摘要

能源与环境是当今世界的两大热点问题，越来越受到世人的瞩目。我国是能源消费大国，能源利用率低，人均能源储备很少，节能工作成为了国家当务之急的重任。本文借助新兴的数据挖掘技术来研究和预测建筑能耗，为本专业在建筑能耗领域的研究提供新方法和新思路，做出跨学科研究的新探索。

本文首先建立上海市商用建筑信息数据库，包含历史能耗数据、建筑基本信息、空调系统信息和建筑使用情况等数据，并以问卷调查的形式收集 95 幢商用建筑的信息作为本数据库的基本数据。通过对这些基本数据的整理、分析和研究，得出这些商用建筑的运行使用情况和能源消耗特点，进一步找出影响建筑能耗的主要因素，并以此作为数据挖掘研究的基础和依据。

本文的数据挖掘过程是由 SAS 软件的 EM 功能模块来实现的。在上海市商用建筑信息数据库的基本数据的基础上，按照 SAS 软件提出的 SEMMA 数据挖掘方法论，建立一个完整的数据挖掘流程，并借助多元线性回归和主成分分析的数据挖掘算法得到两个回归模型。通过对两个模型的解释、说明、比较和验证，最终确定回归模型 I 作为本研究阶段的商用建筑能耗预测模型。

本文首次将数据挖掘技术引入建筑能耗的研究领域，对这一新方法的可行性和可用性做了深入研究，并取得了较好的结果，为数据挖掘技术在暖通专业，尤其是在有关建筑能耗的研究领域中的应用提供了有益的参考和借鉴。

关键词：数据挖掘，商用建筑，建筑能耗，预测模型

ABSTRACT

The two hot topics – Energy and Environment — become more and more concerned by the people around the world. China is a country of huge energy consumption. The energy efficiency and the energy resource per capita are low. Therefore, energy conservation has become one of the most important tasks of China. This thesis employs the rising data mining technique on the analysis and prediction of building energy, which not only provides a new method and idea for the research in the field of building energy but also makes a new explore in multi-discipline research.

First of all, the Shanghai information database of commercial buildings is built up. It contains historical energy consumption data, basic building information, data of HVAC (Heating, Ventilation and Air Conditioning) system and building operational information. The data and information of 95 Shanghai commercial buildings are collected as the basic data of the database via questionnaires. The status of operation and the characteristics of energy consumption of these commercial buildings are worked out by summarizing, analyzing and studying these basic data. Furthermore, the main factors that may affect the building energy consumption are searched as the basis of the study of data mining.

Then the data mining process is carried out by the EM function module of SAS. On the basis of the Shanghai information database of commercial buildings, a complete data mining diagram is established by the SEMMA methodology of SAS. Two regression models are made by multiple variable linear regression analysis and principal component analysis. Through the explanation, illumination, comparison and verification of the two models, regression model I is defined as the prediction model of the energy consumption of commercial buildings during this phase of the research.

This thesis introduces the technique of data mining into the research field of buildings energy consumption, doing a further study on the feasibility and usability of the new method and obtaining good results. It provides a valuable reference for the application of data mining in the domain of HVAC and especially buildings energy

research.

Key Words: data mining, commercial buildings, building energy consumption, prediction model

目录

第 1 章 引言

1.1 课题背景

1.1.1 我国的能源形势

自 1973 年世界能源危机以来，能源短缺已经引起了各国政府的普遍关注，都十分重视节约能源的问题，兴起了世界性的节能运动，并把节能称为五大能源之一，与煤炭、石油、天然气、水电四大常规能源相提并论，联合国维也纳科学技术促进发展会议也把能源列为人类未来将面临的四大问题之一。把能源消费降低到最低限度已经成为当前世界各国共同致力追求的目标[1][2]。

随着我国社会经济的发展和社会生产力的不断提高，我国的能源消费也在逐年增加。2003 年，我国的能源生产总量达到了 160300 万吨标准煤，消费总量达到了 167800 万吨标准煤[3]，一次能源消费总量列世界第二位。同时，我国的能源利用率却比较低，能源消费的增长率已经远远高于 GDP 的增长率。我国是一个能源资源比较贫乏的国家，在不久的将来我国能源消费的相当一部分就必须依赖于进口来解决。能源紧缺和环境污染问题已经成为制约我国国民经济长期、稳定、高速发展的重要因素。因此，节能作为我国国民经济发展的一项基本国策，必须在我们的经济活动的各个方面和领域中得到贯彻和落实。

