摘　要

采购优化包括在企业生产中对原材料的消耗量进行预测，以及根据预测到的原材料消耗量来制订最优的采购方案。采购优化的目标是降低采购占用的资金，同时降低采购成本。在现代企业管理中，大多数企业的采购资金在周转资金中占用重要的比例，因此，制订合理的采购方案，对企业的经营具有重要的现实意义。

制订合理的采购方案，首先必须从大量业务数据中分析原材料的消耗规律，进而预测原材料的消耗量。从20世纪90年代以来，数据挖掘技术开始逐渐应用于解决类似的商业问题。近年来，得益于计算机运算能力的不断提高，数据挖掘的研究方向开始侧重于如何使用搜索优化算法来对数据进行分析，并挖掘出所需要的信息。本文的主要工作是如何利用基于搜索优化算法的数据挖掘方法，来对采购进行优化。

本文中将采购优化问题分解为两个子问题。一是对原材料消耗量的预测问题，二是根据所预测的消耗量制订最优采购方案。对于第一个问题，把原材料的消耗看作是受产品影响的概率模型。根据统计出来的消耗量数据，使用极大似然法计算原材料消耗概率模型的参数。在求解概率模型参数时，把极大似然参数估计问题转换成为约束优化问题，并应用自适应复合形法对参数进行求解。对于采购方案的制订问题，在对数据分析的基础上，找出采购方案的影响因素，并依据这些因素，给出了一个采购成本计算模型，最后通过对该模型的最优化，得到了最优采购方案。

本文中的方法可以在线使用，而且适用于多种概率模型，易于扩展，具有一定的通用性。通过在某印染企业中半年以来的应用，表明这种方法能够依据历史数据来推断未来原材料的消耗规律，并能根据原材料的消耗制订合理的采购方案，实现了采购优化的目标。

关键词：数据挖掘；采购优化；极大似然法；自适应复合形法

The study and application of data mining method on the optimization of procurement process

Abstract

The optimization of procurement includes the prediction on the raw materials in the production of the enterprise and subsequent decision making of procurement. The aim of the procurement optimization is to reduce the occupied capital for procurement and the cost of products. In modern enterprise management, the procurement capital takes great part of the turnover capital, so it is of great importance to make reasonable purchasing decisions.

First the consuming amount of raw materials must be predicted. To be correct, the rule must be analyzed from large volume of data. From 1990s, data mining technology began to be applied to solve the business problem. In these years, due to the ceaseless advancement in computers' computing ability, the research direction of data mining starts to emphasize particularly on how to analyze data and get needed information by optimal searching algorithms. The work of this paper is how to optimize the procurement by the combination of data mining method and optimal searching algorithms.

The problem is divided into two sub problems. One is the prediction of the consuming of raw materials, and the other is make optimal purchasing plan. For the first one, the consuming of the raw materials is considered as a probability model swayed by products. From the historic data of products' producing records, the parameters of the assistant materials' probability model could be worked out by the method of maximum likelihood. The problem of the solution-finding for the parameters is changed to a problem of constraint optimization. In this case the improved "complex method" algorithm is applied considering that it fits kinds of probability models. For the purchasing plan-making problem, on the basis of data analysis, the influencing elements are found and a procurement cost model is constructed. Then through the optimization for the model, the purchasing plan is made.

The method referred in this paper can be used online, and independent of the statistics model, so that it is extendable and universal. By applying in the printing and dying enterprise for nearly half years, this method is proved to be helpful to deduce the rule of the consuming of raw materials from historic data, and the purchasing plan can be made, so that the purpose is realized.

Key Words: Data Mining; Procurement Process Optimization; Maximum Likelihood Principle; Adaptive "Complex" Method

独创性说明

作者郑重声明：本硕士学位论文是我个人在导师指导下进行的研究工作及取得研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写的研究成果，也不包含为获得大连理工大学或者其他单位的学位或证书所使用过的材料。与我一同工作的同志对本研究所做的贡献均已在论文中做了明确的说明并表示了谢意。

作者签名：丁育艳 日期：2008.01.06

大连理工大学学位论文版权使用授权书

本学位论文作者及指导教师完全了解“大连理工大学硕士、博士学位论文版权使用规定”，同意大连理工大学保留并向国家有关部门或机构送交学位论文的复印件和电子版，允许论文被查阅和借阅。本人授权大连理工大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，也可采用影印、缩印或扫描等复制手段保存和汇编学位论文。

作者签名：隋艳苏

导师签名：郭禾

2008 年 01 月 06 日

1 绪论

1.1 研究背景及意义

印染企业每年需要花费近亿的资金用来采购生产和工程中所需的各种原材料。在采购过程中由于不能全面、及时、准确地了解生产上的用料需求、供应商的资质、采购历史情况等信息，不能及时对库存储备定额进行科学准确地测算，使得在采购过程中可能造成部分采购决策不准确或不科学。如果每年形成 1%的失误就会为企业带来近百万的损失。因此印染企业的采购迫切需要一套帮助指导工作的采购优化系统。

采购优化是大多数生产型企业所面临的问题。以印染企业为例，采购工作分为原料的采购和物料的采购。统计结果表明，其原料的采购成本占其总成本的 70%以上。而物料种类繁多，在生产中也占有一定的地位，其占用库存较多，库存维护成本较高。但是目前物料没有科学的依据来指导采购，采购人员凭经验根据季节性来估计需求，主观判断的不确定性经常会使库存过剩，造成资金流失，或者偶尔出现缺货现象，对生产造成一定的滞后影响。研究物料的消耗规律可以帮助采购部门根据生产计划预测其消耗量，确定采购数量。从多个供应商中选择合适的供应商并按适当的比例分配采购量，以达到采购费用最少，从而制订出合理的采购计划，最终实现控制企业采购所需资金，即达到采购优化的目的。

本文中面临的两个问题为：物料消耗量的预测和供应商的选择与采购量分配问题。这类有着明确的商业目标、同时需要根据数据挖掘结果对决策结果在商业应用中进行优化部署的问题，正适合用最优化方法进行解决。尽管在最优化方法和数据挖掘技术结合方面的研究已取得一定成果，但在实际应用中，目前还没有一个面向实际管理决策问题，将最优化方法与数据挖掘技术有机地结合起来应用的理论方法体系和问题求解模型。因此，本文以数据挖掘的统计方法分析历史数据，建立适合企业原材料库存与采购的模型，然后利用最优化方法求解模型中的参数，最后解决上述两个问题。

本文在论述了已有数据挖掘的过程模型的基础上，基于数据挖掘与最优化结合对支持最终采购决策分析的管理问题模型进行求解，目标是有效地将数据挖掘技术与最优化方法在实际应用中有机地结合起来，并为复杂的采购管理决策分析问题的求解和决策实施提供一个可以依赖的参考模型。其中的自适应复合形法求解模型中参数的方法适用于多种模型，当概率模型有一定的变化时，同样可以自适应复合形法对模型的参数求解。

1.2 国内外研究现状

1.2.1 数据挖掘的国内外研究现状

数据挖掘是一个利用各种分析工具在海量数据中发现模型和数据间关系的过程，这些模型和关系可以用来做出预测[1]。数据挖掘涉及到许多其它的研究领域，包括多元统计(主要有组件分析、聚簇分析和多维缩放)，数据库接口(协作数据库接口、模糊查询接口、数据智能浏览)，和信息检索(近似匹配算法)。数据挖掘的工作，主要是使用半自动化的方式来进行知识获取。知识发现和数据挖掘(Data mining)是将为决策者提供重要的、前所未料的信息或知识，从而产生不可估量的效益。研究 KDD 和 DM 技术的重大意义已被人们广泛地认识到，并且被列为数据库研究领域中最重要的课题之一。例如美国政府开发 Sequoia 2000 项目作为大规模数据库中先进的数据分析工具，许多商业公司也充分认识到了深层次地分析本公司业务数据库中的数据能够带来更多的商业利润，例如银行和零售商店通过分析他们的业务数据，进一步掌握和了解客户的信誉、习惯和消费心理，相应地调整他们的市场决策，以拓宽更广泛的市场。

目前，数据挖掘的研究热点包括网站的数据挖掘(Web site data mining)、生物信息或基因(Bioinformatics/genomics)的数据挖掘及其文本的数据挖掘(Textual mining)等等。

(1) 网站的数据挖掘

随着 Web 技术的发展，各类电子商务网站风起云涌，如何让电子商务网站有效益是一个关键问题。电子商务网站每天都生成大量的记录文件和登记表，如果能对这些数据进行分析和挖掘，充分了解客户的喜好、购买模式，并设计出满足于不同客户群体需要的个性化网站，必能增加商家的竞争力。在对网站进行数据挖掘时，所需要的数据主要来自于两个方面：一方面是客户的背景信息，主要来自于客户的登记表；另外一部分数据主要来自浏览者的点击流，主要用于考察客户的行为表现。就分析和建立模型的技术和算法而言，网站的数据挖掘和原来的数据挖掘差别不大，很多方法和分析思想都可以运用。所不同的是网站的数据格式有很大一部分来自于点击流，和传统的数据库格式有区别。因而对电子商务网站进行数据挖掘所做的主要工作是数据准备。目前，有很多厂商正在致力于开发专门用于网站挖掘的软件。

(2) 生物信息或基因的数据挖掘

生物信息或基因数据挖掘在商业上很难讲有多大的价值，但对于人类却受益非浅。无论在数据的复杂程度、数据量还有分析和建立模型的算法而言，生物信息或基因的数据挖掘比通常的数据挖掘要复杂得多。从分析算法上讲，更需要一些新的和好的算法。

现在很多厂商正在致力于这方面的研究。但就技术和软件而合一，还远没有达到成熟的地步。

(3) 文本的数据挖掘

无论是在数据结构还是在分析处理方法方面，文本数据挖掘与数据库中的数据挖掘相差很大。文本数据挖掘并不是一件容易的事情，尤其是在分析方法方面，还有很多需要研究的问题。目前市场上有一些类似的软件，但大部分方法只是把文本移来移去，或简单地计算一下某些词汇的出现频率，并没有真正的分析功能。

随着计算机计算能力的发展和业务复杂性的提高，数据的类型会越来越多、越来越复杂，数据挖掘将发挥出越来越大的作用。当前，数据挖掘研究与开发的总体水平相当于数据库技术在20世纪70年代所处的地位，迫切需要类似于关系模式、DBMS系统和SQL查询语言等理论和方法的指导，才能使其应用得以普遍推广。预计在本世纪，数据挖掘的研究焦点可能会集中到以下几个方面[2]：

(1) 发现语言的形式化描述，即研究专门用于知识发现的数据挖掘语言，也许会像SQL语言一样走向形式化和标准化；

(2) 寻求数据挖掘过程中的可视化方法，使知识发现的过程能够被用户理解，也便于在知识发现的过程中进行人机交互；

(3) 研究在网络环境下的数据挖掘技术(Web mining)，特别是在因特网上建之数据挖掘服务器，并且与数据库服务器配合，实现Web mining；

(4) 加强对各种非结构化数据的挖掘(Data mining for audio & video)，如对文本数据、图形数据、视频图像数据、声音数据乃至综合多媒体数据的挖掘；

(5) 处理的数据将会涉及到更多的数据类型，这些数据类型或者比较复杂，或者是结构比较独特。为了处理这些复杂的数据，就需要一些新的和更好的分析方法和模型，同时还会涉及到为处理这些复杂或独特数据所做的准备的一些工具和软件。

1.2.2 数据挖掘与优化方法结合的研究现状

数据挖掘技术的应用十分广泛，尤其在商业经营和企业决策支持等领域。数据挖掘的最终目标是使决策者根据挖掘得到的分析结果，优化商业决策和行为，进而增加企业的效益。遗憾的是，在实际应用中，传统的数据挖掘技术只是提供了一套数据挖掘应用的方法，对于领域知识表示和数据挖掘过程的集成、挖掘结果的解释、部署等方面很少涉及，缺乏对商业决策和行为优化的能力，限制了数据挖掘的大量应用。而这类有着明确的商业目标同时需要根据数据挖掘结果对决策结果在商业应用中进行优化部署的问题，正适合用最优化技术进行解决[3]。

基于此，将最优化方法和数据挖掘技术相结合，二者互相补充，互相利用，则可有效解决目前存在的这一问题。目前，二者结合的方式主要有两种：一种是在一个复杂的决策分析过程中，最优化方法和数据挖掘方法相互替代，利用各自的优势，从而得到更好的决策结果，这类方式具有较强的通用性和灵活性；另一种是直接将数据挖掘和最优化混杂的过程所要解决的问题定义为一个复杂的最优化问题，通过遗传算法等技术求取近似解，从而达到较好的优化效果。这种方式和问题域强相关，且最后的优化问题会十分复杂，其优点是可能带来非常好的结果。尽管在最优化方法和数据挖掘技术结合方面的研究已取得一定成果，但在实际应用中，目前还没有一个面向实际管理决策问题，将最优化方法与数据挖掘技术有机地结合起来应用的理论方法体系和问题求解模型。

1.2.3 约束优化问题求解方面的研究现状

在模型求解方面，由于模型中目标函数和约束条件是非线性的，因此在求解上的算法也是多种多样的，一般针对模型的特点采用不同的求解方法，非线性规划模型一般是采用迭代算法求解。而从数学的角度，求解非线性规划模型的方法，随着近年来计算机技术的飞速发展，不断的有所创新。

(1) 基于单纯形的方法，此法的主要思想是通过将非线性目标规划转化为近似的线性目标规划，以便使用线性规划的单纯型法。

(2) 直接搜索方法，这种方法是把给定的非线性多目标规划问题转化为一组单目标非线性规划问题，然后，使用解决单目标非线性规划的直接搜索方法加以解决。

(3) 基于梯度的方法，该方法主要思想就是利用梯度来确定一个求解的可行方向，以可行方向为基础求解目标规划。

(4) 进化算法。如遗传算法、模拟退火算法、微粒群算法等，这些算法的可以处理结构复杂的非线性规划模型，优点在于在求解过程中，无须我们考虑函数的导数和连续性问题。

在非线性规划模型求解方法上，发展最快的就是进化算法，尤其是遗传算法，遗传算法在求解非线性规划问题方面，具有很高的鲁棒性，自上个世纪 70 年代初问世以来，发展极为迅速。遗传算法(Genetic algorithms，AG)研究的历史比较短，20 世纪 60 年代末期到 70 年代初期，主要由美国 Michigan 大学的 John Holland 与其同事、学生们研究形成了一个较为完整的理论与方法[4]，从试图解释自然系统中生物的复杂适应过程入手，模拟生物进化的机制来构造人工系统的模型。随着 20 世纪余年的发展，取得了丰硕的应用成果和理论研究的进展，特别是近年来世界范围形成的进化计算热潮，计算智能已作为人工智能研究的一个重要方向，以及后来的人工生命研究的兴起，使得遗传算法受

到广泛的关注。从 1985 年在美国卡耐基.梅隆大学召开的第一届国际遗传算法会议(International conference on genetic algorithms: ICGA'85)，到 1997 年 5 月 IEEE 的 Transactions on evolutionary computation 创刊，遗传算法作为具有系统优化、适应与学习的高性能计算和建模方法的研究渐趋成熟。另外，其他的进化算法如模拟退火算法和微粒群算法方面，近年来发表的文章也很多，取得的理论成果及应用成果也是有目共睹的。

