摘　要

随着企业业务的发展，企业积累了大量的客户的历史消费数据资料，如何从这些大量的数据中发现对公司有价值的信息，成为公司将来决策的一个重要的部分。数据挖掘技术已经逐渐应用到了银行、保险公司以及零售行业的数据库销售中，并且取得了不错的业绩。

本文主要从数据库营销的角度来讨论数据挖掘技术，介绍了当前数据挖掘的在数据营销中的应用，总结了数据挖掘技术在数据库营销中的一般的处理流程、数据挖掘算法在数据库营销应用当中出现的问题、主要应用在数据库营销中的算法以及数据库营销的数据挖掘的一般的评价方法。本文提出了一种改进了的决策树算法，并且通过大量的实验验证了市场值函数算法在数据库营销中的有效性。改进的决策树方法通过统计决策树的叶子节点的基本信息，解决数据库营销中因为类分布不平衡而不能生成合适的决策树的问题，同时解决了传统的决策树算法只能对客户分类，不能对客户进行排序的问题，使之可以很好的用于数据库营销。市场值函数算法是起源信息检索并由它扩展而来的一种应用于数据库营销中客户选择方法，它是一种线性模式。这种方法的好处是可以根据市场值对客户进行排序。市场值函数由两部分组成：效用函数和属性权值，通过两者的线性组合可以计算出每个客户的市场值，从而可以对每一个客户进行排序，对客户进行数据库营销。通过在现实数据上的实验，结果证明市场值函数方法是一种非常适合于数据库营销的数据挖掘方法。在市场值函数方法实验的基础上我们建立了一个具有推荐功能的电子商务网站，企业可以通过该系统对客户进行数据库营销。

数据库营销是营销业的一次革命，相信它会随着信息技术的发展能够发挥越来越重要的作用。

关键词：数据库营销 数据挖掘 市场值函数 决策树

Abstract

Nowadays, enterprises have accumulated many customers' information. It becomes more and more important how to get valuable information from the mass data for enterprise to make decision in database marketing. Now many of data mining techniques have been applied in the database marketing such as bank, insurance and retail increasingly.

This paper focuses on database marketing in the data mining techniques. Firstly, this paper introduces the general process of data mining for database marketing, the problems of data mining in database marketing, the major data mining algorithms and the evaluation methods in database marketing. Secondly, this paper proposes a modified decision tree. The decision tree is constructed by the statistical information of leaf nodes of the tree. It can solve the problem that decision tree cannot construct a suitable decision tree because of the extremely imbalanced class distributions in marketing database, and it can rank the objects according to the probability instead of classifying objects. Thirdly, this paper indicates that market value function is good for database marketing by the result of many experiments. Market value function is a linear model to solve the target selection problem of database marketing by drawing and extending result from information retrieval. A market value function is a linear combination of utility functions on attribute values. The value is used to rank individuals. The main advantage of this model is that it can rank objects according to their market value rather than classify. This paper has constructed an E-commerce website with recommendation function based on market value function algorithm. It can help the enterprise to implement

database marketing by network.

Database marketing is a new revolution for the traditional marketing. It will be more and more important for enterprise with the development of information technology.

Keywords: Database Marketing, Data Mining, Market Value Function, Decision Tree

第1章 绪论

随着 IT 技术的迅速发展，社会信息量的不断增加，使得很多公司企业的数据库的规模不断的扩大，产生了海量的数据。为了给决策者提供一个统一的全局视角，在很多的领域都建立了数据仓库，但是大量的数据往往使人们无法辨别隐藏在数据中的、能对决策提供支持的信息。传统的查询、报表工具无法满足挖掘这些决策信息的要求，数据挖掘技术在这种需求下产生了。

数据挖掘是一门新兴的交叉学科，自 20 世纪末提出以来，引起了许多专家学者的广泛关注。数据库中的知识发现 KDD（Knowledge Discovery In Database）是指从大型数据库或数据仓库中提取隐含的、未知的、非平凡的及有潜在应用价值的信息或模式。它是数据库研究中的一个很有应用价值的新领域，融合了数据库、人工智能、机器学习、统计学等多个领域的理论和技术。它不仅能从历史数据中建立回顾型模型，而且还能够建立预测型模型，为我们从大规模的数据库中提取有用信息提供了强有力的解决工具。数据挖掘不但能够学习已有的知识，而且能够发现未知的知识。通过数据挖掘得到的知识是“显式”的，既能为人所理解，又便于存储和应用，因此一出现就得到广泛的重视。计算机中能够存储已知结果的大量不同事实，然后由数据挖掘工具从这些信息里面沙里淘金，将能够产生模型的信息提取出来，并将模型以图、表、公式等人们易于理解的方式表达出来。数据挖掘有广义和狭义之分，广义的数据挖掘指从大量的数据中发现隐藏的、内在的和有用的知识或信息的过程。狭义的数据挖掘是指知识发现中的一个关键步骤，是一个抽取有用模式或建立模型的重要环节。数据挖掘广泛地应用于零售、营销、银行、保险、交通、电信、医疗及故障诊断等许多领域，在市场预测、股票分析、客户行为分析及决策支持等许多方面取得了可喜的成果。随着 CRM 经营理念的迅速发展和数据挖掘技术所带来的经济效益正越来越受到企业的关注，其应用前景也越来越广阔。

CRM 源于“以客户为中心”的新型商业模式，是一种旨在改善企业与客户之

间关系的新型管理机制。通过向企业的销售、市场和服务等部门和人员提供全面、个性化的客户资料，并强化跟踪服务和信息分析能力，使他们能够协同建立和维护一系列与客户以及生意伙伴之间卓有成效的“一对一关系”，从而使企业得以提供更快捷和周到的优质服务，提高客户满意度，吸引和保护更多的客户，从而增加营业额，并通过信息共享和优化商业流程，有效地降低企业经营成本。

数据库营销是客户关系管理（CRM）中的一个比较关键的部分。数据库营销，是在企业通过收集和积累消费者大量的信息，经过处理后预测消费者有多大可能去购买某种产品，以及利用这些信息给产品以精确定位，有针对性地制作营销信息达到说服消费者去购买产品地目的。通过数据库的建立和分析，各个部门都对顾客的资料有详细全面的了解，可以给予顾客更加个性化的服务支持和营销设计，使“一对一的客户关系管理”成为可能。数据库营销是一个“信息双向交流”的体系，它为每一位目标顾客提供了及时做出反馈的机会，并且这种反馈是可测定和度量的。

数据库营销的成功应用离不开数据挖掘技术，我们可以把数据挖掘算法作用在客户数据上，分析客户的购买习性和趋向，预测客户对相应营销方式的响应率，发现有利顾客的特征，有目的性的开展广告和销售业务。通过对顾客的忠诚度分析，相应调整商品的价格和类型，改进销售服务，有利于保留现有客户，寻找潜在的客户。扩大销售的范围和规模，从而增加销售量。通过在线销售的数据，得出产品关联的商用信息和客户的购买习惯，使进货的选择与搭配更具科学性。

1.1 课题背景

近代营销在历经种种概念变换之后，关注的焦点终于回到了营销活动的主体——人与人的关系上。关系营销是指建立维系和发展顾客关系的营销过程，目标是致力建立顾客的忠诚度，它有别于传统的交易营销，要为顾客增加各种服务的附加值。在这种营销方式下，营销者就必须花费精力对每个顾客进行研究，力求进行“一对一的沟通”，这就要求企业要建立一个先进的顾客数据库，以便更好地了解顾客，为顾客提供其所需要的产品设计和劳务，加强同顾客的忠诚关系。

特别是当市场竞争日趋激烈时，顾客成为企业关注的焦点，如何争取和留住顾客将是企业营销工作的主题。这就需要营销者站在顾客的立场上及时了解顾客的需求及其变化。要依照消费者的价值观念来设计、生产、定位产品，在很多情况下，无法吸引到顾客或失去顾客往往不是产品的质量问题，而是顾客对服务的不满，因此，产品的服务化和服务的产品化应该是高度融合在一起的。提供优良的服务，建立起顾客对企业的忠诚度，就需要把消费者的价值观念贯穿于企业的整个经营过程中，企业的各个部门将被高度地整合起来，以顾客为中心开展工作，另一方面，消费者的需求、价值观念又会在与市场环境的互动中不断的改变着，当这种变化的频率越来越高，那种传统的单向沟通的营销方式已经力不从心，就需要新的双向沟通的营销方式取而代之，建立起顾客与企业间的长期稳定的互动关系。而信息技术的发展为这种双向沟通的方式提供了强有力的支持，畅通的信息沟通与共享使企业的各个部门、顾客以及各种环境因素融为一体，这就使得数据库营销应时而生。

1.2 基于数据库营销的数据挖掘的现状

基于数据挖掘的数据库营销，常常可以向消费者发出与其以前的消费行为相关的营销材料。目前，将数据挖掘应用到数据库营销上已经有了很多成功应用的案例。在市场经济比较发达的国家和地区，许多公司都开始在原有信息系统的基础上通过数据挖掘对业务信息进行深加工，以构筑自己的竞争优势，扩大自己的营业额。数据库营销在西方发达国家的企业里已相当普及，在美国，1994 年 Donnelley Marketing 公司的调查显示，56%的零售商和制造商有营销数据库，10%的零售商和制造商正在计划建设营销数据库，85%的零售商和制造商认为在本世纪末，他们将需要一个强大的营销数据库来支持他们的竞争实力。从全球来看，数据库营销作为市场营销的一种形式，正越来越受到企业管理者的青睐，在维系顾客、提高销售额中扮演着越来越重要的作用。

而在国内基于数据库营销的数据挖掘的研究才刚刚起步，目前还有很多的事情要做，这也是我们从事这个研究课题的出发点。

1.3 本课题的主要研究内容

本课题的主要出发点是为企业决策者提供销售的决策支持，从大量的企业的销售历史数据中发现潜在的客户群，使企业可以采用适当的销售方式。我们的主要工作是根据不同营销方式选用合适的数据挖掘算法，分析客户的购买习性和趋向，预测客户对相应营销方式的回应率，进而为企业数据库营销提供决策支持。

第2章 数据挖掘与数据库营销

2.1 介绍

数据库营销是营销领域的一次重要变革，是一个全新的营销概念。所谓数据库营销（Database Marketing），就是企业通过搜集和积累消费者的大量信息，经过处理后预测消费者有多大可能去购买某种产品，以及利用这些信息给产品以精确定位，有针对性地制作营销信息，以达到说服消费者去购买产品的目的。

随着数据量的急剧增长，现在的用户很难再像从前那样，自己根据数据的分布找出规律，并根据此规律进行分析决策。因此必须借助于相应的数据挖掘工具，自动发现数据中隐藏的规律或模式，为决策提供支持。数据挖掘技术主要用于从大量的数据中发现隐藏于其后的规律或数据间的关系。

作为一种新的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性数据。数据挖掘技术在企业市场营销中得到了比较普遍的应用，它是以市场营销学的市场细分原理为基础，其基本假定是"消费者过去的行为是其今后消费倾向的最好说明"。

通过收集、加工和处理涉及消费者消费行为的大量信息，确定特定消费群体或个体的兴趣、消费习惯、消费倾向和消费需求，进而推断出相应消费群体或个体下一步的消费行为，然后以此为基础，对所识别出来的消费群体进行特定内容的定向营销，这与传统的不区分消费者对象特征的大规模营销手段相比，大大节省了营销成本，提高了营销效果，从而为企业带来更多的利润。

从现在的情况来看很多的企业都建立起了自己的数据仓库系统，数据仓库中包含了大量的客户和企业业务的历史数据。如何从大量的历史数据中挖掘出有用的信息并为企业的决策支持服务，对于企业未来的发展具有很重要的意义。面向市场营销的数据挖掘技术，已经逐渐应用到银行、保险公司以及零售行业。随着

商品种类增多以及人们需求多样化的市场趋势，传统的销售方式不能完全适应市场的发展了，而且随着市场竞争的加剧，浪费的问题变得越来越严重，节约成本减少浪费成为商家必须考虑的一个重要问题。正是由于上述问题，越来越多的人参与到了面向市场的数据挖掘系统的研究，数据挖掘技术和数据仓库技术的发展正好给该问题的解决提供了条件，两者的结合可以很好的解决目标市场中的数据库营销的问题。

基于数据挖掘的营销，常常可以向消费者发出与其以前的消费行为相关的营销材料。目前，将数据库营销技术应用到目标市场问题上已经有了一些成功应用的案例。在市场经济比较发达的国家和地区，许多公司都开始在原有信息系统的基础上通过数据挖掘对业务信息进行深加工，以构筑自己的竞争优势，扩大自己的营业额。美国运通公司(American Express)有一个用于记录信用卡业务的数据库，数据量达到 54 亿字符，并仍在随着业务进展不断更新，运通公司通过对这些数据进行挖掘，制定了“关联结算(Relation ship Billing)优惠”的促销策略，增加了商店的销售量。卡夫(Kraft)食品公司通过收集对公司发出的优惠券等其他促销手段做出积极反应的客户和销售记录建立了一个拥有 3000 万客户资料的数据库，通过数据挖掘了解特定客户的兴趣和口味，并以此为基础向他们发送特定产品的优惠券，并为他们推荐符合客户口味和健康状况的卡夫产品食谱。美国的读者文摘(Reader's Digest)出版公司运行着一个积累了 40 年的业务数据库，正是基于对客户资料数据库进行数据挖掘的优势，使读者文摘出版公司能够从通俗杂志扩展到专业杂志、书刊和声像制品的出版和发行业务，极大地扩展了自己的业务。

2.2 数据库营销中的数据挖掘的处理过程

一般来说数据库中的真正购买商品的客户的比例是很小的[1, 2]，通常为 1%~5%，由于数据库营销的特殊的情况，因此在进行数据库营销选用算法、数据以及数据挖掘处理的时候都有特殊的要求。通常面向数据库营销的数据挖掘系统的处理过程如下（见图 2-1）：

数据预处理 从数据仓库中获取的历史数据在很大程度上存在着不完整、有噪音以及不一致的问题，不能直接应用在数据挖掘算法中。数据的预处理除了要解决数据的不完整、有噪音和不一致的问题还包括转换地址和区域代码以及处理属性值丢失的问题，同时要求从数据的属性集中找出关键属性集，以精简数据，提高运算速度。如果选用的算法不能处理连续的属性，还要对连续属性进行离散化操作。

生成模式 把经过预处理的数据分成训练例集合和测试例集合，在实验中我们可以把训练例集合和测试例集合的大小按照 1:1 的比例分配，将选用的学习算法作用在训练例集合上，得到一个新的模式，再把生成的模式作用在测试例集合上，对模式的结果做评估，如果模式不能满足要求，则修改算法并重新作用在训练例集合上，生成新的模式并再次在测试例集合上作评估，直到找到满意的模式为止。

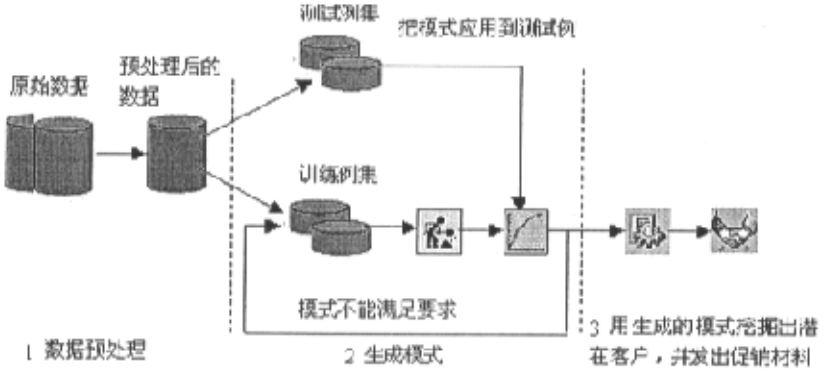

图 2-1 面向市场数据库营销的数据挖掘过程

Fig2-1. The Process of Data Mining in Database Marketing

应用模式 把生成的模式作用在企业的客户数据库中，依次对客户数据进行处理，按照客户成为潜在客户的概率的大小自高到低进行排序，一般如果生成的模式有较好的效果的话，具有高概率的客户会集中在序列的上部。企业可以根据客户的概率大小，灵活的选择前面的客户，并且可以采用不同的联系方式，如前面的 10%客户采用电话联系，对后面 20%的客户采用发邮件的方式联系。

2.3 数据挖掘在数据库营销中表现出的主要问题

数据库营销作为一种新的营销方式有它自己的特点：一般来说营销数据库中客户对商家所推销的产品真正感兴趣人是比较少的（通常 1%），我们需要从大量的客户中找出这一少部分的客户；营销数据库中存在海量的客户的消费历史数据，要求系统能较快的处理这些数据；需要最大限度的考虑到公司的最大营销利润；公司需要针对不同消费特点的客户采用不同的销售策略。由于数据库营销有这些特性，很多的传统数据挖掘方法，在应用到目标市场销售的时候出现了一些问题，这些问题主要表现在以下几个方面：

第一 最明显的问题就是营销数据库中类（包括购买者和非购买者两类）的分布的极端不均衡[1]。最典型的就是数据集中只有 1%的正例（购买者），而其余的则均为负例（非购买者）。很多的学习算法对这种数据集表现得都不是很好，有些算法只是发现一条简单的规则或者将全部的例子判为负例，而这条规则在训练例和测试例上都可达到 99%的精度，而所发现的这样的规则在实际应用中作用比较小。在最近几年很多的数据挖掘和机器学习的研究者都意识到了并且也开始研究这个问题。

第二 即使找到一种合理的模式，传统的数据挖掘中采用的预测精度这种标准已不合适数据库的营销模式的评价[1]。其中一个原因是对误判的处理方式应不一样：把负例判成正例实际上是我们的目标，而把正例误判成负例却是不能接受的。另一个原因就是预测精度对于挖掘客户的作用太弱，二元分类法只能预测购买者和非购买者这两类，而不能在预测的购买者和非购买者之间作一个更好的区分。并且在从预测的购买者中选定一定比例的最可能的购买者作为促销对象时，该方法不够灵活。比如我们选择前 100 位最可能的购买者采用电话的方式促销，而对后面的 1000 位采用发 Email 的方式促销。

第三 数据库营销中数据挖掘算法的效率是一个需要重视的问题。当我们把整个数据集分成大小相等的训练例集和测试例集的时候，训练例集的数据量对于选定的学习算法可能过大，需要消耗较多的时间，在实际应用当中，也可能会因

为算法效率太低而错过最佳的营销时间，给企业造成不必要的经济损失。因此对于相应的数据集应该选用高效率的学习算法。

第四 在已知的算法中，几乎所有的算法都只是考虑了回应率，而在实际生活中从事营销的商人所感兴趣的是获得最大的利润，而不只是回应率[3]。实际生活中回应率在某些时候并不代表利润值，商家从某些回应者的销售中获得的利润可能是其他回应者的好几倍，因此从实际情况考虑，算法不应只考虑回应率，而应更多的考虑企业的利润。

第五 大部分的算法都没有考虑回应的不同方式，一概认为做出回应的客户的回应方式都是一样的，而在实际生活中这是不现实的[3]。不同的客户对同一个促销的材料所做出的反应方式是各异的，有的可能购买廉价的商品，而有的就可能购买奢侈品，两种不同的回应方式给商家带来的利润也是不一样的，而实际中很多算法都没有考虑这一点。

第六 一般给出的关于客户的信息的数据集的属性的太多，不便于选取，而这又是因为正负例分布的极端不均匀的结果[4]。例如如果采用信息增益的方法来选取属性，负例占据 99%的比例，初始的熵值就比较小，因此在选用不同的属性求信息增益时，信息增益值的大小差别不大，导致选不出关键的属性。在实际中一般常用的认为比较重要的几个属性是 RFM(Recent of the last purchase, Frequency of purchase, Monetary Value)，但是这几个属性对于还不是顾户的客户来说，这些客户的资料在公司的数据库里是不存在的。