1.1.2 我国的建筑能耗现状

在能源消费中，建筑是一个能耗大户。从广义的角度来说，建筑能耗包括：建筑材料的生产和运输的能耗、建筑物建造时的能耗、建筑物寿命周期内的逐年运行能耗，甚至包括建筑设备的生产和运输能耗。其中，建筑物的使用运行能耗约占建筑总能耗的 80%～90%，因此我们也常以建筑物建成后，在使用过程中一年所消耗的能量总和作为建筑能耗[4]。

目前，我国建筑能耗已经占社会总能耗的 27%左右，建筑能耗的逐年上升对我国国民经济的影响是非常显著的，主要表现在[5]：

（1）能源消耗巨大。不仅既有的近 400 亿平方米建筑中 99%为高耗能建筑，新建建筑中 95%以上仍属于高能耗建筑，单位建筑面积采暖能耗为发达国家新建建筑的 3 倍以上。按照目前的建筑能耗水平发展，到 2020 年，我国建筑能耗将达到 10.89 亿 tce（吨标准煤），超过 2000 年的 3 倍，空调高峰负荷将相当于 10 个三峡电站满负荷出力。

（2）夏季高峰供电严重短缺，电厂、变电站、电网全部处于负荷极限，致使许多地区不得不采取拉闸限电的极端方式来解决这一问题。全国空调负荷已达到 4500 万 kW,相当于 2.5 个三峡电站满负荷出力。

（3）建筑能耗随建筑物的不同使用功能呈周期性变化，它的不稳定性对电网有潜在的威胁。

（4）燃煤电厂和锅炉是城市的主要污染源和温室气体的排放源。

并且，随着国民经济的发展，人们生活、工作环境的不断改善，建筑能耗还将会保持较高的增长速度，主要原因有[5][6]：

（1）房屋建筑继续增加。21 世纪头 20 年，将是我国建筑业的鼎盛期，2020 年全国建筑面积将接近 2000 年的 2 倍。目前我国每年建成的房屋达 16 亿～20 亿平方米，每年人均新增房屋面积 1.3～1.5 平方米。

（2）城镇化不断加快，农村人口大量向城市转移。城市人口人均能耗为农村人口的 3.5 倍。同时，我国人口总量也在不断增加，每年增加约 900 万人。

（3）人们对建筑热舒适性的要求越来越高。冬天室温由 12℃、16℃提高到 18℃甚至 20℃；热天的室温由 32℃降低至 28℃、26℃，甚至 22℃。

（4）采暖区大大向南扩展，并且空调制冷范围已从公共建筑扩展到居住建筑，从南方扩展到北方。使用采暖和空调的时间也在延长。

我国严峻的能源形势和建筑能耗的现状，决定了建筑节能是我国可持续发展战略的—部分，也是国家的重大战略问题。必须有效地利用能源，积极提高一次能源效率，调整能源结构，在保证和提高建筑舒适性的前提下，尽量降低建筑物使用过程中的能耗。如果继续放任自流，错过当前的大好机遇，不给予高度重视，不采取坚决有效的措施，则将对我国社会经济的可持续发展产生严重阻碍，也会对国家的能源安全和大气环境造成重大威胁。

1.1.3 国内外研究现状

能源与环境已经成为了全球关注的热点问题，建筑节能自然也就成为了本行业的一个研究热点。建筑节能工作的顺利进行离不开对建筑能耗的准确把握、科学分析和合理预测。同时，对于新建建筑的设计、建筑用能的评估、楼宇设备的运行管理等工作也都需要了解建筑的用能特点和规律，并做出合理的分析和预测。近些年来国内外的专业人士在这一领域进行了不懈的探索，取得了许多重大成果。

