基于朴素贝叶斯算法的证券业客户价值细分研究

王 园，李少峰，王永梅，欧冰臻，王秋明，林巧明

（集美大学 工商管理学院，福建 厦门 361021）

摘要：客户价值是客户关系管理的基本依据。在客户细分的基础上运用朴素贝叶斯模型对证券业客户交易数据和客户基本信息进行数据挖掘分析，并利用 Weka 数据挖掘工具中的 Naïve Bayes simple 模型对客户交易数据进行训练、测试和验证，构建客户价值分析模型，同时对模型进行测试和评估，验证了客户价值分析模型的准确率。该模型能将现有的客户进行细分和价值分析，识别出具有不同特征、不同价值的客户群，以便针对不同价值等级的客户提供个性化服务策略。

关键词：证券业；客户价值；数据挖掘；朴素贝叶斯模型

中图分类号：F723 **文献标志码**：A **文章编号**：1671－1807(2013)05－0047－04

随着国际互联网的高速发展和电子商务的广泛应用，在激烈的证券企业行业竞争当中，谁能快速把握住客户的需求，谁能吸引新客户、保持老客户，谁能从客户数据中挖掘出有价值的信息，谁就能获得最大的竞争优势。以客户为中心成为证券公司打造稳定的企业根基、维系长久的客户关系的首要条件。大量的客户数据存储在联机数据库中，证券业关心的是如何从巨大的信息海洋中找到合适的数据分析方法，获取有用的知识模式来帮助证券公司做出正确的决策。

在金融证券行业中，对客户相关数据的深层次挖掘工作在近年来越来越被重视。数据挖掘技术从一开始就是面向应用的，它通过对大量数据进行统计、分析、综合与推理，来指导实际问题的求解，并对未来的发展趋势进行预测。它被广泛应用于金融、电信、医疗、保险等行业。在证券行业中，由于市场竞争日趋激烈，采用数据挖掘技术，为证券企业决策提供技术支持变得越来越重要。许多学者针对这一问题进行了深入研究。合肥工业大学的梁敏君在其学位论文《分形聚类分析在证券客户细分中的应用研究》中，研究了客户细分理论和方法，并提出基于网格和分形维数的聚类算法（GFDC），实验验证结果表明，该算法具有可行性并能取得较好的结果[1]。西安电子科技大学的李君锋硕士在其学位论文《数据挖掘在证券业 CRM 中的应用研究》中，运用数据挖掘的聚类分析，以某证券公司营业部为例研究数据挖掘在客户细分中的应用[2]。合肥工业大学管理学院的赵裕啸、倪志伟等其论文《SQL Server 2005 数据挖掘技术在证券客户忠诚度的应用》中，研究了我国证券客户忠诚度分类和表现形式，提出一种证券客户忠诚度评估的有效方法。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[3]。大连交通大学管理学院信息管理教研室的李会彬，莫东艳在其论文《基于关联规则和模式发现的证券客户行为模式挖掘》中，运用关联规则和模式研究了证券客户行为模式挖掘[4]。兴业证券股份有限公司的刘斌、邱华勇在其论文《证券公司客户综合分析系统的设计与实现》中，运用 K-means 聚类算法构建了客户偏好细分模型；利用决策树以及 Logistic 回归结合构建了客户流失预警模型[5]。浙江大学管理学院谢芳在其论文《数据挖掘在证券客户流失管理中的应用》中，运用客户流失分析系统预测客户流失情况，对潜在客户进行预警，为营业部提前进行客户挽留提供帮助[6]。

1 朴素贝叶斯分类模型

收稿日期：2013－03－19

基金项目：集美大学国家级大学生创新创业训练计划项目（Z81234）

作者简介：王园（1980—），女，山东无棣人，集美大学工商管理学院，讲师，厦门大学在读博士生，研究方向：决策理论与技术；李少峰（1991—），男，福建泉州人，集美大学工商管理学院本科生，研究方向：数据挖掘，企业管理；王永梅（1990—），男，福建三明人，集美大学工商管理学院本科生，研究方向：数据挖掘，计算机编程；欧冰臻（1990—），女，福建南平人，集美大学工商管理学院本科生，研究方向：数据挖掘；王秋明（1990—），女，福建泉州人，集美大学工商管理学院本科生，研究方向：数据挖掘，计算机编程；林巧明（1989—），女，福建泉州人，集美大学工商管理学院本科生，研究方向：数据挖掘。

