文章编号:1001 - 9081(2006)08 - 1996 - 02

数据挖掘在客户关系管理中的应用

郭 炜,何丕廉,王 中
(天津大学 计算机科学与技术学院, 天津 300072)
(guowei_phd@ yahoo. com. cn)

摘 要:利用标准化客户数据,确定了聚类相似度公式和评价指标,使用层次凝聚方法和 K - 平均算法实现了客户的自动聚类;并且在权衡算法效率和聚类精度的基础之上提出了改进的聚类距离公式和 K - 平均算法,达到了较好效果。

关键词:聚类; 相似度; 层次凝聚法; K - 平均算法
中图分类号: TP311. 13 **文献标识码**:A

Application of data mining in customer relation management

GUO Wei[1], HE Pi-lian[1], WANG Zhong[1]
(*College of Computer Science and Technology, Tianjin University, Tianjin* 300072, *China*)

Abstract: The similarity formula of clustering and assessment criteria were determined by using standardized customer data. The automatic clustering of customers was also realized by using the layer-agglomeration approach and K-average approach. On the basis of balancing algorithm efficiency and accuracy of clustering, improved formula of clustering distance and K average algorithm approach were presented. Satisfactory results were achieved.

Key words: clustering; similarity; Layer-agglomeration; K-average algorithm

0 引言

通过对积累的交易数据进行分析,可以按各种客户指标(如自然属性、收入贡献、交易额、价值度等)对客户分类,然后确定不同类型客户的行为模式,以便采取相应的营销措施,促使企业利润的最大化。因此,数据挖掘技术在客户关系管理(Custom Relationship Management,CRM)的客户分析中具有广阔的应用前景。

本文应用层次凝聚法与 K - 平均算法对证券业中的客户进行自动聚类,提出了一种对聚类的评价方法,并基于该评价方法对 K - 平均算法进行了改进,在效率和准确率上有一定的提高。

1 层次凝聚法与 K - 平均算法对客户自动聚类

1.1 数据的预处理

表 1 最终生成的 Customer Clustering 事实表

字段名	字段说明	类型	长度
ZJZH	资金账号	CHAR	12
ZJXM	资金姓名	CHAR	8
KHLB	客户类别	CHAR	2
ZJGM	资金规模	MONEY	NOT NULL
XCSR	息差收入	MONEY	NOT NULL
SXF	手续费	MONEY	NOT NULL
CZPL	操作频率	INT	NOT NULL

在证券公司的日常运营实践中,客户属性一般包括静态属性和动态属性两种。动态属性是指与客户交易行为相关的属性,是进行数据挖掘的主要对象。客户动态属性体现了客户对证券公司的价值和贡献度。动态属性主要包括:资金规模(ZJGM),手续费收入(SXFSR),息差收入(XCSR)和操作频率(CZPL)。其中,资金规模 = 资金余额 + 证券市值,该指标体现了客户从事证券交易的所投入的总体资金的规模。经过对客户数据仓库进行数据清洗、转换生成了客户聚类分析所需要的 Customer Clustering 事实表如表 1 所示。

将所得数据进行标准化,再利用相关性分析将客户聚类相似度公式构造为:

$$d(i,j) = \sqrt{\alpha(z_i - z_j)^2 + \beta(s_i - s_j)^2 + \lambda(x_i - x_j)^2 + \delta(p_i - p_j)^2} \quad (1)$$

在式(1)中,z_i 表示资金规模指标,α 表示该指标的影响因子,值为 0.2324;s_i 表示手续费收入指标,β 表示该指标的影响因子,值为 0.4948;x_i 表示息差收入指标,λ 表示该指标的影响因子,0.0052;p_i 表示操作频率指标,δ 表示该指标的影响因子,值为 0.2676。下面根据上述内容对客户数据进行聚类。

1.2 层次凝聚法与评价指标的提出

层次凝聚聚类方法是一种典型的无导师学习方法,也称为自底向上的方法。我们首先使用该方法进行客户聚类分析。具体算法如下:

设定:客户集合 $G = \{g_1, g_2, \cdots, g_i, \cdots, g_n\}$,其中,$g_i$ 为单个客户,以资金账号字段(ZJZH)标示。

1)将 G 中的每个客户看作是一个具有单个成员的类 $c_i = \{g_i\}$,这些类构成 G 的一个聚类 $C = \{c_1, \cdots, c_i, \cdots, c_n\}$;

2)根据加权的聚类相似度公式(式(1))计算 C 中每对类 (c_i, c_j) 之间的相似度 $d(c_i, c_j)$;