1．计算机能耗模拟

计算机能耗模拟方法是最常用的建筑能耗分析方法。目前可采用的建筑物能耗分析方法有很多，根据所依据的数学模型，可将计算方法分为两大类：一类是建立在稳定传热理论基础上的静态能耗分析法，另一类是建立在不稳定传热理论基础上的动态能耗模拟法。静态能耗分析法没有考虑各部分围护结构的蓄热效应，适用于只需知道整个建筑物或单位建筑面积在一个空调期的传热量，并不需要详细掌握传热量随时间变化的具体情况。静态能耗分析的方法主要有：有效传热系数法、度日法、BIN 方法(变基准温度的度日法)、当量满负荷运行时间法等[8]。由于建筑能耗是时刻变化的，并且对于节能工作更有指导意义的是建筑能耗的瞬时值，因此传统的静态能耗分析法已不能满足现实的需要，必须采用动态的能耗分析方法[9]。动态能耗模拟法对室内外各种扰量考虑较细，得到的结果也比较准确，但是动态模拟的过程非常复杂。随着计算机技术的迅猛发展，现在已可以方便地对不断变化的室外参数作用下建筑物的冷热负荷进行动态计算，以及对建筑物的全年能耗进行动态模拟计算。目前世界上许多国家都开发出具有不同特点的建筑物能耗模拟计算软件[10]~[14]，比较有代表性的有 DOE-2、EnergyPlus、Energy-10、HAP、TRNSYS、TRACE、HASP、Seri-Res、ESP-r、DeST 等等。

计算机模拟技术在各个方面都获得了很好的应用。这些软件采用各种不同的数学模型，充分考虑影响建筑能耗的各种因素，对建筑物进行动态的能耗模拟计算，并得到了较精确的结果[15]~[19]。借助计算机模拟软件，可以再现建筑物实际的运行情况[20]~[22]，得到其能耗特点，分析各种因素对建筑能耗的影响[23][24]，并对一些节能措施进行研究和评价[25][26]。这为研究人员和建筑管理者提供了很大的便利。但是这种方法需要研究人员对模拟软件有相当程度的了解，并

且需要输入大量的建筑及其系统的详细数据资料，有的还需要借助大型电子计算机。它对操作人员的专业技能要求较高，一般的建筑管理人员和非专业研究人员很难完成。另外，建模和校正过程要耗费大量时间，建立的模型只适用于某一特定建筑，不能用于其它建筑，因此具有一定的局限性。

2．建筑能耗统计调查

借助数据库技术和统计调查的方法来研究建筑能耗也是一项很有意义的工作。进行建筑能耗调查统计，可以全面了解国家的建筑能耗水平、建筑终端商品能耗结构、建筑用能模式，积累建筑能耗基础数据，为国家能源结构调整，制定与检验相关能源政策，挖掘建筑节能潜力提供有力的数据支持，对国家建筑节能工作的全面推进有积极现实意义，许多学者在此领域获得了很宝贵的科研成果[27]。1976 年英国开始对建筑物的能耗进行调查，而美国那时已经由华盛顿的国家标准局开始对建筑能耗进行调查。这种详细调查的目的主要是对已经存在的建筑进行建筑节能改造，也就是最原始的节能潜力研究，这种技术被称为能源审计(Energy Audit)。Energy Audit 涉及的内容非常详细，主要调查的项目有：建筑物的分类、所属关系、特征、围护结构状况、采暖/空调系统情况、人员、照明、太阳辐射状况、燃料、年总能耗、成本，运行维护的程序以及推荐的节能措施与潜力。近些年来 Energy Audit 在其他国家也非常的普遍，它几乎成了建筑节能的关键所在[28]~[31]。

我国的建筑能耗调查统计属于能源统计的范畴。一直以来都是以工业和交通为主，终端能耗按一、二、三产业和生活分类，以及对能源种类等方面进行统计。建筑能耗的调查统计作为能源统计中的一个消费环节，长期被分割汇杂在能源消耗的各个领域，比如住宅的能耗被归入城乡人民生活能源消费，而其他各类建筑能耗被归入非物质生产部门的能源消费。因此，大力开展以建筑能耗为研究对象的调查与统计就显得尤为重要。再加上近些年来建筑能耗在全国总能耗中的比例逐年增加，建筑节能被广泛重视，不少业内人士进行了许多建筑能耗的调查统计以及相关的节能潜力研究工作[32]~[34]。