贝叶斯方法提供了推理的一种概率手段。它假定待考察的变量遵循某种概率分布,且可根据这些概率及已观察到的数据进行推理,从而作出最优的决策。即通过给定的训练样本集预测未知样本的类别,它的预测依据就是取验后概率。贝叶斯方法不仅能够计算显式的假设概率,还能为理解多数其他方法提供一种有效的手段。

朴素贝叶斯分类模型的算法基于贝叶斯定理。贝叶斯定理是通过给定的训练样本集预测未知样本的类别,它的预测依据就是取后验概率。贝叶斯分类模型是一种典型的基于统计方法的分类模型。

朴素贝叶斯分类算法将训练实例集分解成属性向量A和决策类别变量H,假定属性向量的各分量相对于决策变量是相对独立的,也就是说各个分量独立地作用于决策变量。尽管这一假设在一定程度上限制了朴素贝叶斯模型的适用范围,但是在实际的应用中,大大降低了贝叶斯网络的构建复杂性。朴素贝叶斯分类模型已经成功地应用到聚类,分类等数据挖掘任务中。

朴素贝叶斯分类模型的计算过程是:

1) 给定一个没有标号的数据样本 X,用 n 维特征向量 $X = X\{x_1, x_2, \cdots, x_n\}$ 表示,分别描述 x 在 n 个属性 $\{a_1, a_2, \cdots, a_n\}$ 上的属性值。假定有 M 个类 $\{C_1, C_2, \cdots, C_m\}$,那么将样本 x 分配给 C_m 的条件是

$$P(c_i/x) > P(c_j/x) \quad (1 \leqslant j \leqslant m,\ j \neq i) \quad (1)$$

即假定样本为类 c_i 的概率大于假定为其他类的概率。

根据贝叶斯定理:$P(c_i/x) = \dfrac{P(x/c_i) \cdot P(c_i)}{P(x)}$,其中 $P(x)$ 指的是任意一个对象符合样本 X 的概率。对于所有类来说,它是一个常数,由公式可以看出,只要使 $P(x/c_i) \cdot P(c_i)$ 最大即可。$P(c_i)$ 为任意一个对象为 c_i 的概率,可以用 $P(c_i) = s_i/s$ 来计算,其中 s_i 是类 c_i 中训练样本数,S 是训练样本总数。

2) 给定样本的类标号,假定各属性值相互条件独立,这样 $P(x/c_i)$ 可以用计算公式:

$$P(x/c_i) = \prod_{k=1}^{n} P(x_k/c_i) \quad (2)$$

概率 $P(x_k/c_i)$ 可以用训练样本估算:如果 a_k 是离散属性,则 $P(x_k/c_i) = s_{ik}/s_i$ 其中 s_{ik} 是属性 a_k 上值为 x_k 的类中的 c_i 训练样本数,s_i 为 c_i 中的训练样本数。如果 a_k 是连续值属性,通常该属性服从正态分布,并把类条件概率密度函数

$$P(x/c_i)P(c_i) > P(x/c_j)P(c_j),\ 1 \leqslant j \leqslant m,\ j \neq 1 \quad (3)$$

其中 μ, σ 分别为属性 a_i 取值的平均差和标准差。

3) 对未知的数据项 X 进行分类,对于每个 C,计算 $P(x/c_i)P(c_i)$,当且仅当 $P(x/c_i)P(c_i) > P(x/c_j)P(c_j)$,$1 \leqslant j \leqslant m$,$j \neq 1$,经过运算即可得到分类结果。

2 数据预处理

2.1 数据理解

初始数据来自于某证券公司营业部2011年的客户历史交易,通过变量计算分别得到第一季度的22 151条数据,第二季度的13 792条数据,第三季度的14 783条数据。

在数据预处理之前,首先要熟悉数据,识别数据的质量问题,并对数据进行描述,生成数据属性报告。

2.2 数据清理

1)2011年第一季度的数据清理。2011年第一季度的初始数据为22 151条。其中部分数据对模型的建立没有价值,在建立模型时删除这些数据。首先,删除9 620条重复的客户记录,删除2 858条销户日期为空的销户客户记录,并将数值中的所有null值赋值为0;其次,针对销户日期早于此季度的过期客户数据,进行删除操作,处理了380条数据;增加字段"交易期间"标记季度时间,增加字段"年龄",增加字段"在网时间"以年为单位计算公式为(当前年份-开户日期),删除客户基本信息外的所有交易信息皆为空的记录。最终获得2011年第一季度的数据,共5 730条。