3)选取具有最大相似度的类对 arg max $sim(c_i, c_j)$,并将

收稿日期:2006 - 02 - 13;**修订日期**:2006 - 04 - 14
基金项目:天津市科技发展计划资助项目(04310941R);天津市应用基础研究计划资助项目(05YFJMJC11700)
作者简介:郭炜 (1978 -),男,天津人,博士研究生,主要研究方向:本体、数据挖掘; 何丕廉(1942 -),男,天津人,教授,博士生导师,主要研究方向:自然语言处理、Web 检索与挖掘、人工智能与计算机辅助教育; 王中(1968 -),男,天津人,博士,主要研究方向:数据挖掘、信息检索.

c_i 和 c_j 合并为一个新类 $c_k = c_i \cap c_j$,从而构成 G 的一个新聚类 $C = \{c_1, \cdots, c_{n-1}\}$;

4) 重复上述步骤,直至 C 中剩下最后一个类为止。

通过上述计算,建立了一个表示客户聚集关系的生成树。经过对聚类过程的进一步分析,取第 $n-3$ 层的四个类别作为公司的客户分类。聚类结果如表 2 所示。

表 2　层次凝聚方法客户聚类结果

	账户数	平均资金规模	平均手续费收入	平均息差收入	平均操作频率
特户	116	5962415.83	95896.30	2035.69	613.49
大户	293	163843.20	23082.41	79.56	225.63
中户	572	96253.29	10567.46	71.89	168.21
散户	20808	18023.82	653.25	6.09	13.37

通过将层次凝聚方法客户聚类结果与人工划分的训练集比较,表明经过聚类之后,客户分类发生如下几方面的变化:

1)在不同类别之间发生了客户类别迁移现象,例如有 9 个特户经过聚类后有 6 个户被划分到大户,3 个被划分到中户;大户中有 29 户被划分到散户和中户类别中去。其余的类别之间也存在此现象。经过数据统计,总共有 529 个帐户发生了类别迁移。

2)经过聚类后,新类别的平均资金规模、平均手续费收入、平均息差收入和平均操作频率指标在不同的客户类别有所变化,体现了以均值为代表的聚类中心的移动。

综合上述两个因素的变化,引入一个新的评价指标来评价客户聚类的质量:

$$Q = \frac{d_{nf}}{d_{of}} \tag{2}$$

在式(2)中,d_{nf} 表示经过聚类之后某个发生类别变化的账户与新类别的聚类均值中心点的距离;d_{of} 表示该账户与原有类别的聚类均值中心点的距离。距离的计算公式采用加权相似度公式计算。Q 表示某一个点(账户)到新旧类别均值中心的比值。Q 越小,表示该点的聚类效果越好。

取经过聚类之后所有发生类别变化账户到新旧类别均值中心的比值 $\overline{Q}$ 作为评价客户聚类效果的指标。在本文中,称该聚类评价指标为 $\overline{Q}$ 指标:

$$\overline{Q} = \frac{\sum_{i=1}^{n} Q}{n} \tag{3}$$

经过计算,使用层次凝聚方法进行聚类的 $\overline{Q}$ 值为 0.2152。在总计 529 个发生类别迁移的帐户中,$Q > 1$ 的账户为 56 个,占 10.59%;$Q < 1$ 的账户为 473 个,占 89.41%。

1.3　K-平均算法

层次凝聚方法经常会遇到合并点选择的困难[2~4],因为一旦一组对象被合并,下一步的处理将在新生成的簇上进行。已经合并的类别不可能撤销,聚类之间也不能交换对象。假如在某一步的簇之间合并出现异常,则可能会导致聚类质量的降低。另外,这种聚类方法不具备很好的可伸缩性,而且算法的复杂度较高。假定在开始的时候数据仓库中有 n 条记录,在结束时生成一个类别,那么在主循环中有 n 次迭代,而在第 i 次迭代中必须在 $n-i+1$ 个簇中找到最靠近的两个聚类,而且算法必须计算所有两两记录之间的距离,因此该算法的复杂度为 $O(n^2)$。

K-平均算法是一种基于质心的划分聚类方法。K-平均算法以 k 为参数,把 n 个对象分为 k 个簇,使簇内具有较高的相似度,而且簇间的相似度较低。相似度的计算根据一个簇中对象的平均值(被看作簇的重心)来进行。

K-平均算法的处理流程如下:首先,随机地选择 k 个对象,每个对象初始地代表了一个簇的平均值或中心。对剩余的每个对象,根据其与各个簇中心的距离将它赋给最近的簇。然后重新计算每个簇的平均值。这个过程不断重复,直到准则函数收敛。通常,采用平方误差准则,其定义如下:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i| \tag{4}$$

其中 E 是数据库中所有对象的平方误差的总和,p 是空间中的点,表示给定的数据对象,m_i 是簇 C_i 的平均值,p 和 m_i 都是多维的。这个准则试图使生成的结果簇尽可能地紧凑和独立。