从 1989 年开始，以涂逢祥为首的“中国建筑节能经济技术政策研究”组，开展了以“系统掌握我国建筑能耗、建筑热环境、建筑节能工作进展的实际情况”为目的的调查工作[35]，调查覆盖我国北方采暖地区和长江沿岸的重庆、宜昌、武汉、南京四城市的各种建筑类型，旨在了解我国城市热环境与能耗状况，完善我国建筑节能政策、计划。通过本次调查得到了被调查城市单位建筑面积

能耗数据，为建筑节能法规政策的制定提供了强有力的数据基础，是我国历史上最早的全方位建筑能耗调查。龙惟定教授等学者对上海市 200 多幢公共建筑的能耗现状进行了调查分析，提出用系统能量效率作为建筑节能的评价指标较之用单位面积平均一次能耗更为合理[36]。清华大学的武海斌等曾对北京市的 410 户城市居民家用空调器的耗电量进行了调查与研究，得出了以建筑面积为自变量的空调用电量的概算指标计算公式，并对未来家用空调器耗电量进行了分析预测[37]。翟超勤等根据其调查结果，在合理简化的基础上，利用建筑热环境模拟分析软件 DeST 给出了全国各省市家用空调器耗电情况的估计[38]。建设部科技发展中心与哈尔滨工业大学合作对哈尔滨市建筑物基本情况、单体建筑能耗、居民住宅的能耗总量、采暖区煤耗量进行了调查，并编制出建筑能耗统计软件[39]，为建筑能耗统计在全国范围内开展提供了有力的工具。

通过大量的统计调查，可以得到不同区域、不同类型的建筑的各种信息和能耗数据。这种方法可以用于能源统计和调查，了解宏观的用能规律，科学地制定能源政策；还可以用于了解区域建筑能耗的特点，研究气候、生活习惯、经济发展水平等因素对建筑能耗的影响；也可以用于掌握同类建筑的用能特点，研究建筑类型、功能特点、系统类型、运行方式等因素对建筑能耗的影响；以及进行能耗组成、能耗指标、能源方式、用能效率等内容的研究。但是，这种研究方法也有其局限性。首先，要想获得准确、可靠的统计结果必须拥有一定数量的统计样本，这就需要大量的人力、物力的支持；其次，样本数据的质量也是重要因素，容易因主观和客观条件的制约而影响数据的准确性；再次，统计抽样方法和调查变量的选取也会对统计结果产生很大的影响。

3．统计分析方法应用于建筑能耗研究

多元回归、多变量分析等统计手段也被用于建筑能耗的研究中。借助回归的方法可以建立起各个影响因素与建筑能耗之间的关系，并在此基础上对建筑能耗进行深入研究[41]。也可以建立 Baseline（基准）模型[42]，将其它建筑与之进行比较，用于进行能耗分析和评价，以及节能评估等等。还可以开发出 Benchmarking Tool（评价工具）[43]，它可以用于建筑设计方案和已建建筑，为它们提供建筑能耗分析、用能水平定位、节能措施的评估等。

尽管统计分析方法有较强的理论基础和数学背景，也能给出比较直观、易懂的结论，但是统计分析结果的好坏取决于数学模型的选择和对客观规律的定性把握。在建筑能耗的研究领域中，客观规律比较复杂，各个方面的因素对建

筑能耗的影响都不是线性的，并且有时在各个因素之间也有复杂的联系。因此，单纯采用统计分析的方法来研究建筑能耗有一定的局限性，它只能在某些特定领域取得较理想的结论。

4．暖通空调领域的专家系统

专家系统[40]是人工智能三大分支之一，兴起于上世纪六十年代中期，它是指在某一特定领域内利用人工智能技术、汇集专家的知识及经验的一种计算机软件系统，简单来说就是借助计算机来模拟专家解决问题。专家系统必须具有专家领域的专业知识，还可以不断从专家那里获得知识，和专家一样能处理十分复杂的现实问题。专家系统自上世纪八十年代中期起已应用于暖通制冷领域，如自 1986 年起，美国的 ASHRAE 年会都会举行专家系统应用专题讨论会。该课题的开发也由(美)Colorado 大学扩展至美国各大学，将专家系统应用到建筑能耗分析、系统诊断、建筑物空气渗透、系统控制和维护管理、能耗模拟等方面。专家系统以其独特的实用性可在如下几个方面得到广泛应用：

