SQL Server 2005 数据挖掘技术在证券客户忠诚度的应用

赵裕啸,倪志伟,王园园,伍章俊
(合肥工业大学 管理学院,安徽 合肥 230009;
合肥工业大学 过程优化与智能决策教育部重点实验室,安徽 合肥 230009)

摘 要:文中主要研究了我国证券业客户忠诚度分类和表现形式,提出了一种证券业客户忠诚度评估的有效方法。依据RFM客户评价方法,结合数据挖掘的一般流程将SQL Server 2005中的数据挖掘技术应用于证券业客户忠诚度模型系统中,并结合某证券公司客户交易数据,对其客户忠诚度进行了准确合理的分类,对其不同忠诚度类型的客户提出相应个性化营销建议,最后通过使用DMX语言在客户端运用数据挖掘产生的分类规则对其客户进行了准确预测。

关键词:RFM客户评价;数据挖掘;客户忠诚度;DMX语言
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2010)02-0229-04

Application of Data Mining Technology of SQL Server 2005 in Customer Loyalty Model in Securities Industry

ZHAO Yu-xiao, NI Zhi-wei, WANG Yuan-yuan, WU Zhang-jun
(School of Management, Hefei University of Technology, Hefei 230009, China;
Ministry of Education Key Lab. of Process Optimization and Intelligent Decision-making,
Hefei Univ. of Tech., Hefei 230009, China)

Abstract: Studied on classification and expression forms of customer fidelity in securities industry and proposed an effective method to evaluate customer fidelity in securities industry. Based on RFM customer evaluation method, integrated the normal process of data-mining and applied data-mining technology of SQL Server 2005 to the customer fidelity model in security, combining customers' transaction data in a security company, made an accurate and rational classification, and proposed corresponding personalized marketing method for customers with different fidelity. At last it used DMX language in client end to exert classifying rules and predict the class of customers.

Key words: RFM customer evaluation; data-mining; customer loyalty; DMX language

0 引 言

以客户为中心的现代证券业经营模式要求券商对客户进行细分,针对不同类别的客户开展个性化、差异化的服务,不断提高客户的忠诚度,实现客户价值的最大化。同时近几年商务智能技术飞速发展,出现了许多像Business Objects, SQL Server 2005等优秀的商务智能软件,运用这些软件对业务数据进行分析,真实地反映出客户的各种行为特征和属性,为券商的经营决策和操作行为提供客观依据。

收稿日期:2009-06-20;**修回日期**:2009-09-06
基金项目:国家高技术研究发展计划(863)(2007AA04Z116);国家自然科学基金(70871033)
作者简介:赵裕啸(1984-),男,硕士研究生,研究方向为数据挖掘;倪志伟,教授,博士生导师,研究方向为人工智能、机器学习。

1 相关概念

数据挖掘扩展插件(DMX)查询语言[1]是一种面向SQL Server 2005数据挖掘模型的查询语言,由DDL数据定义语句和DML数据操纵语句组成。其中DDL语句用于创建和维护数据挖掘结构和模型,而DML语句则用于处理和浏览挖掘模型以及进行模型预测。

数据挖掘又称为数据库中知识发现(Knowledge Discovery from Database, KDD),即从大量的、有噪声的、模糊的、随机的实际数据中发现规律性的、人们事先未知的,但又是潜在有用的并且最终可理解的信息和知识的非平凡过程[2]。

聚类算法是一种无指导的分类方法。研究人员已经提出了不少数据聚类算法,比较著名的有k-Means[3]、k-Center[4]、DBSCAN[5]和OPTICS[6]等。

其中 k－Means 聚类算法由 Mac Queen[7]提出，具有算法结构简单、收敛速度快的优点，十分适用于大规模的数据分析。但是该聚类算法具有两大突出缺点：一是必须事先已知或者给定簇的个数；二是该算法聚类结果受初始聚类中心影响。