在数据库营销当中我们应根据营销数据库的不同情况选用相应的数据挖掘算法，尽量避免陷入这些问题当中，达到最佳的营销效果。

2.4 数据库营销中常用的数据挖掘算法

数据挖掘的核心是采用机器学习、统计等方法进行知识学习的阶段。数据挖掘算法的好坏将直接影响到所发现知识的好坏，因此选取适当的算法或算法组合至关重要。

根据对数据挖掘的认识的不同，人们对数据挖掘技术有不同的划分。第一种

根据学科的不同，把数据挖掘方法分为两类[5]：统计模型和机器学习技术。统计模型应用于数据挖掘主要是进行评估，常用的统计技术有概率分布、关联分析、回归、聚类分析和判别分析等；机器学习是人工智能的一个分支，通过学习训练数据集，发现模型的参数，并找出隐含的规则。第二种是根据客户选择方式的不同，把数据挖掘划分为两类：分割技术（Segmentation）和回应模式(Response Modeling)[3]。分割模式是通过算法把个体分成不同的群体，群体内部的个体尽量相似，群体与群体之间尽量具有最大的不同，从群体中选取较高的回应概率的群体作为发送促销材料的客户。回应模式是采用算法对每一个个体计算出一个目标值（Target Scoring），选择具有最高的目标值的个体作为发送促销材料的客户。两种不同划分出发点不同，各有合理的一面，不过因为面向数据库营销的数据挖掘方法更注重潜在客户的选择，第二种划分更适合面向数据库营销的环境。

下面主要从分割模式和回应模式两种不同的角度介绍目前数据库营销中的几种主要的数据挖掘研究方法。

2.4.1 分割模式

最常使用的分割技术是聚类分析技术，如 AID、CHAID 和 CARD 方法。通常这三种方法是构造一棵决策树，决策树是定义布尔函数的一种方法，其输入为一组属性描述的对象，输出为 yes/no 的决策。树的每一个内部节点（包括根节点）是对输入的某个属性的测试，此节点下面的各个分支被标记为该属性性质的各个值。每个叶子节点表示达到该节点时布尔函数应返回的 yes/no 值。决策树的每一个节点上，就是对客户的一个划分。

单纯的决策树方法并不能完全的解决数据库营销的问题，因为它只能求出一些分类规则，而不能按照客户成为潜在客户的概率值对客户排序，以便选择潜在的客户。为此，有人提出带确信因子（Certainty Factor）的 C4.5 算法。确信因子是指符合每个分类的训练例在整个训练例上的比例。Ling 也使用了带有确信因子的 C4.5 算法作为解决数据库营销该问题的方法[1]。另外，由于决策属性的分布的绝对不平衡，很难构造出一棵合适的决策树，并且构造出来的规则可能很简单，不具有实用价值[4]。

2.4.2 回应模式

回应模式通常是计算每一个的客户的可能做出回应的概率值，从而每一个客户都有一个目标值（Target Scoring），可以根据目标值大小从高到低进行排序，然后选取前面的概率值高的客户作为发送促销资料的客户。在数据库销售中很多数据挖掘方法都是属于回应模式，下面介绍几种常见的回应模式数据挖掘算法。

Market Value Functions[6, 7] 该方法是源于信息检索（Information Retrieval）并由它发展而来的一种应用于目标市场销售的客户选择方法，它是一种线性模式。这种方法的好处是可以根据 Market Value 的值得大小对客户进行排序。市场值函数(Market Value Function)由两部分组成：效用函数(Utility Function）和属性的权值。效用函数源于信息检索的概率模型，属性的权值是基于信息理论的属性重要性的衡量。在挖掘市场值函数方面有很多的方法。一般市场值函数表示为：

$$r(x) = \sum_{a \in Al} \omega_a u_a(I_a(x)) \qquad (2\text{-}1)$$

在这里ω_a是属性 a 的权值，权值ω_a可能为正、为负或者为零，效用函数为$u_a(I_a(x))$。如果属性具有较大权值就说明该属性比较重要，属性的权值如果接近零或为负就说明该属性不太重要，个体 x 的 Market Value 值就是所有属性的 u 函数与ω_a乘积的加权和，在这里假设所有个体的属性都是独立的，最终可以根据每一个客户的市场值大小对客户进行排序。

该方法的优点是：可以对客户按照市场值的大小排序，而不是简单的分类；市场值函数具有可解释性；系统的执行不需要专家的指导。但是它也有它的不足之处：只是考虑了最大的回应率。

Logit/Probit 模式 Logit/Probit 模式主要是用来处理二元（1/0）类型，因此很自然的它适合于应用在数据库销售中的 yes/no 的回应模式。这两种模式的不同点在于 ε 在等式中的分布方式不一样，Logit 模式假设它是对数分布，Probit 模式假设为正态分布。公式中假设每一个个体 i 在时间 t 时刻对邮件都有一个做出回应的趋势r_{it}^{*}，这个趋势受 $x_{k,i,t}$影响：

$$r_{it}^{*} = \beta_0 + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + ... + \beta_k x_{k,i,t} + \varepsilon_{it} \qquad (2\text{-}2)$$

如果r_{it}^{*}>0 表示个体 i 将做出回应，r_{it}^{*}<0 则不做出回应.

上述的公式假设回应的方式都是相同的，通常实际情况并不是那样。例如那些通常不做出回应的客户，一旦作出回应时，他们花费的数目可能比那些经常做出回应但每次花费很少的人要多得多。另外 Logit 和 Probit 模式的缺点是他们只是一种回应模式，选择潜在客户是基于最大的回应，而不是收入。后来 Bult 和 Wansbeek 在 logit 模式中加入了一个断点，通过断点只有选择这些个体时才能获得最大的利润。利润函数定义如下：

$$\Pi_i = rR_i - c \qquad (2\text{-}3)$$

Π_i为由个体 I 产生的纯利润，r 为做出积极的回应所获得的收入，c 为邮资，R_i是个体 i 的二分参数的回应值。通过将个体的回应概率乘以估计的收入，可以获得总的收入的估计。

遗传算法(GMAX) [8, 9] 遗传算法 GMAX 模式的主要目的是建立一个模式（等式）来解决最优化的问题，每一个模式都有他相关的适应度值，表明该模式对于解决问题的适应程度，具有高的适应度值的模式在解决问题上比低适应度值的模式要好。一般遗传算法（GMAX）都是按照以下步骤执行的：

- 定义适应度函数，适应度函数应该能描述什么是好的模式，并且能够识别和修改好的或者坏的模式；
- 选择函数的集合（例如算术操作符{加、减、乘、除}；对数和指数）和你确信与解决手头的问题有关的变量，如用于预测的变量 X_1、X_2...X_n以及数字常数，用预先选择的函数集和变量采用随机模式生成初始种群；
- 把模式应用到训练例集中计算种群中每一个模式的适应度值；

按以下三种操作建立一个新的种群，这些操作是根据当前的种群中模式的适应度的大小来选取的，也就是说最可能被选择到的模式，就是最适应解决问题的模式。三种操作包括：

- 复制 把已经在种群中存在的模式复制到新的种群中；
- 交叉 随机选择两个已经存在的父辈模式通过遗传重组为新的种群创建两个后代；
- 变异 对某些模式作随机的改变。

在产生的新一代中具有最高适应度值的模式被认为是最好的模式，它可能能够解决或者大致能解决要解决的问题。适应度函数选择了 R 平方公式。在建立初始种群的时候通常采用公平的函数一变量轮盘赌模型，也就是说函数和变量具有相同的被选取概率值。在复制到下一代的时候，模式复制到下一代的概率根据它的 PTF（PTF=Fitness Value/Total Fitness）值来获得，也就是不公平的轮盘赌模式。交叉和变异操作模式的选取概率同样基于 PTF。其他按照传统的遗传算法操作即可。

跟许多方法一样，遗传模式也有它的优点和它的局限。遗传算法的优点是具有较好的鲁棒性、不需要假设、非参数模式并且对大的和小的数据样本都有好的效果，只需要定义一个适应度函数。遗传模式对解决大的优化问题和对大的数据集查询上都显示了它的有效性。同时遗传算法在设置遗传模式的参数：种群的规模、复制的比例、交叉和变异的比例上有它的潜在的局限性。这些参数的设置在很大程度上依赖于解决的问题和数据，合适的设置需要一定的经验。

神经网络算法[10] 神经网络与决策树的方法不同，它不是把数据集分成不相交的几部分，由于目标选择的问题是设法从非回应者中区分回应者，因此可以把这个问题归为分类问题，通过设计网络把每一个输入的客户分为回应者和非回应者。在方法中定义了一种回应的可能性的度量标准，亦即在模式中设置了一个阈值，只有输出值高于阈值的客户才给发邮件。将这类方法表述为目标值（Target Scoring），回归模型和神经网络算法都归为这一类，从实用的目的来看，这种可能性被解释成支持者的回应概率，在这里假设支持者的顺序是按照回应的可能性的估计的度量值来决定的，神经网络的输出因此是一个回应概率值的度量。在训练例上主要是建立神经网络的各个层次的参数，神经网络方法实质上是一种非线性的回归模型。神经网络的复杂程度是和所解决的问题的非线性程度相关的，由

于不要考虑到动态性，因此采用周期性的神经网络。从实用的角度来说，前馈神经网络就足以解决目标选择的问题。决定前馈神经网络的参数是隐藏层的数目以及每一层神经元的个数，对于这些参数可以采用很多种方法来选择，使这些参数保持在正确的范围内，如生长和剪枝、启发式搜索以及采用遗传算法进行优化。

算法的优化：由于神经网络算法通常采用梯度下降法，导致算法终止在局部最小而不是全局最小，因此神经网络算法中的一个很大的问题就是解决局部最小问题。对于一个给定的问题为了让算法终止在全局最小值，很难初始化神经网络的权值。为了解决这个问题其中的一个方法是采用遗传算法来决定神经网络的初始权值[11]。

贝叶斯方法 该方法是基于贝叶斯定理，有一种常用的方法叫朴素贝叶斯方法，之所以被称作朴素贝叶斯方法是因为在采用该方法的时候假定每个属性之间是相互独立，互不影响的。Elkan 的 BNB（Boosted Naïve Bayesian）算法，则将推进（Boosting）技术和朴素贝叶斯方法相结合[12]。其基本思想是对每个训练例赋予一个权值，利用朴素贝叶斯方法在该训练例集上得到一个分类器 C1，将 C1 分类错误的训练例的权值增大，用新权值的训练例集通过朴素贝叶斯方法得到新的分类器 C2，显然 C2 更加注重 C1 分类错误的训练例，如此循环 T 次，得到 T 个分类器。预测测试例时用每个分类器的准确度来做为权值，将 T 个分类器的输出加权求和作为最后的输出。Ling 使用了 BNB 方法来解决目标市场的数据库销售的问题。此外，在属性之间有一定联系的情况下，可以使用贝叶斯网（Bayesian Networks）的方法来解决。贝叶斯网（Bayesian Network）由两部份组成，第一部分是有向无环图，图上每个节点代表一个随机变量，节点与节点之间的弧代表一个概率依赖；第二部分是每个属性一个条件概率表，以此来表示其相应的概率分布。Paola 使用了一个可以构建贝叶斯网络的软件 Bayesware Discoverer 来向一个慈善公司分析其邮件响应率。他们的方法是：从训练例中建立一个贝叶斯网来求出顾客对邮件响应的概率 P；然后从记录已经捐过款的人的信息的数据库中建立另一个贝叶斯网来求出顾客可能捐款的数量 E(D)，最后根据公式 $P=-0.68\times(1-P)+P*E(D)$ 来判断是否给该顾客发邮件。P 为正值就表示

可以为该顾客发邮件。

粗糙集方法 粗糙集方法提供了很多挖掘隐藏在数据中的模式有用的工具，它可以用于知识发现中不同阶段的很多的方面，如属性选择、属性提取、数据约简、决策模式的产生和模式的提取。粗糙集中的 ProbRough 系统结合了粗糙集中的基本的方法[13]，它可以用来分析潜在的客户和购买的预测。粗糙集算法分类器的生成包括两个步骤:第一个步骤是在属性空间上的全局分割；第二个步骤是约简决策规则的数量。ProbRough 系统在有冗余属性的训练数据上表现很好，它可以解决两个主要的问题：先验概率和不相等的错误分类代价的问题。

从上面所述的数据挖掘算法来看，每种算法都有他们的最佳的适应范围，在应用时，我们应该根据所选用的数据集和市场营销的具体环境来选择最佳的算法来达到最佳的营销效果。

2.5 数据库营销中常用的评价方法

数据库营销的主要目的就是希望通过挖掘出最好的客户来设法提高销售的效用，因此在设计模式的时候总是希望能从营销中最大可能地识别回应者或者最大利润，所提供的模式都提供了每一个个体的回应概率以及利润分布的估计值。前面提到的在目标市场营销中已经不能再采用通常的预测的精确度作为数据挖掘算法的评价标准，因此提出了几种新的评价标准，一般我们采用 Decile 和 Lift 评价方法[4]。

2.5.1 Decile 分析方法

利用数据挖掘算法对测试数据集进行预测后，我们会得到每个实例即每个消费者可能做出响应的概率。我们引入 Decile 分析方法，Decile 分析是采用一种表的方式来显示生成的模式的性能。回应模式的 Decile 分析通常由以下步骤组成：

1. 用选择的模式计算样本的分值，每一个个体将获得一个估计回应的概率值；
2. 根据测试结果集中的个体概率对测试例进行降序排序，将排序后的结果分成数量相等的 10 组，从顶部到底部分别为 1，2，3，4，5，6，7，8，9，10，

Decile 表的第一组包括最可能做出响应的前面的 10%的消费者个体。第二组是接下来最可能响应的第二部个 10%的消费者个体，对剩余的消费者个体依照购买可能性的顺序继续以整体的 10%的数量分组，从而得到 10 组数据。

表 2-1. Decile 表

Table2-1. Decile Table

组	每组人数	每组平均回应人数	每组实际回应人数	每组回应率	累计回应率	Lift 值
1	139	32	115	82.20157	82.20157	353
2	139	32	75	53.60972	67.90564	291
3	139	32	52	37.16940	57.66023	247
4	139	32	30	21.44389	48.60614	209
5	139	32	28	20.01429	42.88777	184
6	139	32	20	14.29592	38.12247	164
7	139	32	5	3.57398	33.18697	142
8	139	32	1	0.71479	29.12795	125
9	139	32	0	0	25.89151	111
10	139	32	0	0	23.30236	100

3. 每组人数 表示的是消费者的数量，消费者数量是指每层消费者个体的总和，该数量为全部消费者个体的 10%取整；
4. 每组平均回应人数 指测试数据集在不采用数据挖掘算法时回应客户在每个组中的平均客户的数量；
5. 每组实际回应人数 指每组中在采用数据挖掘算法后实际做出回应的客户的数量。我们可以看到在 Decile 表中的第一组累计有 115 名消费者做出了响应，在第二组累计有 75 人做出了响应；
6. 每组回应率 表示的是每组的实际响应率，该数值表示 Decile 表每组的实际响应的人数与该层消费者的数量的比率。即：每组实际回应人数/每组平均人数。如在第一组中，该比率为 82.20157%，他的计算方法为实际响应人数 115 除以该层客户的数量 139，同理在第二层中，该比率为 53.60972%，他的计算方法为该层实际响应人数 75 除以该层的消费者的数量 139。其他各层计算方法一样。

7. **累计响应率** 是指从该层到顶层的响应率的累积，该比率等于该层到顶层的实际响应人数的和除以该层到顶层的总的消费者的数量，即该层到顶层的实际回应人数的和除以该层到顶层的每组人数的和。例如第一组的累积响应率为 82.20157%，它的计算方法为该层到顶层的实际响应人数 115 除以该层到顶层的全部客户的数量 139；第二层的累积响应率值为 67.90564%，其计算方法为该层到顶层的实际响应人数的和(115+75)除以该层到顶层的全部消费者的数量(139+139)。其他各层计算方法类似。

8. **累计 Lift 值** 该比率的计算方法为该层的累积响应率除以该测试数据集响应率后乘 100。该测试数据集响应率即该集合中全部响应者的数量与全部消费者的数量的比值。此值是应用该数据挖掘得到的模式进行目标市场与不应用该模式而进行销售的效果比较的体现，它能够反映出如果您使用该模式的优势。例如，Decile 分析方法表格的第一层中此值为 353，这个数值意味着如果我们向该层中这 10%的消费者进行销售活动，我们得到的响应人数会是不对该数据集使用任何模式而对其中任意 10%的消费者进行消费活动而得到的响应人数的 3.53 倍。又如在示例表格第二层中此数值为 291，此数值是由该层到顶层整体决定的，他意味着当向此层到顶层的 20%的消费者进行销售活动所得到的响应人数会是不应用模式而任选该测试数据集中 20%的消费者进行销售活动而得到的响应人数的 2.91 倍。其他各层该数值的意义也就不难得出，当到达第 10 层的时候我们发现该数值变为 100，也就是说采用和不采用模式得到的响应人数是一样的，这是显然的，因为在我们的测试数据集中购买者的数量是一定的，对所有人都进行数据库营销和对所有人都进行大规模销售后响应者的数值当然是一样的。不过这样就失去了数据库营销的意义。但是我们仍计算出所有层的 Lift 响应指数，这样我们能更直观更清晰的看到数据挖掘带给市场营销带来的益处。

如果该数据挖掘生成的模式经过测试数据集的测试后生成的 Decile 表呈现出顶部层的反映人数较低层的反映人数多，并且累计 Lift 值从顶层到底层呈递减的趋势，那么我们可以认定该模式确实适合应用在该类数据集上，应用该种数

据挖掘算法在数据库销售中是成功的。

2.5.2 Lift 方法

在数据库营销中我们希望越靠近顶层的响应人数越多越好，通过 Lift 评价方法可以反应这响应人数的分布情况[1]。具体来说，Lift 值是通过 Decile 表中的每组实际回应人数进行带权值的求和获得的，我们把排在最前面的组赋予一个大的权值，权值的大小依次递减，通过这样求和，可以体现出实际回应人数在排序的表中的分布情况。假设 Decile 表格中从顶层到底层的实际响应人数为 $S_1, S_2, \cdots S_{10}$，则 Lift 值的计算公式如下：

$$Lift = (1\times S_1+0.9\times S_2+... + 0.1\times S_{10}) /(S_1+S_2+...+S_{10}) \qquad (2-4)$$

例如上面的 Decile 表的 Lift 值可以计算如下：

Lift=(1×115+0.9×75+0.8×52+0.7×30+0.6×28+0.5×20+0.4×5+0.3×1+0.2×0+0.1×0)/(115+75+52+30+28+20+5+1+0+0)= 0.8411043

Lift 值的大小反应的生成的模式的优劣程度。如果分布在上层的人数越多，则 Lift 值越高，模式越好，反之，则模式不好。另外根据 Decile 表我们可以得到累计 Lift 曲线图（如图 2-2）。

