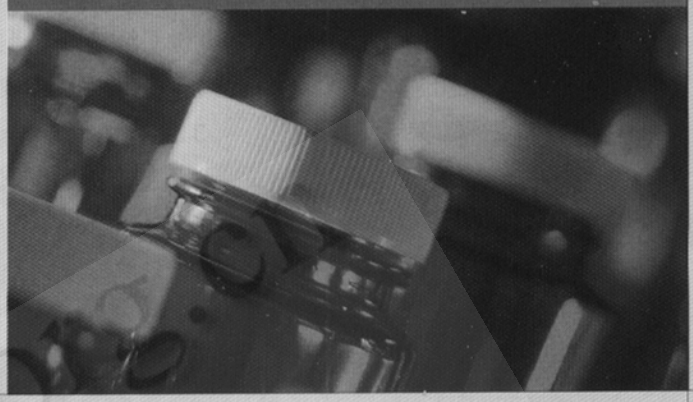
基于数据挖掘技术的药品营销分析系统

曲朝阳 （吉林 东北电力学院信息工程系 132012）
沈 晶 （哈尔滨工程大学计算机科学与技术学院 140001）
宋桂娟 （吉林东北电力学院信息工程系 132012）



Application of Data Mining Technology in Analysis of Medicine Marketing

摘 要：本文介绍了一个应用数据挖掘技术开发的医药营销平台DMMA，从体系结构、系统实现、应用等方面论述了DMMA的设计，对数据挖掘技术的应用进行了有益的探索。

关键词：数据挖掘 数据集市 药品营销

1 引言

随着中国加入WTO，药品行业对外开放迫在眉睫，我国医药商业企业将面临美国等先进国家同行的严峻挑战。企业的经营模式必须从以产品为中心，以销售为中心转变为以客户为中心的模式，产品的购进、生产、销售以客户需求为导向。企业的信息化管理也应成为经营模式改变的支撑，为企业提供客户分析、市场分析、产品销售分析和决策分析的支撑，为提升企业的竞争力服务，为客户提供优质服务。

本文对数据挖掘技术在药品营销分析方面的应用进行了有益的尝试和探索，设计了一个基于数据挖掘技术的药品营销分析平台DMMA(Data Mining for Medicine marketing Analysis).下面从体系结构、系统实现和应用等方面对DMMA进行了详细的介绍。

2 系统体系结构

DMMA以医药营销部门的客户及其交易的数据为源，通过数据清洗、转换、汇总、抽取等技术手段，构建营销数据集市（MADataMart: Medicine Marketing Analysis Data Mart），将业务数据与分析数据隔离。在数据集市MADataMart的基础上，利用数据挖掘技术，构造营销部门的描述模型和预言性模型，以方便企业了解客户消费模式，预测客户行为，发现客户消费趋势，进而来帮助企业管理和决策。

如图1所示，DMMA的体系结构分三个部分。第一部分是医药企业的客户及其交易数据，该层为DMMA提供原始数据；第二部分是DMMA应用服务器，该层首先将第一层的数据进行处理，装载到以客户数据为中心的数据集市MADataMart，然后以数据集市MADataMart为基础，以数据挖掘技术为核心，将分析的结果存放在商业模型库中，应用服务器向客户端提供模型算法的二次开发接口；第三部分是DMMA客户端软件，它通

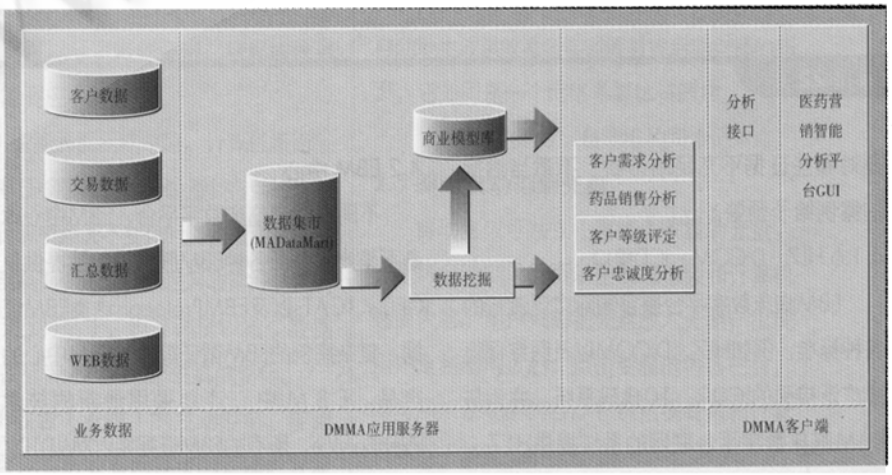


图1 DMMA体系结构图

过对接口的调用创建用户图形接口。

3 系统实现

3.1 系统软件结构

系统采用基于Internet/Intranet的Browser-Server的体系结构，如图2所示。Web服务器采用Windows2000操作系统和Tomcat服务器。数据库服务器采用Windows2000操作系统和Microsoft公司的数据库软件Sqlserver2000。客户端则可根据用户的实际情况选择合适的操作系统和浏览器软件。开发工具是sun公司的jsp（java server pages）。