决策树应用于数据挖掘分成两个阶段：一是模型训练阶段，通过对训练集训练而获得树的模式；其二是使用模型阶段，实际上就是用获取的模型对未知的数据进行分析，比如分类或预测。

2 基于 RFM 计算客户忠诚度指标

最近购买时间(recency)、购买频率(frequency)和总购买金额(monetary value)综合分析(简称 RFM 分析)是一种重要的评价客户忠诚度的方法。Goodman 的研究表明采用 RFM 分析方法可以使企业更多地关注高忠诚度客户，从而利用有限的资源获得最佳的效益[8]。RFM 的分析基础是 3 个重要的客户行为指标，Bult 和 Wansbeek 给出了关于这三个指标的定义：①最近购买时间 R(recency)，即从上次购买到当前的时间间隔，该值越小意味着客户再次购买的可能性越大；②购买频率 F(frequency)，即客户在某一时间段内总的购买次数，购买频率越高表示客户越忠诚；③总购买金额 M(monetary value)，即某一时间段内客户的购买行为也可以包含所有的购买行为[9]。

结合 RFM 综合分析法，总结得出中国证券业客户交易行为的客户忠诚度指标包括以下内容：客户购买频率，客户交易时间，客户交易金额。客户交易金额表明了客户已经建立的忠诚度；客户交易时间在更大程度上可以预测未来客户将建立的忠诚度；客户购买频率既反映了已经建立的客户忠诚度级别，又可以预测今后客户的股票交易频率。

3 SQL Server 2005 的数据挖掘技术在证券业客户忠诚度研究中的应用过程

3.1 应用目标

运用 SQL Server 2005 的数据挖掘技术对根据某证券公司数据库系统中的 2007 年客户历史交易数据(人民币部分)处理出的客户忠诚度指标进行聚类分析，通过对客户行为分析得出各类客户行为特征及忠诚度水平。然后运用决策树方法得出各类客户忠诚度特征与忠诚度指标之间的规则，运用规则根据新客户短期忠诚度指标对该客户将来忠诚度做出预测。公司可以根据客户特征采取不同的服务策略提高客户的忠诚度水平。

3.2 客户忠诚度指标数据采集与处理

经过数据采集、数据清理、数据预处理的步骤得到了便于挖掘分析的数据仓库，内容主要与客户忠诚度指标体系一致。本案例在进行聚类时使用的指标如表 1 所示。

表 1 聚类指标

客户购买频率	
Lastmonth－frequency	上个月交易股票次数
Lasthalfyear－frequency	前六个月交易股票次数
Lastyear－frequency	前一年交易股票次数客户交易时间
客户交易时间	
Lastbuy－time	距离前一次买卖股票的时间
Avgtrade－intervaldays	平均买卖股票的间隔天数客户交易金额
客户交易金额	
Lastmonth－fund	上个月交易金额
Lasthalfyear－fund	前六个月交易金额
Lastyear－fund	前一年交易金额
Lastyearavg－fund	前一年每次平均交易金额
Yjgx	客户佣金贡献

将参加聚类的 9 个属性进行数据变换，对每条记录的每个属性进行标准差标准化变换。

$$X_{ij} = \frac{X_{ij} - X_j}{S_i}$$

$i = 1,2,\cdots,n,\ j = 1,2,\cdots,11$

其中：$X_j = 1/n\sum X_{ij}$ $S_i = [1/(n-1)\sum(X_{ij} - X_i)^2]^{\frac{1}{2}}$

经过变换后各属性的均值为 0，标准差均为 1。

3.3 数据挖掘过程

在本案例中数据挖掘过程分为两步，分别采用 K－means 聚类算法和决策树算法。

挖掘步骤如下：

Step1 确定聚类个数，对数据进行聚类并输出聚类结果；

Step2 对聚类情况进行分析并保存聚类结果；

Step3 为客户忠诚度预测准备输入项；

Step4 运用决策树算法生成分类规则；

Step5 根据分类规则对新客户进行分类预测。

3.3.1 k－Means 聚类方法的运用

k－Means 算法描述：

输入：聚类个数 k，以及包含 n 个数据对象的数据库。

输出：满足方差最小标准的 k 个聚类。

Step1 assign initial value for means;// 任意选择 k 个对象作为初始的簇中心

Step2 REPEAT

Step3 FOR $j = 1$ to n DO assign each X_j to the cluster which has the closest mean;// 根据簇中对象的平均值，将每个对象赋给最类似的簇