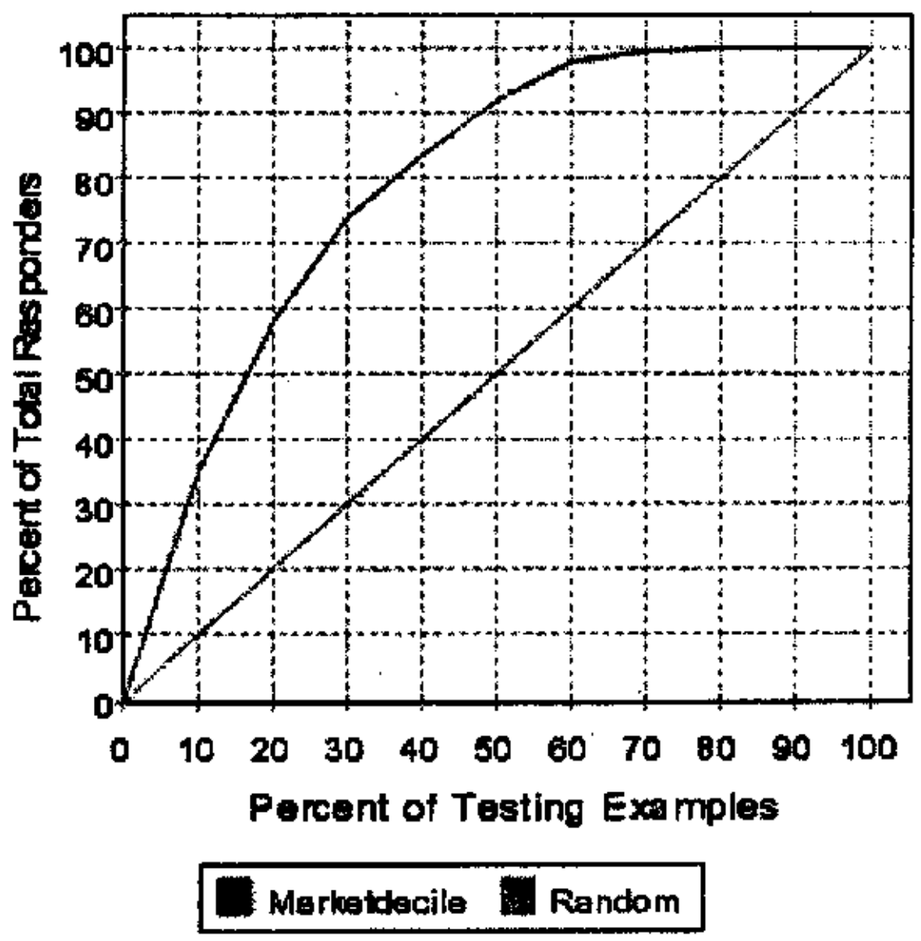

图 2-2 累计 Lift 曲线图

Fig2-2. Accumulative Lift Curve

如果购买者在这 10 组中随机分布（没有发现任何规律），那么这条递增曲线就会与对角线重合，Lift 值就是 0.5。当然，如果这任意分布的 10 组有可能是最好的情况即当$S_1 = \sum_i S_i < 10\%$，此时 Lift 指数为 1，最坏的情况是$S_{10} = \sum_i S_i$，其他的 S_i 都为 0，此时 Lift 指数为 0.1，但是针对数据库营销，不可能向最后一层的消费者进行销售活动，所以如果 Lift 指数值为 0.1 时即认为其值为 0。Lift 指数的优点就在于它不依赖于回应人数的多少。在任意分布的情况下，Lift 指数的值为 0.5，大于 0.5 就是比任意分布的情况好，小于 0.5 就是比任意分布时的情况差。Lift 指数值较高的算法会更适合解决数据库营销中的问题。

2.6 本章小结

本章从总体上介绍了数据库营销和数据挖掘之间的研究现状。介绍了目前数据挖掘技术在数据库营销中的主要处理过程以及在应用过程中出现的主要问题。介绍了目前数据库营销中的几种主要的数据挖掘算法和应用于数据库营销中的两种主要的评价方法：Decile 分析和 Lift 值评价方法。通过本章的介绍，可以从总体上对数据库营销和数据挖掘有一个比较直观的理解。

第3章 决策树方法在数据库营销中的研究

3.1 引言

目标市场的营销是企业采用的一种客户和企业之间的一对一的销售方式，现在目标市场营销中要解决的一个主要问题就是客户的选择的问题，目前有比较多的数据挖掘的方法应用在目标市场营销的客户的选择上，但是由于目标市场的数据集的特殊性，一般来说目标市场数据库中的数据只有 1%~5%的客户才是真正的购买者[1,2]，类分布非常不平衡，很多的传统的方法都很难适应这种数据类型。传统的决策树方法在这种数据集上也出现了很多的问题，主要表现在：对于类分布不平衡的情况，传统的决策树方法很难建立一棵好的决策树，生成的规则可能过于简单而不具有实用价值，从数据集的情况来说，例如它可以生成一条规则并认为所有的客户都是负例，尽管这条规则没有实用价值但精度可以到达95%~99%；对于属性较多的数据集，生成的决策树可能过大，规则过多，不便于应用；传统决策树算法只是对客户进行分类（yes/no），不能计算客户的潜在概率，因此不能对客户按概率值进行排序，不能给决策者提供灵活选择客户的自由。为了解决传统决策树在目标市场营销中的出现的问题，我们提出了一种改进了的适应于数据库营销的决策树的方法。

3.2 改进的决策树算法描述

3.2.1 符号约定

As：属性集　　Es：训练集　　Ts：测试集　　Rs：规则集　　r：规则

P（r）：　规则 r 覆盖的正例数与覆盖的训练例数所占的百分比

S（r）：　规则 r 覆盖的训练例数与整个训练例数所占的百分比

几个阈值：

p_{min}：叶子节点对应的规则 r 覆盖的正例数与覆盖的训练例数的比例的最小值

p_{max}：叶子节点对应的规则 r 覆盖的正例数与覆盖的训练例数的比例的最大值

s：叶子节点对应的规则 r 覆盖的训练例数与整个训练例数的比例的最小值

3.2.2 算法概要

算法：MarketingTree

输入：*As*：属性集

Es：训练集

输出：能正确预测 *Ts* 回应概率的决策树

递归算法：MarketingTree(*Es*, *As*)

(1) 建立决策树的根节点 Root

(2) 统计 Es 中正例的比例 *P (r)*和 *Es* 在整个训练集中的比例 *S (r)*

I. 若 $P(r) \geqslant p_{max}$，则返回单节点树 Root，并将 Root 标记为 *P (r)*

II. 若 $P(r) \leqslant p_{min}$，则返回单节点树 Root，并将 Root 标记为 *P (r)*

III. 若 $S(r) \leqslant s$，则返回单节点树 Root，并将 Root 标记为 *P (r)*

(3) 若 As 为空，则返回单节点树 Root，并将 Root 标记为 *P (r)*

(4) 否则开始

I. *A*←*As* 中可对 Es 进行最佳分类的属性

II. Root 的决策属性←*A*

III. 对 *A* 的每一个可能的值 v_i

a) 在 Root 下加一个新的分支对应测试 $A = v_i$

b) Esv_i←*Es* 中属性 *A* 取值为 v_i 的例子组成的子集

c) 若 Esv_i 为空，则这个新分支下加一个叶子节点，节点标记为 *P (r)* 否则，在这个新的分支下添加一棵子树 MarketingTree(Esv_i, *As*-{*A*})。

(5) 结束

(6) 返回以 Root 为根的决策树

在算法中最佳分类属性的选取是采用信息增益（Information Gain）理论[14]。信息论中通常用熵度量一组实例在某方面的纯度。给定训练集 *Es*，其中正例的比例为 p_+，负例的比例为 $p_- = 1 - p_+$。*Es* 的关于这个布尔分类的纯度可用下式度

量：

$$Entropy(Es) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

设 *Es* 当前的熵为 *E*。若用一个属性 *A* 将 *Es* 分组（*A=v* 的实例分在同一组 *Esv*），*E* 将会降低（因为这是一个从无序向有序的转变过程）。预计 *E* 降低的数量称为属性 *A* 相对于实例集 *Es* 的*信息增益* *Gain(Es, A)*，定义为：

$$Gain(Es, A) = Entropy(Es) - \Sigma_{v \in Value(A)} (|Esv| / |Es|) Entropy(Esv)$$

信息增益越大的属性对训练例的分类越有利，因此在本算法中选用信息增益最大的属性作为决策树的决策属性。

3.3 实验数据与分析

我们在 NEC 公司的商业数据上对该改进的算法进行了试验。共选取了 50000 条记录作为数据集，其中 20000 条作为训练集，用于生成决策树和规则集，30000 条作为测试集（正例 3032 条），用于评价算法。本算法采用不同的阈值在训练集上生成决策树和规则集 R，把 R 作用在测试集上，获得每一个测试例的回应概率 *P (r)*，按照 *P (r)*值的大小将测试例由高到底进行排序。根据 DECILE 分析，我们把排序后的测试集分成 10 份，统计每一份 Decile 中的真正做出回应客户的人数，如 Tree01 试验中 Decile 1 回应人数为 608，Decile 2 回应人数为 400，…，Decile 10 回应人数为 73。根据决策树的不同的阈值的设置我们获得了 4 份决策树的试验结果集（表 3-1）。

表 3-1. 决策树算法试验 Decile 数据

Table3-1. The Decile Data of Decision Tree

	Dec 1	Dec 2	Dec 3	Dec 4	Dec5	Dec 6	Dec 7	Dec 8	Dec 9	Dec10
Tree01	608	400	391	350	329	187	323	214	157	73
Tree005	674	436	385	347	333	254	268	159	138	38
Tree002	838	509	399	380	265	255	184	182	44	56
Tree001	967	668	400	333	259	170	104	45	0	86

各试验结果集试验条件：

Tree01 结果集：p_{min} =5%，p_{max} =50%，s=1%

Tree005 结果集：p_{min} =5%； p_{max} =50%， s=0.5%

Tree002 结果集：p_{min} =5%， p_{max} =50%， s=0.2%

Tree001 结果集：p_{min} =5%， p_{max} =50%， s=0.1%

3.4 决策树算法性能评估

首先，根据试验数据对决策树算法的可行性进行评估。如果算法能够在数据集上发现模式，则在进行 Decile 分析的时候，我们将会发现做出回应的客户更集中在 Decile 的顶部，并且做出回应的客户人数按照递减的规律分布在 Decile 表中。算法效果越好，则越多的回应客户集中在 Decile 的顶部。从 decile 分析(图 3-1)中我们可以看到做出回应的客户主要集中在 decile 的前面几部分，从总体上来说，回应客户的人数按照从 decile 1 到 decile 10 依次递减。从 lift 曲线（图 3-2）可以看到 4 个决策树数据集中的客户累计回应百分比在前面的几个部分都远远高于随机的客户回应百分比。从 decile 分析和 lift 曲线可以看出决策树算法在该类市场数据集上可以很好的生成模式。

其次，不同阈值的设置决策树算法的性能评价。四个不同的实验数据集是用不同的阈值的决策树而获得的，在 decile 分析（图 3-1）中，回应客户的人数在顶部的三个 decile 中按照 Tree001、Tree002、Tree005、Tree01 的顺序依次递减，同样在 lift 曲线中，客户的累计回应百分比在相同的测试例百分比时，也是按照同样的次序依次递减，从实验条件来看四个试验是在 p_{min} =5%和 p_{max} =50%的条件下，s 值是依次增大的。随着 s 值的增大，决策树的性能将逐渐降低。

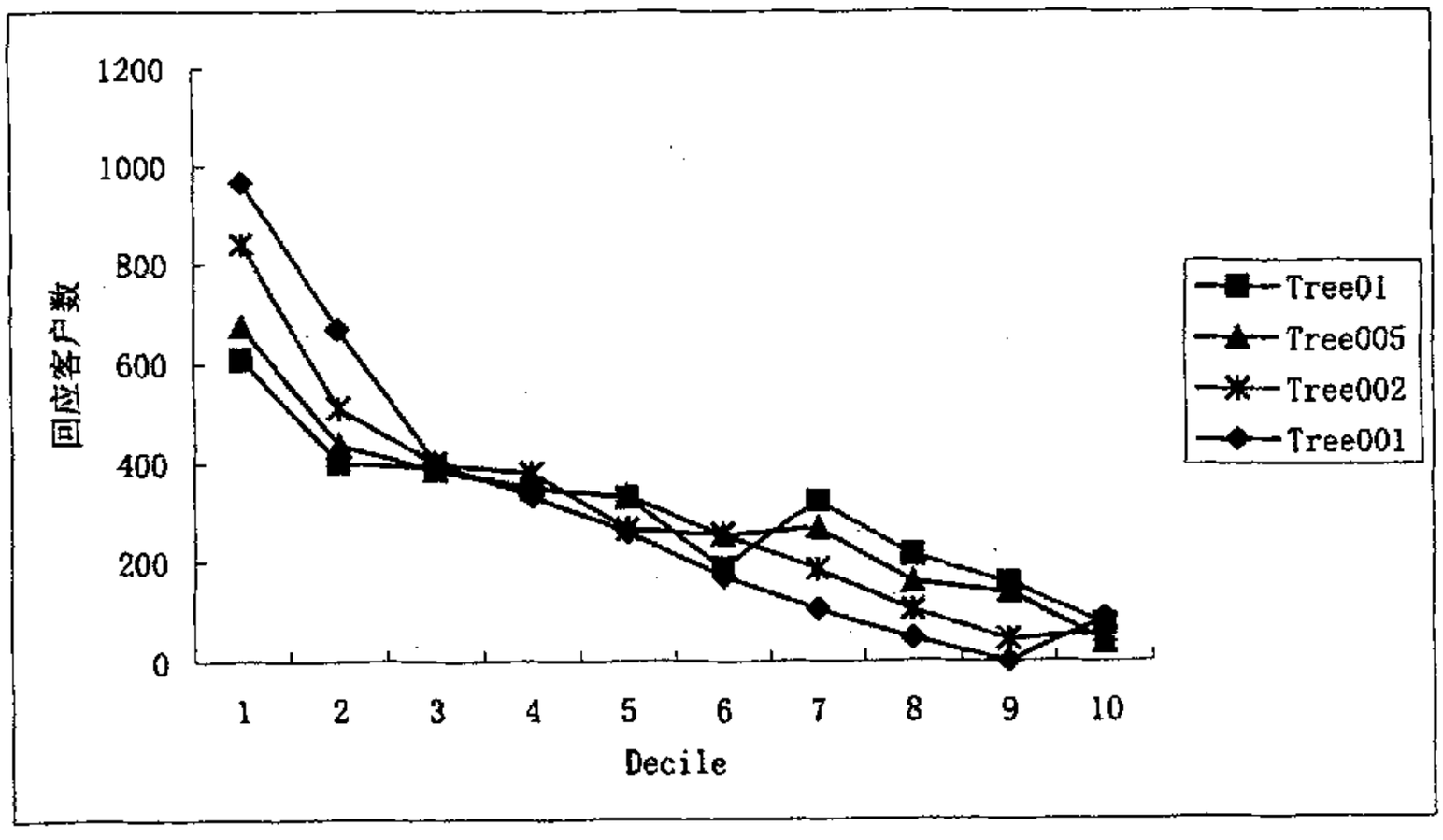

图 3-1 不同阈值决策树之间的回应客户的分布

Fig3-1. The customer distributions in different threshold values decision tree

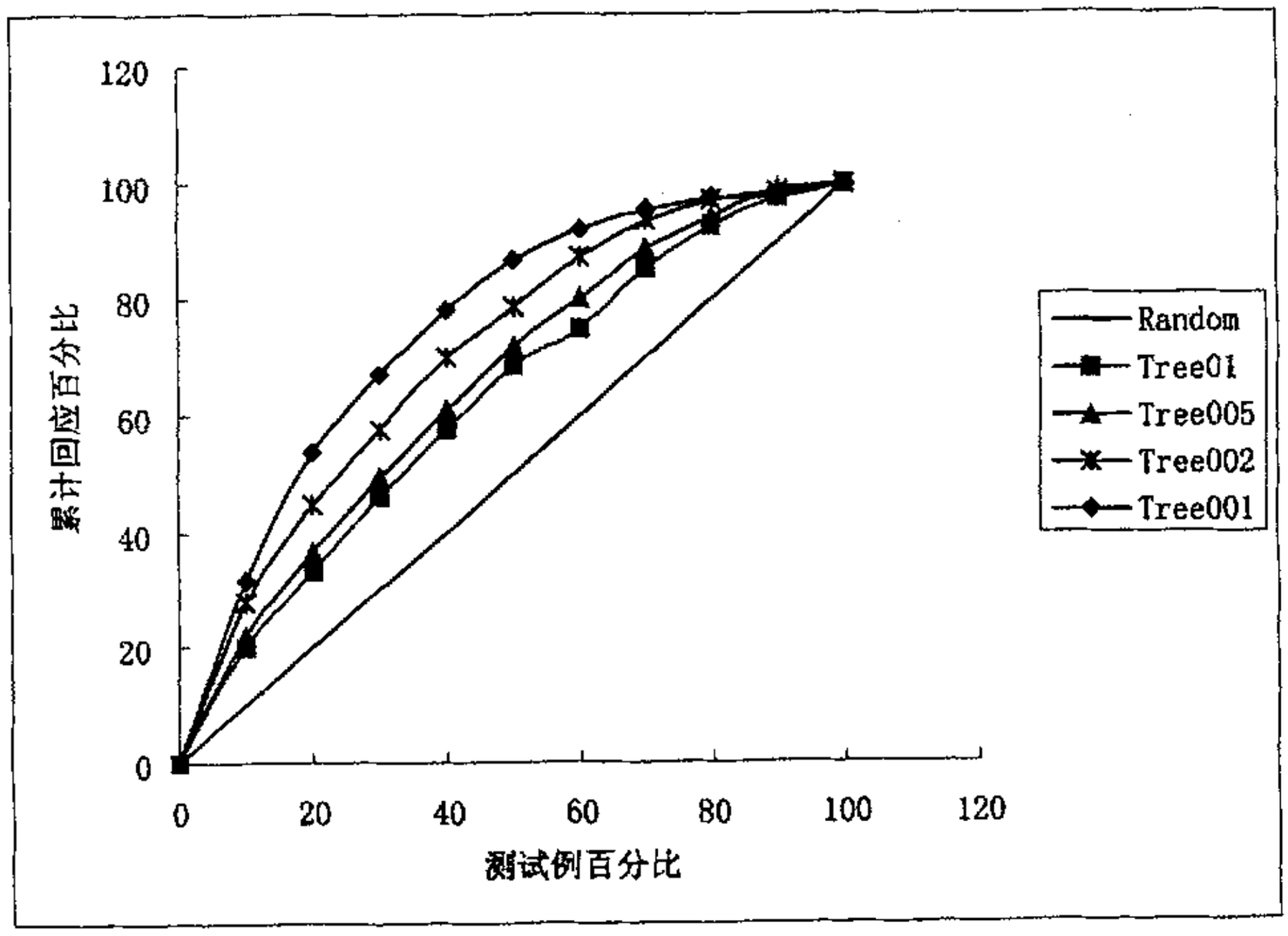

图 3-2 不同阈值决策树之间的累计 lift 曲线

Fig3-2. The accumulative lift curves in different threshold values decision tree

3.5 算法讨论

改进后的决策树算法可以很好地解决传统的决策树方法在目标市场数据营销中出现的问题，并且也有它自己的新的特点：

- 通过设置 p_{min} 、p_{max} 和 s 可以自由的控制决策树的大小，只要节点的阈值 *P(r)*和 *S(r)*超过预先设定的阈值，就停止决策树的生长，从而达到控制决策树大小的目的。
- 可以解决市场数据中的类分布不平衡而不能生成合适的决策树的问题。用户可以设置不同的 p_{min} 、p_{max} 和 s 值生成多棵决策树，根据性能的优劣从中选择最佳性能的决策树，生成规则集 R，并作用市场数据集中。
- 可以解决传统决策树的只能分类而不能对客户进行排序的问题。改进的决策树算法可以按照回应概率 *P(r)*的大小对客户进行排序，可以给决策者提供灵活选择客户的自由。

