

基于背景的个性化客户行为模型研究

何红洲^{1,2}, 周明天¹

(1. 电子科技大学 计算机科学与工程学院, 成都 610054; 2. 绵阳师范学院 数学与计算机科学学院, 四川 绵阳 621000)
(zmoonmoonlm@yahoo.com.cn)

摘要:针对商业客户的背景信息对其消费行为的影响问题,提出了一种基于背景的商业客户行为模型的构建方法。该方法不仅收集了包含三级粒度背景信息的某大学学生客户的网上交易数据并按依赖于交易数据项的统计量对客户进行分组,还利用朴素贝叶斯(NB和TAN)及分组和统计关系数据库(GAC-RDB)分类器学习了各客户分组的背景和非背景预测函数,同时使用各预测变量的受试者运行特征曲线下面积(AUC)值,对客户背景在预测客户购买行为时的作用进行了定量的比较和分析。研究表明:背景信息对客户特别是个性化客户(单一客户)的消费决策具有良好的预测效果。

关键词:客户行为模型;客户背景粒度;朴素贝叶斯分类;受试者运行特征曲线下面积;统计显著性

中图分类号: TP391 **文献标志码:** A

Research of context-based personalized customer behavior model

HE Hong-zhou^{1,2}, ZHOU Ming-tian¹

(1. School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu Sichuan 610054, China;
2. College of Mathematics and Computer Science, Mianyang Normal University, Mianyang Sichuan 621000, China)

Abstract: To investigate the context's influence on business customer consumption behavior, a sort of construction approach of context-based customer behavior model was proposed. One undergraduate customer transaction data online with three-level context granularity was collected and grouped on statistic based transaction data item. The classifiers including Naive Bayesian (NB), Tree Augmented NB (TAN) and Grouping and Counting-relational Database (GAC-RDB) were used to learn context and non-context predicating functions of each customer group. Based on the Area under a Receiver Operating Characteristic Curve (AUC) of predicating variable, the paper compared and analyzed quantitatively the effect of customer context when predicating his buying behavior. The experimental results demonstrate that the context information has preferable predication performance on the consumption decision of the customers especially the personalized customers.

Key words: customer behavior model; customer context granularity; Naive Bayesian Classification (NBC); Area Under a Receiver Operating Characteristic Curve (AUC); statistic significance

0 引言

在商业市场营销活动中,客户的购买行为受到诸多因素的影响,例如特定的付款方式、变化了的经济条件(如退休、疾病等)、转变了的家庭角色(如当上了父母、结婚、离异等)和特定的地理位置等^[1],这些影响因素统称为客户的背景信息。文献[2]将客户的购买意向(使用主体)和对商品的选择方式(按价目、按货柜或导购)作为客户背景。文献[3]探讨了在商品推荐系统中包含背景信息的重要性。文献[4-5]定性地分析了背景信息对客户购买行为的影响。但还没有文献对该问题展开定量地研究。本文利用几种典型的分类算法^[6-7]对客户的交易行为进行建模,进而定量和较系统地研究了客户背景信息对客户购买行为的影响能力。

1 数学描述

设有 n 个客户 C_1, C_2, \dots, C_n , 各客户 $C_i (i = 1, 2, \dots, n)$ 最多有 r (定值) 笔交易记录 $TC_{i1}, TC_{i2}, \dots, TC_{ir}$, 各交易记录 $TC_{ij} (i = 1, 2, \dots, n; j = 1, 2, \dots, r)$ 用如下的几类属性来描述:统计属性 S (用来表示客户基本特征如编号、姓名、性别和年龄

等的的数据项), 交易属性 T (用来表示客户某次交易特征如品号、货柜名称、价格、持续时间和购买与否等的的数据项) 和一些表示客户交易背景的数据项 G 。

由于背景信息的复杂性, G 应该具有较复杂的结构。然而本文中, G 被定义为一个单一的离散变量, 并按图 1 所示的二叉树层次结构取值, 即按商品的使用主体(为谁买)逐层细分, 层次越低, 背景越详细。

