基于贝叶斯分类的网上书店潜在用户挖掘*

董 倩,王克俭,韩宪忠,苑迎春
(河北农业大学 信息科学与技术学院,河北 保定 071001)

摘 要:以网上书店为例,利用贝叶斯分类预测技术,进行了发现潜在客户群体的研究,用随机选取的10组样本进行试验预测,预测准确率达96.5%,表明了该算法是有效的。

关键词:数据挖掘;贝叶斯分类;潜在用户;网上书店

中图分类号:TP301 **文献标识码**:A **文章编号**:1674-7720(2011)01-0047-03

Mining potential users of the online bookstore based on Bayesian classification

Dong Qian, Wang Kejian, Han Xianzhong, Yuan Yingchun
(College of Information Science and Technology, Agricultural University of Hebei, Baoding 071001, China)

Abstract: This article takes online bookstore as the example to study how to find the potential users based on Bayesian classification technique. Ten samples are selected randomly which are used in the test. And the forecast accuracy is 96.5%, which proves the algorithm is effective.

Key words: data mining; Bayesian classification; potential users; online bookstore

当前信息时代发展迅猛,电子商务的兴起使得用户和商家都借助于Internet这个平台进行交流,方便用户购书的网上书店也随之发展起来。在服务成本加大、而收效甚微的前提下,商家面临着拓宽客户的问题。挖掘潜在顾客群体,为网站经营者在激烈的市场竞争中洞察先机、调整有效的顾客服务策略,提供准确的参考信息及科学的决策依据,最终达到识别潜在顾客、吸引新顾客、真正做到以顾客价值为中心,全方位为其提供整体服务,从而提升品牌、促进消费,在总体上减少商业成本并增加利润。

国内外关于面向Web日志挖掘用户行为及潜在顾客信息的研究发现,其包括三个过程:数据预处理、模式识别及模式分析[1]。在国外,Ngu D S W和Wu X等人也研究了SiteHelper系统,其主要方法是使用信息提取的方法提取页面信息,并且结合用户访问历史、用户个人资料提供的线索,向用户动态推荐访问的页面,缺点是涉及了比较敏感的用户个人隐私问题[2]。参考文献[3]根据用户的查询与目标页面的并发关系,分析聚类用户的存取事务,发现用户的个性化搜索模式,对其所需服务进行主动定制。在国内,参考文献[4]提出利用数据挖掘中的分类方法,根据已有用户的访问信息,训练分类器,其贡献在于能够量化地推断匿名用户的访问特性;其不足在于访问特性本身需要人工定义,存在着缺漏。郭新涛等人提出了一种新的支持站点设计优化的Web使用挖掘方案,该方案基于Web日志中的搜寻路径统计用户寻找目标花费的平均时间,以量化Web页面的搜寻费用,在此基础上提出了一种数据挖掘方法,寻找一组能够有效压缩搜寻路径(降低时间费用)的超链接,以便挖掘用户[5]。

基于上述不足,本文利用数据挖掘中贝叶斯分类技术来研究网上书店中的有关挖掘潜在用户的问题。贝叶斯算法作为处理不确定性信息的重要工具,已成功运用在统计决策、医疗诊断、零售业[6]、考试成绩检测机制等领域[7]。最为成熟的是,采用贝叶斯算法对邮件进行判断,建立了最优化的垃圾邮件过滤技术[8]。而本文所说的潜在用户也是具有不确定性,基于这个相似点,而选择使用贝叶斯算法[9]。

1 贝叶斯分类预测方法

分类分析就是通过分析示例数据库中的数据,为每

*基金项目:河北省自然科学基金资助项目(F2009000653);河北省科技厅计划资助项目(072135126);河北省教育厅计划资助项目(Z2009122)

个类别做出准确的描述、建立分析模型或挖掘出分类规则，然后用这个分类规则对新的数据记录进行分类，其中贝叶斯分类方法是一种易于使用并且具有最小错误率的概率分类法，它以完善的贝叶斯理论为基础，有较强的模型示、学习和推理能力，是一种很受欢迎的数据挖掘分类方法。贝叶斯分类是统计学分类方法，可以预测类成员关系的可能性，如给定数据项属于一个特定类的概率。