3.6 应用

在建立一棵决策树后，可以根据决策树生成规则集 *R*，规则集 *R* 的每一条规则 *r* 对应从根节点 Root 到每一个叶子节点的属性的排列，并且每一条规则 *r* 标记为对应的叶子节点的 *P(r)*值，若存在实例与规则 *r* 匹配，则该实例的回应的概率为 *P(r)*。对客户的市场数据集，我们可以将每一个客户与规则匹配，从而得到客户的回应的概率 *P(r)*，根据得到的 *P(r)*值对客户进行排序，选取概率值最大的客户作为潜在的客户，投送促销材料。

3.7 本章小结

本章提出了一种改进的可以应用在目标市场数据库销售的决策树算法。该算法的主要创新点在于：它是通过统计决策树的叶子节点的 P(r)和 S(r)的大小来限制决策树的生长，并采用统计的方式标注叶子节点，可以自由的控制决策树的生长，生成合适的决策树。主要贡献是它可以解决市场数据中的类分布不平衡而不能生成合适的决策树的问题以及传统决策树不能排序的问题。

第4章 市场值函数方法在数据库营销中的研究

4.1 介绍

在目标市场销售中需要解决的一个的重要问题是如何找出具有潜在市场值的客户和产品。市场值函数方法可以用来解决上面提到的问题。它假设每一个对象都可以使用一定数量的属性来描述。该方法是在属性值上的效用函数(Utility Functions）和属性权值的一个线性组合。市场值函数的一个主要的优点是它可以按照对象的市场值的大小对对象进行排序，而不只是对对象进行分类。在这里我们提出了多种不同的方法来挖掘市场值函数。

识别具有潜在市场值的客户和产品是数据挖掘应用中的一个重要的领域，许多标准的数据挖掘方法可以应用在市场的客户挖掘中，例如，我们可以用关联规则从大量的市场交易数据库中挖掘客户[17]。如果在两个集合的元素 A 和 B 之间发现了一个强关联，也就是说如果客户购买了 A，那么他就可能购买 B，这样我们可以用 B 作广告或者把 B 推荐给购买 A 产品的客户，这可能就很有效果。实际上在目标市场的销售中用关联规则推荐相关的产品，这是一种常用的方式。

现在考虑两种其它类型的目标市场的问题。设想一个健康俱乐部需要吸引更多的成员来扩张他们的业务。假设每一个成员都可以使用一组有限的属性来描述，我们则可以根据这些属性来发现他们的共同特征，从而把企业的信息发送给那些与他们具有相同的或者相似特征的非成员。另外一个例子是推荐特殊类型的电话服务和不同类型的信用卡的销售服务，对于这种情况我们需要发现对象之间基于属性值的联系或者相似性，其中的假设就是：*相同类型的人将做出相似的决定和选择相同的服务*。单纯的关联规则方法则不能解决这种问题，而其他基于规则的机器学习和数据挖掘方法（如特征规则学习、分类规则、判别规则和特异规则[19, 23, 35, 38]等），虽然可以应用到目标市场的客户挖掘中，但是这些方法可能产生太多或者太少的规则，选择出一个好的规则集则不是一件容易的事情，另外用这些生成的规则有可能选择出太多的或者太少的客户。

通过扩展信息检索的技术，我们提出了市场值函数的方法来解决上面的问题[7, 27, 28]。信息检索技术的主要目的是识别出有用的信息元素，这项技术可以用来发现具有潜在市场值的对象，它不同于一般使用的数据挖掘技术，它是用来找到一个市场值函数或者判别式函数，而不是用来发现规则的方法。以上述健康俱乐部为例，市场值函数的主要目的是为了衡量潜在的新客户和已存在的客户的相似的程度，从这一点来看，它的思想和基于案例推理的方法很相似[21]，现有的成员作为一个案例用来和潜在的成员作对比。根据客户的市场值，我们可以按照客户加入俱乐部的可能性进行排序，同时我们可以根据多种标准如邮资，人力等，给排序的列表设置一个断点，只取断点前的客户。从某种程度上来说，市场值函数可以解决很多基于规则的方法解决不了的难题，并且当有新的可用信息时，市场值函数可以很容易的更新。

我们所提出来的方法有如下的优点：目标市场的特殊的问题可以公式化成经典的信息检索的问题，很多的信息检索的理论的结果可以马上用到这个领域；市场值的方法可以让我们产生一个排序的队列，可以提供更灵活的方式来解决目标市场中的问题；具有很好的解释性。

4.2 一种目标市场的线性模式

我们主要从知识表示和构建市场值函数模型两方面来描述这个模型。假设每一个对象都是由一个有限的属性集合的属性值来表示，并且可以用一个线性的市场值函数来计算对象的市场值，我们可以认为这种模式是一种线性模式，一种与信息检索的线性模式有关但是却又不同的模式。

假设 U 是一个有限的对象空间，U 的元素可以是在作市场决策时我们感兴趣的客户或者产品。U 被分成三个不相交的类，也就是 $U=P\cup N\cup D$，其中 P、N、D 三个集合分别称为：*正例*、*负例*和*未知例*。以上面提到的健康俱乐部为例：P 是俱乐部当前成员的集合；N 是原来拒绝加入俱乐部的人员的集合；D 是剩余的集合，其中 N 集合可能是空集。目标市场的问题可以定义为，从 D 中也可能是从 N 中找出和 P 中元素相似的或者和 N 中元素不相似的元素。也就是说，我们从 D 中

或者 N中找出可能成为 P中新成员的元素，我们的目的就是为了找出一个市场值函数，从而按照市场值的大小排序 D中的元素。

有限空间的对象的信息可以由一个信息表表示出来[22,34]，表中行对应空间中的对象，列对应属性，每一个单元对应相关属性的属性值。在形式上，信息表是四维的：

$$S = (U,\ A_t, \{V_a | a \in A_t\}, \{I_a | a \in A_t\}),$$

在这里：

U 表示有限非空的对象集合，

A_t 表示有限非空的属性集合，

V_a 表示属于属性 $a \in A_t$ 值的集合，

I_a:U->V_a 是 $a \in A_t$ 的信息函数。

每一个信息函数 I_a是 U 到 V_a中的值映射函数。

解决目标市场问题的一个直接方法是根据 P和 N中元素的特征挖掘出能够描述或者区分 P 和 N 中元素的规则，利用这些规则来分类 D中的元素，这种技术已经得到了广泛的研究。该方法的特点以及缺点在前面的内容中我们已经讨论过了，因此我们的重点主要集中在其他的方法上。

市场值函数是一个从对象空间到实数集合的实数值函数 r:U->R。在信息检索的环境中，r 的值表示文档与查询请求之间的关联度。这些文档可以根据 r 的值进行排序，而对目标市场而言，则可以按照客户的潜在的市场值的大小进行排序。以健康俱乐部的为例，我们可以按照客户成为俱乐部的成员的可能性的大小对客户进行排序，这些可能性可以根据与 P 中成员的相似性来估计。

本文研究了市场值函数的最简单的形式之一--线性判别函数。设 u_a: V_a->R 是一个定义在属性 $a \in A_t$ 上值为 Va 的 Utility 函数，$u_a(\cdot)$ 的值可以是正数、负数或者零，对于 $v \in V_a$，如果 $u_a(v) > 0$ 并且 $I_a(x)=v$，也就是说 $u_a(I_a(x))>0$，说明对象 x 在属性 a 上相似于 P；如果 $u_a(I_a(x))<0$，说明对象 x 在属性 a 上相似于 N；如果 $u_a(I_a(x))=0$，那么说明对象 x 在属性 a 上既不相似于 P也不相似于 N。我们可以用一个线性的市场值函数的形式来计算对象 x 对 P的相似程度：

$$r(x)=\sum_{a\in At}\omega_a u_a(I_a(x)) \tag{4-1}$$

在这里：ω_a是属性 a 的权值。类似的，权值ω_a可以是正值、负值和零。权值越高（绝对值）的属性就越重要，权值接近 0 说明该属性不重要。对象 x 的市场值就是由所有属性值的 Utility 和权值组合而成的。通过使用线性的市场值函数，我们隐含的假设各个属性之间是相互独立的，这个假设通常认为是 Utility 的独立性假设。

4.3 挖掘市场值函数

从上面的分析中，我们可以看出该模式的效果在很大的程度上依赖于 Utility 函数和属性的权值。下面我们介绍多种估计和挖掘市场值函数的方法。在这些方法中 Utility 函数源于信息检索的概率模型，而属性的权值则基于信息论中属性重要性的衡量[24, 29, 30, 32, 33]。

4.3.1 Utility 函数

Utility 函数既可以在正例中定义也可以同时在正例和负例中定义。

4.3.1.1 正例上估计 Utility 函数

只从正例样本来估计市场值函数需要考虑多种情况，因为在数据集中有可能不出现负例，有时即使出现负例，也不一定所有的负例都是可以用来构建市场值函数模型的。例如原来不是健康俱乐部成员或者原来拒绝加入俱乐部的成员将来可能会加入俱乐部，因此我们不应排除那些和负例相似但是最后加入俱乐部的可能。换句话说就是我们从正例出发来加入潜在的新的客户，而不是用负例来排除潜在的新的成员。在现实中，我们一般可以找出一个规则来解释为什么一个对象属于 *P*，但是却很难解释为什么该对象不属于 *P*。

考虑一个属性 a∈*At* 从 Va 上取值，对 v∈Va，设 P 的子集 m(a=v|P)=m(v|P) 定义成如下形式：

$$m(v|P)=\{x\in U|x\in P, I_a(x)=v\} \tag{4-2}$$

它包括了在 P 中属性值为 v 的成员，设|m（v|P）|代表集合 m（v|P）的元素的个数，对于两个值 v，v’ ∈Va，如果|m（v|P）|>|m（v’|P）|表明在 P 中具有 v 值的元素的数目要大于 P 中具有 v’的元素的数目。直观的，属性值取 v 的元素比属性值取 v’的元素更有可能属于 P。如果只是从属性 a 上考虑的话，对于两个元素 x，y∈U，如果 $I_a(x)=v$，$I_a(y)=v'$ 并且|m（v|P）|>|m（v’|P）|，那么我们可以说 x 的市场值要大于 y 的市场值。这就隐含了在 v∈Va 中的 Utility 函数 u_a：V_a->R 的值和 m（v|P）集合的大小是成比例的。因此我们可以采用如下的 Utility 函数：

$$u_a^{\ 1}(v)=|m(v|P)| \tag{4-3}$$

$u_a(\cdot)$ 的值在 0 和|P|之间，这些都是简单的统计 P 中具有属性值 v 的元素的数目。P 的元素的集合可以认为是 U 的一个子集，Utility 函数可以用概率方式表示：

$$u_a^{\ 2}(v)=\Pr(a=v|P)=\Pr(v|P)=\frac{|m(v|P)|}{|P|}=\frac{u_a^{\ 1}(v)}{|P|} \tag{4-4}$$

由于|P|是一个独立于任何属性的常数，u_a^1、u_a^2 将在线性模式中得到相同的结果。

通常如果一个元素的值主要集中在子集 P 中，那么我们可以期望属性对市场值做出更多的贡献。这些可以通过条件概率 Pr(v|P)和非条件概率的比较来得到，其中非条件概率为：

$$\Pr(a=v)=\Pr(v)=\frac{|m(v)|}{|U|} \tag{4-5}$$

在这里：
$$m(v)=\{x\in U|\ I_a(x)=v\}. \tag{4-6}$$

从简洁性考虑，我们假设 $m(v)\neq\emptyset$；否则我们就把 v 从 Va 中删除，相应的 Utility 函数可以定义成：

$$u_a^{\ 3}(v)=\frac{\Pr(v|P)}{\Pr(v)}=\frac{|m(v|P)||U|}{|m(v)||P|} \tag{4-7}$$

如果 $u_a^3(v)>1$，那么属性值 v 在子集 P 出现的概率大于在整个 U 中出现的概率；反之则 $u_a^3(v)<1$。我们用对数的方法把 u_a^3 转换成如下形式：

$$u_a^4(v)=\log u_a^3(v)=\log\frac{|m(v|P)||U|}{|m(v)||P|} \tag{4-8}$$

当且仅当 $u_a^3>1$ 时得到 $u_a^4>0$, 当且仅当 $u_a^3<1$ 时得到 $u_a^4>0$, 当且仅当 $u_a^3=1$ 时得到 $u_a^4=0$。在实际的应用当中，可能发生 $m(v|P)=\emptyset$ 的情况。在这种情况下，我们可以使用信息检索中常用的公式来解决：

$$u_a^4(v)=\log u_a^3(v)=\log\frac{(|m(v|P)|+0.5)(|U|+1.0)}{(|m(V)|+0.5)(|P|+1.0)} \tag{4-9}$$

这样隐含的假设了零样本空间被平等的划分到了 P 和 N。

|U|/|P|的数值是独立于属性的一个常数，它不影响排序。因此它可以从 Utility 函数中去掉，而只采用|m（v|P）|/|m(V)|。

4.3.1.2 从正例和负例上估计 Utility 函数

如果考虑正例和负例的话，我们有两个子集 P 和 N。我们修改前面提到的方法来考虑属性在正例和负例上的情况。

如果属性值 v 在子集 P 出现的概率比在 N 中大，那么对象在属性 a 上相似于 P, 反之则相似于 N。相似的，我们对 Utility 函数 u_a^3 、u_a^4 定义了两个新的 Utility 函数：

$$u_a^5(v)=\frac{\Pr(v|P)}{\Pr(v|N)}=\frac{|m(v|P)||N|}{|m(v|N)||P|} \tag{4-10}$$

$$u_a^6(v)=\log u_a^5(v)=\log\frac{|m(v|P)||N|}{|m(v|N)||P|} \tag{4-11}$$

在这里：
$$m(v|N)=\{x\in U, I_a(x)=v\}。 \tag{4-12}$$

u_a^6 的公式如下：

$$u_a^6(v)=\log\frac{(|m(v|P)|+0.5)(|N|+0.5)}{(|m(v|N)|+0.5)(|P|+0.5)} \tag{4-13}$$

由于 P 和 N 是 U 的不相交的子集，因此新的 Utility 函数不是简单的在 u_a^3 和 u_a^4 中用 N 替代 U。|N|/|P|的比例是独立于任何属性的常数，可以从 Utility 函数

中删除。

4.3.2 属性的权值

为了计算属性的权值，我们引入了信息论的方法[32,33]。对于属性 a，它在 P 中的熵 Hp(a)被定义成这样：

$$H_p(a)=H_p(Pr(\cdot|P))=-\sum_{v\in Va}Pr(v|P)\log Pr(v|P) \qquad (4\text{-}14)$$

在这里，Pr(·|P)表示属性值在 P 中的概率分布。定义 0log0 为 0。熵值是一个非负的函数，也就是说 Hp(a)≧0。熵值越低表示结构化程度越高，如果属性有低的熵值，说明它的属性值在 P 中分布是不均匀的。因此，该属性在预测对象是否属于 P 中可以提供更多的信息；相反，属性熵值越高表明该属性在 P 中分布得越均匀，在预测中提供的信息就越少。因此在计算属性的权值时，应该和属性的熵值成反比关系。按照信息检索理论[32]，我们给出了以下的权值公式：

$$\omega_a^{\ 1}=1-\frac{H_p(a)}{\log|V_a|} \qquad (4\text{-}15)$$

很明显，$0\le\omega_a^{\ 1}(v)\le 1$。同样我们假设|Va|>1。另外，任何对象在属性 a 上都有相同的值。

在全部集合 U 上的属性的熵的公式如下：

$$H(a)=H(Pr(\cdot))=-\sum_{v\in Va}Pr(v)\log Pr(v) \qquad (4\text{-}16)$$

它反映了属性 a 的值在 U 中分布的结构化程度，对于能提供更多信息的属性，我们可以看到在它在 U 上面结构化的程度要比在 P 上的程度要更小。我们可以用到其他的涉及到 $H_p(a)$ 和 H(a) 的公式：

$$\omega_a^{\ 2}=\frac{H(a)-H_p(a)}{\log|V_a|} \qquad (4\text{-}17)$$

权值为正值表明属性 a 在 P 上比在 U 上结构化的程度越高，负值则相反。$\omega_a^{\ 1}$ 是 $\omega_a^{\ 2}$

的一个特例，在这里 H(a) 取得最大值 $\log|V_a|$。

著名的 Kullback Leibler 理论给出了其他的属性权值的计算公式[29]：

$$\omega_a^{\ 3} = D(\Pr(\cdot|P)\|\Pr(\cdot)) = \sum_{v\in V_a}\Pr(v|P)\log\frac{\Pr(v|P)}{\Pr(v)} \qquad (4\text{-}18)$$

我们同样也可以从正例和负例上来计算属性的权值。从属性的观点来看，如果子集 P 和 N 彼此不同的话，那么属性能提供更多的信息。在这种条件下，我们有三个子集 P、N 和 $P\cup N$。设 $H_P(a)$，$H_N(a)$ 和 $H_{P\cup N}(a)$ 分别代表属性 a 在三个子集中的熵值，如果属性值在两个子集中的分布是一样的话，两者应该在 $P\cup N$ 中应该也有相同的分布。我们可以看到属性 a 的熵值在 P、N 和 $P\cup N$ 中只有稍微的不同。相反，如果属性值在 P 和 N 中的分布不一样的话，那么就有大的不同。从这一方面考虑，我们引入了如下的权值计算公式：

$$\begin{aligned}\omega_a^{\ 4} &= H_{P\cup N}(a) - \left[\frac{|P|}{|P\cup N|}H_p(a) + \frac{|N|}{|P\cup N|}H_N(a)\right] \qquad (4\text{-}19)\\ &= H_{P\cup N}(a) - \left[\lambda_P H_P(a) + \lambda_N H_N(a)\right]\end{aligned}$$

在这里 $\lambda_p + \lambda_N = 1$，在公式中我们需要考虑的另外一个问题是两个子集 P 和 N 的熵的平均值，对任何属性值 v 有：

$$\Pr(v|P\cup N) = \lambda_P \Pr(v|P) + \lambda_N \Pr(v|N) \qquad (4\text{-}20)$$

由于-xlogx 是凹函数，可以由 Jesen 不等式马上得到 $\omega_a^{\ 4}$ 的下限是 0，也就是 $\omega_a^{\ 4} \geq 0$。当在两个子集 P 和 N 的分布是一样的时候，$\omega_a^{\ 4}$ 可以达到最小值 0。如果分布完全不同，也就是说只要 $\Pr(v|N)=0$ 就有 $\Pr(v|P)\neq 0$ 并且只要 $\Pr(v|N)\neq 0$ 就有 $\Pr(v|P)=0$，此时 $\omega_a^{\ 4}$ 可以达到最大值 $-\left[\lambda_P\log\lambda_P + \lambda_N\log\lambda_N\right]$。

根据 Kullback Leibler 理论，$\omega_a^{\ 4}$ 可以表示：

$$\omega_a^{\ 4} = \lambda_P D(\Pr(\cdot|P)\|\Pr(\cdot)) + \lambda_N D(\Pr(\cdot|N)\|\Pr(\cdot)) \qquad (4\text{-}21)$$

公式ω_a^3是公式ω_a^4的前面一部分，ω_a^4考虑了两个子集的情况，因此它比ω_a^3更具有通用性。熵函数只是由概率分布的概率值来决定，它不依赖于这些概率值是怎样分配给不同属性的，不同的概率分布可以产生同样的熵值的。例如，如下的分布尽管具有完全不同的分布，但是它们有同样的熵值：

$$\Pr(v_1|P)=0.5 \quad \Pr(v_2|P)=0.5$$
$$\Pr(v_3|P)=0.0 \quad \Pr(v_4|P)=0.0$$
$$\Pr(v_1|N)=0.0 \quad \Pr(v_2|N)=0.0$$
$$\Pr(v_3|N)=0.5 \quad \Pr(v_4|N)=0.5$$