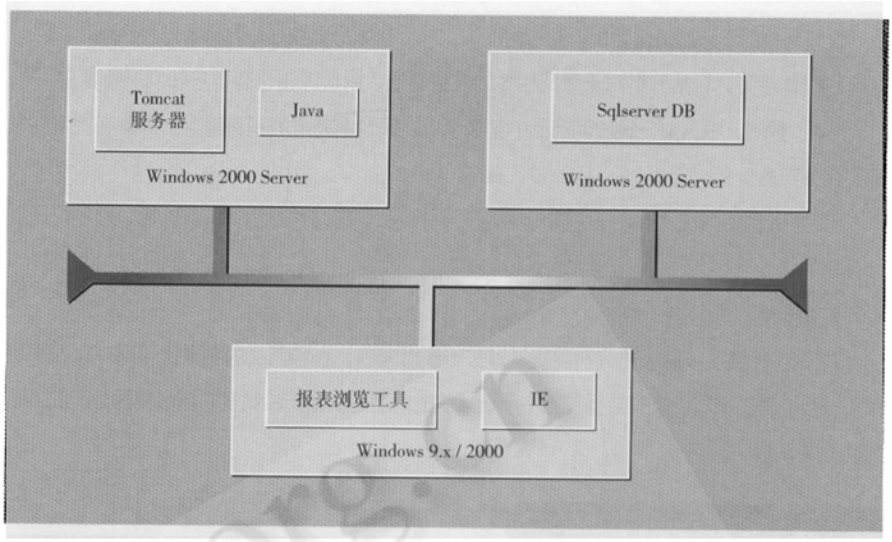


图 2　系统软件结构图

3.2 数据挖掘的实现过程

系统数据挖掘的实现过程如图3所示。主要包括3个部分：

（1）信息来源：数据挖掘的信息来源主要是web数据库，web日志文件，web数据仓库和企业的后台交易信息。

（2）数据过滤：根据这些数据信息，去除一些多余的、无用的信息，以排除不相关数据，增强数据挖掘的准确性。

（3）综合数据挖掘：根据挖掘要求，在数据挖掘算法库中选择合适的数据挖掘算法，并利用这些算法去执行相应的挖掘任务。

4 系统应用实例分析

下面是数据挖掘技术在药品营销分析中应用的一个实例。

4.1 药品营销分析需求

将药品营销分析应用需求进行技术型整理，可以归纳出以下几种数据需求：

（1）数据描述和总结。如根据历史数据分析某段时间某医院对药品需求的变化，确定各医院对药品的需求比平时分别增加多少等等。

（2）运用数据特征化的方法汇总。即归纳出目标类，如：对于企业销售量最大，创造利润最多的重点客户分析其销售（或销售额）占40％以上的药品或销售量增加20％以上的药品分别作为目标类，确定重点客户的偏好，以提供客户服务。

（3）聚类分析。在不预先确定特征化目标的前提下，将重点客户的销售数据进行适当的聚类分析（数据聚类标准在数理统计中体现为不同距离方法，如Markov距离，欧式距离等，不同类型的数据应选择适当的聚类标准），可以获得许多信息，如：哪些客户，哪些地区偏好哪种药品？特别要重视对孤立点分析：对于偏离聚点的对象，即孤立