使用 K-平均方法对 Customer Clustering 事实表进行聚类分析,根据证券公司的实际分类需求,同时为了和训练集的人工分类结果进行有效的比较,取 $k = 4$,目标是聚成四个类。聚类程序运行结果如图 1 所示。

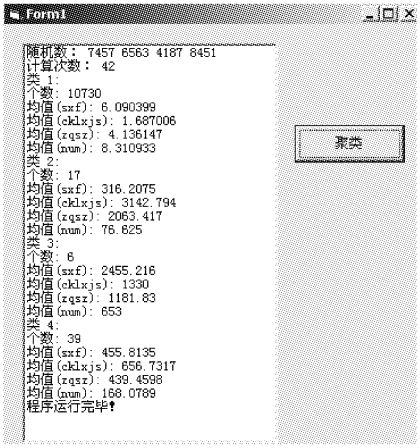

图 1　K-平均算法运行结果

同样,采用经过聚类之后发生类别变化账户到新旧类别均值中心的比值 $\overline{Q}$ 作为评价客户聚类效果的指标。

经过计算,使用 K-平均方法进行聚类的 $\overline{Q}$ 值为 0.3952。总计有 861 个帐户发生类别迁移,$Q > 1$ 的账户为 141 个,占 16.37%;$Q < 1$ 的账户为 720 个,占 83.63%。从 $\overline{Q}$ 指标上分析,使用 K-平均方法较之层次凝聚方法的聚类效果要差一些。经过对各项指标的分析,各个簇之间大小差别相对较大是造成这种结果的重要原因。

2　改进的 K-平均算法

通过对 K-平均方法聚类结果中帐户类别迁移错误的点(条件 $Q > 1$)进行统计分析发现,对数据库中两点 $i(x_{i1}, x_{i2}, \cdots, x_{ip})$ 和 $j(x_{j1}, x_{j2}, \cdots, x_{jp})$,其中 p 为点的维数,传统的聚类算法计算这两点的距离时所采用的公式一般为欧几里得距离:

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{ip} - x_{jp}|^2} \tag{5}$$

假设数据库中存在这样两点:$i(1000,1000,1,1)$ 和 $j(1,1,1000,1000)$。用上面的公式计算这两点的距离,得出的结论是:它们不属于同一个类,因为它们之间的距离很大。而在

(下转第 2000 页)

不精确,不可避免的仍然会有一些嵌入的隐藏信息无法读取,而且作为防伪识别的数字水印信息也不必需要如此大的数据量,于是就需要使用差错控制编码技术[6,7]。差错控制编码属信道编码,要求在满足有效性前提下,尽可能提高数字通信的可靠性。即通过降低有效信息量来换取可靠性提高。

在本文的应用中使用 Reed Solomon erasure code(RS)。RS code 是使用最广泛的一种差错控制编码。它是一类有很强纠错能力的多进制 BCH 码,也是一类典型的几何码。它不仅可以纠正突发错误,还可以纠正随机错误,特别适用于纠正信号的突发错误。

RS code 是一类线性码(利用线性代数性质的编码)的实现,并且具有 MDS 编码(Maximal Distance Separable,编码后的消息中任意长度等于原始消息的子消息都可以解码得到原始消息的编码称为 MDS 编码)性质,即 $r = m$。Erasure code 的线性码实现关键是定义合适的生成矩阵(与源数据进行某种操作实现编码的矩阵)。生成矩阵 $\boldsymbol{G}$ 的性质直接影响到 RS code 的编码和译码效率。为了满足编码的 MDS 性质,生成矩阵 $\boldsymbol{G}$ 必须具有下面的两个性质:1)生成矩阵 $\boldsymbol{G}$ 为 $m \times n$ 矩阵;2)任意 m 列均为线性无关,即 $\boldsymbol{G}$ 的任意 $m \times m$ 子矩阵可逆。线性码可以表示为 $y = x\boldsymbol{G}$,其中 $x = (x_1, x_2, \cdots, x_m)$ 为源数据,编码之后得到的数据为 $y = (y_1, y_2, \cdots, y_n)$。设 y' 由 y 中任意 m 个数据项组成,$\boldsymbol{G}'$ 为 $\boldsymbol{G}$ 中相应的 m 列组成的子矩阵,则有 $y' = x\boldsymbol{G}' \Rightarrow x = y'\boldsymbol{G}'^{-1}$,完成解码过程得到源数据 x。对于此基于二维图形码的数字水印的应用,只需要嵌入的 n 位水印信息,编码成 m 位嵌入信息,使用上文中的水印嵌入算法嵌入到二维条形码中;解码时只要能够正确读取其中的任意 n 位信息,即可解码得到原水印信息。

5 结语

随着信息化进程的快速发展,二维图形码在各行各业中有着越来越广泛的应用。本文所提出的基于二维图形码的信息隐藏技术,可以用作在二维图形码中嵌入一个不可见的数字水印,此技术可以用于鉴定二维图形码的真伪等方面。提高了二维图形码的安全性和适用性。

参考文献:

[1] INGEMAR JC, MATTHEW LM, JEFFREY AB. Digital watermarking [M]. Morgan Kaufmann Publishers, 2002.