（1）建立暖通空调制冷领域内专家系统知识库——知识工程；

（2）监视、预测，并控制系统处于最佳状态下运行．并实现节能目标和智能控制；

（3）诊断与预测故障，指导管理、维修、保养等工作；

（4）实现计算机辅助设计、分析系统，制定规划和指导安装调试设备。

专家系统在建筑能耗研究领域的应用还有很多工作要做。有关建筑能耗的规律和特点是非常复杂的，需要专家系统汇集非常多的专家以获取足够的专业知识和经验；建筑能耗研究领域的问题也是非常复杂的，并带有随机性，这对专家系统的处理问题的能力提出了很高的要求。

1.1.4 相关课题研究进展

数据挖掘（Data Mining）技术是近年来新兴的研究领域——知识发现（Knowledge Discovery in Database, KDD）的一个重要部分。它是“对数据库中蕴涵的、未知的、有潜在应用价值的、非平凡的模式的提取”[44]，它是将过去累积的大量繁杂的历史数据进行分析、归纳和整合等工作，提取有用的信息，找出有意义的模式，其基本过程为：定义问题、创建清理数据和数据预处理、数据挖掘、模式解释和知识评价。数据挖掘技术在近几年得到了迅猛发展，它

所表现出的广阔应用前景吸引了众多的研究人员和商业公司，目前数据挖掘技术已在商业、经济、金融、管理、医学等领域都取得了应用性成果[45]。

数据挖掘主要有以下三个特点[46]：

（1）数据挖掘得到的知识以多种方式表达出来，比传统的单纯用数字表示的方式更直观，更易接受，也更有助于实时快速决策。数据挖掘只需用到对其有用的或有意义的输入，而不需要对系统做完全的描述，故提高了计算效率。

（2）数据挖掘可通过对历史和当前数据的分析来得到对未来的预测，从而使人们对系统潜在的问题有更明确的认识。

（3）针对系统中可能存在的不确定因素的影响，数据挖掘通过放宽对模型的假设来进行仿真，避免了建立精确的数字模型的困难。

由于数据挖掘具有以上的特点，该技术在暖通空调领域也有着非常广阔的应用前景。目前，在本专业领域内已经有所尝试[47]，利用数据挖掘技术来分析各种数据，以及确定室内热环境与通风方式的关系等等，取得了很好的研究成果。本文将进一步扩大数据挖掘技术的应用范围，探寻数据挖掘方法在建筑能耗的分析和预测中的应用。

1.2 研究内容及方法

本文首先建立上海市商用建筑信息数据库，该数据库是基于互联网的实时数据库，包括商用建筑的基本信息和历史能耗数据等内容；

通过问卷调查的形式收集数据库的基本数据，并对这些基本数据进行统计分析，得出上海市商用建筑的一些概况和能耗特点，然后将这些数据作为数据挖掘的基础；

使用专业的统计分析软件 SAS 的 EM 模块对目前收集的基本数据进行数据挖掘研究，遵循其 SEMMA 的方法论完成从数据预处理到建立模型以及模型评价的完整的数据挖掘流程。得到各个影响因素与建筑能耗的关系，并借助回归、决策树、神经网络、主成分分析等多种挖掘算法进行能耗预测模型的研究；

对数据挖掘的结果进行比较分析，并对各个预测模型进行评价，确定最终的商用建筑能耗预测模型。

1.3 研究目的及意义

本文将新兴的数据挖掘技术引入暖通空调的研究领域，借助数据挖掘技术进行商用建筑能耗的分析与预测研究。

本文将建立目前为止国内第一个商用建筑的能耗信息数据库—上海市商用建筑信息数据库，以此来了解上海市商用建筑的基本状况和能耗情况，积累建筑能耗的基础数据，并为建筑节能工作和能源政策提供数据支持。

在此数据库的基础上使用数据挖掘技术确定商用建筑能耗与各个影响因素之间的关系，从而建立商用建筑的能耗预测模型，为商用建筑能耗的分析和预测提供新方法、新思路。

本课题运用数据挖掘技术进行建筑能耗的研究，是跨学科研究的一种尝试，为本专业的应用研究做出了新的探索，验证了数据挖掘技术在暖通空调领域的可用性，并展现出很好的应用前景。