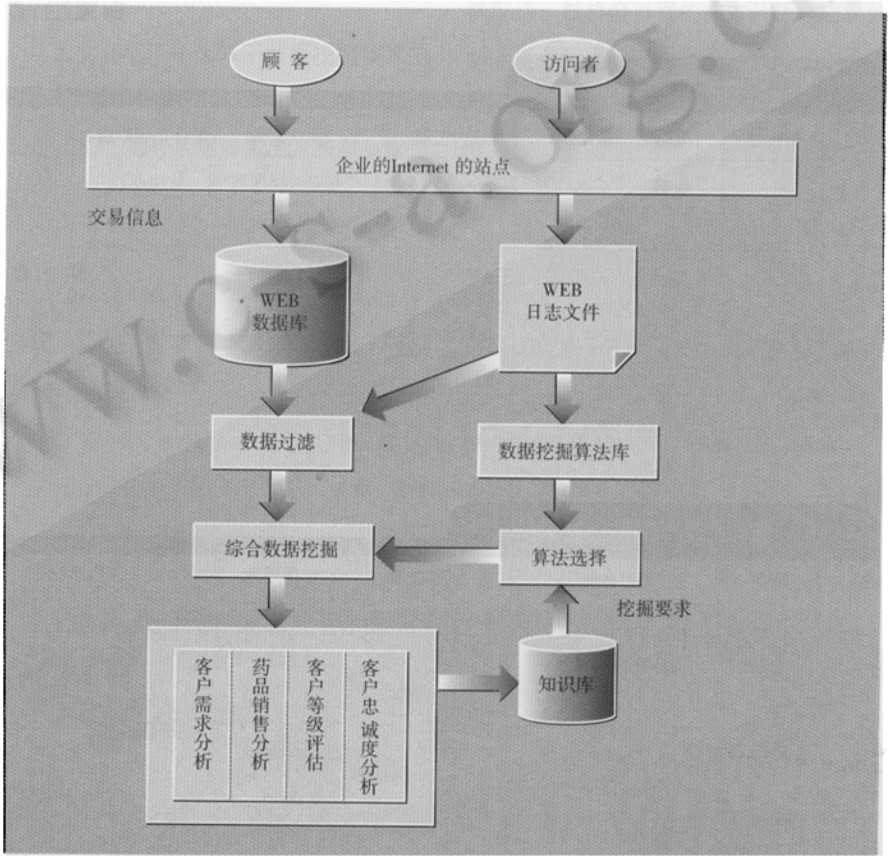


图 3　数据挖掘框图

点，这些客户的忠诚度值得警惕，有可能流失。同时根据历史数据，分析客户和产品销售随时变化的群体趋势，包括分析对于时间相关的数据特征，以便保留已有客户和发展新客户。

（4）预测。根据前一段时间药品的销售情况，来预测将来的销售情况。以此来决定将来购买什么药品。

（5）数据相关性发现。一个客户购买了一种药品，由此判断他得了什么病，他还可能买什么药品。

4.2 鉴别问题

根据实际情况，某医药公司要统计过去一段时间内，哪种药品的销量比较好，以此来决策下一步的进货情况。本例数据挖掘中所要解决的问题是找出药品销量与药品属性之间的关系并建立关系模式。

表1 药品基本信息数据

ID	品名	规格	计量单位	生产企业	效期	剂类	剂型	作用用途	…
1	金樽	3T×6袋	盒	深圳海王药业有限公司	2002-10-10 10	中药饮片	片剂	其它类用	…
2	尼莫地平片	20mg×50	瓶	郑州化学药业有限公司	-6	中药饮片	片剂	心血管脑循环与促智类	… …
…	…	…	…	…	…	…	…		…

表2 药品销售信息数据

ID	销售ID	品名	批号	销售日期	购货单位	销售数量	…
1	1	金樽	030309	2001-5-15	长春大格医药有限公司	30	…
2	2	尼莫地平片	020201	2001-4-106	郑州化学药业有限公司	15	…
…	…	…	…	…	…	…	…

表3 重新构建的数据挖掘数据库

ID	品名	生产企业	销售数量	作用用途	购货单位
1	金樽	深圳海王药业有限公司	30	其他类用	长春大格医药有限责任公司
2	尼莫地平片	郑州化学药业有限公司	15	心血管脑循环与促智类	郑州化学药业有限公司
…	…	…	…	…	…

在本例数据挖掘中，采用了选择树分类器来建立销量与药品作用用途之间的关系模式。

4.3 数据整理

在此模型中，我们主要用到药品基本信息和药品的销售信息，药品的基本信息中保存的是药品的属性，包括名称、规格、计量单位、产地、批准文号、效期、剂别、剂型等等。药品的销售信息包括销售ID、销售单位、销售数量、销售日期、批发价、零售价等。…

此外，要从原始的药品信息数据库提取出数据挖掘感兴趣的相关字段和信息，用两个表的ID连接，重新组合一个数据挖掘数据库，以便在应用数据挖掘时操作方便高效。所提取出来的挖掘数据库信息见表3所示。

4.4 数据分析

在众多的特征之中，我们比较关心的是药品的品名，生产企业，作用用途，购货单位等这些属性，而诸如效期、剂类、剂型、批号等信息与药品的销售数量之间的关系并不是我们所要关注的。因此在数据挖掘工作中可以去掉，以提高数据挖掘的效率和准确性。这样就可以确定数据挖掘中的几个基本的变数：“品名”，“生产企业”，“作用用途”，“购货单位”。