1.3 本文的主要工作和论文的组织结构

本文详细讨论了企业采购优化的问题，在无法直接得到原料与物料消耗关系的情况下，把物料的消耗看作是受原料影响的概率模型。应用统计出来的原料消耗数据，使用极大似然法得出物料消耗概率模型参数。在此基础上根据未来生产计划中预定的原料数量来预测未来物料消耗量，从而对采购数量进行优化控制。在求解物料消耗概率模型时，把极大似然参数估计转换成为约束优化问题，并应用改进的复合形法进行求解，便于在多种概率模型上计算，得到的结果在企业生产中进行了验证。在得到预测需求量以后，对企业原材料的供应商选择与采购分配量问题进行了分析并建模，并同样应用自适应复合形法对模型的参数进行求解。最终，给出了系统的整体实现结果。

全文的组织结构如下：

第 1 章绪论部分，主要介绍了本文的研究背景、意义及应用领域，数据挖掘、优化方法等国内外的研究现状。

第 2 章数据挖掘理论部分，主要介绍了数据挖掘的主要功能、常用技术和数据挖掘的应用领域及研究成果，并介绍了数理统计与数据挖掘的关系以及数理统计在预测型数据挖掘中的应用。

第 3 章介绍了数值优化理论与用于数值计算的常用优化方法，详细介绍了约束优化方法的概念，及用于求解约束优化问题的常用方法。

第 4 章物料的概率模型及问题求解是本文的重点之一，该部分以某企业为背景，详细描述了产品与物料消耗的关系及物料消耗模型的建立，将模型参数估计转化为最优化问题，应用自适应复合形法求解参数，最后给出了试验的结果及实现效果。

第 5 章采购优化部分也是本文的重点之一，该部分讨论了供应商选择和订购量分配问题，建立了基于确定需求的采购优化模型。模型中的参数同样以自适应复合形法进行求解，最终实现了采购优化系统，说明了所提出方法的可行性与有效性。

第 6 章对全文做总结，并给出了对未来工作的设想与展望。

2 数据挖掘理论与方法

2.1 数据挖掘概述及应用

2.1.1 数据挖掘的概念

所谓数据挖掘就是设计一套数学模型、算法或软件系统，用以从数据库中找出某一类特有性质数据的分布规律，也就是归纳现有海量数据中某类数据分布的知识[5]。在已有的数据挖掘技术中，最常用的是统计分析、回归分析、聚类分析等。应用这些技术的基础是设计数学模式，对大量的数据进行过滤检查，也就是挖掘[6]。

无论是商业企业、科研机构或者政府部门，在过去若干年的时间里都积累了海量的、以不同形式存储的数据资料。由于这些资料十分复杂，要从中发现有价值的信息或知识，达到为决策服务的目的，成为非常艰巨的任务。数据挖掘方法的提出，让人们有能力最终认识数据的真正价值，即蕴藏在数据中的信息和知识。目前的数据库系统可以高效地实现数据的录入、查询、统计等功能，但无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势。缺乏挖掘数据背后隐藏的知识的手段，导致了“数据爆炸但知识贫乏”的现象。这就需要新的技术和工具来帮助人们自动地提取和分析隐藏在这些数据中的知识，数据挖掘(Data mining)，也称知识发现(KDD，Knowledge discovery in database)就是致力于这门学科。

数据挖掘(Data mining)，一种比较公认的定义是 W.J.Frawley, G.Piatetsky Shapiro 等人提出的[7,8]：数据挖掘，就是从大型数据库的数据中提取人们感兴趣的知识。这些知识是隐含的、事先未知的潜在有用信息，提取的知识表示为概念(Concepts)、规则(Rules)、规律(Regularities)、模式(Patterns)等形式。这种定义把数据挖掘的对象定义为数据库。而更为广义的说法是：数据挖掘意味着在一些事实或观察数据的集合中寻找模式的决策支持过程。数据挖掘的对象不仅是数据库，也可以是文件系统，或其它任何组织在一起的数据集合。数据挖掘确切地讲是一种决策支持过程，它主要基于人工智能、机器学习、统计学等技术，高度自动化地分析企业原有的数据，做出归纳性的推理，从中挖掘出潜在的模式，预测客户的行为，帮助企业的决策者调整市场策略，减少风险，做出正确的决策。

2.1.2 数据挖掘的主要功能及模式

数据挖掘的主要功能是确定数据挖掘任务中要找的模式类型，数据挖掘任务一般可以分为描述和预测两大类，描述性挖掘任务主要是刻画数据库中数据的一般特性，预测

性挖掘任务是在当前数据上进行推断，以进行预测。如图2.1所示，每类模型下都包含一些需要用到该类模型的最常用的数据挖掘任务[9]。

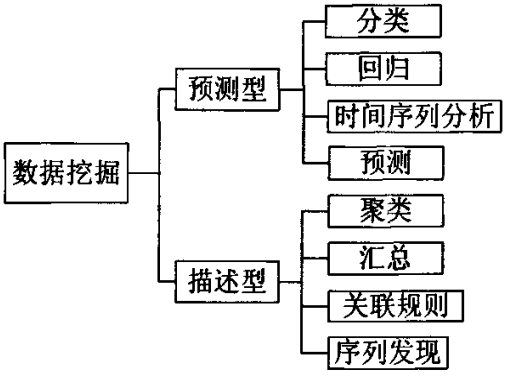

图 2.1　数据挖掘类型及任务图

Fig. 2.1　Components of data acquisition system

数据挖掘功能以及它们可以发现的模式类型介绍如下：

(1) 分类、预测

分类就是通过研究已分类的样本集的特征，分析样本集的属性，建立一个分类函数或分类模型，通过这个分类模型，未分类的或新的数据就可以分派到不同的类别中，达到分类的目的。分类可以用决策树归纳、贝叶斯网络、人工神经元网络(如 BP 网络等)、粗糙集、遗传算法、K—最临近分类和支持向量机等方法。分类可以预测对象的类标记，当要预测的数据是数值数据(连续值)，而不是离散的类别标志时，我们可以称之为预测。预测主要使用回归方法，当然也可以使用人工神经元网络、遗传算法、支持向量机等机器学习方法。

(2) 关联规则

数据库中的数据之间一般都存在某种关联关系，即变量之间可能存在某种规律，关联规则挖掘的任务就是找出数据库中哪些事物或属性共同出现的条件。最有影响力的关联规则挖掘的算法是 Rakesh Agrwal 等人提出的 Apriori 算法，近年来，也出现了很多 Apriori 的改进算法，如 Edith Cohen 等人提出的不需要剪枝的改进算法，Mohammed J.Zaki 提出的可伸缩的改进算法等。

(3) 聚类分析

聚类是将对象集合按照相似性归为若干类别，属于无指导分类，属于同一类的对象具有较高的某种相似性，而不同类的对象之间的差别较大。通过聚类，识别密集和稀疏

的区域，发现全局的分布模式及数据属性之间的相互关系，帮助人们建立宏观概念。聚类的方法主要可以分为：划分方法(Partitioning method)、层次方法(Hierarchical method)、基于密度的方法(Density-based method)、基于网格的方法(Grid-based method)和基于模型的方法。其中，划分方法中用的比较多的是K—平均算法和K—中心点算法。BIRCH和CURE就是比较典型的层次方法，DBSCAN是比较有代表意义的基于密度的方法，STING算法是典型的基于网格的方法，基于模型的方法有统计学方法、人工神经元网络方法(如Kohonen网络)等。

(4) 类/概念描述

数据可以与类或概念相关联，用汇总的、简洁的、精确的方式描述每个类和概念是有用的，目的是对数据进行浓缩，给出它的总体的综合描述，实现对原始数据的总体把握。这种类或概念的描述称为类/概念描述。通过类/概念描述使得人们能够在复杂数据库中了解数据的意义以及产生数据的过程。这种描述可以通过汇总所研究类的数据来获得(这个过程也叫数据特征化)或将所研究类与其它的比较类进行比较来获得，或采用上面两种方法的结合。基于数据立方体的OLAP上卷操作来执行指定维的数据汇总就是一种很有效的数据特征化的方法，数据特征化的输出通常采用如饼图、柱状图、多维数据立方体等形式来形象的表现出来。

(5) 孤立点分析

数据库中经常存在这样一些数据对象，它们与数据的一般行为或模型不一致，这些数据对象我们就称之为孤立点。在一般情况下，数据挖掘方法会将孤立点视为噪声或异常而丢弃，但是在特殊场合，如在电子商务领域，探测和分析孤立点显得比正常数据还来的重要。

(6) 演变分析

数据演变分析(Evolution analysis)用来描述行为随时间变化的对象的规律或趋势，并对其建模。时间趋势分析考虑时间上的变化趋势，空间趋势则要根据某空间维找出变化趋势。

2.1.3 数据挖掘常用技术

(1) 决策树

代表着决策集的树形结构。决策树是对分类问题进行深入分析的一种方法，在实际问题中，按算法生成的决策树往往复杂而庞大，令用户难以理解。这就告诉我们在重分类精确性的同时，也要加强对树修剪的研究。

数据挖掘中决策树是一种经常要用到的技术，可以用于分析数据，同样也可以用来作预测。常用的算法有 CHAID, CART, QUEST 等。决策树提供了一种展示类似在什么条件下会得到什么值这类规则的方法。

决策树中最上面的节点称为根节点，是整个决策树的开始。决策树的每个节点子节点的个数与决策树所用的算法有关。每个分支要么是一个新的决策节点，要么是树的结尾，称为叶子。在沿着决策树从上到下遍历的过程中，在每个节点都会遇到一个问题，对每个节点上问题的不同回答导致不同的分支最后会到达一个叶子节点。这个过程就是利用决策树进行分类的过程，利用几个变量(每个变量对应一个问题)来判断所属的类别(最后每个叶子会对应一个类别)。假如负责借贷的银行官员利用上面这棵决策树来决定支持哪些贷款和拒绝哪些贷款，那么他就可以用贷款申请表来运行这棵决策树，用决策树来判断风险的大小。“年收入>￥40.000”和“高负债”的用户被认为是“高风险”，同时“收入<￥40.000”但“工作时间>5 年”的申请，则被认为“低风险”而建议贷款给用户。

(2) 人工神经网络

神经网络建立在自学习的数学模型基础之上。它可以对大量复杂的数据进行分析，并可以完成对人脑或其他计算机来说极为复杂的模式抽取及趋势分析。神经网络系统由一系列类似于人脑神经元一样的处理单元组成，我们称之为节点(Node)。这些节点通过网络彼此互连，如果有数据输入，它们便可以进行确定数据模式的工作。神经网络有相互连接的输入层、中间层(或隐藏层)、输出层组成。中间层由多个节点组成，完成大部分网络工作。输出层输出数据分析的执行结果。例如：我们可以指定输入层为代表过去的销售情况、价格及季节等因素，输出层便可输出判断本季度的销售情况的数据。

神经网络近来越来越受到人们的关注，因为它为解决大复杂度问题提供了一种相对来说比较有效的简单方法。神经网络可以很容易的解决具有上百个参数的问题(当然实际生物体中存在的神经网络要比我们这里所说的程序模拟的神经网络要复杂的多)。神经网络常用于两类问题：分类和回归。在结构上，可以把一个神经网络划分为输入层、输出层和隐含层。输入层的每个节点对应一个个的预测变量。输出层的节点对应目标变量，可有多个。在输入层和输出层之间是隐含层(对神经网络使用者来说不可见)，隐含层的层数和每层节点的个数决定了神经网络的复杂度。

神经元网络和统计方法在本质上有很多差别。神经网络的参数可以比统计方法多很多。由于多，参数通过各种各样的组合方式来影响输出结果，以至于很难对一个神经网络表示的模型做出直观的解释。实际上神经网络也正是当作“黑盒”来用的，不用去管“盒子”里面是什么，只管用就行了。在大部分情况下，这种限制条件是可以接受的。

比如银行可能需要一个笔迹识别软件，但他没必要知道为什么这些线条组合在一起就是一个人的签名，而另外一个相似的则不是。在很多复杂度很高的问题如化学试验、机器人、金融市场的模拟、和语言图像的识别，等领域神经网络都取得了很好的效果。神经网络的另一个优点是很容易在并行计算机上实现，可以把他的节点分配到不同的 CPU 上并行计算。

(3) 遗传算法

遗传算法(Genetic algorithm, GA)是近几年发展起来的一种崭新的全局优化算法，是一种基于生物进化论和分子遗传学的搜索优化算法。它借用了生物遗传学的观点，通过自然选择、遗传、变异等作用机制，实现各个个体的适应性的提高。这一点体现了自然界中“物竞天择、适者生存”进化过程。1962 年 Holland 教授首次提出了 GA 算法的思想，从而吸引了大批的研究者，迅速推广到优化、搜索、机器学习等方面，并奠定了坚实的理论基础。

遗传算法是首先将问题可能的解按某种形式进行编码，编码后的解称为染色体；随机选取 N 个染色体作为初始种群，再根据预定的评价函数对每个染色体计算适应值，性能较好的染色体有较高的适应值；选择适应值较高的染色体进行复制，并通过遗传算子，产生一群新的更适应环境的染色体，形成新的种群，直至最后收敛到一个最适应环境的个体，得到问题的最优化解。

(4) 近邻算法

将数据集合中每一个记录进行分类的方法。依据"Do as your neighbors do"的原则，相邻数据必然有相同的属性或行为。K—Nearest 邻居方法的含义为：K 表示某个特定数据的 K 个邻居，可以通过 K 个邻居的平均数据来预测该特定数据的某个属性或行为。

(5) 规则推导

从统计意义上对数据中的“如果—那么”规则进行寻找和推导。采用上述技术的某些专门的分析工具已经发展了大约 10 年的历史，不过这些工具所面对的数据量通常较小。现在这些技术已经被直接集成到许多大型的工业标准的数据仓库和联机分析系统中去了。

2.1.4 数据挖掘的应用

数据挖掘所要处理的问题，就是在庞大的数据库中找出有价值的隐藏事件，并且加以分析，获取有意义的信息，归纳出有用的结构，作为企业进行决策的依据。其应用非常广泛，只要该产业有分析价值与需求的数据库，皆可利用 Mining 工具进行有目的的发掘分析。

(1) 数据挖掘解决的典型商业问题

需要强调的是，数据挖掘技术从一开始就是面向应用的。目前，在很多领域，数据挖掘都是一个很时髦的词，尤其是在如银行、电信、保险、交通、零售(如超级市场)等商业领域。数据挖掘所能解决的典型商业问题包括：数据库营销(Database marketing)、客户群体划分(Customer segmentation & classification)、背景分析(Profile analysis)、交叉销售(Cross-selling)等市场分析行为，以及客户流失性分析(Churn analysis)、客户信用记分(Credit scoring)、欺诈发现(Fraud detection)等等。

(2) 数据挖掘在市场营销的应用

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。

商业消费信息来自市场中的各种渠道。例如，每当我们用信用卡消费时，商业企业就可以在信用卡结算过程收集商业消费信息，记录下我们进行消费的时间、地点、感兴趣的商品或服务、愿意接收的价格水平和支付能力等数据；当我们在申办信用卡、办理汽车驾驶执照、填写商品保修单等其他需要填写表格的场合时，我们的个人信息就存入了相应的业务数据库；企业除了自行收集相关业务信息之外，甚至可以从其他公司或机构购买此类信息为自己所用。

这些来自各种渠道的数据信息被组合，应用超级计算机、并行处理、神经元网络、模型化算法和其他信息处理技术手段进行处理，从中得到商家用于向特定消费群体或个体进行定向营销的决策信息。这种数据信息是如何应用的呢?举一个简单的例子，当银行通过对业务数据进行挖掘后，发现一个银行账户持有者突然要求申请双人联合账户时，并且确认该消费者是第一次申请联合账户，银行会推断该用户可能要结婚了，它就会向该用户定向推销用于购买房屋、支付子女学费等长期投资业务，银行甚至可能将该信息卖给专营婚庆商品和服务的公司。