第 2 章 上海市商用建筑信息数据库

2.1 上海市商用建筑信息数据库简介

2.1.1 建立数据库的目的和意义

长期以来，我国的能源统计一直是以工业和交通为主，终端能耗按一、二、三产业和生活分类，以及针对能源种类等方面的统计，建筑能耗只是其中的一个消费环节，分割混杂在能源消耗的各个领域，没有形成独立的统计对象，这就无法全面、准确地了解我国的建筑能耗的状况，不能给建筑节能工作的顺利开展提供详实、可靠的数据资源。同时，数据挖掘工作也需要大量的基础数据作为其研究依据。因此，有必要建立一个包含建筑的各种基本信息和建筑能耗的综合数据库。

通过建立上海市商用建筑信息数据库，可以较为全面、准确地了解上海市商用建筑的基本状况和能耗情况，积累建筑能耗的基础数据。利用这些数据可以制定明确的节能目标；可以提出节能的相对值和绝对值，为具体的节能措施提供参考标准；可以了解本地的建筑能耗特点，分析地域、气候、生活习惯、建筑形式等因素对建筑能耗的影响，有针对性地进行建筑节能技术的研究；还可以为政府调整能源结构、制定相关能源政策提供有力的数据支持，为其他城市的建筑能耗统计工作提供参考。

2.1.2 数据库的构建

由于目前上海市还没有一个比较全面的有关建筑能耗的数据库可供借鉴，因此本文先以商用建筑为试点建立起商用建筑的信息数据库，在取得研究成果之后再推广到其他类型的建筑。这里的“商用建筑”是指纯办公建筑和包含一部分餐饮、娱乐、商场等功能的综合性公共建筑，这类建筑在上海的数量比较多，也是比较典型的公共建筑。

目前国内的一些建筑能耗数据库有的仅包含了能耗方面的一些数据，而没

有收集其它对建筑能耗有影响的信息；有的数据库虽然包含了各方面的数据，但是大多以统计调查为目的，收集的数据可用性不强。本文建立的商用建筑信息数据库是以建筑能耗的预测为最终目的，因此既要包含建筑物的能耗数据，又要包含影响建筑能耗的各种因素的数据。此外，因为本文的研究手段是数据挖掘技术，所以建立的数据库应能满足数据挖掘的需要，可以为数据挖掘提供可靠的数据支持。

在构建数据库时，除了考虑上述因素还参考了国内外已有的建筑能耗方面的数据库，以确定上海市商用建筑信息数据库所应包含的数据信息。如美国环保署（EPA）的建筑物能耗评测工具就是基于建筑物的规模、租户密度、运行时间、插头负载、当地气候条件等因素来对建筑物的用能情况进行比较和排序。但是该工具是以美国的建筑物能源消耗、运行特点以及管理实践等因素作为假定条件，所以如果我国的建筑物直接使用该工具得到的结果就会有较大偏差。因此，本数据库以上海市商用建筑一般的运行、管理情况为基本条件，这就排除了气候、地域等影响因素。另外，结合大多数商用建筑的实际运行情况，将建筑基本信息、围护结构情况、能源系统状况、空调系统及其辅助设备信息、建筑使用情况等信息作为影响建筑能耗的因素引入数据库。

2.1.3 数据库的主要内容和功能

本文所建立的上海市商用建筑信息数据库是与上海市节能监察中心合作完成的。该数据库主要包括建筑基本信息和历史能耗数据两部分内容，如图 2.1、2.2 所示。建筑基本信息部分主要包括基本信息、空调系统信息和建筑使用情况。其中，基本信息包括竣工年份、建筑面积、层高、围护结构材料、玻璃材料、建筑功能分区等信息；空调系统信息包括空调冷热源、空调系统形式、新风系统、冷却塔等相关信息；建筑使用情况包括空调运行时间、建筑使用时间等信息。历史能耗数据部分则包括了该建筑近几年来全年各月的电、气、油等能源的使用情况。