如果 P 和 N 在属性 a 上相似，我们就不能知道$H_P(a)$和$H_N(a)$之间的不同。这主要是因为在$\Pr(\cdot|P)$和$\Pr(\cdot|N)$的概率之间没有继承的关系。另一方面，前面提到的方法ω_a^3和ω_a^4不会遇到这个问题，在这些公式中我们在关联的集合中使用了概率分布。

4.4 试验结果及分析

为了验证市场值函数方法性能和效果，我们选择了 NEC 公司的现实销售数据作为我们的实验数据来对市场值函数方法做两个方面的评价：

- 属性权值和效用函数的不同组合对市场值函数结果的影响；
- 市场值函数方法在现实营销数据中表现。

该数据集包括 124402 条客户的记录，每一条记录包括 96 个基本属性和在某一时间段的销售的产品情况，客户的基本属性包括了每个成员的性别、年龄、收入、爱好等与客户基本特征相关的属性。客户购买的产品共包括 24 种产品，如果客户购买了该产品，则对数据库中对应的产品标记为 1，否则标记为 0。在进行数据预处理后，我们选取了 58102 条客户的记录作为我们训练和测试的数据集。

4.4.1 权值和效用函数的不同组合

4.4.1.1 市场值函数

在下面的评价当中我们选择以下的不同权值和效用函数作不同的组合：权值公式我们选用（4-18）和（4-19）

$$\omega_a^{\ 3} = D(\Pr(\cdot|P)\|\Pr(\cdot)) = \sum_{v \in Va} \Pr(v|P)\log\frac{\Pr(v|P)}{\Pr(v)}$$

$$\omega_a^{\ 4} = H_{P\cup N}(a) - \left[\frac{|P|}{|P\cup N|}H_p(a) + \frac{|N|}{|P\cup N|}H_N(a)\right]$$
$$= H_{P\cup N}(a) - \left[\lambda_P H_P(a) + \lambda_N H_N(a)\right]$$

效用函数公式我们选用（4-7）和（4-10）

$$u_a^{\ 3}(v) = \frac{\Pr(v|P)}{\Pr(v)} = \frac{|m(v|P)\|U|}{|m(v)\|P|}$$

$$u_a^{\ 5}(v) = \frac{\Pr(v|P)}{\Pr(v|N)} = \frac{|m(v|P)\|N|}{|m(v|N)\|P|}$$

通过权值和效用函数的不同组合，可以得到四种不同的市场值函数。针对这四种不同的组合，我们在产品二上来评价这几种不同的组合对市场值函数的性能的影响。

4.4.1.2 数据集描述

我们从 NEC 数据集中选用产品二的客户的购买历史作为数据集的决策属性，购买产品二则标记为 1，否则标记为 0。在数据预处理以后我们得到 58102 条完整的客户记录（见表 4-1），我们选用了前面的 18963 条记录作为训练例，其中包括正例 894 个，24134 条记录作为测试例集，其中包括正例 649 个。

表 4-1. 产品 2 上的数据库描述

Table4-1. Database Description in Product Id2

Dataset Name	Whole Instances	Positive Instances
Usable Dataset	58102	3856
Training Dataset	18963	894
Testing Dataset	24134	649

4.4.1.3 试验结果

表 4-2、表 4-3、表 4-4 和表 4-5 记录了不同效用函数和权值组合的市场值函数的 Decile 结果，图 4-1 记录了不同权值和效用函数组合的市场值函数的 Decile 分析和 Lift 曲线图

表 4-2. 产品 2 上的 Decile 分析表（ω_a^3 和 μ_a^3）

Table4-2. Decile Analysis of Testing Dataset in Product Id2（ω_a^3 and μ_a^3）

Decile	Number of Individuals	Number of Responses	Number of Real Responses	Decile Response Rate	Cumulative Response Rate	Cum Response Lift
1	2413	65	125	5.179415	5.179415	193
2	2413	65	104	4.309273	4.744344	176
3	2413	65	81	3.356261	4.28165	159
4	2413	65	66	2.734731	3.89492	145
5	2413	65	68	2.817602	3.679456	137
6	2413	65	47	1.94746	3.39079	126
7	2413	65	53	2.196072	3.220116	120
8	2413	65	35	1.450236	2.998881	112
9	2413	65	40	1.657413	2.849829	106
10	2413	65	30	1.24306	2.689152	100

表 4-3. 产品 2 上的 Decile 分析表（ω_a^3 和 μ_a^5）

Table4-3. Decile Analysis of Testing Dataset in Product Id2（ω_a^3 and μ_a^5）

Decile	Number of Individuals	Number of Responses	Number of Real Responses	Decile Response Rate	Cumulative Response Rate	Cum Response Lift
1	2413	65	125	5.179415	5.179415	193
2	2413	65	104	4.309273	4.744344	176
3	2413	65	79	3.27339	4.254026	158
4	2413	65	69	2.859037	3.905279	145
5	2413	65	67	2.776166	3.679456	137
6	2413	65	47	1.94746	3.39079	126
7	2413	65	52	2.154636	3.214197	120
8	2413	65	36	1.491671	2.998881	112
9	2413	65	41	1.698848	2.854433	106
10	2413	65	29	1.201624	2.689152	100

表 4-4. 产品 2 上的 Decile 分析表（ω_a^4 和 μ_a^3）

Table4-4. Decile Analysis of Testing Dataset in Product Id2（ω_a^4 and μ_a^3）

Decile	Number of Individuals	Number of Responses	Number of Real Responses	Decile Response Rate	Cumulative Response Rate	Cum Response Lift
1	2413	65	125	5.179415	5.179415	193
2	2413	65	104	4.309273	4.744344	176
3	2413	65	80	3.314825	4.267838	159
4	2413	65	67	2.776166	3.89492	145
5	2413	65	68	2.817602	3.679456	137
6	2413	65	47	1.94746	3.39079	126
7	2413	65	53	2.196072	3.220116	120
8	2413	65	35	1.450236	2.998881	112
9	2413	65	42	1.740283	2.859037	106
10	2413	65	28	1.160189	2.689152	100

表 4-5.产品 2 上的 Decile 分析表（ω_a^4 和 μ_a^5）

Table4-5. Decile Analysis of Testing Dataset in Product Id2（ω_a^4 and μ_a^5）

Decile	Number of Individuals	Number of Responses	Number of Real Responses	Decile Response Rate	Cumulative Response Rate	Cum Response Lift
1	2413	65	125	5.179415	5.179415	193
2	2413	65	104	4.309273	4.744344	176
3	2413	65	79	3.27339	4.254026	158
4	2413	65	70	2.900472	3.915637	146
5	2413	65	66	2.734731	3.679456	137
6	2413	65	48	1.988895	3.397696	126
7	2413	65	51	2.113201	3.214197	120
8	2413	65	36	1.491671	2.998881	112
9	2413	65	41	1.698848	2.854433	106
10	2413	65	29	1.201624	2.689152	100

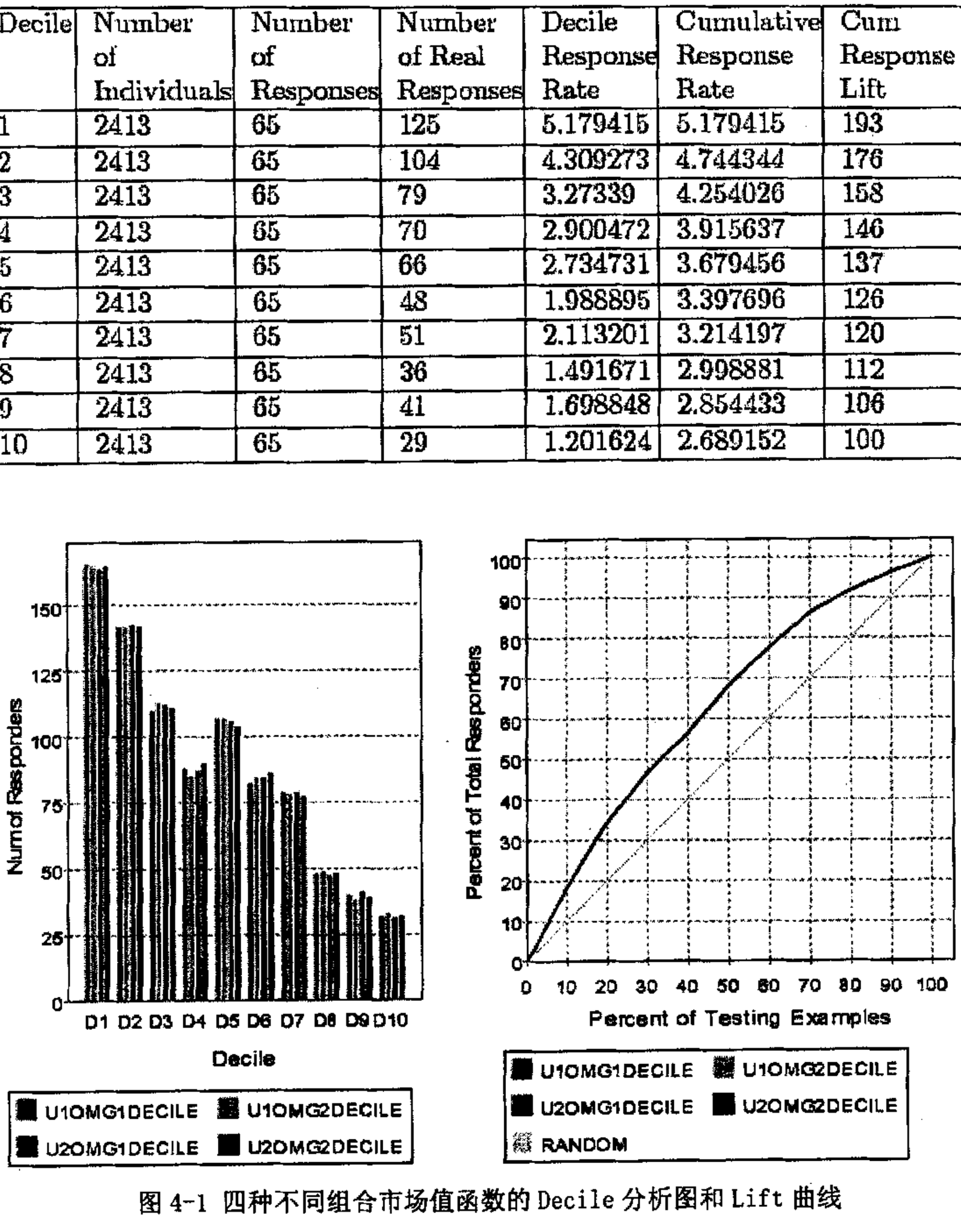

图 4-1 四种不同组合市场值函数的 Decile 分析图和 Lift 曲线

Fig4-1. The Decile Analysis Bar and Lift Curves in Four Different Combinations of Market Value Functions

4.4.1.4 结论

从表 4-2、表 4-3、表 4-4 和表 4-5 可以看到，在每一个 Decile 中分布的真正做出回应的人数基本一致，自上而下 Decile 中各回应人数的下降趋势也基本一致。而且从图 4-1 中可以看到各种不同组合的市场值函数的客户的 lift 曲线基本重合。从上面的实验数据来看，权值公式（4-18）、（4-19）和效用函数（4-7）、（4-10）的不同结合产生的市场值函数，最终所得到的实验结果基本一致。这种情况可以从表 4-6 和表 4-7 中的权值和效用函数值得到答案，我们从训练例中生成的模式中选择了权值排在前面的 17 位的属性的权值（表 4-7）和 6 个属性的效用函数值（表 4-6）。从表 4-7 中可以看出各属性的权值 ω_a^3 和 ω_a^4 的权值大小排序基本一致，同样我们也可以在表 4-6 中看到效用函数 μ_a^3 和 μ_a^5 值的大小排序也基本一致。因此两者不同组合所得到的市场值函数虽然因为权值和效用函数值的不同结合而值的大小不一样，但是每个实例的市场值的最终排序是基本一致，因此最终所得到的 Decile 分析的结果基本一致。这些结果可以从表 4-2、表 4-3、表 4-4、表 4-5 和图 4-1 中看到。

表 4-6. 产品 2 上相应的属性值的效应函数值

Table4-6. Utility Value of the Corresponding Attribute Value on Id2

Attribute Name	Attribute Value	μ_a^3	μ_a^5
WANTS31	0	0.8070314	.7993991
	1	1.248332(s)	1.263861
WANTS30	0	0.7864021	.7781782
	1	1.188952 (s)	1.200172
WANTS29	0	0.8755524	.8701944
	1	1.11742 (s)	1.123949
WANTS27	0	0.8985494	.8940617
	1	1.107886 (s)	1.113832
WANTS6	0	0.9417834 (s)	.9390785
	1	1.307697	1.327913
ADDRESS	1	1.265173	1.281992
	2	0.7377882	0.7283391
	3	0.8417226	0.8351822
	4	0.8054966	0.7978188
	5	2.121141	2.245712
	6	1.719844	1.78336
	7	0.666327	0.6555052
	8	0.4820775	0.4700328
	9	1.287836	1.306441
	10	0.8461998	0.8398093
	11	1.647488	1.702013
	12	1.641359	1.69515
	13	0.7876916	0.7795034
	14	0.8400558	0.8334602
	15	0.6060403	0.5944532
	16	1.24773	1.263213
	17	1.021864	1.02297
	18	1.131275	1.138671
	19	1.235616	1.25019
	20	1.362201	1.387058
	21	1.387662	1.414799
	22	1.220801	1.234285
	23	0.6526588	0.641632
	24	1.223735	1.237433
	25	1.275003	1.29259
	26	0.8550979	0.8490111
	27	1.293379	1.312429
	28	1.161046	1.170371
	29	1.515101	1.554724
	30	1.071283	1.075075

表 4-6 记录了从训练例集中生成的模式中的 4 个属性的效用函数值 μ_a^3 和 μ_a^5。

表 4-7. 产品 2 上相应的属性的权重

Table4-7. Attribute Weights of Training Dataset in Product Id2

No.	Attribute	ω_a^3	ω_a^4	Description
1	WORK	0.0354138	0.001756435	
2	HOBBY2	0.02521296	0.00124318	Computer
3	WANTS31	0.02371247	0.001173834	Personal Computer Software
4	HP_GENRE3	0.02361109	0.001167616	Computer Software
5	HP_GENRE17	0.02222864	0.001105044	Adult Belongings
6	SEX	0.02212581	0.001089783	
7	ADDRESS	0.02171585	0.001076463	
8	WANTS30	0.02048995	0.001013022	Personal Computer Peripheral
9	BIRTH_DATE	0.01726366	0.000857008	
10	HP_GENRE16	0.01461868	0.000723922	Search Engine
11	HOBBY4	0.01248269	0.000619547	Communication Instrument
12	INCOME	0.01082302	0.000536787	
13	HOBBY6	0.01024668	0.000508163	Camera and Video Tape Recorder
14	WANTS6	0.008313148	0.000412886	MIDI
15	WANTS29	0.007341529	0.000363135	Personal Computer
16	HP_GENRE15	0.00598366	0.000296921	Online Shopping
17	WANTS27	0.005470819	0.000270667	Digit Camera

表 4-7 记录了从训练例集中生成的模式中的 17 个属性的权值 ω_a^3 和 ω_a^4。

4.4.2 市场值函数在不同数据集上的评价

4.4.2.1 市场值函数

通过前面对不同权值和不同效用函数组合的市场值函数的分析，我们发现不同权值和效用函数组合的市场值函数，实验结果几乎完全一致，因此在后面的评价中我们只选用了一种效率较高市场值函数，其中权值公式（4-18）

$$\omega_a^3 = D(\Pr(\cdot \mid P) \| \Pr(\cdot)) = \sum_{v \in Va} \Pr(v \mid P) \log \frac{\Pr(v \mid P)}{\Pr(v)}$$

效用函数公式（4-7）

$$u_a^3(v) = \frac{\Pr(v \mid P)}{\Pr(v)} = \frac{|m(v \mid P) \| U|}{|m(v) \| P|}$$

作为市场值函数作用在不同的产品上，来评价市场值函数。

4.4.2.2 数据集描述

同样，我们也选择了 NEC 公司的现实销售数据作为我们的实验数据，该数据集包括 124402 条客户的记录，其中只有 58102 条客户的记录是完整的，因此我们从整个数据集中选取了其中的 58102 条完整记录作为训练和测试数据集。在数据集中客户属性作为数据集的依赖属性，购买产品历史作为决策属性，因此客户属性和不同产品的购买属性结合可以构成多种不同的实验数据集。在本次实验中我们选用了产品号 6、10、12 和 20 的产品的购买属性和客户属性组成四个不同的实验数据集，数据集的描述见各自的数据集描述。

4.4.2.3 产品六上的实验

数据集描述

在数据预处理以后我们得到 58102 条完整的客户记录，选用了前面的 26803 条记录作为训练例，其中包括正例 2781 条，余下的 31298 条记录作为测试例集，其中包括正例 351 条（见表 4-8）。

表 4-8. 产品 6 上的数据库描述

Table4-8. Database Description in Product Id6

Dataset Name	Whole Instances	Positive Instances
Usable Dataset	58102	3132
Training Dataset	26803	2781
Testing Dataset	31298	351

实验结果 见表 4-9 和图 4-2。

表 4-9. 产品 6 测试集上的 Decile 分析

Table4-9. Decile Analysis of Testing Dataset in Product Id6

Decile	Number of Individuals	Number of Responses	Number of Real Responses	Decile Response Rate	Cumulative Response Rate	Cum Response Lift
1	3130	35	100	3.195092	3.195092	285
2	3130	35	55	1.757301	2.476197	221
3	3130	35	41	1.309988	2.08746	186
4	3130	35	37	1.182184	1.861141	166
5	3130	35	29	.9265768	1.674228	149
6	3130	35	20	.6390185	1.501693	134
7	3130	35	19	.6070675	1.37389	123
8	3130	35	16	.5112148	1.266055	113
9	3130	35	18	.5751166	1.189284	106
10	3130	35	16	.5112148	1.121477	100