数据挖掘构筑竞争优势。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(Relationship billing)优惠”的促销策略，即如果一个顾客在一个商店用运通卡购买一套时装，那么在同一个商店再买一双鞋，就可以得到比较大的折扣，这样既可以增加商店的销售量，也可以增加运通卡在该商店的使用率。再如，居住在伦敦的持卡消费者如果最近刚刚乘英国航空公司的航班去过巴黎，那么他可能会得到一个周末前往纽约的机票打折优惠卡。

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推销材料。卡夫(Kraft)食品公司建立了一个拥有 3000 万客户资料的数据库，数据库是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而建立起来的，卡夫公司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，其中容纳有遍布全球的一亿多个订户的资料，数据库每天 24 小时连续运行，保证数据不断得到实时的更新，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。

基于数据挖掘的营销对我国当前的市场竞争中也很具有启发意义，我们经常可以看到繁华商业街上一些厂商对来往行人不分对象地散发大量商品宣传广告，其结果是不需要的人随手丢弃资料，而需要的人并不一定能够得到。如果搞家电维修服务的公司向在商店中刚刚购买家电的消费者邮寄维修服务广告，卖特效药品的厂商向医院特定门诊就医的病人邮寄广告，肯定会比漫无目的的营销效果要好得多。

(3) 成功案例

AutoTrader.com 是世界上最大的汽车销售站点，每天都会有大量的用户对网站上的信息点击，寻求信息，其运用了 SAS 软件进行数据挖掘，每天对数据进行分析，找出用户的访问模式，对产品的喜欢程度进行判断，并设特定服务，取得了成功。Reuteres 是世界著名的金融信息服务公司，其利用的数据大都是外部的数据，这样数据的质量就是公司生存的关键所在，必须从数据中检测出错误的成分。Reuteres 用 SPS 的数据挖掘工具 SPSS/Clementine，建立数据挖掘模型，极大地提高了错误的检测，保证了信息的正确和权威性。Bass Export 是世界最大的啤酒进出口商之一，在海外 80 多个市场从事交易，每个星期传送 23000 份定单，这就需要了解每个客户的习惯，如品牌的喜好等，Bass Export 用 IBM 的 Intelligent Miner 很好的解决了上述问题。在科学研究方面，一个天文学上的著名应用系统 SKICAT 就是相当成功的数据挖掘应用，利用该系统，天文学家已发现 16 个新的极其遥远的星群。在生物医学和 DNA 数据分析上，数据挖掘可以完成异构、分布式基因数据库的语义集成，用关联规则分析同时出现的基因序列，用路径

分析发现在疾病不同阶段的致病基因等。NBA 教练就运用 Advanced scout 来挖掘信息，安排阵型，提高了获胜的机率。在金融投资方面，FALCON 系统是信用卡欺诈估测系统，已被相当数量的银行采用，FASI 是一个用于识别与洗钱有关的金融交易系统，LBS Capital management 则使用了专家系统、神经网络和基因算法技术来辅助管理多达 6 亿美元的有价证券。

2.2 数理统计方法在数据挖掘中的应用

2.2.1 数据挖掘与数理统计的关系

数据挖掘利用人工智能和统计分析的进步带来了许多好处。这两门学科都致力于模式发现和预测。数据挖掘把统计和人工智能的算法及技术封装起来，使人们不用了解这些技术的细节就能实现许多有价值的功能，从而使人们把更多的精力去专注于所要解决的问题[10]。

数据挖掘不能替代传统的统计分析技术。相反，它是统计分析方法学的延伸和扩展。大多数统计分析技术都基于完善的数学理论和高超的技巧，预测的准确度还是令人满意的，但对使用者有很高的要求。而随着计算机能力的不断增强，有可能利用计算强大的计算能力只通过相对简单和固定的方法完成同样的功能。数据挖掘作为一门新型学科，是多学科交叉、渗透、结合的产物。数据挖掘就其算法本身而言，很大一部分可能从数理统计中获得理论解释。但是作为一个整体的研究方向，应该从计算机的层面进行全局的考虑。即从系统的角度进行分析。

在现代统计学中，模型是核心，计算、模型选择准则等往往都被认为是次要的，是建立模型的枝节。但数据挖掘却不同，它的核心是算法，当然也考虑模型和可解释性问题，但算法及可实现性是第一位的。它所强调的首先是发现，其次才是解释。因而，数据挖掘并不过分依赖于严格的逻辑推理，而是大量采用很多“黑箱”方法和本质上是探索性的方法。从目前数据挖掘的实践看，几乎所有方法都集中于计算，忽视了事实上统计才是核心问题[11]。为进一步了解统计学在数据挖掘中的重要性以及统计学和数据挖掘的相互联系，下面从三个方面再作进一步论述[12]：

(1) 概率分布

了解分布族的性质(如共轭族在条件化下封闭，多元正态族关于线性组合封闭)对分析数据和统计推断都是非常有用的。统计推断主要涉及以下几个方面：估计、假设、相合性、稳健性、不确定性、模型平均。许多推断方法可看作估计量或从数据到某个评估对象的函数，这个对象可以是参数值、区间值、结构、决策树等。当数据集是由概率分布描述的(实际或潜在)总体的样本时，在任何给定容量的样本上估计量值有概率分布。

统计学通过考察估计量的这种分布来识别估计量与信息、可靠性和不确定性相关的特征。

估计量的一个重要特征是相合性，即随着样本容量的无限增加，估计量应几乎必然收敛到要估计的正确值。在机器学习中存在大量的启发式方法并不能保证结果收敛到正确的答案。同样重要的一个特征是用有限样本作估计所产生的不确定性。这个不确定性可用概率分布来描述。统计学理论提供了许多度量不确定性的方法和算法以及评价估计量不确定性的再抽样方法和模拟方法。一般而言，在其它条件(如相合性)都相同的情况下，优先选择使不确定性度量达到极小的估计量。数据挖掘算法也需要解决类似的相合性和不确定性问题，显然，概率分布描述的方法值得借鉴和参考。

统计学研究的一个目的就是用尽可能少的假设作足够好的估计。不过，由于很多原因这些假设并非严格成立。如果假设不成立，模型就不正确，那么，基于假设的估计自然就不可靠。这样的问题在数据挖掘中尤为突出。一个可能的解决途径就是应用现代稳健统计的思想，开发适宜于海量数据挖掘的新型稳健方法。

Bayes“模型平均”方法(BMA)强调的不是单个模型，而是同时考虑几个竞争模型，然后，以各个模型所得估计的加权平均作为最终的估计量。这种的做法在平均意义上改进了估计的性能。因为在数据挖掘中得到的模型通常是使用某个自动搜索方法的结果，所以，如果这个搜索方法的误差频率已知的话，就可更好地利用模型平均的优势改进搜索结果。

(2) 假设检验

虽然统计假设检验应用广泛，但也应注意，一般地，一个水平α的假设检验和另一个水平α的假设检验并不能对二者的联合假设提供水平α的检验，除非检验的水平α随样本容量的增加适当减少。因此，在数据挖掘中，一个水平为α的检验与搜寻多个假设检验的方法所犯错误的概率之间并没有什么直接的联系。例如，对变量组中的每一对变量来说，若在$\alpha = 0.5$的水平上检验独立性假设，那么，0.5 并不是所有变量对都独立时却错误发现某些相依性的概率。这样，在使用一系列假设检验的数据挖掘方法中，检验的水平一般不能作为有关搜索结果犯错概率的估计。进一步地，检验犯错误的概率依赖于假设的正确性，而不是近似正确性，因为即使假设与实际情况非常近似，在大样本情形下也可能被拒绝。例如，线性模型的检验在大样本情形一般就拒绝这些假设，不管它们与数据拟合得多么接近。

另外，需要注意的是数据挖掘算法通常也将整个数据总体作为分析处理的出发点，如，分析一家公司逐年完成的交易量。在这种情形下，显著性检验失去效用，因为观察到的统计量值(如所有年交易量的平均值)就是总体的参数值。

(3) 因果推断

在统计学发展初期，Bernoulli 和 Laplace 就把两个变量之间不存在因果关系的现象看作是概率独立的；在试验设计理论中也有同样的思想。1934 年，生物学家 Wright S 引入了用有向图来表示因果假设(顶点表示随机变量，边表示直接影响)的方法。1982 年，统计学家 Hkiiveri 和 Speed T P.把独立性与无因果性之间的广义联系和有向图结合起来，引入了 Bayes 网模型，后来，这种模型已成为社会科学、生物学、计算机科学和工程界表示因果假设的普遍方法。对方便样本作因果推断易产生错误的主要根源有三，它们是潜变量(或混杂者)、样本选择偏差和模型等价性。所谓潜变量是指在记录单元之间变化且其变化影响记录特征的任何未记录到的特征。结果是记录特征之间的联系，这种联系实际上并不是由记录特征本身的任何因果关系产生的。潜变量在数据挖掘中要有的话很少能够忽略掉。当任何两个变量的值本身对数据库是否记录一个特征会有影响时，就可能发生样本选择偏差的问题。有丢失值的数据集也会引起样本选择偏差的问题。另外，不同的 Bayes 网模型可能对概率分布产生同样的约束，因而无法区分实验干扰，这就是模型等价性问题。

从数据挖掘在国际上的发展来看，数据挖掘的研究重点已从提出概念和发现方法，转向系统应用和方法创新上，研究注重多种发现策略和技术的集成，以及多种学科之间的相互渗透，数据挖掘技术迫切需要系统、科学的理论体系作为其发展的有力支撑。最近，经验统计方法和人工智能相结合而产生的衍生技术，如分类回归树(Classification and regression tree, 简称 CART)，卡方自动交互探测法(Chi-square automatic interaction detector, 简称 CHAID)等前沿方法，以算法的形式展示了统计和信息技术结合发展的新方向。这些都预示着数据挖掘技术与统计学的集成已成为必然的趋势。

2.2.2 数理统计在数据挖掘中的应用

(1) 市场研究。在市场经济的今天，开展市场研究尤其重要。市场研究是为某一特定的市场营销问题的决策而开发和提供其所需的信息的一种系统的、有目的的活动或过程。这里所说的信息，不仅是市场调查所得的数据资料，还包括市场研究人员对资料进行分析所得的结果(如结论、建议等)。市场研究的范围包括：产品研究、销售研究、市场与销售潜量的估计、价格研究、购买行为研究、竞争分析、广告及促销研究、销售成本和利润分析、营销环境研究。这些研究又可归结为四类：

① 探测性研究，适用于需要研究的问题和范围不是很明确，无法确定调查内容，引起问题的原因不很清楚的情形。通常的做法：收集与分析第二手资料，集中专家或相关人员的意见，进行小规模的试点调查、定性研究、实例分析。

② 描述性研究，回答事先计划好的问题，如“什么”、“何时”、“怎样”。通常的做法是：给出统计图；计算基本统计量(如相对指标、平均指标、变异指标等)；进行相关分析，确定市场营销中有关问题的相关因素、相关关系及关联程度。

③ 因果关系研究，用于寻找问题的原因，或确定变量之间有无因果关系。通常的做法是先确定分析对象及其影响因素，然后建立具有单方程或联立方程形式的经济计量模型，应用回归分析方法进行测定。

④ 预测性研究，用于对市场潜力、企业产品前景等进行预测，以此作为市场营销决策的依据。通常的方法包括：专家意见法，移动平均法，指数平滑法，Box-Jenk-ins法，经济计量模型法，投人产出模型法等。

在市场研究中，常用且有效的方法就是应用数理统计方法，比较基本的有列联表分析、相关分析、方差分析；比较高级的有多元统计分析中的聚类分析、判别分析、主成份分析、因子分析、多特性模型等。

聚类分析的目的在于辨别在某些特性上相似的事物，并按这些特性将样本划分成若干类(群)，使在同一类内的事物具有高度的同质性，而不同类的事物则有高度的异质性。在市场研究中，聚类分析主要用于对消费者群进行市场细分，对产品进行分类，选择试验市场，确定分层抽样的层次，分析消费者的性格特征和行为形态等方面。

判别分析的原理是：依据样本的某些特性，以判别样本所属类型。与聚类分析不同的是，判别分析是在已知研究对象用某种方法分成若干类的情况下，建立判别函数，用以判定未知对象属于已知分类中的那一类。在市场研究中，判别分析主要用于对一个企业进行市场细分，以选择目标市场，有针对性地进行广告、促销等活动。

主成份分析是把多个指标化为少数几个综合指标的一种多元统计方法。它采用一种降维的方法，即通过适当的变换，找出几个综合因子来代表原来众多的变量，使这些综合因子能尽可能地反映原来变量的信息量，而且彼此之间互不相关。在市场研究中，主成份分析主要用于分析消费者的嗜好，以此对消费品进行分类。

因子分析是这样的一种多元统计技术：从众多变量中提取出少数共同“因子”，用最少的因素综合解释大量资料，简化变量间的关系，从而分析影响变量、支配变量的共同因素(又称公共因子)有几个，各因素的本质如何，以由表及里地探索事物之间的本质联系。在市场研究中，因子分析常用来分析消费者对各种消费品的态度，研究消费者选择消费品的因素，以对制定营销策略和拟定广告宣传主题提供参考依据。

多特性模型就是利用与产品有关的多种特性来选择产品的一种统计模型，是一种总合所有消费者意见的多元统计分析方法。

(2) 经济活动分析。多元统计分析中的聚类分析、判别分析、主成份分析，因子分析等方法除了应用于市场研究之外，还可用于进行企业或部门的经济活动分析。例如应用聚类分析方法按经营状况对企业或部门进行分类；根据已有的关于企业经营状况的分类及某一企业的一些经济效益指标，应用判别分析方法确定该企业的经营状况的类属；根据一些经济效益指标，应用主成份分析方法进行综合评价，确定企业的排名；根据反映企业经营状况的大量指标，应用因子分析方法确定影响经济效益的潜在因素。

经济活动分析本质上是事后分析，或称为影响因素分析，即在明确分析对象的影响因素的基础上，要确定各因素影响的显著性及影响程度，并从中找出主要的影响因素。此时，有统计软件包帮忙，经济计量学的经济计量模型就可以使经济活动分析变得容易得多了。

2.3 数理统计在预测型数据挖掘中的应用

数理统计在数据挖掘中具有重要的地位。文[11]从统计学的角度来透视数据挖掘中相关的统计问题，提出了传统统计学面临的挑战，以及在这个领域将带来的一些新的研究方向。文[12]介绍了一些常用的统计分析方法在数据挖掘中的应用。本文中将在强调统计学是数据挖掘的重要理论基础的同时，从最优化的观点出发，论述预测型数据挖掘中的优化问题。

关于预测型数据挖掘还没有一种被共同接受的理论框架，其实现方法大致可以分为[9]：基于逻辑的方法，基于距离的方法和基于数学的方法。其中被广泛使用的是基于数学的方法，具体地有如下三种方法[10]：第一种是经典的参数估计方法，例如线性回归方法。在这种方法中，用带参数的代数方程来描述输入输出间的关系，方程中的待定参数由训练样本确定。尽管参数模型有很好的理论基础，但它常常过于简单化或者对涉及的数据要求过多的、无法获得的知识。因此，对于现实世界中的问题来说，这些参数模型可能是不实用的。第二种方法是经验非线性方法，如前向人工神经网络。这种方法利用已知样本建立非线性模型，克服了传统参数估计方法的困难。但是，这种方法也遇到了许多难以解决的问题(比如过拟合问题、局部极小问题等)，并且缺乏一种统一的数学理论。第三种是基于统计学习理论的实现方法。统计学习理论(Statistical learning theory, SLT)是一种专门研究有限样本情况下的统计理论[11]。该理论针对有限样本统计问题建立了一套新的理论体系，在这种体系下的统计推理规则不仅考虑了对渐近性能的要求，而且追求在现有有限信息的条件下得到最优结果。Vapnik V.等人从 20 世纪 70 年代开始致力于此方面研究，到 20 世纪 90 年代中期，随着其理论的不断发展和成熟，也由于神经网络等方法在理论上缺乏实质性进展，SLT 开始受到越来越广泛的重视。这一理论基