[2] 朱卫东,张树艳. 二维条码技术与应用[J]. 北方交通大学学报, 1997,(3):371 -374.

[3] 张基宏,肖薇薇,纪震. 基于二维条码 PDF417 的数字图像水印算法[J]. 深圳大学学报(理工版),2002,19(1):3 -8.

[4] 石睿. 基于二维条码的数字水印系统开发[J]. 武汉船舶职业技术学院学报,2005,(1):15 -19.

[5] 周琳娜,杨义先,郭云彪,等. 基于二值图像的信息隐藏研究综述[J]. 中山大学学报(自然科学版),2004,43(S2):71 -75.

[6] PETER S. Error Control Coding, From Theory to Practice[M]. 北京:清华大学出版社,2004.

[7] 王新梅,肖国镇. 纠错码——原理与方法[M]. 西安:西安电子科技大学出版社,2001.

(上接第 1997 页)

实际的业务分析中,这两点各自所代表的客户对券商赢利的贡献都很大,所以这两个客户应该属于同一类的概率较大。据此,本文提出一种计算这种类型数据的距离的方法,对于点 $i(x_{i1}, x_{i2}, \cdots, x_{ip})$ 和 $j(x_{j1}, x_{j2}, \cdots, x_{jp})$,其距离公式为:

$$d(i,j) = \frac{\sqrt{\alpha_1 x_{i1}^2 + \alpha_2 x_{i2}^2 + \cdots + \alpha_p x_{ip}^2}}{\sqrt{\alpha_1 x_{j1}^2 + \alpha_2 x_{j2}^2 + \cdots + \alpha_p x_{jp}^2}} \quad (6)$$

在式(6)中,$\alpha_1, \alpha_2, \cdots, \alpha_p$ 为各项指标的权重,满足:

$$\alpha_1 + \alpha_2 + \cdots + \alpha_n = 1 \quad (7)$$

点 i 与点 j 距离的大小取决于 $d(i,j)$ 与 1 相差的大小:$|d(i,j) - 1|$ 越大,则 i 与 j 距离越远;反之,则 i 与 j 距离越近。

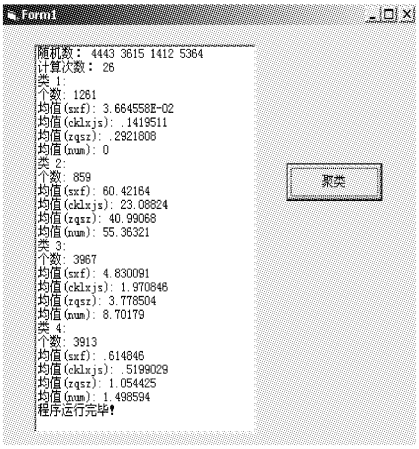

图 2 改进的 K - 平均算法运行结果

在输入数据与其他条件不变的情况之下,采用上述计算距离方法的 K - 平均算法的结果如图 2 所示。

改进的 K - 平均算法的 $\overline{Q}$ 指标值为 0.2311,基本接近了层次凝聚方法的 $\overline{Q}$ 指标。在总计 553 个发生类别迁移的帐户中,$Q > 1$ 的账户为 65 个,占 11.75%;$Q < 1$ 的账户为 488 个,占 88.25%。从聚类结果看出,这个聚类结果相对于改进前有很大的改善。因此,改进的 K - 平均算法既提高了运行效率又提升了聚类的精度。

3 结语

本文在客户自动聚类的实现方面,建立了资金规模、手续费收入、息差收入和操作频率等四项指标体系,生成了数据仓库并对数据进行了标准化处理,确定了各项指标的权重,从而定义了加权的聚类相似度公式和评价指标;在算法实现方面,使用层次凝聚方法和 K - 平均算法实现了客户的自动聚类,并在权衡算法效率和聚类精度的基础之上提出了改进的聚类距离公式和 K - 平均算法。

参考文献:

[1] FAN JN, LI DY. An overview of datamining and knowledge discovery [J]. Journal of Computer Science and Technology, 1998, 13(4):348 -368.

[2] 张孝令. 贝叶斯动态模型及其预测[M]. 济南:山东科学技术出版社,1992.

[3] 姚妙新. 非线性理论的数学基础[M]. 天津:天津大学出版社,2005.

[4] 张尧庭. 人工智能中的概率统计方法[M]. 北京:科学出版社,1998.