上海市商用建筑信息数据库是基于互联网的实时在线数据库，主要面向上海市的各个商用建筑，这些建筑的业主或物业管理部门只要通过互联网登录到数据库网站就可以使用该数据库。除了提交建筑信息和能耗数据以外，用户还可以使用数据库提供的建筑能耗排序功能，了解自己建筑的能耗情况以及在上

海市商用建筑中的排名。借助本数据库，商用建筑的业主或物业管理者可以根据自己的建筑在上海市的定位进行大楼的能源管理，还可以根据对不同周期的能耗数据的比较进行节能改进和评估工作，既取得了降低运行能耗、提升物业管理水平的经济效益，又获得了减少污染物和温室气体排放的社会效益。

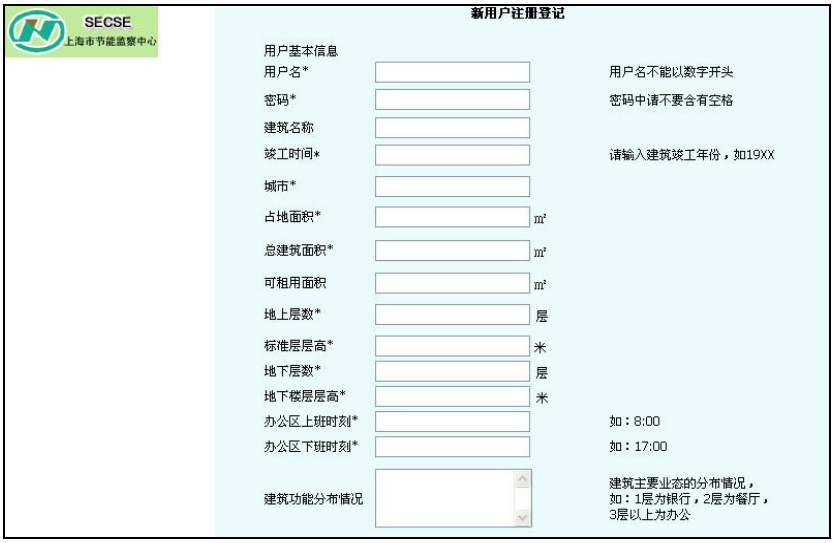

图 2.1 基本信息输入界面

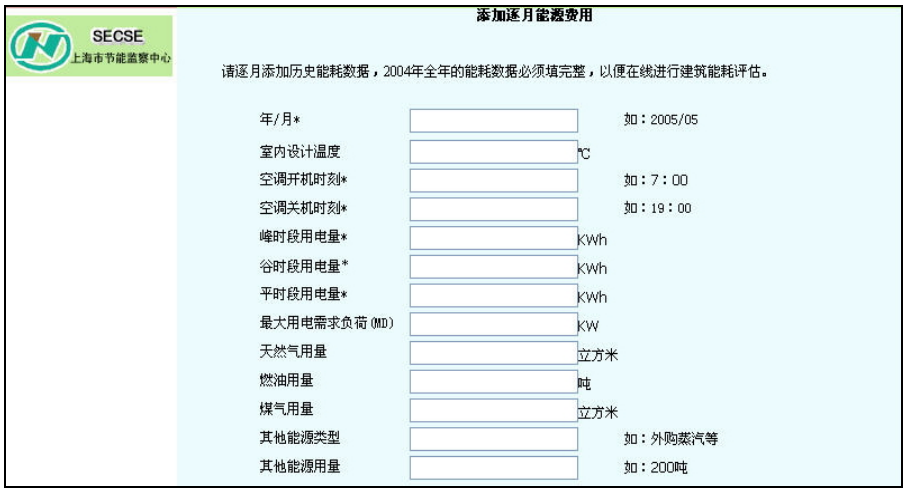

图 2.2 能耗数据输入界面

2.2 上海市商用建筑信息数据库的统计分析研究

由于开始时数据库处于调试阶段，不能通过互联网直接向数据库输入信息，因此在 2005 年 5 月至 2005 年 7 月间，笔者与上海市节能监察中心相关研究人

员一起，通过问卷调查的形式收集了 95 幢商用建筑的信息作为本数据库的基本数据。虽然目前数据库中只有这 95 幢建筑的信息，相比上海市的几千幢高层建筑比例很小，但是由于选择的这些建筑是随机抽取的，涵盖了全市各个城区、各种样式的商用建筑，因此具有一定的代表性，在此基础上进行的分析和研究具有一定的意义。