2)2011年第二季度的数据清理。2011年二季度的初始数据为13 792条。第一,删除5 110条销户日期为空的销户客户记录,并将数值中的所有null值赋值为0;然后,对销户日期早于此季度的客户数据,进行删除操作,处理了346条数据;增加字段"交易期间"标记季度时间,增加字段"年龄",增加字段"在网时间"以年为单位计算公式为(当前年份-开户日期),删除客户基本信息外的所有交易信息皆为空的记录。这些数据对模型的建立没有意义所以在建立模型时删除这些数据。结果获得2011年第二季度的数据,共6 485条。

3)2011年第三季度的数据清理。2011年三季度的初始数据为14 783条。开始,删除5 110条销户日期为空的销户客户记录,并将数值中的所有null值赋值为0;接着,对销户日期早于此季度的客户数据,进行删除操作,处理了380条数据;增加字段"交易期间"标记季度时间,增加字段"年龄",增加字段"在网时间"以年为单位计算公式为(当前年份-开户日

期)，删除客户基本信息外的所有交易信息皆为空的记录。这些数据对模型的建立没有意义所以在建立模型时删除这些数据。结果获得 2011 年第二季度的数据，共 6 887 条。

3 实证结果分析

通过 Weka 数据挖掘系统中的朴素贝叶斯模型(Naïve Bayesian Model，NBC)分类器实现对客户价值分类。选择某证券公司营业部 2011 年的客户历史交易中的客户等级作为决策变量属性，分量属性如表 1 所示。

表 1 属性选择表

序号	变量	含义及计算标准
1.	期初市值	股票市值
2.	期末市值	股票市值
3.	期末基金	基金市值
4.	交易次数	成交次数
5.	总佣金	A 股＋权证＋基金
6.	总交易量	买卖交易金额
7.	平均每次交易量	总交易量/交易次数
8.	期初总资产	期初资金＋期初市值＋期初场外基金
9.	期末总资产	期末资金＋期末市值＋期末基金
10.	盈亏金额	期末总资产－期初总资产－市值转入(余额入账转入)＋市值转出(余额入账转出)－银行转入＋银行转出
11.	盈亏率 1	盈亏金额/累加资产
12.	换手率 1	总交易量/累加资产

表 2 客户群属性均值对比

客户等级 / 平均值	CLASS2＝明星型	CLASS3＝萎缩型	CLASS1＝潜力型	CLASS0＝退出型
交易次数	343	2 328	9	95
佣金	18 780.08	145 630.56	169.03	4 251.23
盈亏金额	188 186.102	－3 154 278.36	－20 049.41	－382 828.10
客户数量	53	169	6 640	9

根据计算结果表 2，可以发现：

CLASS2 客户群的特点是，客户交易量大，且盈利情况非常好的，由于交易量大，对于赚取佣金的证券企业来说，其贡献值也是相对较高的，我们将此类用户定义为明星型客户。

CLASS3 客户群的特点是，客户交易量大，但是亏损严重。由于此类客户交易量大，对于企业来说贡献高，但是又因为持续的亏损，很可能因为这样的结果导致他们对股市的信心受到影响。我们将此类客户定义为萎缩型客户。

CLASS1 客户群的特点是，客户属于交易量小，但亏损不多的客户。由于交易量较小，对于企业的贡献不多，亏损不大。可见其成长潜力较大。我们将此类客户定义为潜力型客户。

对于 CLASS0 客户群的特点是，交易量小，亏损严重。他们很可能因为交易失败而丧失对股市的信心，因此也较容易退出股市，进行其他投资，因此我们将其定义为退出型客户。

由于证券公司的竞争手段同质化还是比较明显的，通常为了吸引客户，公司营业部常常使用例如开户送炒股软件、股票机、手机、上网流量甚至电脑等办法，在同一市场中大部分营业部采取的服务模式、服务品种十分相似。然而为了在竞争中取得优势，证券公司面对以上不同客户群必须有不同的服务需求，这种差异不仅要体现在金融产品的类型和档次的需求上，还体现在服务方式、服务手段和服务内容等方面。企业需要采取个性化的服务，有针对性的制定出合适特定细分客户群的服务策略和产品策略；通过客户经理、理财经理的及时沟通，根据每个阶段的行情特点把握客户需求的变化，不断改善服务并及时调整市场营销策略；集中优势资源投放到目标市场去的差异化和个性化竞争优势，才能提高客户满意度，维持长久的客户关系，从而取得最佳的经营效果[7－8]。