Step4 FOR $i = 1$ to k DO $\overline{x_i} = \sum_{X \in C_i} X / | C_i |$;//更新簇的平均值,即计算每个对象簇中对象的平均值

Step5 Computer E;//计算准则函数 E

Step6 UNTIL E 不再明显发生变化

在 Visual Studio 中新建一个 Analysis Services 项目,创建聚类分析挖掘模型,设置 Microsoft 聚类分析算法参数,CLUSTERING_METHOD 参数选择 k - Means 方法,将聚类个数 MODELLING_CARDINALITY 设置不同数值对数据集进行多次聚类,然后分析聚类结果特征,发现当设置聚类个数 $k = 5$ 时,各个聚类类别特征非常明显,类别之间的区别非常大,因此确定聚类个数 $k = 5$。

经过挖掘聚类后,数据库中的客户分为 5 类。为方便分析,将挖掘模型生成的每类具体属性值统计后制成表,如表 2 所示。

表 2　聚类后各类属性均值表

	Total (均值)	Cluster1 (均值)	Cluster2 (均值)	Cluster3 (均值)	Cluster4 (均值)	Cluster5 (均值)
Lastmonth - frequency	2.79641	0	0	7.254265	2.708859	16.22027
Lasthalfyear - frequency	31.0742	19.41007	2.360039	67.96627	15.90483	164.1921
Lastyear - frequency	64.2791	50.28724	6.523770	131.1143	26.31933	336.9045
Lastmonth - fund	54208.5	0	0	90372.38	20846.30	626481.8
Lasthalfyear - fund	740814.125	332316.002	16726.5516	967504.170	125573.779	8290389.935
Lastyear - fund	1592616.79	908216.266	51516.7949	1962157.2946	216922.6407	17336640.51
Lastyearavg - fund	18842.5	21880.38	9521.964	17415.22	9009.602	94739.64
Lastbuy - time	59.5084	69.17391	136.5759	3.607799	5.297595	15.21185
Avgtrade - intervaldays	50.0217	14.78570	134.8102	4.254098	39.01300	6.335296
Yjgx	14065.6	8781.888	509.1584	12034.60	1413.309	170818.6
人数	354826	99576	97588	69936	69418	18308

由实验结果解读数据获取各类客户忠诚度特征如下:

(1)第一类客户人数占总人数的 28.07%,该类客户虽然交易频率比较低,但是每次交易金额都较大,给公司的佣金贡献也比较可观,但是这类客户在下半年尤其是最后一个月交易频率下降很多,因此这类客户忠诚度一般,随时有流失的风险。因此,可以及时向这类客户提供各种股市信息,提高他们对股市的关注程度,从而提高其交易频率,客户忠诚度也就随着提高。

(2)第二类客户人数占总人数的 27.50%,该类客户无论是交易频率还是交易金额都很低,给公司的佣金贡献最少,该类客户应该是初入股市并且对股市不关注,仅是抱着试试的态度跟风。因此,该类客户忠诚度最差,有一部分客户很可能已经流失,这类客户比重比较大,应该是和 2007 年股市行情大好,许多人盲目跟风入市有关。

(3)第三类客户人数占总人数的 19.71%,该类客户虽然每次交易金额不是很大,但是交易频率非常高,客户异常活跃,给公司的佣金贡献也很高,应该都是公司的老客户,属于忠诚度比较高的客户。公司应该把服务重点偏向于此类客户,防止此类客户的流失。