础上发展了一种新的预测方法——支持向量机(Support vector machine, SVM)，已初步表现出很多优于已有方法的性能[13,14]。SLT 和 SVM 正在成为继神经网络研究之后新的研究热点，并将推动预测型数据挖掘理论和技术的重大发展。由于分类和回归都可以形式化地表述为风险最小化问题，因此优化方法在预测型数据挖掘算法中起着重要作用[3]。事实上，所谓预测型数据挖掘算法，其本质就是一种优化算法，例如参数估计的最小二乘法、前向人工神经网络的 BP 算法以及统计学习理论中的 SVM 方法。

2.4 本章小结

本文在介绍数据挖掘理论和数理统计在数据挖掘中的应用的基础上，进一步论述了数理统计方法在预测型数据挖掘方面的研究，了解到算法是数据挖掘的核心。应用优化方法进行数据挖掘中的问题求解是一种重要的思路，如将数学优化模型转化为参数估计问题，然后将参数估计转化为复杂函数的优化问题。这样本文所要解决的问题可以通过统计方法建立模型，转化为模型参数求解的问题，然后寻找合适的最优化方法对模型中的参数进行估算。

3 数值优化理论与方法

3.1 优化方法概述

求解优化模型的方法称为优化方法。优化方法在国民经济如国防、工农业生产、交通运输、金融、贸易、管理、科学研究中有着广泛的应用[15]。最优化是人们在工程技术、科学研究和经济管理等诸多领域中经常遇到的问题。结构设计要在满足强度要求等条件下使所用材料的总重量最轻资源分配要使各用户利用有限资源产生的总效益最大；安排运输方案要在满足物资需求和装载条件下使运输总费用最低；编制生产计划要按照产品工艺流程和顾客需求，尽量降低人力、设备、原材料等成本使总利润最高。可以预料，随着科学技术尤其是计算机技术的不断发展，以及数学理论与方法向各门学科和各个应用领域的更广泛、更深入的渗透，在这个新世纪信息时代，最优化理论和技术必将在社会的诸多方面起着越来越大的作用。从而也不断推动着最优化理论的不断发展。为解决各种优化计算问题，人们提出了各种各样的优化算法，如单纯形法、梯度法、动态规划法、分枝定界法等。这些优化算法各有各的长处，各有各的适用范围，也各有各的限制。

一个好的优化方法满足：总的计算量小、存储量小、精度高、逻辑结构简单[16]。优化方法的分类与其针对的优化问题相关。一般从不同的角度出发，优化方法可分为不同类别。从优化问题是否带约束条件分为无约束优化方法和有约束优化方法；从优化问题是否为线性分为线性优化方法和非线性优化方法等。最优化理论是一门实用性很强的学科，优化技术是最优化理论与方法的实际应用，是用于求解各种工程问题优化解的应用技术。首先介绍一些与优化方法相关的概念。

3.2 约束优化方法

3.2.1 约束优化方法概念

约束优化问题的模型来源于生产实践，现实生活中许多问题经过建模都可归结为求解如下形式的数学模型(NLP)[16]：

$$\min \quad f(x) \tag{3.1}$$

$$\begin{aligned} s.t \quad & g_i(x) \geq 0, \ i=1,2,\cdots,m. \\ & g_i(x) = 0, \ i=m+1,m+2,\cdots,m+p. \end{aligned} \tag{3.2}$$

其中 $x=(x_1,x_2,\cdots x_n)^T \in R^n$，即 x 是 n 维列向量，$f(x)$, $g_i(x)$, $i=1,2,\cdots,m+p$ 为 x 的函数，s.t.是 subject to 的缩写，表示 x 受不等式条件及等式条件的限制；$f(x)$ 称为目标函数。

若 $m+p=0$，则称其为无约束优化问题，若 $m+p>0$，即如果有不等式和/或等式约束存在，则称为约束最优化问题。若目标函数 $f(x)$ 与约束函数 $g_i(x)$，$i=1,2,\cdots,m+p$，都是线性函数，则称之为线性规划问题，否则称之为非线性约束优化问题。我们研究的对象主要是非线性约束优化问题。

记 $X=\{x\mid g_i(x)\le 0,i=1,2,\cdots,m.g_i(x)=0,i=m+1,\cdots,m+p\}$，称 X 为上述非线性约束优化问题的可行域。若 $x\in X$，则 x 称为可行解。若对某一 $x^*\in X$，存在常数 $\xi>0$，使得对 $\forall x\in X\bigcap\{x\|x-x^*\|<\xi\}$，有 $f(x)\ge f(x^*)$（或 $f(x)>f(x^*)$）则称 x^* 为局部最优解（或严格局部最优解）；若对一切 $x\in X$ 都有 $f(x)\ge f(x^*)$（或 $f(x)>f(x^*)$），则称 x*为全局最优解（或严格全局最优解）。

求解最优化问题，就是要求目标函数 $f(x)$ 在约束条件下的极小点，即求出其全局最优解，但在一般情况下，我们往往只能求出它的一个局部最优解。一般说来，非线性约束优化问题比起线性规划无论从理论还是到算法的研究都要困难些。不论何种情形，在讨论算法得到的点列 $\{x_k\}$ 的收敛性及收敛速度时，都放宽到求得问题的 KKT 点。

3.2.2 常用约束优化方法

按照对约束条件的处理方法的不同，对有约束非线性最优化问题的求解，可以分为以下两类方法：间接法和直接法[17]。

(1) 间接法

将约束条件转化成无约束最优化问题的求解方法。主要有罚函数法、增广乘子法、序列二次规划算法等。

(2) 直接法

不作约束条件的转化，利用约束条件参与对问题求解的方法。主要有消元法、复合形法、沿约束边界寻优法、线性逼近法、可行方向法、投影梯度法等。

有约束非线性最优化问题求解，是在无约束非线性最优化问题以及经典解析法与运筹学的理论基础上发展的。然而，由于有约束条件的存在，使非线性最优化问题变得比较复杂，在求极小点时会遇到更多的问题。它可能造成求解约束非线性最优化问题的极小点不是可行域的边界上；或增加新的局部极小点（指无约束时不存在的点）；或所有极小点中没有一个是无约束问题的全局极小点。所以，处理约束非线性最优化问题有时更加麻烦。当然，这也是促使非线性最优化问题不断发展的原因之一。

解约束非线性优化问题的最常用的方法有复合形法[16]，惩罚函数法，逼近规划法等。近年来出现了通用性的方法如：模拟退火算法，遗传算法等，但这些通用算法存在算法

比较复杂，而且计算量大，优化过程消耗的时间长等问题。本文采用算法结构较为简单，而且适用于求解库存优化模型的复合形法。

3.3 本章小结

数据挖掘算法可以认为是由模型结构、评分函数、搜索方法和数据管理技术构成的组合。这其中包含了大量的可能算法，因为只需把不同的模型或模式结构，不同的评分函数、搜索方法和数据管理策略组合起来。但是在实际操作时，我们需要针对数据集和挖掘任务找到组成部分的最优组合方式。在现有数据和评分函数的指导下搜索并优化参数和结构是所有数据挖掘算法中最核心的部分。

通常模型或模式是以各种形式的结构来描述的，有时还带有未知的参数。优化和搜索的目标就是决定这些结构和参数值，以使评分函数达到最优。发现模型中的最佳参数值的任务通常被称为最优问题。本章介绍了优化理论与方法，和几种常用约束优化算法特点，为本文中的模型参数求解做理论铺垫。

4 物料消耗概率模型

4.1 物料消耗的分析与模型的建立

4.1.1 影响物料消耗量的因素的分析

数据挖掘本质上就是建模,即发现客观事务的规律。这里的建模当然是指广义的,例如,描述式建模。数据挖掘中的建模是为了捕捉数据中存在的关系。模型或模式结构决定了从数据中寻找的潜在结构或函数形式。模型是对一个数据集的高层次、全局性的描述，可以是描述性的，也可以是推理性的。而模式则是数据的局部特征，或许只支持几条记录或者几个变量。在确定模型或模式结构时，首先要明确挖掘任务，针对特定的任务选择最适合的模型或模式结构。大部分数据挖掘的算法使用了一个或几个目标函数、使用若干搜索方法(如启发式算法、梯度下降方法、最大最小值法、网络推演法等),去找出在数据体中或建立了距离关系的数据空间中的一个点或小区域。这个最优点和小区域有时不是最重要的结果,伴随它建立起的模型才是人们所需要的东西。

本章以印染企业为例，对企业的原材料消耗问题进行分析、建模并求解。

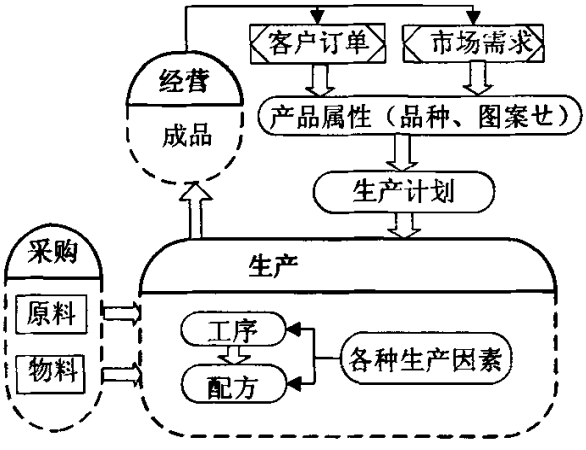

图 4.1 企业生产经营活动关系图

Fig. 4.1 Graph of major activities of enterprise

印染企业中的生产原材料可以分为原料与物料。原料是指需要进行印染的各种坯布，物料是指坯布印染过程中使用的各种染料及其它辅助性材料。如图 4.1 所示，印染企业主要有采购、生产、经营三项业务。采购为生产准备所需的原料和物料，生产上将

通过各种工序将原材料加工处理，按照某种比例消耗原料和物料，最终生产出成品花布交付经营部门。生产上同时接受来自经营部门的客户订单，并同时与市场需求结合制订生产计划。三大业务关系密切，不可分割。

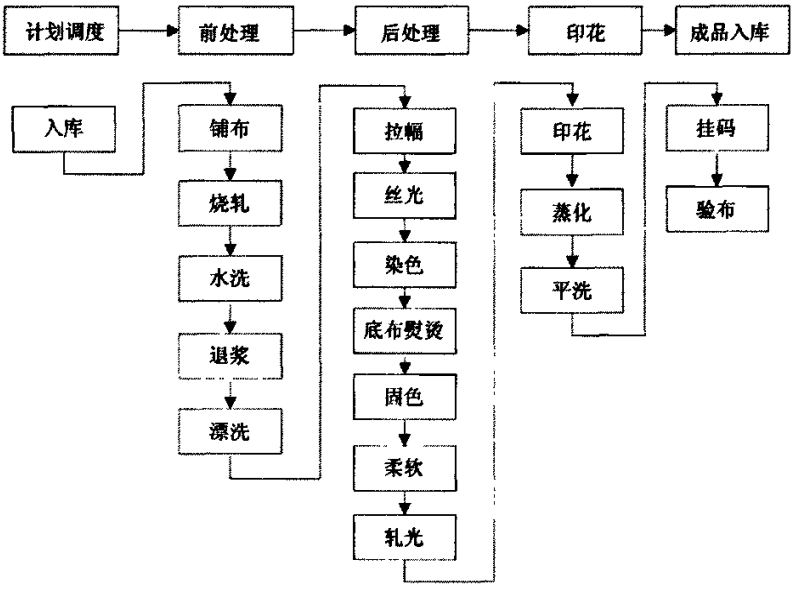

图 4.2　生产流程图

Fig. 4.2　Graph of production process

图 4.2 为印染企业生产主要流程图。从前处理到印花，有十余种复杂的工艺，不同的工艺对于原料都有不同的处理，而每一道工艺和工序都有相应的物料消耗。物料的消耗量同时受到生产方式以及原料消耗量的影响。但是在生产方式固定的情况下，相关物料的消耗量会呈现为与对应原料消耗量相关的某种概率分布。此外，在多种原料同时生产时，物料的消耗量与他们的理论消耗量之和也会呈现出较大的偏差。

对于物料和每一类产品的消耗关系，在定性分析基础上，运用定量方法，对未来生产计划可能消耗的物料数量作判断和推测。通过历史数据来计算单位产品消耗的物料数量的概率分布模型，来推测下一次产品订单中的各种产品消耗物料的最可能的值。这种推测是根据历史数据来分析产品和物料之间的某种关系，利用数据挖掘技术已成为必然的趋势，结合数据挖掘中的预测技术可以达到我们的目的。对于物料历史消耗的概率分布，可以通过数据挖掘的预测技术给出一个近似值。这种预测是由可靠数据得出的，因

此可以利用结果指导采购。

这种比例受工序、配方以及其它各种生产因素的影响。原料对物料的作用关系是由很多复杂因素决定的。生产型企业虽然有产品生产的基本配方，但同于原料类别较多，每一种产品的工艺路线，每一道工序及其工艺配方都不一样，物料采购员对整体的数量把握仅仅凭往年的经验估计一段时间内所需的物料用量，并一次性购买远大于生产需要的物料作为储备，直到物料少于一定库存时继续采购。这样，虽然每种物料只多出生产需要看似微乎其微的数量，上万种物料的库存却很容易造成多余库存占用的浪费和库存积压问题。如何有效地掌握物料的库存量并保证生产需要是一个比较直接的问题。如果能发现出物料消耗量随不同原料投产量组合而变化的大概趋势，就能够帮助采购员获得一段时间内的物料采购量依据，从而达到库存优化的目的。而如何寻找原料投产量与物料消耗量之间的关系，可以用统计分析方法来进行分析，提供可靠的数据为采购决策做基础。

在实际生产过程中，原料的采购量由客户订单和市场的需求来决定。原料有百余种类别，物料有上万种，生产计划员从客户订单的产品档案制定产品的生产工艺路线，由于每种产品的花色图案等属性和所需的原料种类不同，使其所需要的各种工序所用的加工方式不同，该生产企业的每个工序的工艺配方都有几百种甚至上千种，且每个工艺配方所需的各类物料的用量理论定额都不同，这样，物料的实际消耗量可能会与理论值有很大差距。因此，往往需要采购人员估计未来一段时间内所需的物料用量并在此基础上进行采购，在采购数量不准确的时候，可能会影响到企业的正常生产。因此，如何根据历史数据来预测物料的消耗规律是一个重要而又困难的问题。

为简化问题的计算，对实际生产中物料的消耗情况进行分析后，我们把多种生产方式按照生产计划分为 4 个类别，分别是冷色系、暖色系、中色系以及杂色系。这样，在每种色系下，物料的消耗量可以看作是有坯布种类和数量决定的概率模型。这样，采购优化问题就可以简化为根据生产数据来预测相关材料的消耗数量的问题。本文中我们来寻找物料消耗量随不同类产品生产量组合而变化的大概趋势，帮助采购员获得一段时间内的物料采购量依据。由色调分类和产品的原料类别来统计过去生产中各产品分类所消耗的物料概率分布。这里可以用统计分析方法来进行分析，提供可靠的数据依据为采购决策做基础。

物料的消耗有多种，原料的消耗也有多种。对某种物料来说，不同的工艺决定了其消耗量，也并非每种原料都会影响到它的消耗，如果预先提取出具有相互影响关系的原料和物料，则能够减少物料消耗概率模型的计算量。这里使用相关系数法，认为只有相关系数大于某个阈值的原料和物料相互之间存在影响关系。当我们认为原料和物料之间