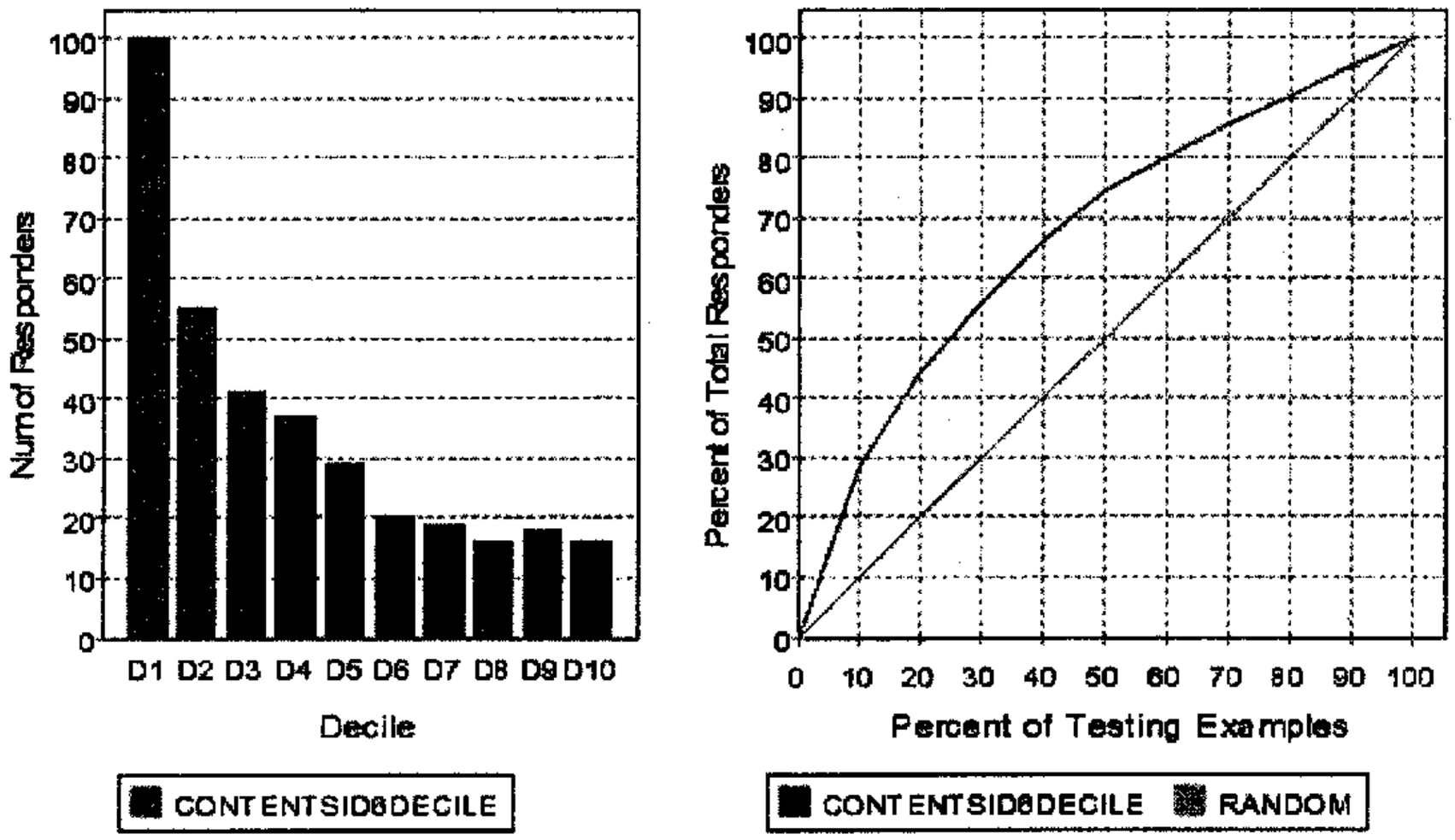

图 4-2 产品六上的市场值函数 Decile 分析图和 Lift 曲线

Fig4-2. The Decile Analysis Bar and Lift Curve of Market Value Function in Product Id6

4.4.2.4 产品十上的实验

数据集描述

在数据预处理以后我们得到 58102 条完整的客户记录，我们选用了前面的 26803 条记录作为训练例，其中包括正例 7527 条，余下的 31298 条记录作为测

试例集，包括正例 4594 条（见表 4-10）。

表 4-10. 产品 10 上的数据库描述

Table4-10. Database Description in Product Id10

Dataset Name	Whole Instances	Positive Instances
Usable Dataset	58102	12121
Training Dataset	26803	7527
Testing Dataset	31298	4594

试验结果　见表 4-11 和图 4-3。

表 4-11. 产品 10 测试集上的 Decile 分析

Table4-11. Decile Analysis of Testing Dataset in Product Id10

Decile	Number of Individuals	Number of Responses	Number of Real Responses	Decile Response Rate	Cumulative Response Rate	Cum Response Lift
1	3130	459	1199	38.30916	38.30916	261
2	3130	459	1042	33.29286	35.80101	244
3	3130	459	738	23.57978	31.72727	216
4	3130	459	477	15.24059	27.6056	188
5	3130	459	366	11.69404	24.42329	166
6	3130	459	216	6.901399	21.50297	146
7	3130	459	186	5.942872	19.2801	131
8	3130	459	167	5.335804	17.53706	119
9	3130	459	118	3.770209	16.00741	109
10	3130	459	85	2.715828	14.67825	100

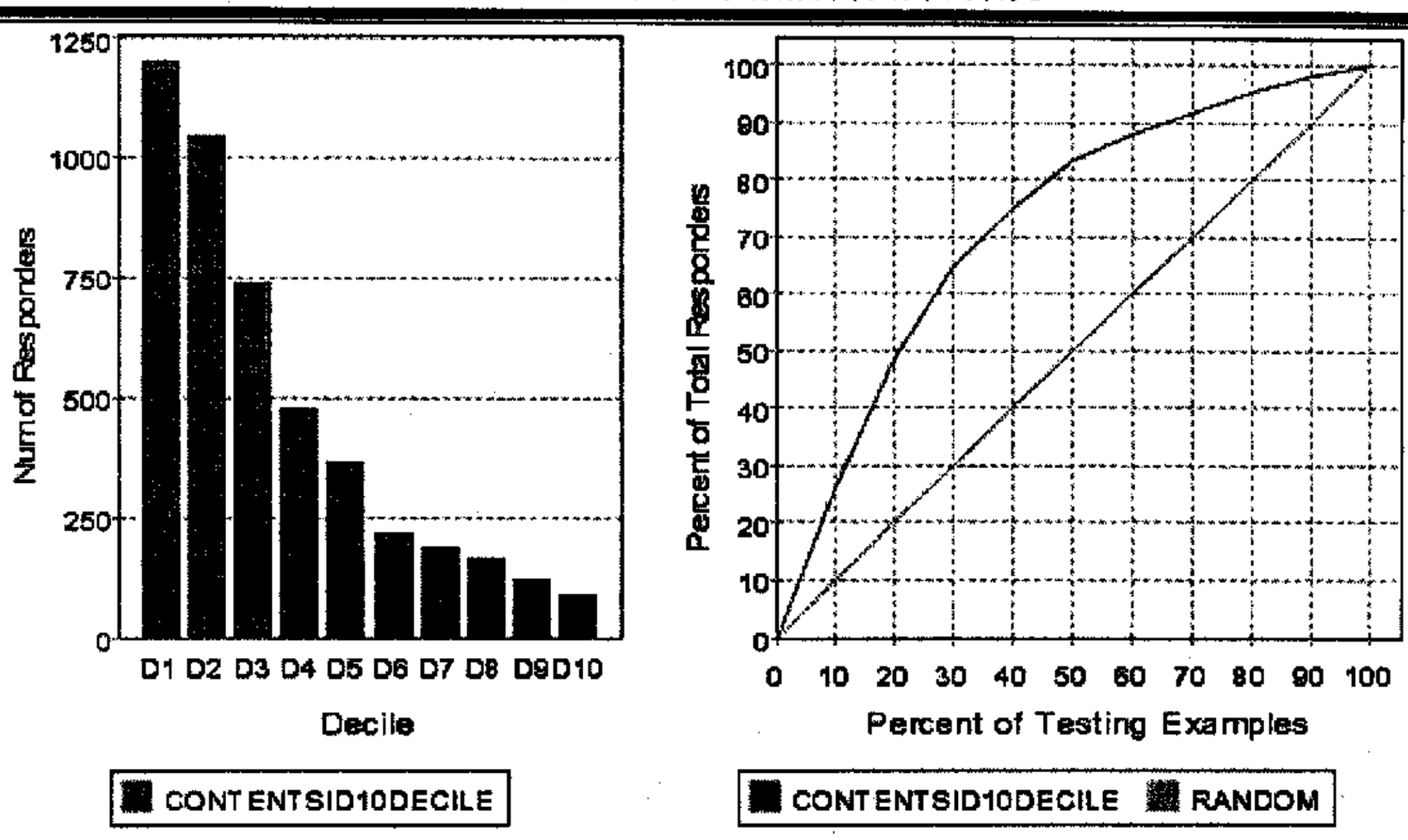

图 4-3 产品十上的市场值函数 Decile 分析图和 Lift 曲线

Fig4-3. The Decile Analysis Bar and Lift Curve of Market Value Function in Product Id10

4.4.2.5 产品十二上的实验

数据集描述

在数据预处理以后我们得到 58102 条完整的客户记录，选用了前面的 26803 条记录作为训练例，其中包括正例 2644 条，余下的 31298 条记录作为测试例集，其中包括正例 1570 条（见表 4-12）。

表 4-12. 产品 12 上的数据库描述

Table4-12. Database Description in Product Id12

Dataset Name	Whole Instances	Positive Instances
Usable Dataset	58102	4214
Training Dataset	26803	2644
Testing Dataset	31298	1570

试验结果 见表 4-13 和图 4-4。

表 4-13. 产品 12 测试集上的 Decile 分析

Table4-13. Decile Analysis of Testing Dataset in Product Id12

Decile	Number of Individuals	Number of Responses	Number of Real Responses	Decile Response Rate	Cumulative Response Rate	Cum Response Lift
1	3130	157	338	10.79941	10.79941	215
2	3130	157	295	9.425522	10.11247	202
3	3130	157	274	8.754553	9.659829	193
4	3130	157	198	6.326283	8.826442	176
5	3130	157	128	4.089718	7.879098	157
6	3130	157	91	2.907534	7.050503	141
7	3130	157	80	2.556074	6.408442	128
8	3130	157	74	2.364368	5.902933	118
9	3130	157	58	1.853154	5.452957	109
10	3130	157	34	1.086331	5.016295	100

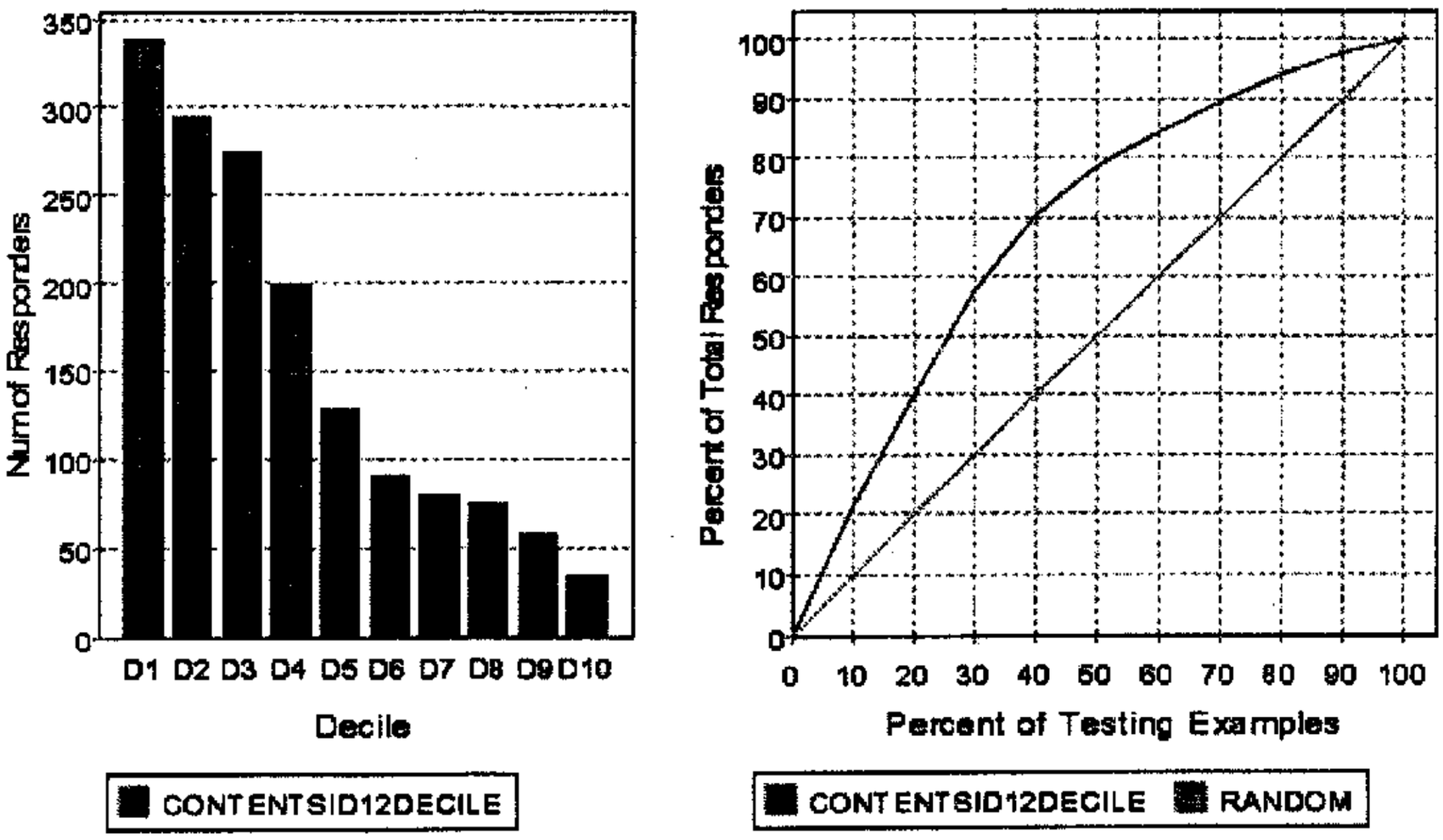

图 4-4 产品十二上的市场值函数 Decile 分析图和 Lift 曲线

Fig4-4. The Decile Analysis Bar and Lift Curve of Market Value Function in Product Id12

4.4.2.6 产品二十上的实验

数据集描述

在数据预处理以后我们得到 58102 条完整的客户记录，选用了前面的 26803 条记录作为训练例，其中包括正例 335 条，余下的 31298 条记录作为测试例集，其中包括正例 180 条（见表 4-14）。

表 4-14. 产品 20 上的数据库描述

Table4-14. Database Description in Product Id20

Dataset Name	Whole Instances	Positive Instances
Usable Dataset	58102	515
Training Dataset	26803	335
Testing Dataset	31298	180

试验结果 见表 4-15 和图 4-5

表 4-15. 产品 20 测试集上的 Decile 分析

Table4-15. Decile Analysis of Testing Dataset in Product Id20

Decile	Number of Individuals	Number of Responses	Number of Real Responses	Decile Response Rate	Cumulative Response Rate	Cum Response Lift
1	3130	18	55	1.757301	1.757301	306
2	3130	18	30	.9585277	1.357914	236
3	3130	18	25	.7987731	1.171534	204
4	3130	18	17	.5431657	1.014442	176
5	3130	18	22	.7029203	.9521375	166
6	3130	18	8	.2556074	.8360491	145
7	3130	18	11	.3514602	.7668222	133
8	3130	18	7	.2236565	.6989264	122
9	3130	18	3	.0958527	.6319182	110
10	3130	18	2	.0639018	.5751166	100

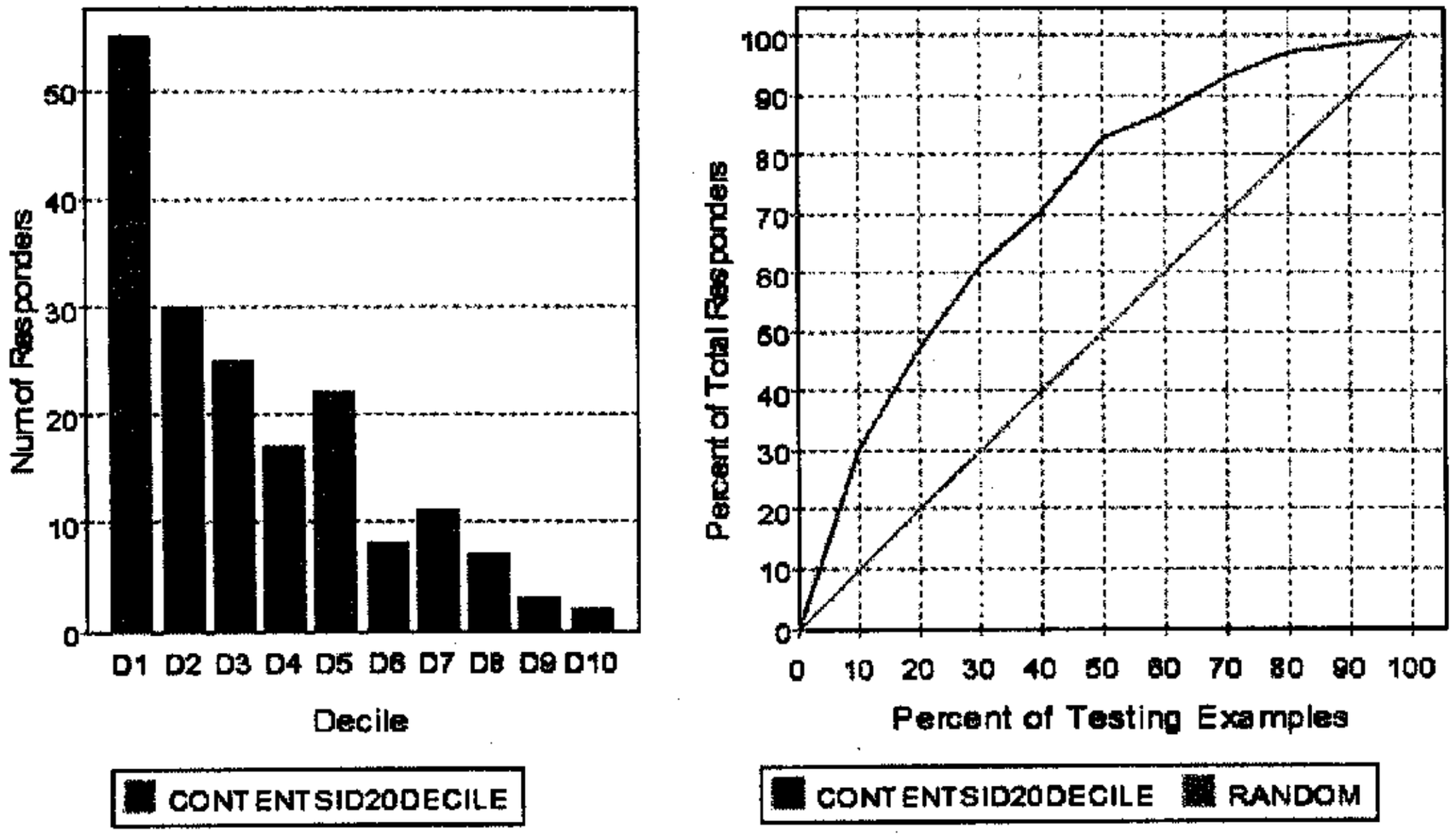

图 4-5 产品二十上的市场值函数 Decile 分析图和 Lift 曲线

Fig4-5. The Decile Analysis Bar and Lift Curve of

Market Value Function in Product Id20

4.4.2.7 各数据集的综合实验结果

各数据集的 Lift 曲线 见图 4-6

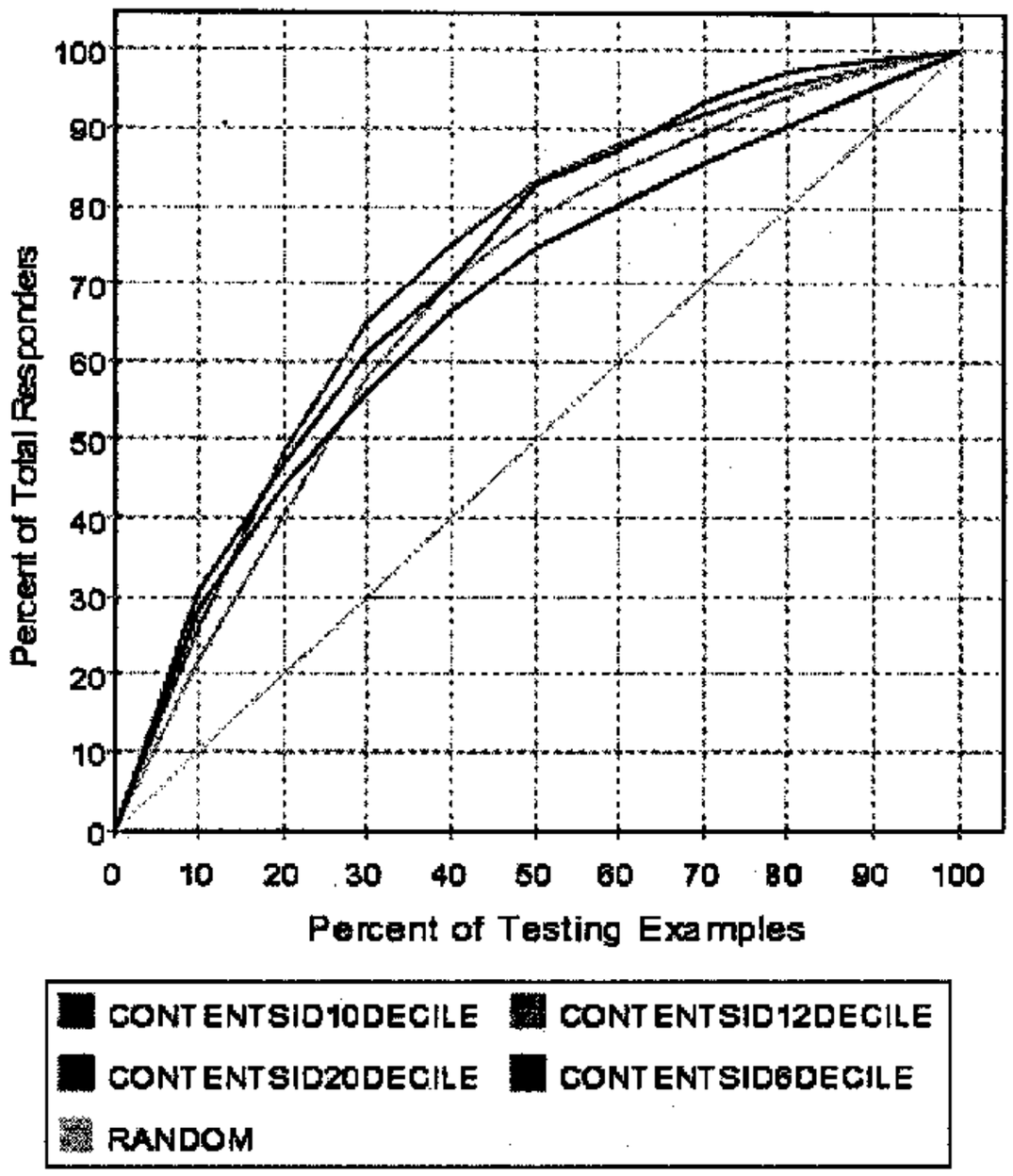

图 4-6 各数据集的 Lift 曲线

Fig4-6. The Lift Curve in Different Dataset

各数据集的市场值函数的 lift 值(见表 4-16)