贝叶斯分类基于贝叶斯定理。数据项是由条件属性值(如：浏览持续时间，一天内的浏览次数，书的销售类型)组成的特征向量(如：0~5 min，2 次，特价书)。此外，数据项还有一个目标属性(如：是否购买该类型小说)。一个具体的数据项形式为：$X\ \{x_1, x_2, \cdots x_n; c\}$。其中，$x_i$ $(1\leqslant i\leqslant n)$表示条件属性值，$c$ 表示目标属性。下面简要介绍贝叶斯分类预测的工作过程。如果样本集有 n 个属性：$A_1, A_2, \cdots, A_n$，属性值构成样本的特征向量，可能的类别有 m 个，具体为$\{C_1, C_2 \cdots C_m\}$。假设待分类样本 X 的特征向量为$\{x_1, x_2, \cdots, x_n\}$，计算 X 分别属于每个类别的概率 $P(C_i|X)$，概率最大的那个就是 X 的预测类别。其中，$P(C_i|X)$由以下的贝叶斯公式计算得到：

$$P(C_i|X)=P(X|C_i)P(C_i)/P(X)$$

可以看到，对所有的 C_i，$P(X)$都相同，因此，只需要比较 $P(X|C_i)\times P(C_i)$即可。而 $P(C_i)$可以由训练集得到，该值等于训练集中类别为 C_i 的样本所占的比例。在属性独立的假设下有 $P(X|C_i)=\prod_{j=1}^{n} P(x_j|C_i)$

2 贝叶斯分类技术在网上书店挖掘潜在用户中的应用

本文以网上书店欲销售小说为案例，以网上书店的顾客 cookies 数据库为对象，用贝叶斯分类的挖掘技术对收集到的已经购买过本产品顾客的浏览持续时间、浏览次数、书的销售类型以及小说类型等数据进行分析，生成对当前数据库有效的用户分类模型，从中识别顾客购买行为，发现顾客购物模式和倾向，挖掘潜在用户，对不同顾客实施不同的推销策略，为该商店调整有效的销售策略提供一些有用的参考依据。

2.1 数据描述

首先把分类结果即目标属性定为两类：购买和不购买。其中数据样本可用一个五维特征向量 $X=\{x_1, x_2, x_3, x_4, x_5\}$分别描述以下属性(浏览持续时间、一天之内的浏览次数、书的销售类型、小说类型、是否购买)，其中各属性的数据泛化过程如下：

浏览持续时间：0 表示 0~5 min，1 表示 5~10 min，2 表示 10~30 min。

一天之内的浏览次数：3 代表浏览 1 次，4 代表浏览 2 次，5 代表浏览 5 次。

书的销售类型：6 代表特价书，7 代表热卖书。

小说类型：8 代表言情小说，9 代表武侠小说。

是否购买：-2 代表购买，-1 代表不购买。

2.2 预处理数据

把 cookies 数据库中的部分信息 (顾客购买的子集，14 人)作为训练样本(可随机抽取)，推断一下网站对未知类别样本的购买情况，以简单说明贝叶斯分类的一般工作流程。

表 1 给出了一个类别标记的数据项的样本，它是商店的 cookies 数据库中抽取的顾客训练集样本。

表 1 商店顾客训练集样本

顾客 ID 号	浏览持续时间/min	浏览次数/次	书的销售类型	小说类型	是否购买
1	0~5	1	特价	言情	否
2	0~5	1	特价	武侠	否
3	5~10	1	特价	言情	是
4	10~30	2	特价	言情	是
5	10~30	5	热卖	言情	是
6	10~30	5	热卖	武侠	否
7	5~10	5	热卖	武侠	是
8	0~5	2	特价	言情	否
9	0~5	5	热卖	言情	是
10	10~30	2	热卖	言情	是
11	5~10	2	热卖	武侠	是
12	5~10	2	特价	武侠	是
13	5~10	1	热卖	言情	是
14	10~30	2	特价	武侠	否

其相应的数据泛化后的顾客样本为：

顾客 1{0，3，6，8，-1}　顾客 2{0，3，6，9，-1}
顾客 3{1，3，6，8，-2}　顾客 4{2，4，6，8，-2}
顾客 5{2，5，7，8，-2}　顾客 6{2，5，7，9，-1}
顾客 7{1，5，7，9，-2}　顾客 8{0，4，6，8，-1}
顾客 9{0，5，7，8，-2}　顾客 10{2，4，7，8，-2}
顾客 11{1，4，7，9，-2}　顾客 12{1，4，6，9，-2}
顾客 13{1，3，7，8，-2}　顾客 14{2，4，6，9，-1}