确实具有某种程度的关系时，可以用从统计的角度，假定投产一定量的某种原料对于这种物料的使用数量服从某种分布，以一年中各种原料的投产情况作为样本，用极大似然估计法和优化算法来计算期望值和方差等重要参数。

4.1.2 物料消耗概率模型

下面讨论印染企业的物料消耗模型。产品有百余种类别，物料有上万种，这里分析单位某类产品的平均消耗物料量。

设产品类别共有 m 种，物料类别有 n 种。一般来说，单独生产千米 S_i 类原料时，物料 M_j 的消耗量服从高斯概率分布，即

$$X_{ij} \sim N(q_{ij}, \sigma_{ij}^2), \quad q_{ij}, \sigma_{ij} > 0 \text{均为未知参数} \tag{4.1}$$

若产品 S_i 类投产量 Y_i，定义

$$C_i = Y_i / 1000, \tag{4.2}$$

则产品 S_i 类投产 Y_i 米对于物料 M_j 的消耗量 $W_{ij} = C_i X_{ij}$ 服从正态分布[18]，即

$$w_{ij} \sim N(C_i q_{ij}, C_i^2 \sigma_{ij}^2) \tag{4.3}$$

因此，在各类产品每日投产量相互独立的条件下，物料 M_j 日消耗总量 W_j 服从如下分布：

$$E(W_j) = \sum_{i=1}^{m} C_i q_{ij}, \quad D(W_j) = \sum_{i=1}^{m} C_i^2 \sigma_{ij}^2 \tag{4.4}$$

以 $x_{1j}, x_{2j}, \cdots x_{nj}$ (n为一年中生产次数)代表一年中该企业实际生产所消耗物料 M_j 的样本观察值，应用参数估计的方法，就可以根据这些样本信息估算公式(4.4)中的模型参数。

4.2 模型的参数估算

4.2.1 应用极大似然法估算参数

在数理统计中，常用的参数估计方法分为点估计法和区间估计法。常用的点估计法有矩估计法和极大似然法等。在实际问题中经常遇到随机变量分布函数的形式已知，但其中参数未知的情况。如果得到了随机变量的一组样本值后，希望利用样本值来估计变

量分布中的参数值，这在工程中是一个比较重要的问题。而常用来进行参数估计的方法有作图法、回归分析法、矩估计法、最小二乘法、极大似然法等[19-20]。

由于极大似然估计的渐进最优性质，它已经成为参数估计的一种常用方法，并且已经在众多领域中广泛得到应用，例如系统辨识、语音处理、图像处理及模式识别等等。它是以观测值出现的概率最大作为准则。设 x 为连续随机变量，其分布密度函数为 $p(x|\theta)$，$\theta=\{\theta_1,\theta_2,\cdots,\theta_M\}$，这个密度函数由参数 θ 完全决定。已知 N 个观测值 $x_1,x_2,\cdots x_N$，假设它们是从分布密度为 $p(x|\theta)$ 的总体中独立抽取的。记 $X=\{x_1,x_2,\cdots x_N\}$，则

$$p(X|\theta)=\prod_{i=1}^{N}p(x_i|\theta)\ \triangleq\ L(\theta|X) \tag{4.5}$$

函数 $L(\theta|X)$ 称为似然函数。当 X 固定时，$L(\theta|X)$ 是 θ 的函数。极大似然参数估计的实质就是求出使 $L(\theta|X)$ 达到极大时 θ 值。为了便于求出使 $L(\theta|X)$ 达到极大的 $\hat{\theta}$，通常对式(4.5)两边取对数，即

$$\ln(L(\theta|X))=\sum_{i=1}^{N}\ln(p(xi|\theta)) \tag{4.6}$$

对式(4.6)求解可以得到极大似然估计值 $\hat{\theta}$。

大量工程实践表明，最大似然估计法精度最高，而且该方法充分利用样本中包含的所有信息及总体分布表达式所提供的关于变量的信息，因而具有许多优良的性质。例如，当样本容量固定时，它在某些分布族中就是最小方差无偏估计(minimum variance unbiased estimate，MVUE)或接近于 MVUE；而且当样本容量充分大时，它也有优良的大样本性质；通常情况下，最大似然估计不仅是相合估计，而且是渐近正态估计。因此，在实际应用中，最大似然估计是最常用的方法，也是最重要的方法。

常用的概率模型的参数估算方法有矩法、最小二乘法以及极大似然法等等[21-25]。极大似然法既具直观性，又有数学严密性，是应用最广的参数估计方法。对于本文提出的问题，由于目标函数是多个概率模型的组合结果，而且与观测到的样本数据没有直接关系，因此，采用极大似然估计法来对参数进行估计。极大似然法使用固定样本观测值，在取值的可能范围内，挑选使似然函数达到最大(从而概率达到最大)的参数值作为参数的估计值，这样将原来求参数的极大似然估计值问题就转化为求似然函数的最大值问题了[25,26]。

上述模型的对数似然函数为

$$L\left(q_{1j},q_{2j},\cdots,q_{mj},\sigma_{1j},\sigma_{2j},\cdots,\sigma_{mj}\right)=\ln\prod_{k=1}^{N_k}f\left(x_{kj},q_{ij},\sigma_{ij}\right)$$

$$=\ln\prod_{k=1}^{N_k}\left(\frac{1}{\sqrt{2\pi}(\sum_{i=1}^{m}c_{ik}{}^2\sigma_{ij}{}^2)^{\frac{1}{2}}}\exp-\frac{(x_{kj}-\sum_{i=1}^{m}c_{ik}q_{ij})^2}{2\sum_{i=1}^{m}c_{ik}{}^2\sigma_{ij}{}^2}\right)$$

$$=-\frac{1}{2}\sum_{k=1}^{N_k}(\ln\sum_{i=1}^{m}c_{ik}{}^2\sigma_{ij}{}^2+\frac{(x_{kj}-\sum_{i=1}^{m}c_{ik}q_{ij})^2}{2\sum_{i=1}^{m}c_{ik}{}^2\sigma_{ij}{}^2})-\frac{1}{2}N_k\cdot\ln(2\pi) \qquad (4.7)$$

为使目标函数 $L\left(q_{1j},q_{2j},\cdots,q_{mj},\sigma_{1j},\sigma_{2j},\cdots,\sigma_{mj}\right)$ 取最大值，求得 $q_{1j},q_{2j},\cdots,q_{mj},\sigma_{1j},\sigma_{2j},\cdots,\sigma_{mj}$ 的极大似然估计量，利用最优化方法可以求得估计值 $\hat{q}_{1j},\hat{q}_{2j},\cdots,\hat{q}_{mj},\hat{q}_{1j},\hat{q}_{2j},\cdots,\hat{q}_{mj}$。

4.2.2 几种约束优化算法的研究与分析

优化方法在数据挖掘算法中占有重要的地位。应用优化方法进行参数估计问题求解是一种重要的思路，本文中将数学优化模型转化为参数估计问题，然后将参数估计转化为函数的优化问题。这样通过以上的统计方法建立的模型，转化为模型参数求解的问题，然后寻找合适的最优化方法对模型中的参数进行估算。目前求解约束优化问题有下面几种常用的方法。

(1) 惩罚函数法

惩罚函数法也称罚函数法，是间接法求解有约束非线性最优化问题的一种具有代表性的解法。该法将有约束问题转化成带有罚项的无约束问题的目标函数，在求解新的目标函数中，随着罚因子取值的变化，最优解也在不断变化，最后收敛于原问题的最优解。因此，这种方法需要求解一系列无约束极值问题，其间可以用解析法或某种数值解法求解。罚函数原理简单、算法易行，目前已发展出有外点法、内点法和混合点法三种方法，且均可与各种有效的无约束优化问题解法结合起来使用。因此，罚函数法是目前国内应用较广泛的最优化方法之一。

(2) 复合形法

复合形法是一种在实用中很有效的最优化搜索算法，概念简单、不需要大的计算机存贮单元，不用计算函数的梯度，适用于多变量优化场合，因此在工程中得到广泛地应用，同样适合本文中所要求解的问题。复合形法是求解无约束优化问题的单纯形法在求解约束优化问题上的推广，它有诸多优点。

① 适用范围广。对目标函数及约束函数无特殊的条件限制，而范围广是使用优化方法追求的目标之一。因为目前解非线性约束优化问题的数值方法虽然很多，但尚没有一种在任何情况下都适用的通用方法；

② 该方法建立在全局观点基础上，追求整体最优解。这是只建立在局部分析观点基础上的所有间接法所不能比拟的；

③ 该方法简单直观，易于控制，出错率低。上述所长恰恰是实际工程优化者最为关心的。另外，复合形与随机试验相比，前者确定性程度高，收敛速度快，是一种确定性意义下的进化算法。

(3) 遗传算法

遗传算法是一类可用于复杂系统优化计算的鲁棒搜索算法，与其他一些优化算法相比，它主要有下述几个特点[27]：

① 遗传算法以决策变量的编码作为运算对象。这种对决策变量的编码处理方式，使我们在优化计算过程中可以借鉴生物学中染色体和基因等概念，可以模仿自然界中生物的遗传和进化等机理，也使得我们可以方便地应用遗传操作算子。特别是对一些无数值概念或很难有数值概念，而只有代码概念的优化问题，编码处理方式更显示出了其独特的优越性。

② 遗传算法直接以目标函数值作为搜索信息。遗传算法使用由目标函数值变换来的适应度函数值，来确定搜索方向和搜索范围，无需目标函数的导数值等其他一些辅助信息。

③ 遗传算法同时使用多个搜索点的搜索信息。遗传算法对群体进行选择、交叉和变异等运算，产生出的是新一代的群体，在这之中包括了很多群体信息。这些信息可以避免搜索一些不必搜索的点，所以实际上相当于搜索了更多的点，这是遗传算法所特有的一种隐含并行性。

④ 遗传算法使用概率搜索技术。遗传算法用各种概率参数来确定搜索方向，因此搜索具有灵活性，是一种鲁棒性算法。

遗传算法提供了一种求解复杂系统优化问题的通用框架，它不依赖于问题的具体领域，对问题的种类有很强的鲁棒性，除了应用于函数优化领域外，还广泛应用于很多学科。函数优化也是对遗传算法进行性能评价的常用算例。很多人构造出了各种各样的复杂形式的测试函数，有连续函数也有离散函数，有凸函数也有凹函数，有低维函数也有高维函数，有确定函数也有随机函数，有单峰值函数也有多峰值函数等。用这些几何特性各具特色的函数来评价遗传算法的性能，更能反映算法的本质效果。而对于一些非线性，多模型，多目标的函数优化问题，用其他优化方法较难求解，而遗传算法却可以方

便地得到较好的结果。遗传算法适合于维数很高、体积很大、环境复杂、问题结构不十分清楚的场合。

通过分析以上几种算法的特点，可以看出，不同算法有各自的优点，但实际应用中，并不是每一种算法都适合所有的问题，不同的算法有其不同的应用场合。惩罚函数法的主要缺陷是罚参数的选取比较困难，而且算法的性能强烈地依赖于参数的选取。遗传算法编码比较困难，在向局部最优解收敛的过程中，运行速度较慢。本文选择一种改进的复合形法进行求解，主要是由于复合形法适用范围广，易用扩展，适用于多种模型，同时，其思想简单，易于实现，而且改进后的复合形法收敛比原来要快很多，在精度要求不是很高的情况下，在允许时间内在可以接受。

4.2.3 复合形法的介绍

复合形法是求解约束非线形优化设计问题的一种应用比较广泛的直接局部搜索算法，它是在 Nelder 和 Mead 提出的单纯形法的基础上通过改进而形成的[28,29]。所谓复合形法是指在 n 维设计空间的约束可行域内，有 $n+1 \le k \le 2n$ 个定点所构成的多面体。复合形法就是在 n 维设计空间的约束可行域内，对复合形各项点的目标函数值逐一进行比较，不断去掉最坏点，代之以既能使目标函数值有所下降，又能满足所有约束条件的新顶点，逐步调向局部最优点。复合形法由于不必保持规则的图形，较之单纯形法更为灵活易变，同时其调优过程始终在可形域内进行，所求结果可靠，有一定的收敛精度，能够有效处理不等式约束的优化设计问题。设约束优化问题为：

$$\begin{array}{lll} \min & f(X) & X \in R^n \\ s.t. & g_n(x) \le 0 & \text{n=1,2,}\cdots\text{,m} \end{array} \tag{4.8}$$

采用复合形法也与其他搜索算法一样，关键是确定每步迭代的搜索步长。只要能找到目标函数下降方向，并使迭代点沿此方向移动一个适当的步长，最终会使迭代点趋向极小点。通常迭代点逐步向最优点逼近的过程在设计空间将形成一条轨迹。而复合形法虽然也是一个逐步逼近的过程，但它利用由若干个定点所构成的复合形，通过定点的不断迭代而发生变形和位移，最终趋向最优点。用复合形法求解需要分为两个步骤进行：

第一步：在设计空间的可行域 $D=\{X \mid g_i(x) \le 0 \text{ (i=1,2,}\cdots\text{,m)}\}$ 内产生 k 个初始点，以 k 个初始点为定点构成一个多面体，一般取 $n+1 \le k \le 2n$。当 n 等于 2 时，即在可行域内构成一个三角形或四边形，见图 4.3。

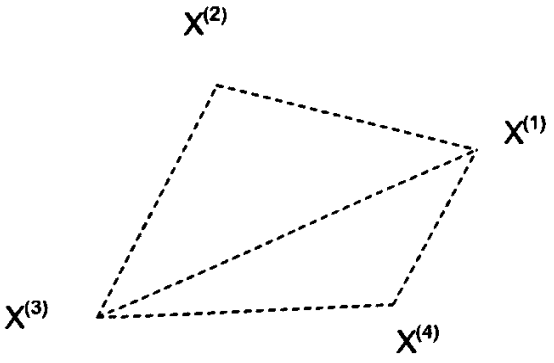

图 4.3 二维问题复合形

Fig. 4.3 Dimensional complex

第二步：对复合形法进行调优迭代计算。也就是利用复合形各顶点函数值的大小关系，判断目标函数值的下降方向，不断丢掉最坏点，使复合形不断向最优点移动和收缩，直到达到一定的收敛精度为止。如图 4.4 所示，取二维问题的复合形为一四边形，其定点分别为 $X^{(1)},X^{(2)},X^{(3)},X^{(4)}$，函数值为 $F(X^{(1)}),F(X^{(2)}),F(X^{(3)}),F(X^{(4)})$，若 $F(X^{(1)})>F(X^{(2)})>F(X^{(3)})>F(X^{(4)})$，则 $X^{(1)}$ 为最坏点，用 $X^{(H)}$ 表示，$X^{(2)}$ 为次坏点，用 $X^{(G)}$ 表示，$X^{(4)}$ 为最好点，用 $X^{(L)}$ 表示。然后求出最坏点外其余各点的几何中心 $X^{(S)}$，大致可以判定出，连接最坏点 $X^{(H)}$ 和中心点 $X^{(S)}$ 的方向为目标函数值下降的方向。

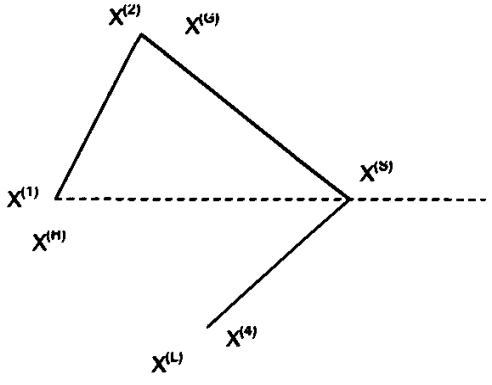

图 4.4 复合形法的调优迭代

Fig. 4.4 Optimizing iteration of complex method

沿此方向取一点为 $X^{(R)}$，并令 $X^{(R)}=X^{(S)}+\alpha(X^{(S)}-X^{(H)})$，称 $X^{(R)}$ 点为 $X^{(H)}$ 关于点的映射点(亦成反射点)，α 称为反射系数，一般先取 $\alpha>1$。不断地进行复合形迭代，即利

