image Database.[2012-03-22]. http://sipi.usc.edu/ services/database/Database.html

(上接第 155 页)

Data Mining and Knowledge Discovery, 2003, 7.

2 Huang ZX. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mi-ning and Knowledge Discovery, 1998,2: 283–304.

3 雷小锋,谢昆青,林帆,夏征义.一种基于 K-Means 局部最优性的高效聚类算法.软件学报,2008,7.

4 赵伟,张姝,李文辉.改进 K-means 的空间聚类算法.计算机应用研究.2008,7.

5 Tzortzis G, Likas A. The Global K-Means Clustering Algorithm. Proc. of the International Joint Conference on Neural Networks,2008.

6 夏宁霞,苏一丹,覃希.一种高效的 K-medoids 聚类算法.计算机应用研究,2010,12.