表 4-16. 不同产品上的 Lift 值

Table4-16. Lift Value in Different Contents

Contents Id	Lift value (%)
6	72.136754
10	77.1724
12	73.426753
20	76.888883

4.4.2.8 结论

从上面四种不同的数据集的实验结果来看，市场值函数在上述数据集上表现良好：4 个数据集的 Lift 值都超过了 70%；在 Decile 分析中第一个 Decile 中累计回应的 Lift 值都在 200 以上，而且真正作出回应的客户的人数在 Decile 中从上到下一次呈递减的规律分布。这些实验结果说明市场值函数方法是一种适合

于目标市场的数据库营销的方法。

4.5 与朴素贝叶斯算法的比较

我们在NEC公司的客户数据集上进行了测试并与朴素贝叶斯方法在该数据集上的结果进行了比较。该数据集中包括58102个成员的完整信息，其中有3856个成员购买了该公司的产品二。我们将数据集中的18963条记录（其中正例为894条）作为训练集生成的模型应用在两个测试例上，其中测试例1有24134条记录，正例为649条，测试例2为整个数据集，表4-17显示了各种方法根据Lift方法评价得到的S_{lift}。图4-7，图4-8分别显示了利用不同的效用函数和权值的线性组合以及朴素贝叶斯方法的结果，横坐标表示Decile评价中的每个Decile的编号，如：1表示第一个Decile，纵坐标表示将排序后的测试例分成10份，每个Decile中正例的数目。

表4-17 两个测试例上各种方法的S_{lift}值

Table4-17 The S_{lift} Value in Different Dataset

算法 / S_{lift} / 测试例	Naive bayes	Wa1 和 ua1	Wa2 和 ua2	Wa3 和 ua3
测试例1	67.6%	67.3%	67.3%	67.1%
测试例2	66.9%	66.0%	66.0%	66.1%

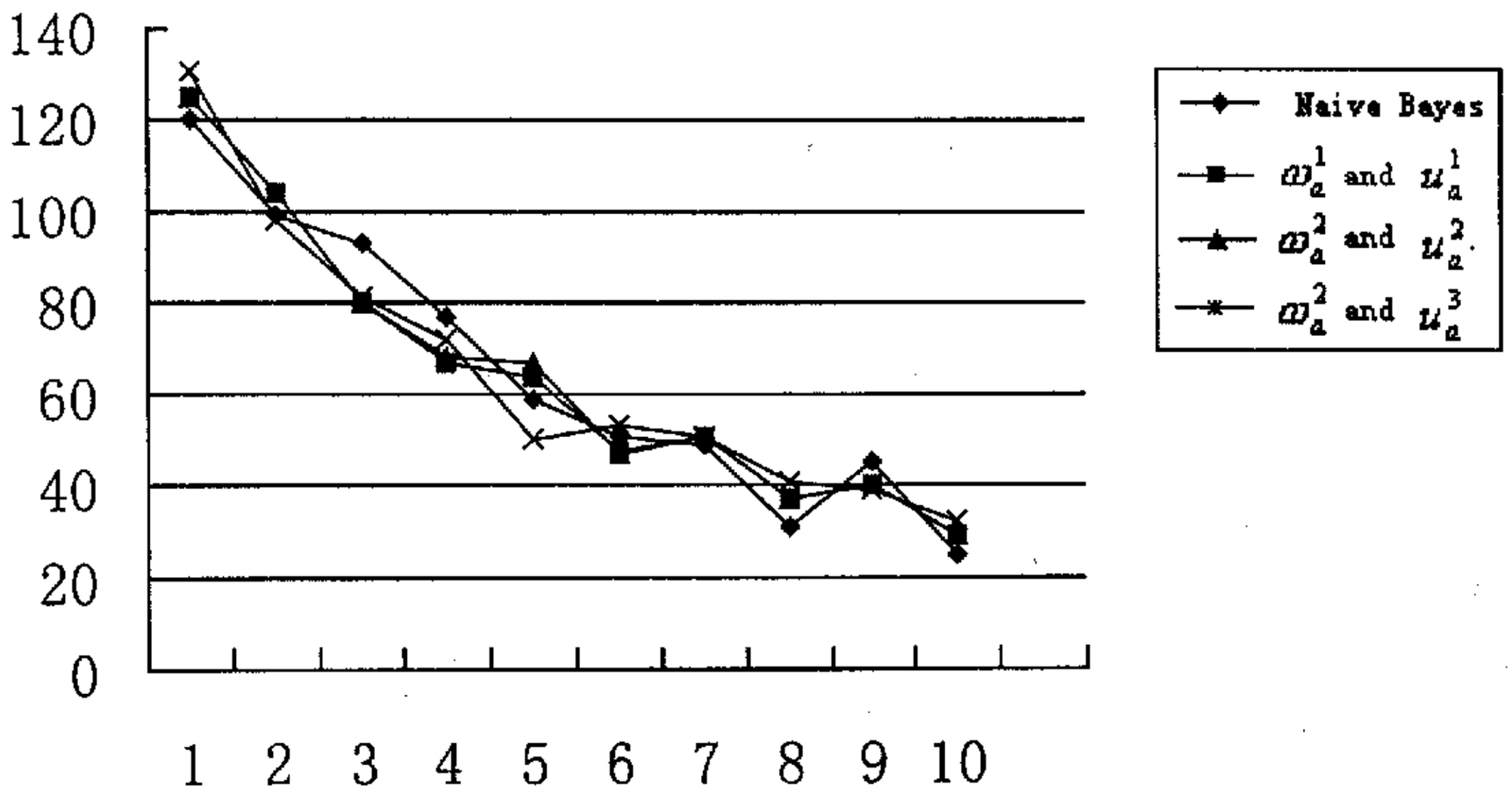

图 4-7　排序后测试例集 1 上回应客户的分布

Fig4-7. The Customer Distribution in Testing Dataset1 after sort

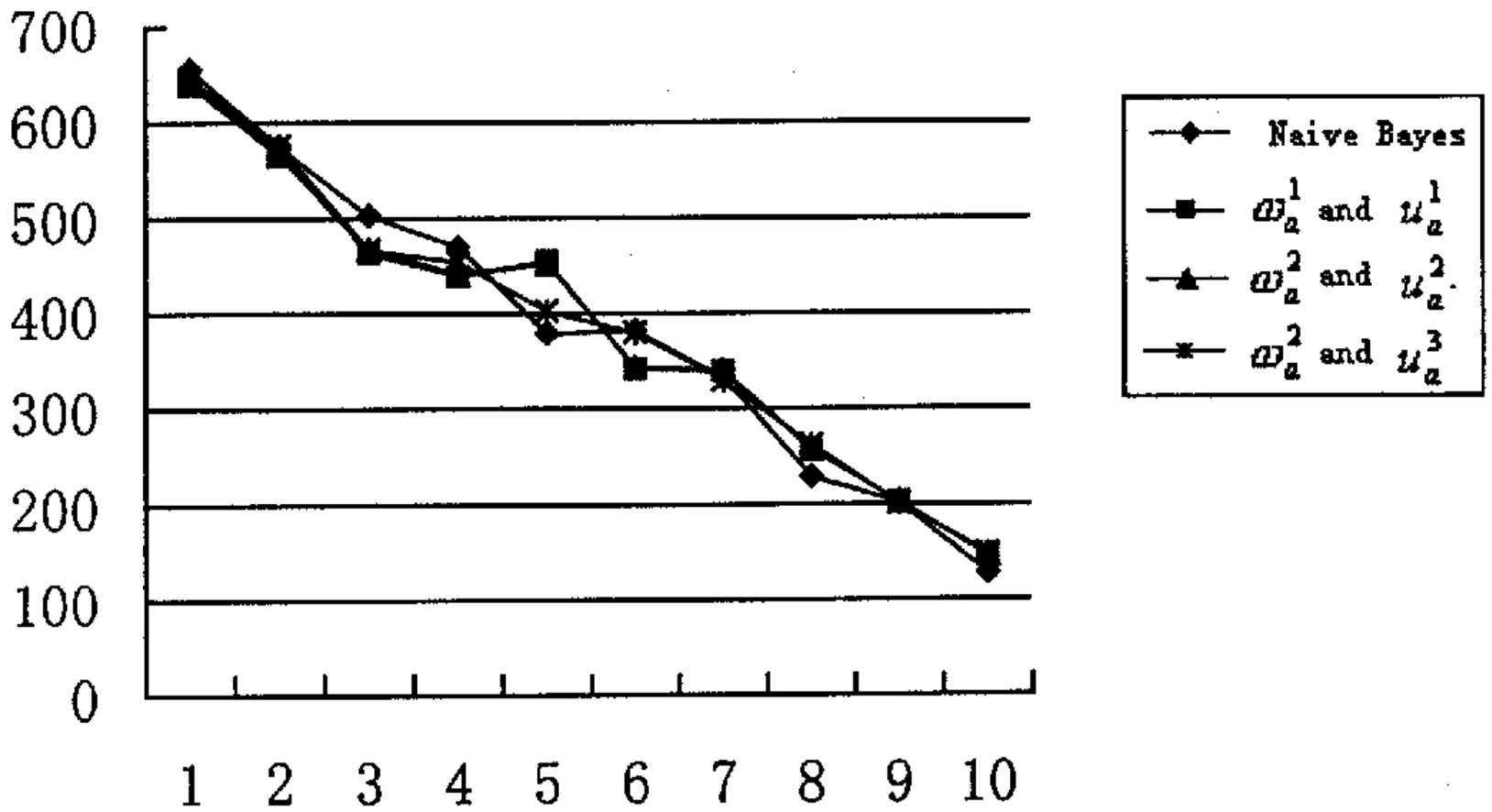

图 4-8 排序后测试例集 2 上回应客户的分布

Fig4-8. The Customer Distribution in Testing Dataset2 after sort

从表 4-17 中我们可以看到，与朴素贝叶斯方法相比，利用该线性模型所得结果还是可以接受的，图 4-7 和图 4-8 更能说明这一点。而该模型较之朴素贝叶斯方

法的优点在于，以信息检索理论依据在完成目标销售中识别顾客问题的同时，还通过计算属性的权值以及各个属性上属性值的效用函数值，向使用者提供了属性的重要性等信息，从而有利于进一步的决策。

4.6 结论

通过对市场值函数在NEC现实数据库的多种不同商品上的实验和与朴素贝叶斯方法的实验结果比较，证明市场值函数的方法可以很好的用于市场数据库的数据库营销，并且能发挥它的优点。在我的后续工作中我们另外又提出了市场值函数的Boosting方法，进一步提高了市场值函数的性能[41]。

4.7 本章小结

本章提出了市场值函数的概念，提出了多种不同的效用函数和权值公式，通过他们的不同组合可以得到不同的市场值函数。它可以把复杂的市场数据库的问题公式化成经典的信息检索的问题，很多的信息检索的理论的结果可以马上用到这个领域。市场值的方法可以让我们产生一个排序的队列，可以提供更灵活的方式来解决目标市场中的客户的选择的问题，并且市场值函数方法具有很好的解释性。通过在NEC现实的数据上的大量试验，表明市场值函数是一种很理想的应用于目标市场数据库营销的方法。

第5章 网络数据库营销

5.1 引言

数据库营销的作用就是利用企业经营过程中收集、形成的各种顾客资料，经分析整理后作为制订营销策略的依据，并作为保持现有顾客资源的重要手段。从理论上说，数据库营销并不是网络营销中特有的手段，在传统营销中，如直邮广告、电话营销等过程中，数据库营销也是一种常用的手段，不过，在网络营销中，数据库营销有着更加独特的优越性，因而成为网络营销的重要策略之一。

在当前的营销模式中，以 Internet 为支撑，在数据库营销基础上发展起来的网络营销日渐盛行，网络数据库营销（Internet Database Marketing）是一种交互式营销处理方法，它通过独特的可记载营销媒体和营销渠道（主要是互联网络，同时还包括电话和销售人员），将公司的目标顾客、潜在顾客的资料，以及进行的交流沟通和商业往来信息存储在计算机的数据库中，对顾客提供更多及时服务，发现顾客新的潜在需求，加强与顾客紧密关系，帮助公司改进营销方法和营销策略，使公司能系统了解市场和把握市场更好满足市场需求。网络数据库营销是从传统的数据库营销发展而来的，它通过利用 Internet 的交互特性直接与顾客进行沟通，顾客通过网络访问企业站点，企业可以直接了解和掌握顾客的数据。因此，利用网络营销企业可以直接与顾客沟通，同时可以简单快捷的收集营销数据，同时网络营销可以在数据库营销的基础上更好了解顾客、服务顾客。网络数据库营销是一种新型、有效的营销方法，目前，有许多大公司对此投入大量资金，如通讯业、计算机业和办公设备供应商中的德尔公司（DELL）、ＩＢＭ公司和施乐公司（Xerox），汽车厂商福特公司（Ford）等。网络数据库营销是近年来，随着计算机技术和网络通讯技术的发展，才逐渐日显威力的，它不仅是现在许多流行营销策略，如电话营销、直复营销等营销策略的有效前提保证和基础，而且意味着以一种新的方法开展业务，新的概念进行营销管理，并产生新型的公司和顾客关系。

5.2 网络数据库营销的特点

当消费者的需求呈现出理性化、个性化和衍生化的特点时，营销者就必须对市场的变化做出及时反应和调整，不断细分目标市场，开发出满足市场需求的产品，以保持在市场竞争中处于有利位置。数字信息技术发展和互联网络的出现，为网络数据库营销的发展奠定了基础，它的主要特点是：

- 营销渠道更多是依赖互联网络，而不是传统的通讯手段和销售人员。
- 网络数据库营销更具效率性和交互性，同时提供的服务和信息交流可以跨越时空限制。特别是由于搜索引擎技术的飞速发展，使互联网络上蕴藏的大量消费者的需求信息和公司的相关信息将会在最短时间内被有效挖掘和利用。
- 网络数据库营销获取信息比较容易，它的重点和难点在于利用信息挖掘知识，即找出有价值的信息，而传统的数据库营销重点和更多的时间是在收集信息和简单分类信息。

5.3 基于数据库营销的商务网站推荐系统

我们在NEC公司的营销数据集上构建了一个基于数据库营销的电子商务网站的推荐系统。

5.3.1 系统功能描述

本系统包括两个主要功能：第一，作为普通的 E-Commerce 网站，面向访问电子商务网站的客户，客户可以访问网站并购买自己喜欢的产品，商家也可以在网站上发布自己的产品；第二，为管理员提供，该部分集中了各种数据挖掘算法和数据库管理操作，管理员可利用这一部分进行数据挖掘操作，利用它在公司的历史数据集上生成有效的数据挖掘模式，并利用它分析数据库中的客户资料并找出回应概率最大的潜在客户群，向他们发送邮件推荐相关的产品，同时系统可以提供各种图表分析工具来评价模式的优劣。用户在登陆 E-Commerce 网站时，系统可以动态的向客户推荐最喜欢的产品。系统的体系结构和功能描

述见图 5-1。

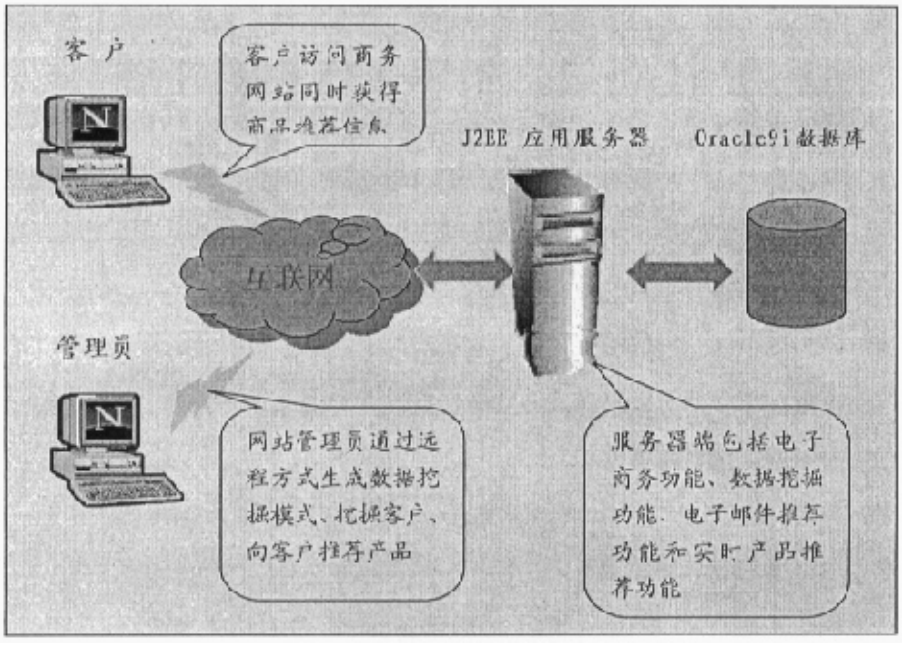

图 5-1 网络推荐系统体系结构和功能描述

Fig5-1. The Architecture and Function Description of Network Recommendation System

5.3.2 系统体系结构

在系统中我们采用了三层的体系结构，即客户层、中间应用层和数据服务层等三个层面。客户层提供用户交互和数据表示，负责收集用户的请求信息提交中间应用层进行业务逻辑处理，并将结果显示给用户；中间应用层接受客户层的服务请求，实现核心业务逻辑处理，将业务逻辑转换为服务器能够执行的命令交服务器执行，并将服务器处理的结果返回给请求者；数据服务层执行中间层提交的数据请求命令，负责应用系统的数据服务。在多层应用结构中，要求层与层之间必须有明确的接口定义，从而保证多层之间可以协作完成应用任务。以上各层间通讯通过公共接口实现，每层只能看到其邻近层的公共接口，一个层的改动对其它层的影响很小，这一规则使体系处于一个合适的位置，可以轻松地扩展并自由升级。