用复合形各个顶点处目标函数值的大小关系，判断目标函数值的下降方向，不断丢掉函数值最大的所谓最差点，代之以既使目标函数值有所下降又能满足所有约束条件的一个新点，从而不断地构成新的复合形，如此重复计算，使新的复合形不断地向可行域最优点移动和收缩，直到得到满足收敛准则的近似解为止。在选择新的比较好的点替换最差点的迭代过程中，复合形的反射、扩张和压缩计算要满足一个条件，即计算中涉及到的复合形所有顶点、复合形形心点和反射点等必须在可行域内。

4.2.4 自适应复合形法在参数估算中的应用

$$\begin{cases} \min f(x)=\sum_{k=1}^{N_k}(\ln\sum_{i=1}^{m}c_{ik}{}^2\sigma_{ij}{}^2+\dfrac{(x_{kj}-\sum_{i=1}^{m}c_{ik}q_{ij})^2}{2\sum_{i=1}^{m}c_{ik}{}^2\sigma_{ij}{}^2}), \quad k=1,2,3,\ldots,N_k & (4.9) \\ S.T. \quad \min_{k\in N_k}\{x_{jk}\}<q_{ij}<\max_{k\in N_k}\{x_{jk}\}, \quad i=1,2,3,\ldots,m & (4.10) \\ 0<\sigma_{ij}<q_{ij}, \quad i=1,2,3,\ldots,m & (4.11) \end{cases}$$

在预测型数据挖掘中被广泛使用的是基于数学的方法。从方法结构上看，基于数学的方法主要由三种成分构成，即预测函数集、择优准则以及搜索和优化方法。在数据挖掘、统计学和机器学习的文献中经常低估了高效搜索和优化的重要性，但是在实践中搜索和优化方法对一个应用成功与否起着关键的作用。

数据挖掘为优化方法提供了一个新的应用领域的同时，也给优化理论和方法的研究提出了严峻的挑战。数据的海量性导致了优化算法设计的困难。优化算法一直是运筹学的重要研究内容，经过几十年众多学者的努力，已经取得了大量成果。然而，面对海量数据挖掘问题，现有的优化算法已不能完全适应需要。因此，针对海量数据分析和处理的需要，设计快速有效的能够解决大规模优化问题的模型与算法已成为数据挖掘迫切需要解决的问题。

目前尽管有一些优化算法已经应用于数据挖掘问题，但大多数算法直接来源于机器学习领域，其设计往往结合启发式信息，技巧多于理论。由于这些算法或多或少地存在着随意性，对于不同的算法难以比较优劣，因此只有对算法进行深入的理论分析，例如收敛性分析、收敛速度估计和计算复杂性分析等，才能有效推动算法的进一步深入发展。

数据挖掘包含很多算法，具体使用哪种数据挖掘算法，要根据具体情况和应用要求。一种算法可能在一种情况下适用，而在另一种情况下就不适用。惩罚函数法的主要缺陷是罚参数的选取比较困难，而且算法的性能强烈地依赖于参数的选取。遗传算法编码比

较困难[30]，收敛速度比较慢，运行时间过长，不适合工程上的要求。复合形法的在阀值较大时，收敛的速度比较快，对于精度不高的场合搜索效率非常高。

对于本文中的模型求解，由于实际生产活动对于估计量的精度要求不高，对于而且难以确定所分析的物料消耗量对应的概率分布模型，为了使这种物料消耗估算方法能应用到多种概率分布模型上，采用自适应复合形法来对似然函数进行求解。所谓自适应就是在寻优过程中随移动次数不同而可以调整步长的自适应变化[32]。复合形法的基本思想来源于尼勒迪尔(Nelder)和米爱德(Mead)提出的解决无约束非线性最优化问题的单纯形法，实际上是求解无约束优化问题的单纯形调优法的推广[29]。它是求解多维非线性有约束最优化问题的一种重要的方法。复合形法是一种迭代算法，迭代过程分三部分：构造初始复合形、调优计算和判断收敛。其基本思想是：在可行集内产生一个K个顶点的初始复合形，然后对其各个顶点的函数值进行比较，不断地剔除最坏点，用既能使目标函数有所下降，又满足约束条件的新点代替，逐步逼近最优点，它的核心是“吐故纳新”。

复合形法是一种试验最优化方法，是处理目标函数梯度难以计算的实用型方法[33]。该方法不受约束条件的多少，约束性质的限制，为企业生产提供指导，具有广泛的意义。用复合形法来求解经济模型，程序简单，使用方便，特别对经济领域中维数高、非线性约束少的问题十分有效[29]。其优点是不用求目标函数的梯度Hesse矩阵，不用进行复杂的矩阵运算，可以代替最小二乘法用于数据处理中，对于大型统计和复杂方程式(如指数曲线)的回归等所涉及函数复杂的问题有着其他方法不可比拟的优越性。由于在迭代计算中不必计算目标函数的一阶或二阶导数，对目标函数和约束函数的性质无特殊要求，所以是工程设计中常用的求解有约束的直接解法。因此被广泛地应用于建筑结构、电子技术、空间技术、生物医学工程等学科。但复合形法也有其自身的缺点[31]，如随着约束条件和设计变量的增多，其计算效率显著下降，而且在迭代过程中有可能出现迭代失败现象，这在复杂优化问题中尤其如此。从已发表的利用复合形法求解优化问题的文献中可以发现，学者们大都利用基本的复合形法进行计算，对出现的关于最坏点映射失败问题，其处理方法较为粗糙[32]。图4.5为最坏点映射示意图。

在解决本文提出的优化问题时，多次出现了震荡现象。表现为：在收缩过程中，新产生的点可能会在中心点处来回震荡，导致无法收敛。针对这一问题，在计算过程中，根据点的移动次数和收缩次数的比值来确定收敛步长[33]。比值接近于1时，说明收敛太慢，可以适当增加步长，比值过大时(大于1.5)，说明出现震荡，需要减小步长。

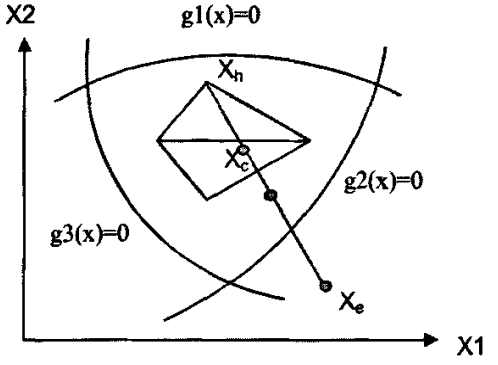

图 4.5 最坏点映射示意图

Fig. 4.5 Graph of reflection of worst point

公式(4.6) (4.7) (4.8)所表示的最优化问题，可以使用自适应性复合形法按照如下步骤求解：

(1) 设置参数$\alpha,\beta,\gamma,MvCnt,k$。这里，设$\alpha$=1.3,$\beta$=0.01,$\gamma$=1000，令顶点移动计数器初始值$MvCnt=1$。确定顶点个数，方程中共有$2m$个变量，故顶点个数$K=2m+2$。随机生成满足约束条件的K个初始可行顶点。

(2) 首先构造初始复合形，每个点用公式$x_i=x_{\min}+r_i\times(x_{\max}-x_{\min})$计算，$r_i$为(0, 1)均匀分布的随机数；计算$2m+2$次，获得$2m+2$个初始点$x_1,x_2\cdots x_{2m+2}$。

(3) 若x_i不可行，首先计算现行点的中心$x_{avg}^i=\frac{1}{i}\sum_{j=1}^{i}x_i$，并向中心$x_{avg}^i$收缩一半，既令$x_i=x_i+\frac{1}{2}(x_{avg}^i-x_i)=\frac{1}{2}(x_{avg}^i+x_i)$，重复上述过程，直至得到$2m+2$个可行点后，组成一个初始复合形。

(4) 对每个可行点分别计算$f(x_i)$，计算K个顶点的目标函数值。

(5) 选取为最大函数值的最好点，和最小函数值的最坏点x_r $f(x_r)=\max(f(x_i))=F_{\max}$，$x_r$将通过反射被丢弃。

(6) 计算除去最坏点的中心点x_{avg}，$x_{avg}^i=\frac{1}{k-1}\sum_{j=1}^{k-1}x_j$。

(7) 除去最坏点，x_{old}，计算反射点$x_{new}=x_{avg}+\alpha(x_{avg}-x_{old})$，并计算$f(x_{new})$。若$x_{new}$是可行点，转向步骤 8；否则，向中心点$x_{avg}$收缩一半并记录移动次数，即

$x_{new}=x_{old}+\frac{1}{2}\left(x_{avg}-x_{old}\right)=\frac{1}{2}\left(x_{avg}+x_{old}\right)$，$MvCnt=MvCnt+1$，直到该点可行，生成新的一组复合形顶点。

(8) 如果K个顶点的所有函数值都在γ迭代内都在β范围内变动则迭代终止，否则转向步骤5。

(9) 当找到全局最优值时停止。

(10) 否则调整α。若$\left(Mt/\gamma\right)>1.5$，$\alpha=0.8\times\alpha$，若$\left(Mt/\gamma\right)<1.1$，$\alpha=1.3\times\alpha$。令$\gamma=1$，$Mt=1$，转向步骤5。

4.3 实验与分析

4.3.1 试验结果

以某一物料为例，根据以往的生产数据计算它的消耗期望数量，以此期望值计算07年上半年内所有生产计划的预计消耗。然后和此期间内物料的实际消耗作对比。

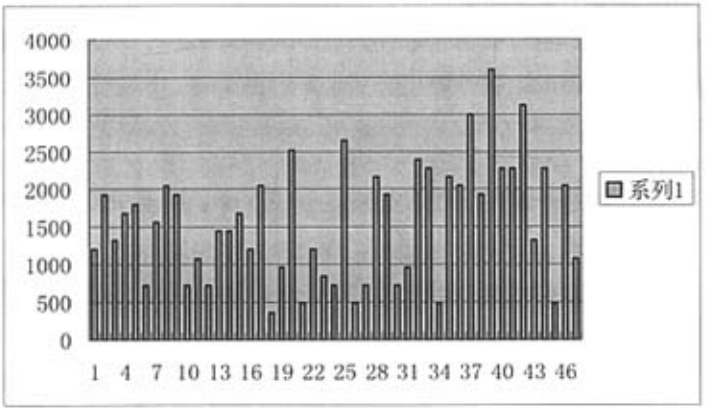

图4.6 2007年上半年模型所得的物料“稳定剂(可稳得CPB)”预测数量

Fig. 4.6 Predicted amounts of material “Kewende CPB” from Jan. to Jun. in 2007

图4.6代表上半年期间根据本文所建立的模型计算出以产品订单的产品图案与数量预测的物料“稳定剂(可稳得CPB)”理论上的消耗数量。图4.7代表上半年期间实际可稳得CPB的消耗数量。其中，横坐标表示物料使用时间，纵坐标表示对应该时间的物料消耗量。

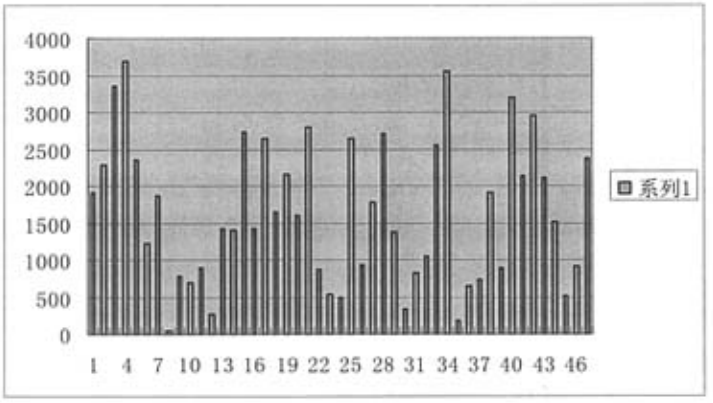

图 4.7 2007 年上半年“稳定剂(可稳得 CPB)”的实际消耗数量

Fig. 4.7 Actual consuming amounts of material “Kewende CPB” from Jan. to Jun. in 2007

4.3.2 结果分析

从图中可以看出，超过一半的数据点与预测值的偏差符合高斯模型中的3δ原则，这表明预测的模型能够在一定程度上满足未来生产的需求。但是也有很多的数据与实际消耗量不吻合，例如在 3，21，35 这些点处，相差比较大。经过分析，可能有如下三个原因导致这些差异：

(1) 模型中假定的是正态分布，有可能实际上是服从其他分布，或者是多种分布的混合分布。

(2) 客户的产品订单会有很多零散化型的产品，这些数量较小的产品没有作为数据被统计到输入范围内。

(3) 样本数据量太少，导致计算出的结果与实际期望值偏差较大。在使用更多的样本数据的情况下，结果可以更符合实际的生产。

将此模型应用到采购系统中，可以帮助指导采购工作，图 4.8 为根据某一期间的产品订单而计算物料预计消耗的实际应用案例。

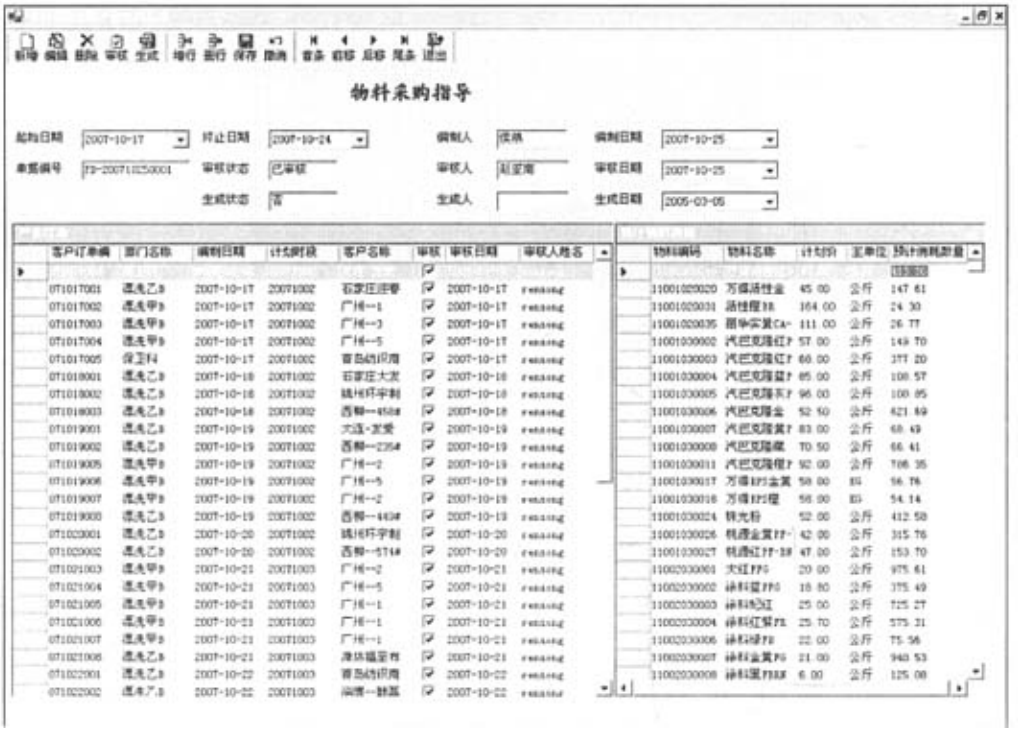

图 4.8　采购决策支持系统物料采购界面

Fig. 4.8　Graph of materials' purchasing direction in stock decision-support system