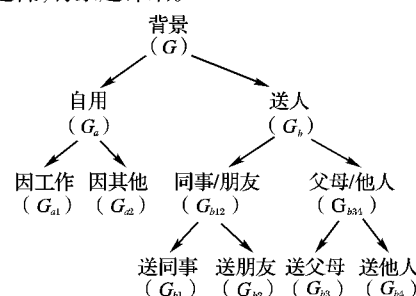


图1 一类购物交易的背景(G)二叉树图

首先将客户集按对某种统计属性或交易属性的统计量(如平均年龄、交易商品平均价格等)进行聚类, 从而定义了相应的统计空间, 然后针对每一类客户(一个分组, 如经常在

收稿日期:2009-06-22;修回日期:2009-08-15。 基金项目:四川省教育厅自然科学研究项目(2005A117)。

作者简介:何红洲(1967-),男,四川绵阳人,副教授,博士研究生,主要研究方向:数据挖掘、复杂网络;周明天(1939-),男,广西容县人,教授,博士生导师,主要研究方向:计算机网络、并行分布处理、信息安全。

同一货柜购买的客户分为一组)的交易数据,建立如下的数学模型(简称预测模型一):

$$Y = f(X_1, X_2, \dots, X_q) \quad (1)$$

其中: $X_i (i = 1, 2, \dots, q)$ 为某个统计属性或某个交易属性, Y 是一个预测变量, f 是通过某种机器学习方法^[6-7]学习的预测函数。利用它们,商家可预测客户购买商品的货柜、购买商品的种类或者购买商品的价格范围等。

因为预测模型一没有考虑背景因素,因此,我们称该类模型为非背景(预测)模型。为了将背景因素考虑在内,我们建立模型一相应的背景模型如下(简称预测模型二):

$$Y = f_{G=G_0}(X_1, X_2, \dots, X_q) \quad (2)$$

即只利用与背景 $G = G_0$ 相关的交易数据构建这个模型。例如,如果该模型用于某大学的计算机科学系的全体人员,并且 $G = \text{"送人"}$ (即取 G_0 为 G_b ,见图 1),就意味着模型二只限于该系人员与礼品相关的交易数据。由于背景变量 G 的层次性,背景(预测)模型的数量可能视背景信息的粒度的不同(体现在 G_0 所处的层次)而不同。

2 试验设置

为了收集相关背景信息数据,我们设计了一个特定的浏览器来帮助客户(某大学的学生)浏览一个著名电子商务零售网站,并在上面模拟购买商品(不花费真正的钱)。从 2008 年 2 月 1 日起,学生在 3 个月之内可以访问该网站的任何页面并能使用网站所有的浏览和导航功能。一旦一件商品被学生选定,浏览器将记录被选定商品的品号、价格及其他相关的交易数据。另外,在每笔交易过程中,要求学生按图 1 的层次填写购买背景。浏览器直接连接到我们创建的关系数据库,并按表 1 记录所有有关客户在该网站上浏览及购买活动的信息及按图 1 记录背景信息。一次成功的提交表示一次交易活动的结束。

表 1 某大学学生交易表的各属性定义

序号	属性	类型	值/范围
1	编号	char(8)	spjy0001 ~ spjy9999
2	姓名	char(10)	汉字或英文字母
3	性别	char(2)	{男,女}
4	年龄	smallint	18 ~ 35
5	家庭月收入	money	1 000 ~ 12 000
6	品号	char(8)	PHa00001 ~ PHz99999
7	货柜名称	varchar(20)	{电子用品,家居/园艺,展销品,婴幼儿用品,图书/音乐,新开货柜}
8	价格	money	1 ~ 1 500 元
9	交易日期	datetime	02/01/08—05/01/08
10	持续时间	numerical	0 ~ 1 800 s
11	点击数	smallint	1 ~ 35
12	购买否	char(2)	{是,否}

注:属性 1~5 为统计属性,属性 6~12 为交易属性。

为了使问题更简单一些,在处理过程中剔除了三个月之内交易笔数不足 90 的学生及相关交易信息,结果交易人数及总交易笔数(即按表 1 属性描述的总交易记录数)分别降为 358 和 45 925。下面详细讨论各种参数的设置情况。

2.1 客户分组粒度

我们分别考虑了以下四级分组^[8-9]:

总客户集 所有客户整体为一个分析单元。

10-聚类 所有客户使用 FFC 聚类方法^[10]按统计向量

(平均价格、交易的平均点击数、交易的平均持续时间)聚为 10 个类(10 个分组),每一分组为一个分析单元。

100 聚类 按与第二类相同的聚类方法和统计向量进行聚类(分组),聚类数为 100,每一分组为一个分析单元。

单一客户 每一客户为一个分析单元。

2.2 客户背景粒度

背景信息 G 是一个如图 1 所示的三层粒度结构:第一层, G 取两种不同的值(“自用”和“送人”)。第二及以下各层,“自用”细分为“因工作”和“因其他”,“送人”细分为“送同事”、“送朋友”、“送父母”及“送他人”。在第二层,送同事或朋友归并为一个中等粒度细目 G_{b12} ,送父母或其他人也归并到另一中等粒度细目 G_{b34} 。背景信息粒度从一到三逐层变细。试验中针对第 2.1 节的每一个分析单元,为每一层中的每一种背景粒度建立了相应的预测模型。

2.3 预测变量

我们从交易属性中选择了如下的预测变量:

购买否 模型预测一个或一组客户是否要购买。

货柜名称 模型预测成功交易最终将发生在哪个货柜。

2.4 性能度量依据 AUC

我们对各分析单元使用非参数估计的受试者运行特征(Receiver Operating Characteristic, ROC)曲线下面积(Area under the ROC Curve, AUC)^[11]来度量其性能。ROC 曲线为临床科研文献中应用最广泛的、国际公认的比较和评价两种或两种以上试验准确度的统计曲线。在平面直角坐标系下,ROC 曲线由 1-特异度为横坐标、以灵敏度为纵坐标绘制而成。其 AUC 的大小可从量上具体表明试验的准确度。非参数 ROC 曲线下面积估计是根据实验结果直接计算绘制 ROC 曲线所需的工作点,由此绘制出经验的 ROC 曲线,其面积可由 Mann-Whitney 检验统计量得到,其计算使用了 SAS 统计软件。非参数法因其没有限制条件,所以适用于任何预测试验 ROC 曲线下面积的估计。我们使用零假设来比较非背景和背景模型的性能差异,看它们之间是否具有统计显著性。

2.5 用分类器学习预测函数

我们使用了三种典型的分类器来建立第 1 章提出的预测模型一和预测模型二:朴素贝叶斯分类(Naive Bayesian Classification, NBC)^[6]是最基本的贝叶斯分类方法,它利用后念概率的最大值点来判别预测变量的取值类别,前提是交易记录各属性相互独立;树增强朴素贝叶斯分类器(Tree Augmented NB, TAN)^[6]分类通过发现属性对之间的关联来降低 NB 中任意属性之间独立性的假设;而分组和统计关系数据库(Grouping and Counting-relational Database, GAC-RDB)^[7]分类使用关系数据库系统提供的聚集运算功能,它首先利用 SQL 语句计算用每个属性进行类别判定的信息含量,从而选择一个最好的分裂属性,并且按照信息含量的大小对属性进行排序,接着循环地选择属性、生成和剪裁候选分类表以及计算分类误差,直到满足结束条件(如最小误差阈值和误差没有改善)为止。选择 NB 和 TAN 分类器是因为它们比较常用也比较简单,选择 GAC-RDB 是因为它利用了关系数据库的成批数据记录的处理特点,提高了试验数据的处理速度。为了适当减少工作量,我们没有考虑其他分类器。

根据以上提出的试验条件模型类型和总个数如下:

背景 1(粗):3(总客户集)+30(10-聚类)+300(100-聚类)+358×3(单一客户)=333+1 074=1 407(个)

背景 2(中): $5(\text{总客户集}) + 50(10\text{-聚类}) + 500(100\text{-聚类}) + 358 \times 5(\text{单一客户}) = 555 + 1790 = 2345(\text{个})$

背景 3(细): $7(\text{总客户集}) + 70(10\text{-聚类}) + 700(100\text{-聚类}) + 358 \times 7(\text{单一客户}) = 777 + 2506 = 3283(\text{个})$

总共有: $(1407 + 2345 + 3283) \times 3(\text{分类器数}) \times 2(\text{预测变量数}) = 42210(\text{个})$ 模型。