多层结构的优点：

- 瘦客户：由于客户端只负责操作界面的表示逻辑，单用浏览器即可胜任此项工作，减轻了客户机的负担，降低了对客户机的硬件配置要求，真正实现瘦客户；
- 易维护：由于将应用系统的业务逻辑迁移到中间层，当事务处理发生变化时只需更新位于应用服务器上的业务组件模块，不必更新客户端，这样可大大降低系统的维护费用；
- 易扩展：某应用层的变化并不影响其他层，给系统的升级带来了极大的方便。多个应用层可以分布在不同的机器上，当业务逻辑比较复杂时，可以均衡负载配置；
- 重用性强：由于应用层提供客户的共享服务，提高中间层的可重用性；
- 开发效率高：多层结构中各层在逻辑上相互独立，可同时进行各层软件的开发，从而提高系统的开发效率；
- 安全性高：在基于分布式计算的多层结构中，所有业务逻辑都位于服务器端，可以进行业务逻辑的封装，便于进行安全控制，确保系统安全可靠性。

5.3.3 系统主要功能模块介绍

作为一个实用的企业的数据库营销推荐系统，其中最主要核心的部分是数据挖掘功能模块，数据挖掘子系统的好坏决定了推荐系统的性能的优劣。一般来说，数据挖掘系统主要包括五个功能模块：数据管理模块、数据预处理模块、训练模块、测试模块和应用模块。

数据管理模块 在本系统中数据管理模块是基于 Oracle 9i 数据库，数据管理模块的主要功能包括：存储和管理训练数据、测试数据以及在应用模块中将要用到的企业的业务数据；存储算法数据；协调算法数据和业务数据的交互；通过系统向用户显示数据库中的数据。

数据预处理模块 数据预处理模块作为连接数据库和训练模块的中间部分，具有比较重要的作用。预处理模块的主要功能包括：去除原始数据中的噪音数据；

解决属性值丢失的问题；根据算法的不同从原始数据的大量属性中选用可以获得最佳的评价结果的属性集合，达到精简数据，提高运算速度的目的，属性集的选取可以选用遗传算法并结合所选用的学习算法来获得，如 ELSA/ANN 算法的结合[37]，粗糙集方法也可以用来选取属性集，具体可根据学习算法不同，采用不同的算法来选择属性集；对于不能处理连续属性的学习算法，必须将连续属性离散化，离散化算法可以选用 Chi2[38]或者 C4.5 基于熵的离散算法[14]。

训练模块 训练模块是整个系统中的最为关键的模块。其主要功能包括：为学习算法生成模式；统一管理各种学习算法。训练模块是一个开放的模块，新的学习算法可以很自由的添加到训练模块中。对于同一挖掘任务各学习算法在相同的训练例集合上进行训练，生成不同的模式，各模式之间相互独立，互不影响。现在应用在目标市场的目标选择算法主要有两种模式：分割模式和回应模式。分割模式主要包括决策树算法[40]，如 CHAID、AID 和 CART 方法。回应模式主要有：Market Value Functions[6,7]、遗传算法 GMAX[8, 9]、神经网络算法[10]、Naïve Bayesian[39,40]方法以及 Logit/Probit 算法。用户可以根据需要，把不同的学习算法添加到训练模块中。

测试模块 测试模块是学习算法在训练模块中生成模式后，将生成的模式作用在测试例集中，并利用评估方法对在训练模块中生成的模式进行评测。测试模块的主要功能为：对模式进行评测，向用户提供模式的评估结果，用户可以根据评估结果，决定是否要修改算法重新生成模式或者根据评估结果决定选用哪一种模式作为应用模块中的模式。目标市场中的常用的几种评估方法是 Decile 和 Lift 方法。

应用模块 应用模块是面向用户的一个模块，用户可以从生成的多种模式中选取最满意的模式应用到业务中。在该模块中选择客户的方式是根据客户成为潜在客户的概率大小从高到低进行排序，选取前面的最有可能成为潜在客户的客户，从数据库中提取他们的联系方式，采用不同的方式向他们发出促销材料，如采用电话联系、信件或者电子邮件的方式，具体的运作方式可以由用户根据自己的需要来选择。

5.4 数据库营销推荐系统

该推荐系统以 Market Value Functions 算法为基本算法，作用在 NEC 公司的历史业务数据库上，从实验结果来说，该系统具有较好的效果。

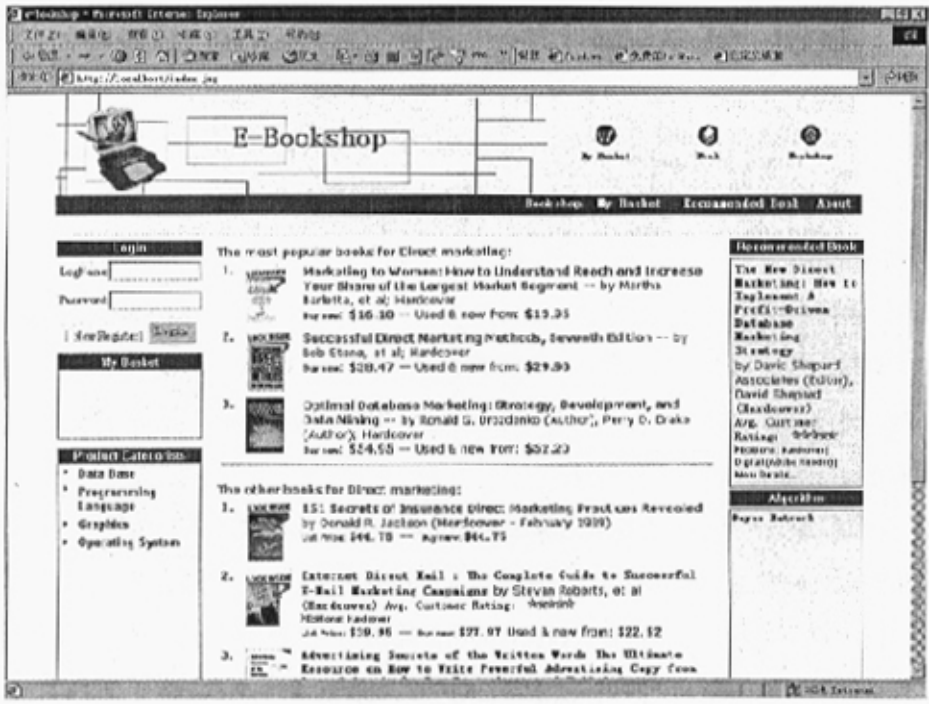

图 5-2 数据库营销推荐系统原型系统（网络版）

Fig5-2. The Prototype of Database Marketing Recommendation System (Network Version)

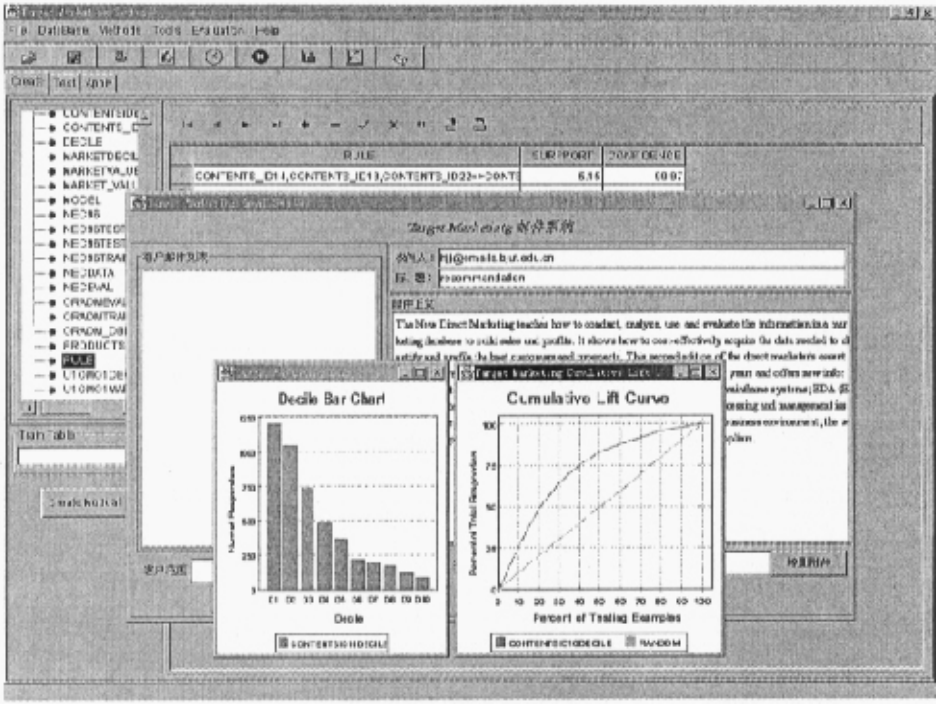

图 5-3 数据库营销推荐系统原型系统（单机版）

Fig5-3. The Prototype of Database Marketing Recommendation System

本系统从公司的客户的历史数据中通过预处理选取了客户资料中的 28 个属性作为基本属性，客户数据中训练例集合和测试例集合各 30000 条记录。首先在 30000 条训练例集合上训练生成模式，然后将生成的模式作用在测试例集中，通过评估方法 decile 分析发现回应者主要分布在 decile 的前面 4 个部分，其中以第一部分居多，并且回应者数量的从 decile 1 到 decile 10 呈递减分布（见图 5-3 中 Decile Bar Chart），可以很好的实现挖掘潜在客户的要求。在应用中用户可以灵活的选取前面的客户分发促销材料，本系统给用户提供了一个邮件系统，用户可以选择客户发送电子邮件。

整个系统集中了训练、测试和应用等几个主要的功能模块，同时系统具有很好的开放性，可以根据需要向系统中添加挖掘算法和功能模块。

5.5 本章小结

面向市场销售的数据挖掘系统的建立对于提升企业竞争力和促进企业的发展具有很重大的现实意义。本章主要探讨了面向市场营销的数据挖掘系统的框架，并以此框架为基础建立了面向市场数据库营销的数据挖掘原型系统。

结论

本文主要从数据库营销的角度来讨论了数据挖掘技术，介绍了当前数据挖掘的在数据营销中的应用，总结了数据挖掘技术在数据库营销中的一般的处理流程、数据挖掘算法在数据库营销应用当中出现的问题、主要应用在数据库营销中的算法以及数据库营销的数据挖掘的一般的评价方法。

在面向市场的数据库营销中，传统的决策树方法遇到了很多的问题。本文所提出的决策树方法可以解决这些问题。该方法的主要创新点在于：它是通过统计决策树的叶子节点的基本信息，通过设置不同的阈值来控制决策树的生长，生成合适的决策树，可以解决市场数据中因为类分布不平衡而不能生成合适的决策树的问题；新的决策树方法解决了传统决策树只能分类而不能排序的问题。通过现实数据的试验，证明这种方法是可行的。

市场值函数算法是一种起源信息检索并由它扩展而来的新的数据挖掘算法，它主要应用于数据库营销的客户选择。它的主要优点有以下：可以根据市场值对客户进行排序而不是简单的分类；具有可解释性；系统的执行不需要专家的指导。市场值函数的主要创新点是：它是一种线性模式，由两部分组成：效用函数和属性权值，通过两者的线性组合可以计算出每个客户的市场值，从而可以对每一个客户进行排序，对客户进行数据库营销。我们在大量的现实销售数据集上进行了多次试验，试验结果证明该方法非常适合于数据库营销。

在数据库营销的应用上，我们也做出了一些尝试，基于市场值函数方法我们建立了一个具有推荐功能的电子商务网站，企业可以通过该系统对客户进行数据库营销。在网络时代，这对于提高企业的营销具有积极的作用。

目前在我国，传统的营销方式仍占据着相当的地位，数据库营销只是对传统营销方式的补充和改变。但数据挖掘应用市场正在逐渐形成，应用前景十分广阔。从长期看，数据库营销必将随着企业管理水平和信息技术的飞速发展而得到创新应用。

参考文献

1 Charles X. Ling and Chenghui Li. Data Mining for Direct Marketing: Problems and Solutions. In Proceeding 4th International Conference on Knowledge Discovery in Databases (KDD-98), New York, 1998.

2 Peter Van Der Putten. Data Mining in Direct Marketing Databases. Walter Baets (ed). Complexity and Management: A Collection of Essays. World Scientific Publishers, (1999) Singapore.

3 J-J. Jonker & P.H. Franses & N. Piersma (2002). Evaluating Direct Marketing Campaigns; Recent Findings and Future Research Topics. Erasmus Research Institute of Management (ERIM), Erasmus University Rotterdam in its series Discussion Paper with number 166.

4 Chuangxin Ou, Chunnian Liu, Jiajing Huang and Ning Zhong: On Data Mining for Direct Marketing. RSFDGrC 2003: 491-498. 64, EE.

5 冯萍等， 数据挖掘技术及其在营销中的应用，北京轻工业学院学报，第 19 卷，第一期，2001.3.

6 Y.Y.Yao, Ning Zhong. Mining Market Value Function for Targeted Marketing. IEEE Computer Society Press (2001) 517-522.

7 Y.Y. Yao, Ning Zhong, Jiajin Huang, Chuangxin Ou and Chunnian Liu: Using Market Value Functions for Targeted Marketing Data Mining, International Journal of Pattern Recognition and Artificial Intelligence. 16(8) (2002) 1117 - 1131.

8 Bruce Ratner, Ph.D. Genetic Modeling in Direct Marketing. Journal of Research Council of Direct Marketing Association, 1999

9 Bruce Ratner, Ph.D. Finding the Best Variables for Direct Marketing Models. Journal of Targeting Measurement and Analysis for Marketing,

2001, vol. 9, no. 3, pp. 270-296

10 Rob Potharst, Uzay Kaymak, Wim Pijls. Neural Networks for Target Selection in Direct Marketing. (Book chapter), in: Kate A. Smith and Jatinder N. D. Gupta (eds.) Neural Networks in Business: Techniques and Applications, Idea Group Publishing, ISBN 1-930708-31-9, pp. 89-110.

11 C. J. van Wieringen and M. D. de Gelder. Genetic Initialization of Neural Networks for Target Selection in Direct Marketing.

12 Elkan C. (1997). Boosting and Naive Bayesian Learning. Technical Report No. CS97-557, September 1997, UCSD. Jim Georges and Anne H. KDD99 Competition: Knowledge Discovery Contest.

13 D. Van den Poel and Z. Piasta:Purchase Prediction in Database Marketing with the ProbRough System, L. Polkowski and A. Skowron (eds.) LNAI 1424 (1998) 593-600

14 Tom M. Mitchell: Machine Learning. The McGraw-Hill Company, Inc. (1997)

15 Han Jiawei. 数据挖掘概念与技术（中文版）. 机械工业出版社. 2001

16 Nils J. Nilsson. 人工智能（中文版）. 机械工业出版社，2000.

17 R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large databases," Proc. ACM SIGMOD Int. Conf. Management of Data, 1993, pp. 207-216.

18 P. C. Fishburn, Seven independence concepts and continuous multi attribute Utility functions. *J. Math. Psychol.* 11 (1974) 294 -327.

19 J. Han, Y. Cai and N. Cercone, Data-driven discovery of quantitative rules in relational databases. IEEE Trans. Knowl. Data Engin. 5 (1993) 29-40.

20 S. Kullback and R. A. Leibler, On information and sufficiency. Ann.

Math. Stat. 22 (1951) 79-86.

21 D. B. Leake, Case-Based Reasoning, AAAI Press, 1996.

22 Z. Pawlak, Rough Sets, Theoretical Aspects of Reasoning about Data, Kluwer, 1991.

23 J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufman, 1993.

24 C. R. Rao, Diversity and dissimilarity coefficients: a unified approach. *Theoret. Popul. Biol.* 21 (1982) 24-43.

25 S.E. Robertson, On relevance weight estimation and query expansion. *J. Document* 42 (1986) 182-188.

26 S.E. Robertson and K. Sparck Jones, Relevance weighting of search terms. *J. Amer. Soc. Inform. Sci.* 27 (1976) 129-146.

27 G. Salton and M. H. McGill, Introduction to Modern information Retrieval, McGraw-Hill, NY, 1983.

28 K. Sparck Jones and P. Willett, Readings in Information Retrieval, Morgan Kaufman, 1997.

29 S. Watanabe, Pattern recognition as a quest for minimum entropy. Patt. Recogn. 13 (1981) 381-387.

30 S. K. M. Wong and Y. Y. Yao, A probability distribution model for information retrieval. Inform. Process. Manag. 25 (1989) 39-53.

31 S. K. M. Wong and Y. Y. Yao, A generalized binary probabilistic independence model. *J. Amer. Soc. Inform. Sci.* 41 (1990) 324-329.

32 S.K.M. Wong and Y. Y. Yao, An information-theoretic measure of term specificity. *J. Amer. Soc. Inform. Sci.* 43 (1992) 54-61.

33 Y. Y. Yao, S. K. M. Wong and C. J. Butz, On information-theoretic measures of attribute importance. Methodologies for Knowledge Discovery and Data Mining, Eds. N. Zhong and L. Zhou, LNAI No. 1574,

Springer, 1999, pp. 479-488.

34 Y. Y. Yao and N. Zhong, Granular computing using information tables. Data Mining, Rough Sets and Granular Computing, Eds. T. Y. Lin, Y. Y. Yao and L. A. Zadeh, Physic-Verlag, Heidelberg, 2002, pp. 102-124.

35 N. Zhong, Y.Y. Yao and S. Ohsuga, Peculiarity oriented multi-database mining. Principles of Data Mining and Knowledge Discovery, Eds. J. Zytkow and J. Rauch LNAI No. 1704, Springer, 1999, pp. 136-146.

36 N. Zhong, J. Z. Dong and S. Ohsuga, Rule discovery by soft induction techniques. Neurocomput. Int. J. 36 (2001) 171-204.

37 YongSeog Kim, W. Nick Street, Filppo Menczer. An Evolutionary Multi-objective Local Selection Algorithm for Customer Targeting. Proceedings of the 2001 Congress on Evolutionary Computation CEC2001. Page 759-766

38 Liu H, & Setiono R (1995) Chi2: Feature selection and discretization of numeric attributes. Proceedings of the 7th IEEE International Conference on Tools with Artificial Intelligence, pp. 388--391.

39 Elkan C. (1997). Boosting and Naive Bayesian Learning. Technical Report No. CS97-557

40 J.R.Quinlan.C4.5: Bagging, boosting, and C4.5. In Proc.13th Natl. Conf. Artificial Intelligence (AAAI' 96), Portland, 1996

41 苌彩卿、刘椿年、黄佳进、欧创新：基于 Boosting 的市场值函数算法及其评价。北京工业大学学报（2004）。

致谢

我的论文得以顺利的完成，在此我要深深地感谢我的导师刘椿年教授和钟宁教授，是他们在这期间给了我悉心的指导，给我指明了课题的研究方向。刘老师和钟老师的严谨的治学作风给我留下了深刻的印象，感谢他们在这三年中对我生活的关怀。

我的这篇论文是在我周围的若干合作者的帮助下完成的。我要特别感谢同一课题组的黄佳进和苌彩卿同学，没有他们的密切合作，我的工作和论文是难以顺利完成的。同时感谢实验室所有的其他同学，感谢他们在这三年里对我学习和生活的关怀与帮助。

我还要感谢我的父母家人和朋友，是他们在这三年的学习中给了我无私的关怀和帮助，使我能顺利完成我的学业。

再次感谢你们！