4.4　本章小结

本文应用基于数据挖掘与最优化结合、支持最终决策分析的管理问题求解模型，为复杂的采购预测分析问题的求解和决策实施提供一个切实可行的参考模型。首先对生产型企业用统计的方法建立了产品与物料的消耗关系模型，并应用改进的复合形法求解最优问题，然后以产品生产计划的投产量预测物料消耗量，实现了物料的采购决策支持，对采购工作进行了指导。理论分析和实验表明，改进的复合形算法在解决此问题上性能较稳定。在采购决策支持系统中，统计模型的适用是非常重要的。如何更好地结合统计理论和优化算法以寻找适合用户要求的预测模型是今后讨论的重点。

5 采购优化系统设计与实现

5.1 采购优化模型

5.1.1 问题分析

在信息系统中，已经积累了大量的有关供应商供货情况的数据，充分利用这些数据，可以帮助采购人员了解供应商的情况，以便更有选择性地进行供应商的评价，使原材料采购更符合生产的要求，同时达到采购成本最低。这些大量的数据中，包含了各种各样的信息，如供应商供货的质量如何，其最终生产出的成品效果及销量如何，其价格与多个供应商相比是否最低，其每次供货是否能一次性满足订单的数量，即供货能力有多大，还有交货时间是否最快等等，这些都是采购方在选择供应商所考虑的重要因素。然而，并不是每一个供应商都在各方面是最有竞争力的，供应市场是多样化的，每一个供应商都有自己的优势和特点，如 A 家的规模比较大，供货能力强，随时订货都能满足采购方的需求；B 家的供货质量最好，生产出的产品质量好，比较畅销等等。因此，如何在这样一个具有多样性的供应市场中，根据客户需求选择一些最符合企业生产需要的，使企业在各方面收益最高，是具有重要的现实意义的。数据挖掘技术已经广泛用于 SCM 中，用于解决类似的问题。

为了确保销售、生产等各部门工作的顺利进行，供应部门往往加大库存量，从而导致库存成本增加，企业产品价格居高不下，使企业失去了市场竞争力。因此，研究以降低采购成本为目标的原材料采购计划对优化企业资源、分配、减少损耗、降低成本、缩短产品周期、提高企业柔性起到关键性的作用。供应商选择的正确与否不但直接影响采购成本水平的高低，而且对企业的产品成本、柔性以及竞争能力等都将产生重要的影响。由此可见，科学合理地选择供应商和分配采购量，企业具有极为重要的现实意义。

采购包括了解需要、选择供应商、协议价格、签订合同、催促交货和保证供应等环节。采购的订货决策就是要确定合理的订货时间和数量，用最少的费用去满足生产和经营的需要，并借以获得最大的经济效益。企业采购优化是根据买方市场的要求，对企业采购行为在品种、价格、批量、时间、空间等方面进行科学分析做出合理决策并付诸实施的过程。其目的是降低企业流动资金占用及保证生产经营的顺利进行。库存控制则是要保证所需物品的数量和质量以及它们存贮的时间和地点，同时要使维持费用尽可能低。我们只能控制商品的输入，另外在实践中只是控制部分库存品，因而库存控制的基本问题是，要求回答“订购多少和何时订购”。所以说采购的订货决策和库存控制在实

质上是相同的，为了寻求合理的解决办法，必须定性与定量相结合，为此必须寻求数学模型并提出合理的解决办法。

对供应商进行评估时，是根据 ERP 系统提供的供应商供货能力分析报表、质量分析报表、价格分析报表提供的数据，再考虑供应商的其他考核指标进行综合考察与评估，最后将评估结果录入系统，系统通过查看或报表的形式反应供应商交货期的准确率，送货的合格率、退货率、总的采购量、金额、产品、价格等信息并以此作为评价供应商的依据。ERP 系统所提供的各类报表为进行供应商的评价提供了较为详细的数据和信息来源，但是还只是停留在定性分析的层次上，并未能对决策提供科学的定量的数据比较。结合数据挖掘的方法，可以从中提取有意义的信息。

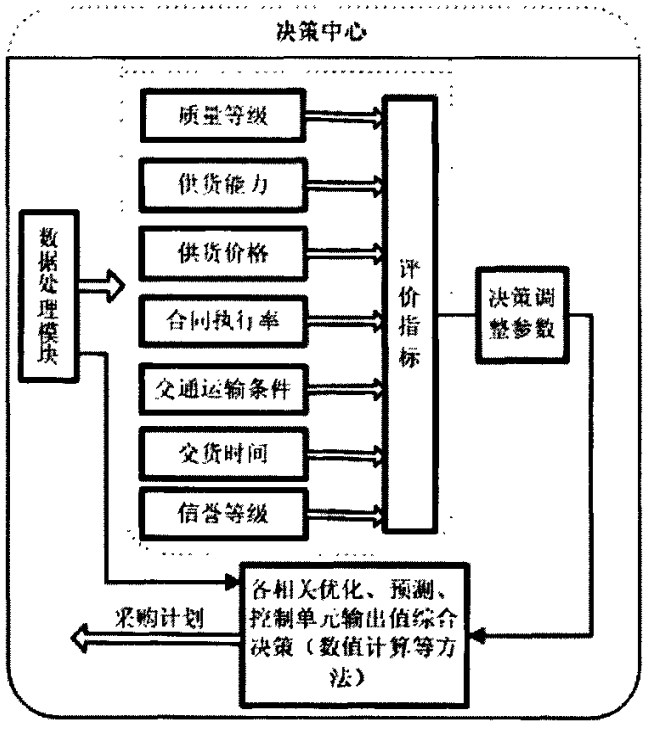

图 5.1　供应商评价指标图

Figure 5.1　Graph of supplier assessment index

企业根据生产情况审查存货，及时调整采购计划是非常必要的。本章将根据第四章所建立物料预测模型计算的结果继续讨论该企业如何实现采购优化，降低采购成本。针对采购方面的问题，根据原材料的特点，建立合适的采购—库存模型，在多个供应商之间确定合理的订货批量，一方面减少采购成本，另一方面又要保证生产需要，优化库存

管理。在此基础上，开发采购优化系统，解决企业在采购到库存管理方面存在的一系列问题。

采购成本法对质量和交货期都能满足要求的供应商，则需要通过计算采购成本来进行比较分析。采购成本一般包括售价、采购费用、运输费用等各项支出的总和。采购成本比较法是针对各个不同供应商的采购成本，计算分析来选择采购成本较低的供应商的一种方法。

统计表明，大多数采购商在选择供应商时，首先考虑的是供货质量、其次是物品价格、交货提前期。结合印染企业的实际情况，建立印染企业的供应商的评价指标体系，如图 5.1，供应商的评价指标一般有：供货的质量等级、供应商的供货能力、供货价格、合同的执行率、距离及交通运输条件、其交货时间以及该供应商的信誉度等等。

(1) 质量等级。原材料的质量是影响供应商选择的一个“极端重要”的因素。原材料的质量直接影响到最终产品的质量，决定了其服务水平。不同类型的产品有着不同的质量标准，不同供应商的产品质量水平也不尽相同。

(2) 供货价格。供货价格是直接影响订单成本的因素之一。供应商提供价格折扣的种类很多，但是其中最重要的是由于订货批量本身的大小而带来的折扣的大小，一般来说，批量越大，价格折扣也就越大。采购时应该优先考虑供应商的提供的价格，但低价格往往导致质量降低，而价格高的原材料生产出的产品质量可能更高，使产品销量增加。价格和质量往往是互相矛盾的。

(3) 供应能力。供货能力是评价供应商的一个重要指标，对于任何一种原材料，由于供应商的生产能力有限，其供应能力不一定完全符合采购商的要求。采购商一次性需求可能远远大于供应商一次供货的能力。采购方在分配采购量时，必须考虑这个因素。

(4) 交货时间。交货时间即提前期，零库存的概念使得采购方特别注意交货时间。交货期的延误会导致其库存缺货，生产不能正常进行，最终客户服务及市场销售。

(5) 信誉等级。信誉等级也是非常重要的一项指标，如供应商在原材料市场是否稳定，交货是否按合同严格履行，是否为长期提供优质服务的老客户。

(6) 交通运输条件。供应商的地理位置、距离及交通运输条件也是不可忽视的因素之一，交通运输条件决定了供货成本，及是否能按期交纳货物。

由于每种原材料有不同的厂家供应，每个厂家的供应能力、供货价格、运输条件及费用等各方面条件都有一定的差别，如何选择最适合的供应商，以及在各供应商一次性供应能力有限的条件下，在不同的供应商之间制订一个合理的订货计划使订货总成本最小且同时满足生产需要，是一个非常重要的问题。将评价指标定量化，调整参数，建立一个适当的模型，然后应用数值计算方法，最终求得最优订货方案。

由于供应商生产能力的相对稳定性和市场需求的多变性，导致在计划期的某时段内，供应商的供应能力与企业生产的需求能力不平衡，时常会出现供应能力短缺或超额现象。当供应能力短缺时，通常的解决办法是提前订货或推迟交货。然而，提前订货占用大量的流动资金且增加存储费用；拖期交货会影响企业生产，需要交付拖欠罚金并且降低供应链的服务水平，增加供应链循环时间。

原材料质量、价格、加工方案和加工流程直接影响到成品布的数量和质量，对企业生产计划的顺利进行和经营目标的实现起到至关重要的作用。所以，当某类原材料发生短缺时，采购计划要综合考虑短缺原材料的品种、数量、质量进行分配以保证企业生产的需求。因此采购优化的目标是：在供应链计划期内，在供应能力许可的条件下，根据企业的需求合理制定采购方案使采购成本、运输成本、超额/拖欠惩罚费用和交货时间得到优化，同时，根据企业对原材料质量和准时交货的要求，合理的选择出供应商及资源的优化配置方案，为采购决策提供可靠的数学依据。

对质量和交货期都能满足要求的供应商，则需要通过计算采购成本来进行比较分析。采购成本一般包括售价、采购费用、运输费用等各项支出的总和。采购成本比较法是针对各个不同供应商的采购成本，计算分析来选择采购成本较低的供应商的一种方法。基于以上的分析，下面建立采购优化模型。

5.1.2 模型建立与求解

在建立模型之前，首先对模型的适用范围做以下假设：

(1) 运输时间与运输距离成正比；供应商收到采购订单便立即发货，即交货时间等于运输时间。不考虑由于原材料品种不同引起运输时间上的差异。

(2) 运输价格与运输距离成正比，运输费用与运输量成正比，运输费用由需求方承担。从一个供应商处运输多种原材料的运输成本是独立的。

(3) 采购量与库存量之和大于预计消耗量。预计消耗量是不确定的。

(4) 考虑数量折扣。

根据以上的假设，分析原材料订货总成本。订货总成本可分为两个部分：第一部分为订单处理成本，第二部分为运输成本。

为了建立相应的模型，我们给出符号与假定如下：

第 i 类原材料的所需采购量为 M_i；

现有库存量为 O_i；

计划期内需求量为 N_i；

对于第 i 类原材料，供应链中的供应商有 J_i 家；

供应商 j 的供应能力为 C_{ij}；

信誉等级为 G_j；

供货质量等级为 Q_{ij}；

交通距离为 d_j；

从第 j 供应商处订购第 i 类原材料数量为 x_{ij}。

供应商的信誉度定为 1~10 个不同的信誉等级，其供货质量有 A、B、C、D、E，可以量化为 1~5 个质量等级。

(1) 订单处理成本

由于供货价格 P_{ij} 是随订货量 x_{ij} 的增加而降低的，其最低价格渐近于该供应商生产此原材料的成本价格 L_{ij}，因此，价格函数可表示为

$$P_{ij}=u_{ij}(x_{ij})=\frac{a_{ij}}{x_{ij}}+L_{ij},\quad x_{ij}\geq Y_{ij}, \tag{5.1}$$

其中 Y_{ij} 为最低采购量。

则原材料购买成本

$$S_i=\sum_{j=1}^{J_i}x_{ij}\cdot P_{ij}=\sum_{j=1}^{J_i}x_{ij}\cdot(\frac{a_{ij}}{x_{ij}}+L_{ij}) \tag{5.2}$$

(2) 运输成本

订货费用主要为运输费用由供应商的地理位置和其他交通运输条件决定，可表示为关系

$$v_{ij}(x_{ij})=b_j\cdot d_j\cdot x_{ij} \tag{5.3}$$

其交货时间为

$$T_j=c_j\cdot d_j,\ \text{其中}\ b_j\ \text{和}\ c_j\ \text{为常数。} \tag{5.4}$$

则原材料运输总成本为

$$TS_i=\sum_{j=1}^{J_i}v_{ij}(x_{ij})=\sum_{j=1}^{J_i}b_j\cdot d_j\cdot x_{ij} \tag{5.5}$$

(3) 约束条件

首先，采购量必须满足生产计划的数量。原料的订货量直接由产品订单上的数量决定，物料的订货量以上述物料概率模型所预测的物料消耗量作为最低订货量。因此，对于订货数量已限定的原料和物料，可以以相同的方式来寻找其最佳订货方案。

$$\sum_{j=1}^{J_i} x_{ij} = M_i, \quad i=1,2,\cdots,m. \tag{5.6}$$

其次，对供应商 i 的订货数量必须满足供应能力条件和最低采购启动量条件。

$$Y_{ij} \le x_{ij} \le C_{ij}, \quad j=1,2,\cdots,J_i, \quad i=1,2,\cdots,m. \tag{5.7}$$

最后，对于采购量

$$M_i + O_i > N_i, \quad i=1,2,\cdots,m. \tag{5.8}$$

(4) 采购成本模型

式(5.2)表示在各个供应商订货的总成本，式(5.5)表示运输总成本，式(5.6)(5.7)(5.8)表示采购量的约束条件。

对于各个评价指标，根据信息系统中的对供应商的评价数据，确定权重系数，量化指标，给出一个目标函数。

$$\begin{cases} f_i = \sum_{j=1}^{J_i} (x_{ij} \cdot (\frac{a_{ij}}{x_{ij}} + L_{ij}) + b_j \cdot d_j \cdot x_{ij}) \bullet T_j \bullet (10 - G_j) \bullet (5 - Q_{ij}) \bullet K1_j \bullet K2_j \bullet K3_{ij}, \\ \qquad i=1,2,\cdots,m \qquad (5.9) \\ \sum_{j=1}^{J_i} x_{ij} = M_i, \quad i=1,2,\cdots,m. \qquad (5.10) \\ Y_{ij} \le x_{ij} \le C_{ij}, \quad j=1,2,\cdots,J_i. \ i=1,2,\cdots,m \qquad (5.11) \\ M_i + O_i > N_i, \quad i=1,2,\cdots,m \qquad (5.12) \end{cases}$$

式(5.9)表示由供应价格、采购数量、运输成本、交货时间、信誉等级、质量等级决定的的目标函数。其中 $K1_j$ 表示交货时间 T_j 的影响因子，$K2_j$ 表示信誉度 G_j 的影响因子，$K3_{ij}$ 表示质量等级 Q_{ij} 的影响因子。可以看出，当供应商的质量等级和信誉度越高，订货总成本越低，函数取最小值。此时，从各供应商采购的订货量为最佳订货方案。利用数值优化方法可以求得此函数的最小值。对于此模型的求解，我们同样利用自适应复合形法来求解。