4 模型评估和验证

k 折交叉验证(k-fold cross-validation)是指将原始的数据随机分为数量均匀的 k 个集合,然后进行 k 次的迭代。每次迭代过程是,从这 k 个集合中选出不同的一个集合作为测试集,剩下的 k－1 个集合用来训练分类器。通过进行了 k 次的训练和测试,然后将 k 次测试的误差率进行平均,得到一个总的综合误差估计。相当于每个集合参与了 k－1 次训练,1 次测试。如果切分后的每个集合的类的比例,跟原始的数据中基本一致,那么称为分层交叉验证。所谓的分层,就是指要求切分后的集合中类的比例一致。通常,使用分层技术可以改进结果。

通过计算分类交叉验证,得到此次实证分析的分类比例和混合矩阵。

表 3　分类比例表

项目名称	数量/个	百分比/%
正确分类实例	6 597	95.789 2
错误分类实例	290	4.210 8

根据分类交叉验证的结果可以看出,本次朴素贝叶斯客户价值分析模型分类的结果中,正确分类实例达到 6 597 个,百分比为 95.789 2% ,而错误的分类实例只为 290 个,百分比为 4.210 8% 。可见经过验证,本次模型分类的正确比例是比较高的。

表 4　混合矩阵表

	CLASS2	CLASS3	CLASS1	CLASS0	总数
明星型	45	2	0	6	53
萎缩型	16	3	0	150	169
潜力型	16	6	6 394	240	6 640
退出型	1	8	0	0	9

由上述混合矩阵中可以看出,明星型客户数量为 5 345个是分类准确的。2 两个被错误分类到萎缩型客户类别当中,有 6 个被分到退出型客户类别中。萎缩型客户数量为 169 个,150 个分类是准确的。有 16 个被错误分类到明星型客户当中,3 个被错误分类到萎缩型客户类别当中。潜力型客户数量为 6 640 个,6 394 是分类准确的。有 16 个客户被错误分类到明星型客户当中,有 6 个客户被错误分类到萎缩型客户类别当中,有 240 个客户被错误分类到退出型客户类别当中。在退出型客户当中有 9 个客户分类是正确的。而 1 个客户被错误分类到明星型客户当中,有 8 个客户被错误分类到萎缩型客户当中。

5　对策建议

针对客户分析数据挖掘模型分析结果,证券公司可以相应的客户关系管理策略。

1)针对明星型客户的 CRM 策略。针对明星型的客户,将其当成企业重量级客户来对待,可以采取追踪式的客户营销。此类客户炒股技术丰富、具备交易经验、交易次数较大,他们大多有自己的见解和对行情的判断,证券企业主要工作是配合客户的投资组合进行信息收集、个股分析,给出配置或交易建议等服务。

2)针对萎缩型客户的 CRM 策略。萎缩型客户,其特征为交易量大,但是亏损较为明显。针对此类客户,企业应当根据其经济背景,风险偏好,收益预期,结合未来行情可能发展的情况为其设计一个投资组合,提供较为专业的投资咨询服务,来提高客户自身的盈利能力,从而与企业维持长久的客户关系。

3)针对潜力型客户的 CRM 策略。潜力型客户交易次数不多,其在股市中很有可能持相对保守的态度,此时证券企业应该增强对企业品牌的塑造和宣传,将公司的产品、价格、服务有效的通过合适的方式传递给客户,采取刺激措施鼓励其进一步交易。

4) 针对退出型客户的 CRM 策略。由于此类客户可能对股市完全丧失信心,证券公司可以考虑为其推荐其他种类的投资方案。

参考文献

[1] 梁敏君. 分形聚类分析在证券客户细分中的应用研究[D]. 合肥:合肥工业大学,2009.

[2] 李君锋. 数据挖掘在证券业 CRM 中的应用研究[D]. 西安:西安电子科技大学,2009.

[3] 赵裕啸,倪志伟,王园园. SQL Server 2005 数据挖掘技术在证券客户忠诚度的应用[J]. 计算机技术与发展,2010,20(2):229－232.

[4] 李会彬,莫东艳 . 基于关联规则和模式发现的证券客户行为模式挖掘[J]. 信息系统工程,2010(5):43－44.