3 结果

按照第 2 章描述的试验条件,使用 3 种不同的分类器得出了客户行为的背景和非背景模型针对各背景粒度和各客户分组粒度对两种预测变量的 AUC 值。但限于篇幅,我们不可能把所有 42210 个模型的结果完全列出来,因而我们首先针对第 2.1 节的各分析单元(各个分组),取各分类器得到的 AUC 值的平均值,以此为基础,再针对每一背景粒度,取同一类客户分析单元(例如 10-聚类中的 10 个分组)的上述平均值的平均值,然后再根据这个平均值代入相对差异公式(式(3))来表达这些结果。

$$(AUC_{\text{背景}} - AUC_{\text{非背景}}) / AUC_{\text{非背景}} \quad (3)$$

式(3)为正(负)值意味着背景模型优(次)于非背景模型,即背景模型有收益(损失)。表 2 分别对两个预测变量“购买否”(上半部分)和“货柜名称”(下半部分)列出了按式(3)计算的 AUC 值的相对差异,表中列表示了各级客户分组粒度,而行表示了各级背景粒度。第 1~2 和 13~14 行记录了背景粒度最粗(即 G 只取两值“自用”和“送人”)时的比较结果,而 7~12 和 19~24 记录了背景粒度最细(即 G 取 6 个值“因工作”,“因其他”,“送同事”,“送朋友”,“送父母”和“送他人”)时的比较结果。图 2 给出了按不同的背景粒度(不同的折线)和不同的客户分组粒度(x 坐标)对两个预测变量用表 2 的数据取均值后的结果。

表 2 在各背景粒度下的 AUC 值的相对差异 %

各级背景	总客户集	10-聚类	100-聚类	单一客户
自用	4.12	5.66	8.33	9.46
送人	7.96	5.66	11.67	13.51
因工作	11.83	9.43	13.33	22.97
因其他	2.05	-3.77	10.08	13.51
送同事/朋友	5.98	0.94	23.33	15.54
送父母/他人	3.99	-1.89	26.67	21.62
因工作	11.83	9.43	13.33	22.97
因其他	2.05	-3.77	10.08	13.51
送同事	14.14	2.83	29.12	18.92
送朋友	8.78	9.43	12.92	22.28
送父母	8.02	0.94	36.67	23.65
送他人	5.16	-3.77	37.50	29.73
自用	5.30	3.85	8.67	8.25
送人	-2.83	2.86	-2.76	-2.44
因工作	28.41	12.25	11.95	23.92
因其他	8.25	3.73	10.68	8.56
送同事/朋友	5.43	2.84	3.24	7.52
送父母/他人	-5.64	-6.43	-3.81	6.38
因工作	28.41	12.25	11.95	23.92
因其他	8.25	3.73	10.68	8.56
送同事	10.38	3.95	17.84	16.78
送朋友	6.23	-2.68	5.62	9.25
送父母	-3.63	-8.62	-4.65	5.37
送他人	-1.81	-3.65	12.53	18.62