5.2 采购优化系统的实现

(1) 物料采购优化

图 5.2 物料采购分配界面

Fig.5.2 Graph of assistant material procurement distribution

图 5.2 为物料采购分配界面图。其中的超始日期和终止日期期间为计划期，界面上方表格内的数据为根据当前计划期内的生产计划自动计算的各物料所需的采购总数量。通过“导入”，可读入物料采购预测数据，根据需求可选择物料进行采购优化计算，或者将读入的数据全部计算。界面下方的表格内中的数据表示上方表格中当前选择的物料的采购由两个供应商提供时，采购总成本最低，并分别显示了由采购优化所计算的各供应商的最佳分配量。

(2) 原料采购优化

原料的采购可由生产计划期内的产品所属原料种类和数量决定，因此，原料采购总量由生产计划数量决定。

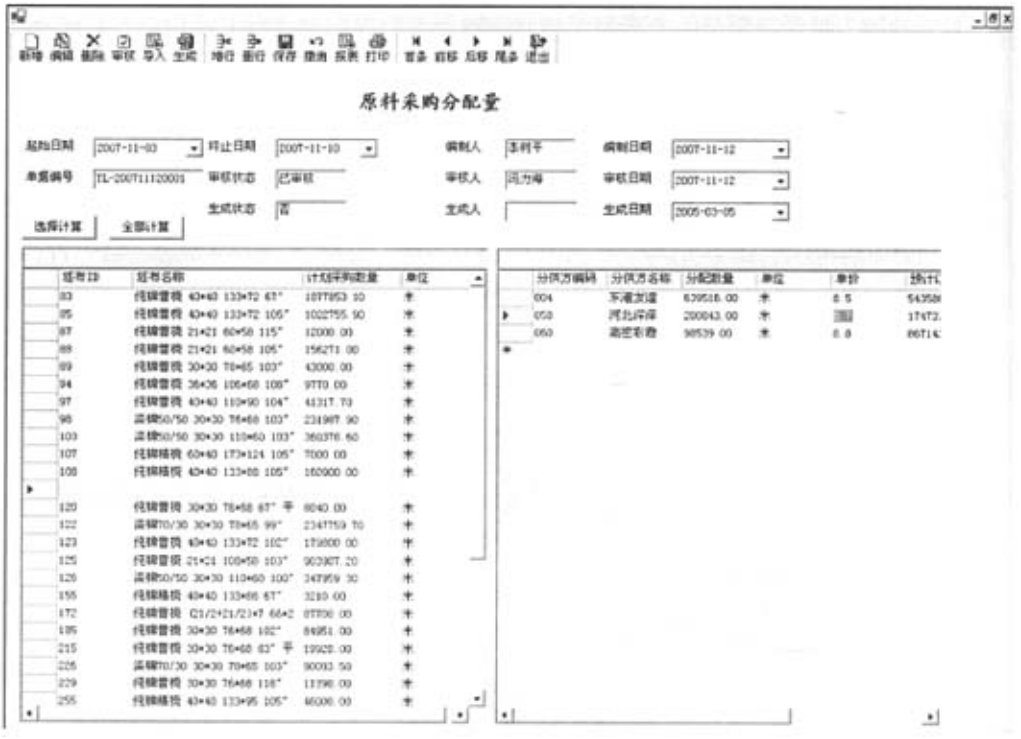

图 5.3 原料采购分配界面

Fig.5.3 Graph of main material procurement distribution

图 5.3 为原料采购分配量指导图。其中的超始日期和终止日期期间为计划期，界面左面表格内的数据为根据生产订单自动计算的当前计划期内各原料所需的采购总数量。界面右面的表格内中的数据表示左表格中当前选择的原料的采购有三个供应商可提供，并分别显示了由采购优化所计算的各供应商的最佳分配量。

图 5.4 为印染企业管理系统整体结构图，图中详细列出了采购管理子系统的主要模块，有采购决策、采购计划、采购订单、采购入库、采购结算、原材料库存管理以及采购退货等主要模块。其中的采购决策即本文中的采购优化子系统，也是本文完成的主要工作。采购决策是为采购计划服务的，它为采购人员提供了参考依据，目前实现的部分模块有物料消耗预测、物料订货方案和原料订货方案，为采购指导提供了基础，其他的功能有待于进一步研究与开发。

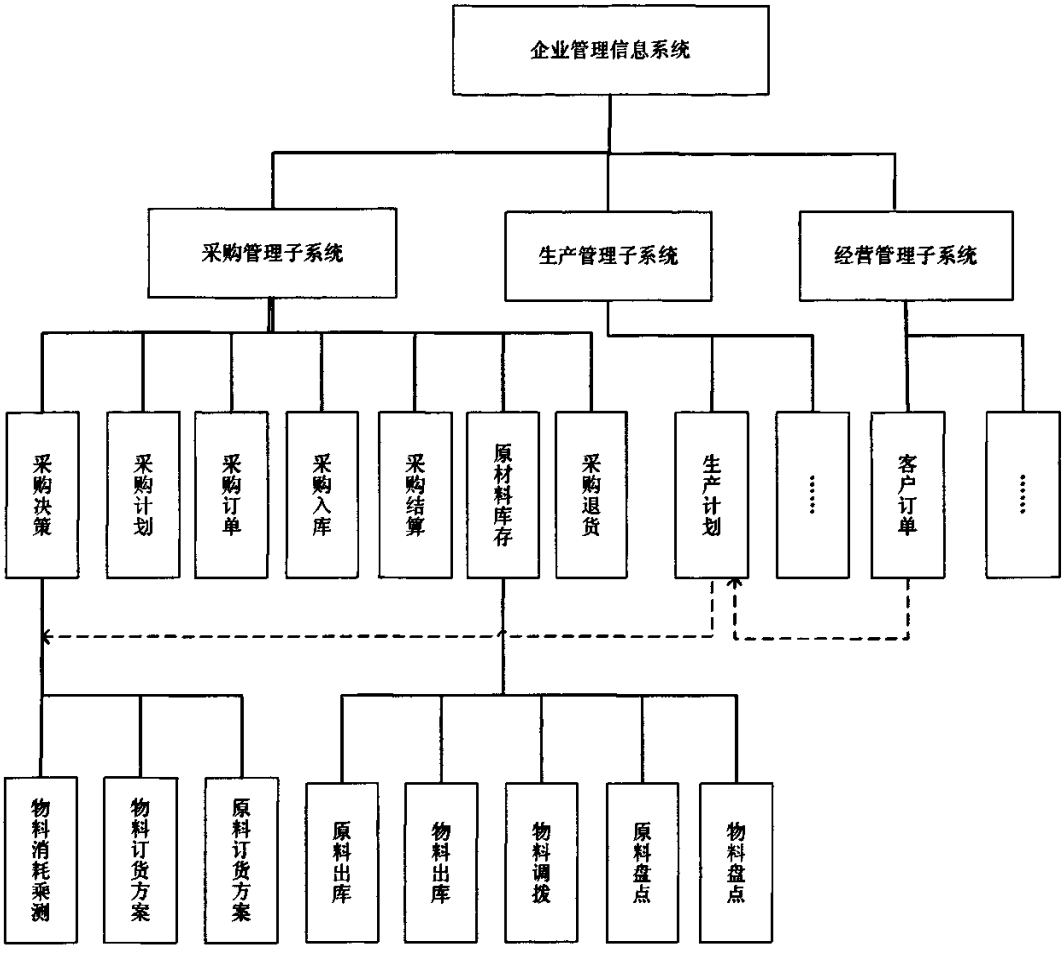

图 5.4 企业信息系统采购优化系统结构图

Fig. 5.4 Graph of procurement optimization system in MIS

图中的虚线部分表示两功能模块之间的数据通信关系，客户订单是生产计划的来源之一，物料消耗预测是根据生产计划而生成的。采购人员可以调整物料消耗预测，作为物料需求，然后根据调整后的物料需求制订物料订货方案。原料订货方案中的总采购量由生产计划与原料库存决定。

5.3 本章小结

本章针对供应商选择和订购量分配问题，建立了基于确定需求的采购优化模型，此采购优化模型同时适用于原料。针对在现实采购活动中，人们对各个目标的重视程度往往并不相同的实际情况，把权重向量值引入到模型中。通过建立此模型，根据以往供应

商供货的数据，可以判断其供货能力、质量、价格、信誉度等信息，这些信息可以帮助用户选择供应商，来推断在多供应商之间如何合理地分配订货量，以达到最低订货总成本，确定最优订货计划。

模型中的参数同样以自适应复合形法进行求解，最终实现了采购优化系统，通过应用算例说明了所提出方法的可行性与有效性，结果表明，自适应复合形法在参数求解方面比较稳定，适用于企业生产中的模型计算。

结　论

本文在论述了数据挖掘和最优化的理论与方法的基础上，分析了利用基于搜索优化算法的数据挖掘过程，首先以印染企业为背景，将采购优化问题分解为两个子问题。一是对原材料消耗量的预测问题，二是根据所预测的消耗量制订采购最优方案。对于第一个问题，通过数据挖掘中的统计方法建立了物料消耗概率模型，并结合最优化算法计算出模型的参数，得到了单位产品分类消耗物料的期望值，并根据生产计划计算出物料的预测消耗量。然后以得到的物料预测消耗量建立了合适的采购优化模型，这种采购优化模型同时适用于原料，通过对信息系统中的供应商的历史评价指标等数据制订最优订货方案，同样使用自适应复合形法求解参数，即实现如何合理地多个供应商之间分配订货量，以期达到最低订货总成本。实验表明，自适应复合形法在参数求解方面比较稳定，适用于企业生产中的模型计算。

本文中所采用的最优化方法——自适应复合形法作为数据挖掘中的算法，对于需要发现模型中参数值精度不高的挖掘任务，可以较快的求解，适用于工程上的应用。基于数据挖掘与最优化结合对支持最终采购决策分析的管理问题模型进行求解，目标是有效地将数据挖掘技术与最优化方法在实际应用中有机地结合起来，并为复杂的采购管理决策分析问题的求解和决策实施提供一个可以依赖的参考模型。

在物料消耗的预测问题上，今后的工作重点之一可以是寻找计算速度更快的最优化方法，或者用多种优化算法的结合对参数求解，使计算结果更符合实际生产活动、预测结果更加可靠。对于采购优化问题，可以加入更多的因素，以使采购优化模型更加合理，使结果更加精确可信。

参 考 文 献

[1] 刘同明. 数据挖掘技术及应用. 北京:国防工业出版社, 2001.

[2] 张尧庭, 谢邦昌, 朱世武. 数据采掘入门及应用. 北京:中国统计出版社, 2001.

[3] Padmanabhan B, Tuzhilin A. On the use of optimization for data mining:theoretical interactions and eCRM opportunities. Management Science, 2003, 49(10):1327-1343.

[4] Goldberg D E. Genetic algorithms in search, optimization and machine learning. Addison-Wesley, 1989. 159-160.

[5] Han J W, Kamber M. Data Mining:Concepts and Techniques. 2000.

[6] Kantardzic M, Data mining concepts and models, methods and algorithms. 2003, 68-72.

[7] Usama M F. Advance in knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.

[8] 胡侃, 夏绍纬. 基于大型数据库的数据采掘:研究综述. 软件学报, 1998, 9(1):4-6.

[9] Weiss S M, Indurkhya N. Predictive data mining:a practical guide. Morgan Kaufmann Publishers, 1998.

[10] Cherkassky V, Mulier F. Learning from Data:Concepts, Theory and Methods. John Wiley and Sons Inc, 1998.

[11] 马江洪, 张文修, 徐宗本. 数据挖掘与数据库知识发现:统计学的观点. 工程数学学报, 2002: 19(1):1-13.

[12] Vapnik V. The Nature of Statistical Learning Theory. Springer-Verlag, 1998:48-55.

[13] Cortes C, Vapnik V. Support-vector networks. Machine Learning, 1995, 20:273-297.

[14] Burges J C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998:2(2):121-167.

[15] 陈宝林. 最优化理论与算法. 北京:清华大学出版社, 1989 年 8 月.

[16] 施光燕, 董加礼. 最优化算法. 高等教育出版社, 1999 年 9 月.

[17] 姜英新, 孙吉贵. 约束满足问题求解及 ILOG SolVer 系统简介. 吉林大学学报(理学版), 2002, (1):53-60.

[18] 刘晓石, 陈鸿健, 何腊梅. 概率论与数理统计. 北京:科学出版社, 2000.

[19] 陈新海. 最佳估计理论. 北京:北京航空航天大学出版社, 1985.

[20] 边肇祺等. 模式识别. 北京:清华大学出版社, 2000.

[21] Dempster A, Laird N, Rubin D. Maximum likelihood estimation from incomplete data via EM algorithms. J. Royal Statistical Society Series B, 1977, 39:1-38.

[22] Ghahramami Z, Jordan M. Learning from incomplete data. Technical Report AI Lab Memo No. 1509, CBCL Paper No. 108, MIT AI Lab, August 1995.

[23] Blimes J. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden markov model. TR-97-021, April, 1998.

[24] Mangasarian O L. Mathematical programming in data mining. Data Mining and Knowledge Discovery, 1997, 1(2): 183-201

[25] 贾沛璋, 朱征桃. 最优估计及其应用. 北京: 科学出版社, 1984.

[26] 王炯琦, 周海银, 吴翊. 基于最优估计的数据融合理论应用数学. 2007, 20(2): 392-399

[27] 杨冰. 实用最优化方法及计算机程序. 哈尔滨船舶工程学院出版社, 1994.

[28] 袁亚湘. 非线性规划数值方法. 上海科技出版社, 1999.

[29] 袁亚湘, 孙文瑜. 最优化理论与算法. 科学出版社, 1997.

[30] 王小平, 曹立明. 遗传算法—理论, 应用与软件实现. 西安: 西安交通大学出版社, 2003.

[31] 徐涛. 数值计算方法. 吉林科学技术出版社, 1998, 7(1): 14-62.

[32] Box M J. A new method of constrained optimization and a comparison with other methods, Computer Journal, Oxford University Press, Oxford, U. K., 1965, 8: 42-52.

[33] Manetsch T J. Towards efficient global optimization in large dynamic systems-The adaptive complex Method. IEEE Transactions on Systems, Man and Cybernetics, IEEE Systems, Man, and Cybernetics Society, Charlottesville USA, 1990: 257-260.

攻读硕士学位期间发表学术论文情况

[1]隋艳茹, 郭禾. 改进的复合形法在库存优化系统中的应用. 大连理工大学网络学刊(录用).

相关章节：第四章.

致　谢

本论文是在我的导师郭禾老师和王宇新老师的悉心指导下完成的。从研究方向到选题立论，从资料收集到程序设计，以及论文的最后完成，自始至终都得到了老师和各位学长无微不至的关怀和帮助。两位老师严谨的治学态度，对事业的执着追求，令我受益匪浅，终身难忘。他们的教诲是我人生经历中最宝贵的财富。在此，谨向郭老师和王老师致以我最诚挚的谢意。

还要感谢计算机系的全体老师，正是他们六年来的谆谆教诲使我能够在计算机领域里有所发展，从基础到前沿对计算机有了深入的学习，为我将来博士阶段的学习和研究打下了坚实的基础。

另外要感谢我所在的教研室的全体成员，正是教研室严谨、团结、友爱的氛围才使我能将全部精力投入到本论文的研究中去。

感谢几年来与我朝夕相处的同学，他们在学习上和生活上都给予我非常大的帮助。感谢师兄刘天阳，师姐张晓莉，同学张洪业、陈鑫，师妹郑琴琴、李丹等在论文内容的讨论，论文撰写中对我的帮助。

最后，我要特别感谢我的父母，正是他们 24 年的养育，支持和鼓励才能使我顺利完成学业，在最困难的时候给予我信心和力量，帮助我不断前进。

各位老师、亲人，同学，朋友，是你们的热忱关心和帮助，使我的论文能够顺利完成。衷心的祝愿你们身体健康，工作顺利，万事如意！