(4)第四类客户人数占总人数的 19.56%,该类客户虽然每次交易金额较少,给公司的佣金贡献也不大,应该是忠诚度较低的客户,但是其交易频率呈现增加的趋势,具有发展潜力。该类客户应该也是初入股市,投资比较谨慎,但是对股市比较关注,如果公司对这类客户多加扶持,该类部分客户就会转变为第三类客户。

(5)第五类客户人数占总人数的 5.16%,该类客户各项指标值综合为最好,交易频率非常高,每次交易金额很大,是公司忠诚度最高的客户,公司的大部分利益来自于此类客户,因此应该对此类客户提供 VIP 服务,继续维持此类客户的忠诚度,杜绝此类客户的流失。

3.3.2　运用决策树方法获得分类规则

Microsoft 决策树算法是 Microsoft 研究院开发的混合型的决策树算法,是由 Microsoft SQL Server 2005 Analysis Services(SSRS)提供的分类和回归算法[10],用于对离散和连续属性进行预测性建模。对于离散属性,该算法根据数据集中输入列之间的关系进行预测。它使用这些列的值或状态预测指定的可预测列的状态,具体的说,该算法标识与可预测列相关的输入列;对于连续属性,该算法使用线性回归确定决策树的拆分位置。如果有多个列设置为可预测列,或输入数据包含设置为可预测的嵌套表,则该算法将为每个预测列分别生成一个决策树。

生成决策树之后就可以提取决策树表示的知识,并以 if - then 形式表示。对从根到树叶的每条路径创建一个规则,沿着给定路径上的每个属性—值对形成规则前件(if 部分)的一个合取项,叶结点包含预测,形成规则后件(then 部分)。If - then 规则易于理解,特别是当给定的树很大时[11]。

在 SQL Server Business Intelligence Development Studio 建立决策树挖掘模型,将聚类结果作为预测列,各忠诚度指标变量作为输入列,设置算法参数,生成决策树,然后依据生成的决策树得到分类规则。由决策树得到客户类别依赖关系网络如表 3 所示。

由于得到的分类规则比较多,在此就不一一列出,仅给出规则的一般表达式:

表 3 客户类别属性依赖关系表

类别标识属性	决策属性	依赖关系强弱次序
客户类别	距离前一次买卖股票的时间(Lastbuy_time)	1
	平均买卖股票的间隔天数(Avgtrade_intervaldays)	2
	上个月交易股票次数(Lastmonth_frequency)	3
	前一年每次平均交易金额(Lastyearavg_fund)	4

If Lastbuy_time = "$N_1 \sim N_2$" THEN 该客户∈某忠诚度客户类别;

If Lastbuy_time = "$N_3 \sim N_4$" AND Avgtrade_intervaldays = "$N_5 \sim N_6$" THEN 该客户∈某忠诚度客户类别;

If Lastbuy_time = "$N_7 \sim N_8$" AND Avgtrade_intervaldays = "$N_9 \sim N_{10}$" AND Lastmonth_frequency = "$N_{11} \sim N_{12}$" THEN 该客户∈某忠诚度客户类别;

If Lastbuy_time = "$N_{13} \sim N_{14}$" AND Avgtrade_intervaldays = "$N_{15} \sim N_{16}$" AND Lastmonth_frequency = "$N_{17} \sim N_{18}$" AND Lastyearavg_fund = "$N_{19} \sim N_{20}$" THEN 该客户∈某忠诚度客户类别。