4.5 建立模型

这里采用的分类器是选择树模型。选择树模型是决策树模型的一种扩充，其基本原理和分析方法，与决策树相同，只是选择树提供了更多的选择分支。

决策树构造的输入是一组带有类别标记的数据，构造的结果是一棵二叉树或多叉树。二叉树的内部结点（非叶节点）一般表示为一个逻辑判断，如形式为（ai=vi）的逻辑判断，其中ai是属性，vi是该属性的某个属性值；树的边是逻辑判断的分支结果。多叉树（ID3）的内部结点是属性，边是该属性的所有取值，有几个属性，就有几条边。树的叶子结点都是类别标记。

构造决策树的方法是采用自上而下的递归构造，以多叉树为例，它的构造思路是：如果训练数据集合中的所有数据是同类的，则将之作为叶子结点，结点的内容即是该类别标记，否则，根据某种策略选择一个属性；按照属性的各个取值，把数据集合化为若干个子集合，使得每个子集上的所有数据在该属性上具有相同的属性值；然后在依次递归处理各个子集。二叉树的原理与此类的差别仅在于要选择一个好的逻辑判断。

选择树分类器将每条记录都归入一个类中。归类的基本结构仍为决策树。虽然生成选择树的代价比决策树大，但选择树有两个显著的优点：可理解性更强，可以选择您认为最容易理解的或在具备一定的背景知识的基础上确认更利于作出预测的分支；准确性更高，在很多情况中，选择树比决策树更准确（更低的错误率），综合使用多项选择通常可以作出更稳定，风险更小的分类预测。

在该例中，之所以采用选择树模型，而不是决策树模型，除了具有以上优点外，还因为根据实际情况，树的某些节点其分支信息属性并不唯一，比如各销售数量范围的结点都包含有品名，生产企业，作用用途，购货单位等信息属性。为便于分析，选取销售数量作为目标属性，其它属性作为独立变量。依据销售数量将所有的记录划分为几类。销售数量有以下几类：0，1~50，50~150，大于150。

以这些分类为基础，利用数据集合来生成一个完整的选择树。在生成的选择树中可以建立一个规则基。一个规则基包含一组规则，每一条规则对应选择树的一条不同路径，这条路径代表它经过节点所表示的条件的一条连接，最后构造出来的选择树模型如图4所示。

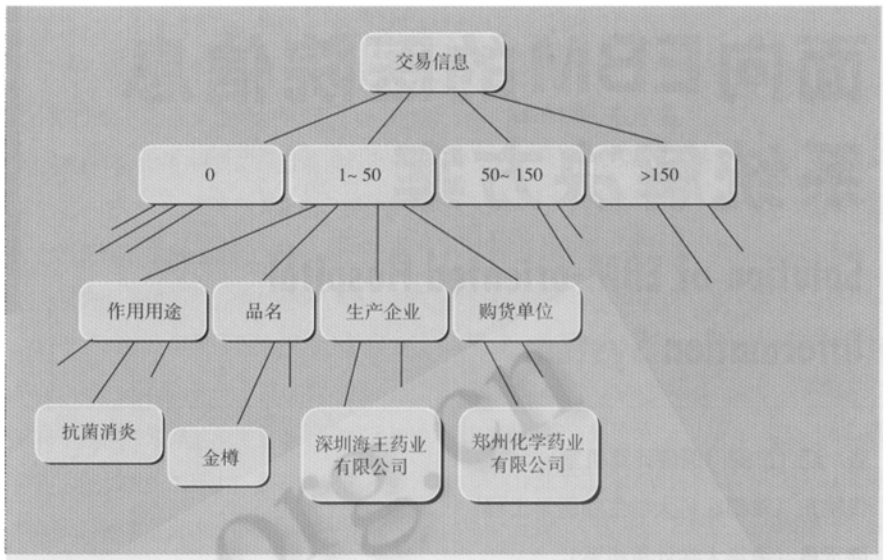


图 4　构造选择树分类器

5 结束语

本文介绍了一个基于数据挖掘技术的医药营销分析系统DMMA的实现，通过给出的应用实例分析可以看出，将数据挖掘技术应用于药品营销分析是必需的和可行的，对药品营销企业来说，建立适合自己企业的各种主题的数据仓库，运用数据挖掘技术深入分析市场，客户，从信息中获取知识，并深入探讨药品购入、仓储、运输配送，销售全过程的成本分析，质量分析，资源配置，对于提高企业管理水平和决策水平也是十分有益的。

参考文献

1 虞文进，数据挖掘技术在烟草企业中的应用，计算机工程，2002.4。

2 U..M.Fayyad, G. Piatetsky-shaperio,P. Smyth and R. Uthurusamy, Advances in Knowledge Discovery and Data Mining , AAAI/MIT Press, 1996。

3 厉小军、朱鸿斌、胡上序，基于B/S结构的第三方物流系统设计与实现，计算机工程，2003.3。

4 何茉勤著，CRM原理、设计、实践，电子工业出版社，2002。