图 3~4 分别按背景粒度和客户分组粒度画出了对表 2 中的正值和负值分别作均值计算后取绝对值的结果,因而可分开讨论背景模型的收益和损失情况。

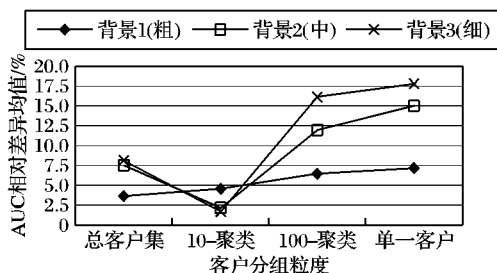


图 2 各背景粒度的平均相对差异随客户分组粒度变化的趋势

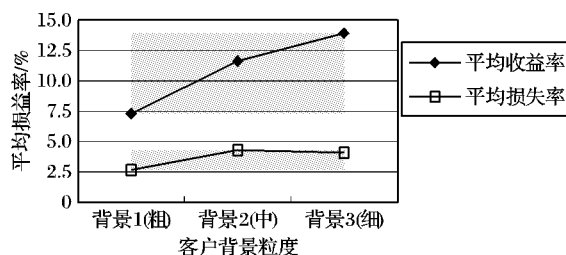


图 3 按背景粒度变化的性能平均收益线和平均损失线

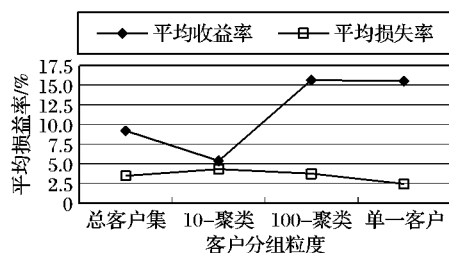


图 4 按客户分组粒变化的性能平均收益线和平均损失线

由于比较量巨大,所以要给出每种比较的统计显著性是不可能的,图 5~6 分别针对背景模型的收益和损失情况给出了按不同客户分组粒度(用不同的折线表示)和不同背景粒度(x 坐标值)的显著比较在 95% 以上的统计显著性事件的百分比。

对以上结果分析如下:

1) 一般而言,客户分组粒度越细,背景信息的效果越明显。如图 2 所示,当分析单元从“总客户集”向“单一客户”移动时,背景模型中可获得的收益增加(从“总客户集”模型到“10-聚类”模型除外)。表 2 显示的结果也表明:越靠右边的列负值越少。更精确地说,“单一客户”模型相比于其他模型,基本上没有性能损失。

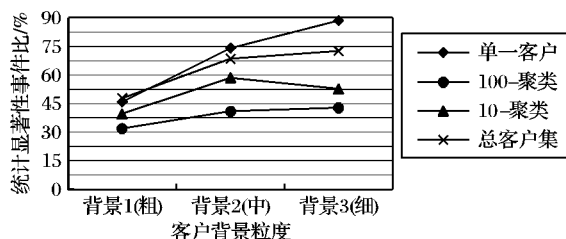


图 5 性能收益时各客户分组粒度的统计显著性事件

2) 对个性化(单一客户)的模型,背景信息效果明显。这时,背景模型的预测性能总是高于非背景模型的预测性能,表 2 最右一列只有一个绝对值很小的负值(-2.44%)和图 4 都表明了这一点。图 4 还表明了个性化背景模型可获得性能

的平均收益接近15.6%,而收益的峰值可达到29.73%(表2的最右一列)。与此相反,建立一个个性化的背景模型也存在性能降低的情况,但如图4所示,平均损失不超过2.5%。

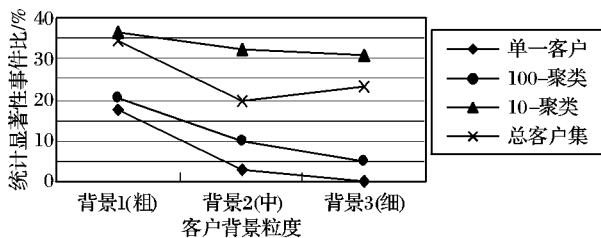


图6 性能损失时各客户分组粒度的统计显著性事件

3) 客户背景粒度影响了对客户行为的预测准确程度。对客户背景知道得越细,对客户行为的预测就越准确。分析表2的最右边一列,唯一的负值仅发生在最粗的背景信息粒度上,背景粒度越细,性能收益也就越高。当分析单元是其他情况(表2向左移)时,虽然负值个数有所增加(特别是对于“10-聚类”的情况),但图3所示的平均收益率和平均损失率表明:背景粒度的性能收益在较宽的中带部分,从7.33%(最粗粒度)到13.94%(最细粒度)不等;而背景粒度导致的性能损失在非常窄的低带部分,绝对值从2.68%到4.31%不等,其最高点仅在中等背景粒度时出现,并且不超过4.5%。图5进一步说明了对于背景模型优于非背景模型的情况,除“10-聚类”模型的中粒度背景到细粒度背景外,统计显著性事件的数目随背景粒度的不断变细而增长。而图6进一步说明了对于相反的情况,除“总客户集”模型的中粒度背景到细粒度背景外,统计显著性事件的数目随着背景信息粒度的不断变细而衰减,特别是在“单一客户”模型的细背景上,其统计显著性事件的数目为0。这一事实说明:在“单一客户”模型的细背景上,没有非背景模型优于背景模型且差异是统计显著的事件。

4 结语

客户资源是商业企业的生命力,定量地分析和研究针对客户背景的预测模型(特别是个性化预测模型)的性能,能更好地预测客户下一步的购买意向。本文就单类型的背景数据

集进行了研究,并证实:不同粒度的客户背景都不同程度地增强了对客户购买行为预测的准确度。

参考文献:

- [1] BERRY M J A, LIOFF G S. Data mining techniques: For marketing, sales, and customer relationship management [M]. 2nd ed. Indianapolis: Wiley Publishing, 2003.
- [2] LILIE G L, KOTLER P, MOORTHY S K. Marketing models [M]. Upper Saddle River: Prentice Hall, 1992.
- [3] ADOMAVICIUS G, SANKARANARAYANAN R, SEN S, *et al.* Incorporating contextual information in recommender systems using a multidimensional approach [J]. ACM Transactions on Information Systems, 2005, 23(1): 103 - 145.
- [4] BETTMAN J R, LUCE M F, PAYNE J W. Constructive consumer choice processes [J]. Journal of Consumer Research, 1998, 25(3): 187 - 217.
- [5] DEY A K, ABOWD G D, SALBER D. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware application [J]. Human Computer Interaction, 2001, 16(2): 97 - 166.
- [6] FRIEDMAN N, GEIGER D, GOLDSZMIDT M. Bayesian network classifier [J]. Machine Learning, 1997, 29(1): 131 - 163.
- [7] LU HONG-JUN, LIU HONG-YAN. Decision tables: Scalable classification exploring RDBMS capabilities [C]// Proceedings of the 26th International Conference on VLDB. Cairo: Morgan Kaufmann, 2000: 373 - 384.
- [8] KOTLER P. Marketing management [M]. 11th ed. New Jersey: Prentice Hall, 2003.
- [9] JIANG T, TUZHILIN A. Segmenting customers from population to individuals: Does 1-to-1 keep your customers forever? [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(10): 1297 - 1311.
- [10] DUNHAM M. 数据挖掘教程[M]. 郭崇慧, 田凤占, 靳晓明, 等译. 北京, 清华大学出版社, 2005.
- [11] HANLEY J A, MCNEIL B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve [J]. Radiology, 1982, 143(1): 29 - 36.

(上接第3282页)

参考文献:

- [1] FERREIRA C. Gene expression programming: A new adaptive algorithm for solving problems [J]. Complex Systems, 2001, 13(2): 87 - 129.
- [2] 周明, 孙树栋. 遗传算法原理及应用[M]. 北京: 国防工业出版社, 1999.
- [3] 李曲, 蔡之华, 朱莉, 等. 基因表达式程序设计方法在采煤工作面瓦斯涌出量预测中的应用[J]. 应用基础与工程科学学报, 2004, 12(1): 49 - 54.
- [4] ZUO J, TANG C J, LI C, *et al.* Time series prediction based on gene expression programming [C]// WAIM 2004: 5th International Conference. Berlin: Springer, 2004: 55 - 64.
- [5] CHI Z, WEI M X, THOMAS T, *et al.* Evolving accurate and compact classification rules with gene expression programming [J]. IEEE Transactions on Evolutionary Computation, 2003, 7(6): 519 - 531.
- [6] 刘大有, 卢奕南, 王飞, 等. 遗传程序设计方法综述[J]. 计算机研究与发展, 2001, 38(2): 412 - 423.
- [7] BASTIEN C, MICHEL D. Cellular automata modeling of physical systems [M]. London: Cambridge University Press, 1998.
- [8] ALBA E, DONRONSORO B. The exploration/exploitation tradeoff in dynamic cellular genetic algorithms [J]. IEEE Transactions on Evolutionary Computation, 2005, 9(2): 126 - 142.
- [9] 田志友, 王浣尘, 吴端明. 区域市场连锁经营选址与布局的元胞自动机模拟[J]. 系统工程理论方法应用, 2005, 14(1): 50 - 54.
- [10] 朱刚, 马良. TSP的元胞蚂蚁算法求解[J]. 计算机工程与应用, 2007, 43(10): 79 - 100.
- [11] XIN L, CHI Z, WEI M X, *et al.* Prefix gene expression programming [C]// GECCO'2005. Washington, DC: [s. n.], 2005.
- [12] 唐丽钰, 李森, 张建, 等. 基于自动定义函数GP的自适应建模研究[J]. 小型微型计算机系统, 2005, 26(6): 1000 - 1005.