[5] 刘斌,邱华勇 . 证券公司客户综合分析系统的设计与实现[J]. 计算机系统应用,2010,19(10): 125－130.

[6] 谢芳. 数据挖掘在证券客户流失管理中的应用[J]. 科技管理研究,2011(10):180－183.

[7] 王园. 证券业客户细分模型构建及实证研究[J]. 上海管理科学,2012,34(2):30－35.

[8] 王园. 客户风险评级管理研究与应用——基于证券 CRM 管理[J]. 哈尔滨商业大学学报:社会科学版,2012(2):10－15.

(下转第 83 页)

3 结语

综上，浙江省中小板市场具备一定的使用会计数据做出理性预期的能力的同时，投机气氛比较浓烈，投资股市的目的在于短期内出售套利。但是，投资者必须注意到，在市盈率的波动性和公司的业绩波动性显著正相关这一现象，选择业绩波动较大的公司作为投资对象的风险也较大。对于长期、稳健的投资决策者，应该选取市盈率较高且波动较为平稳的股票作为投资对象，长期看获利的可能性较大。另一方面，对于浙江省中小上市企业管理者而言，应采取优化贷款质量、改革生产运营等措施，提高公司的流动性水平、突破成本约束的瓶颈，使企业建立起抵御宏观经济周期性风险的长效机制，从而提升公司财务状况的健康稳定性与可持续的盈利性，赢得长期投资者青睐，顺利完成产业升级。另外中小板上市企业稳定可持续的发展，有利于引导投资者形成正确投资理念，从而有助于投资市场的完善和成熟。

参考文献

[1] LINCH P. 黄金投资法则[M]. 1 版. 北京：高等教育出版社，2008：54.

[2] SAHOO P，RUSSELL J，LEX C，HUBERTS，MICHAEL J. The price-earnings ratio and the equity returns in India[J]. Journal of Banking Financial Services and Insurance Research，2011(3)：9－16.

[3] PARK S. What Does the P-E Ratio Mean? [J]. Journal of Investing，2000(9)：27－40.

[4] 宋剑峰. 净资产倍数、市盈率与公司的成长性——来自中国股市的经验证据[J]. 经济研究，2000(8)：15－16.

[5] 陈共荣，刘冉. 市盈率能否成为投资决策分析的有效指标——来自中国 A 股的经验数据[J]. 会计研究，2011(9)：9－13.

[6] 吴世农，李常青，余玮. 我国上市公司成长性的判定分析和实证研究[J]. 战略管理，1999(4)：49－50.

[7] 朱星宇，陈勇强. SPSS 多元统计分析方法及应用[M]. 北京：清华大学出版社，2011.

[8] 古扎拉蒂，波特．经济计量学精要[M]. 北京：机械工业出版社，2010.

Research on the Guiding Value of P/E to Investment Based on Growth Potential

——An empirical study on middle and small scale enterprises in Zhejiang province

LIU Lang，WANG Jing-fang

(School of Management，Northwestern Polytechnical University，Xian 710072，China)

Abstract：This paper takes the middle and small scale enterprises in Zhejiang province for example and select PE ratio as the subject to examine the guiding significance of PE ratio in investment analysis and comes to the following conclusions：A company's growth potential may be reflected in PE ratio，especially for indicators such as incensement indicator and cash return ratio. However，investors may intend to speculate to gain profits in a short time，thus increase its risk. Some suggestions to investors and managers are provided in the end of the paper.

Key words：P/E ratio；value investment；growth potential；stability

(上接第 50 页)

Research of Customer Value Segmentation Based on Naive Bayes Classify Method

WANG Yuan，LI Shao-feng，LI Yong-mei，OU Bing-zhen，WANG Qiu-ming，LIN Qiao-ming

(School of Business Administration，Jimei University，Xiamen Fujian 361021，China)

Abstract：Customer value is the basis of customer relationship management. This paper applies Naïve Bayes model to analysis data mining of clients' transaction data and the basic information. It uses Naïve Bayes model of Weka to train，test and check customer transaction data，which is to construct customer value analysis model，then evaluate itself，finally establish a highly efficient data mining model. The model classifies customer and value analysis so that it can identify different characteristics，different value of the customer base，all that can help provide personalized service according to different levels of customer value.

Key words：securities company；customer value；data mining；naïve Bayes model