2.3 挖掘潜在用户的算法流程

基于贝叶斯的挖掘潜在用户的分类算法流程如图 1 所示。

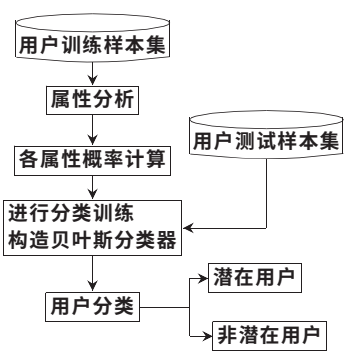

图 1 挖掘潜在用户算法基本流程

2.4 实例分析

推断新样本 $X(2，3，6，9)$的用户类别，用贝叶斯分

类解法挖掘潜在用户的每个步骤的结果为：

(1)P(是否购买="购买")=9/14=0.643，P(是否购买="不购买")=5/14=0.357。

(2) 使用贝叶斯算法计算各属性的所有取值相对于每个类别的概率结果如表 2 所示。

表 2 各属性值的条件概率

各属性值取值	相对于购买的条件概率	相对于不购买的条件概率
浏览持续时间 0~5 min	0.111	0.6
浏览持续时间 5~10 min	0.556	0
浏览持续时间 10~30 min	0.333	0.4
浏览次数 1 次	0.222	0.4
浏览次数 2 次	0.444	0.4
浏览次数 5 次	0.333	0.2
特价书	0.333	0.8
热卖书	0.667	0.2
言情小说	0.333	0.6
武侠小说	0.667	0.4

(3)判断用户类别

P('10~30 min，浏览 1 次，特价书，武侠小说'|'购买')×P('购买')=0.333×0.222×0.333×0.333×0.643=0.005 3

P('10~30 min，浏览 1 次，特价书，武侠小说'|'不购买')×P('不购买')=0.4×0.4×0.8×0.6×0.357=0.027 4

根据上述结果可知，P('不购买')>P('购买')，所以由贝叶斯挖掘技术预测的新样本的用户类为："是否购买=不购买"，也就是具有这种基本信息的顾客有很大的可能性不购买该商店的产品(武侠小说)。

3 实验结果与分析

为了验证贝叶斯分类方法的正确性和有效性，从 cookies 数据库随机抽取 10 组样本，分类结果如表 3 所示。可以看出，每组样本的样本个数不确定，其中有 9 组样本的正确率达到了 95%以上，在这 9 组样本中有 5 组样本的正确率达到了 100%，有一组样本的正确率在 95%以下。同时也可以看出，贝叶斯算法的不足之处在于，对发生频率较低事件的预测效果和对于样本个数较少的样本预测效果不好。从 10 组样本的预测结果中得出平均正确率为 96.5%，说明贝叶斯算法分类的正确率相当高，贝叶斯分类算法具有很强的学习、推理能力，能很好地利用先验知识。

本文研究了贝叶斯分类挖掘技术在购书网站挖掘潜在用户中的运用，基于贝叶斯方法的分类预测具有形式简单、易于解释、预测结果正确率高，且可以很容易从不同的领域进行推广等优点，但是对发生频率较低事件的预测效果不好，在这方面需要进一步改进。

表 3 预测结果

组数	样本个数	预测结果	正确分类	正确率/%
1	10	10	8	80
2	20	20	19	95
3	15	15	15	100
4	25	25	24	96
5	18	18	18	100
6	20	20	20	100
7	12	12	12	100
8	30	30	29	97
9	28	28	28	100
10	35	35	34	97
平均数	21.3	21.3	20.7	96.5

参考文献

[1] 王岚，翟正军.Web 日志挖掘的预处理及路径补全算法的研究[J].微电子学与计算机，2006，23(8)：113-114.

[2] NGU D S T，WU X. Sitehelper：A locall'zed agent that helps incremental exploration of the World Wide Web[C]. 6th International World Wide Web Conference. Santa，Clara，CA，1997：1249-1255.

[3] DOUG B，ADAM B. Agglomerative clustering of a search engine query log[C]. Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston，Massachusetts，United States，2000. New York：ACM Press，2000：407-415.

[4] 张娥，郑斐峰，冯耕中.Web 日志数据挖掘的数据预处理方法研究[J].计算机应用研究，2004，3(2)：58-60.

[5] 郭新涛，梁敏，阮备军，等.挖掘 Web 日志降低信息搜寻的时间费用[J].计算机研究与发展，2004，41(10)：1737-1747.

[6] 魏小琴，刘慧玲，李明东.朴素贝叶斯分类挖掘技术在零售业的应用[J].中国西部科技，2008，27(7)：28-29.

[7] 任喜峰.基于朴素贝叶斯分类的考试成绩监测机制研究[J].统计与决策，2007，59(22)：163-164.

[8] 张付志，伍朝辉，姚芳.基于贝叶斯算法的垃圾邮件过滤技术的研究与改进[J].燕山大学学报，2009，33(1)：47-52.

[9] 李艳，刘信杰，胡学钢.数据挖掘中朴素贝叶斯分类器的应用[J].潍坊学院学报，2007，7(4)：48-50.

(收稿日期：2010-09-10)

作者简介：

董倩，女，1985 年生，硕士研究生，主要研究方向：人工智能应用技术。

王克俭，女，1971 年生，硕士，副教授，主要研究方向：计算机网络与数据库，人工智能。

韩宪忠，男，1964 年生，硕士，教授，主要研究方向：计算机网络与数据库。