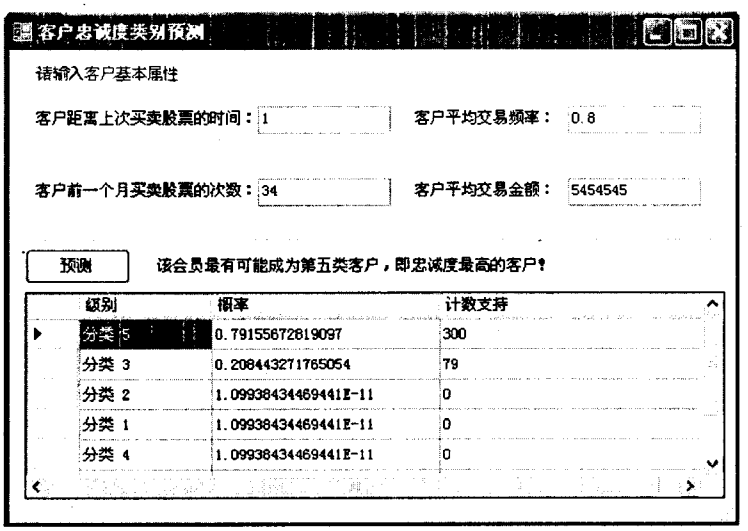

图 1 客户忠诚度类别预测结果图

3.4 客户端运用分类规则进行新客户预测数据挖掘

在 Microsoft SQL Server 2005 Analysis Services (SSRS)中,数据挖掘也有自己独特的语言,即:数据挖掘扩展插件 DMX 语言。用 DMX 语言进行数据挖掘的优势在于:数据挖掘功能可以全部用语句来实现,可以将这些语句嵌入其它的 MIS 或 ERP 系统中,实现有机整合。

基于 Microsoft 决策树算法的挖掘模型中最常用的预测函数是 Predict()和 PredictHistogram( ),Predict( )函数返回预测结果,PredictHistogram( )函数返回预测列的所有可能状态,以及每个状态的支持事例数和概率。使用 SELECT Predict( ), PredictHistogram( ) FROM < model > NATURAL PREDICTION JOIN (MDX)格式实现查询输入客户最有可能的所属类别和客户类别的所有状态,以及每个类别的支持事例和概率。通过调用决策树模型 CustomerPredict 预测客户的忠诚度类别。客户忠诚度预测效果如图 1 所示。

4 结束语

文中从客户关系管理角度出发,依据 RFM 客户评价方法,结合 SQL Server 2005 中的数据挖掘技术,建立了一个关于证券业客户忠诚度分析与预测的模型,以及如何使用 MDX 语言将该数据挖掘模型嵌入到 MIS 系统进行简单的应用。通过对客户忠诚度的分析和预测,采用不同的营销策略,从而使客户价值最大化。

参考文献:

[1] 郑宇军,杜家兴. SQL Server 2005 + Visual C# 2005 专业开发精解[M]. 北京:清华大学出版社,2007:366-373.

[2] 倪志伟,李峰刚,毛雪岷. 智能管理技术与方法[M]. 北京:科学出版社,2007:180-184.

[3] Grabmeier J, Rudolph A. Techniques of Cluster Algorithms in Data Mining[J]. Data Mining and Knowledge Discovery, 2002, 6(4):303-304.

[4] Han Jiawei, Kamber M. Data Mining: Concepts and Techniques [M]. San Francisco: Morgan Kaufmann Publishers, 2000:200-245.

[5] Ester M, Kriegel H P, Sander J, et al. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases [C]//Proc. of 1996 Intl. Conf. on Knowledge Discovery and Data Mining. Portland, OR: [s. n.], 1996: 226-231.

[6] Ankerst M, Breuning M, Kriegel H P, et al. Optics: Ordering Points to Identify the Clustering Structure[C]//Proc. of 1999 ACM-SIGMOD Intl. Conf. on Management of Data. Philadelphia, PA: [s. n.], 1999.

[7] Mac Queen J. Some methods for classification and analysis of multivariate observations [C]//Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: [s. n.], 1967:281-297.

[8] Goodman J. Leveraging the customer database to your competitive advantage[J]. Direct Marketing, 1992, 55(8):26-27.

[9] Bult J R, Wansbeek T J. Optimal selection for direct mail[J]. Marketing Science, 1995, 14(4):378-394.

[10] 王 欣,徐腾飞,唐连章,等. SQL Server 2005 数据挖掘实例分析[M]. 北京:中国水利水电出版社,2008:158-159.

[11] 杨善林,倪志伟. 机器学习与智能决策支持系统[M]. 北京:科学出版社,2004:65-68.