通过对这些基本数据的整理分析和研究，得出这些商用建筑的运行使用情况和能耗特点，可以进一步找出影响建筑能耗的主要因素，以改进和完善数据库的内容和功能，并且可以为下一步的数据挖掘研究提供基础和指导。

2.2.1 商用建筑基本信息的统计与分析

在这 95 幢建筑中，纯办公建筑有 20 幢，其余都包含商场、餐饮、银行等设施。这些建筑最早的建成于 1934 年，最新的竣工于 2004 年 7 月，有 65 幢建成于 1996 至 2000 年间，占总数的 68.4%，这一阶段正是上海房地产市场高速发展的时期，浦东的开发开放又给了上海前所未有的机遇，大批的高层建筑如雨后春笋般拔地而起。

这些建筑的建筑面积从 10000 平方米到 247000 平方米不等，平均为 71000 平方米，分布情况如图 2.3 所示。建筑高度从 34 米到 286 米不等，最低 8 层，最高 66 层。

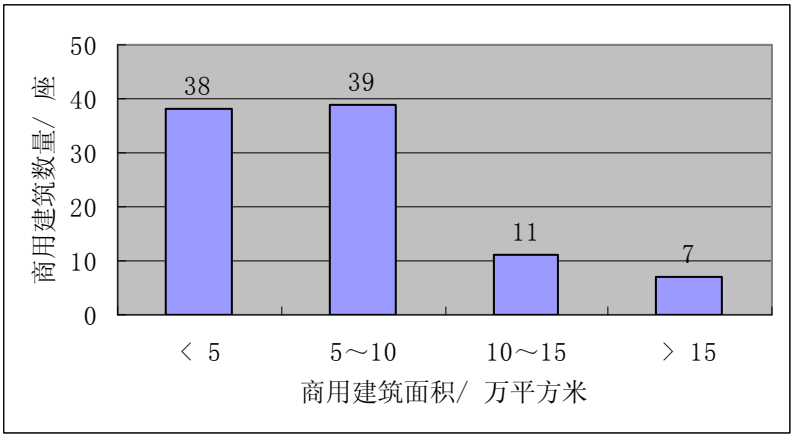

图 2.3 商用建筑总建筑面积分布情况

这些商用建筑的围护结构基本都是钢筋混凝土的框架式结构，钢结构建筑 1 座，多孔砖建筑 2 座。在高层建筑中多孔砖的使用比例越来越小，而钢筋混凝

土结构的保温隔热性能又比较差[48]，这就对建筑节能工作提出了挑战。从另一个角度来看，钢筋混凝土结构建筑的节能潜力也较大[49]，这也可以看作是建筑节能工作的一个机遇。

绝大多数被调查的商用建筑都采用了玻璃幕墙的外立面，其中采用单层玻璃的有 38 座，采用双层中空玻璃的有 29 座。出于对建筑美观的考虑，越来越多的高层建筑采用玻璃幕墙的外立面，这对室内环境的热舒适度影响很大，空调负荷随之增加。同时也对室外环境造成了一定的影响，由于玻璃的反射作用更多的太阳光照射到城市地面，使气温升高。在建筑幕墙的节能研究方面也同样存在着机遇和挑战的问题。在调查的这些建筑里有 36 座大楼采取了对玻璃幕墙进行贴膜、隔热等处理，物业管理人员认为这些措施对改善室内热环境和降低空调负荷起到了一定的作用。

2.2.2 空调系统信息的统计与分析

1．空调系统冷热源形式

上海市商用建筑的空调冷热源形式统计如表 2.1、表 2.2 所示。95 幢商用建筑空调冷源以电驱动式制冷机组（包括离心式、螺杆式、活塞式和空气源热泵）为主，占总数的 84.3%。其中，离心式制冷机组占的比例最大，仅单纯使用离心式制冷机组的建筑就占了 32.6%。空气源热泵也有相当的使用比例，占总建筑数量的 17.9%。另外，吸收式机组（包括各种热源方式）占 6.4%。商用建筑的空调热源则以燃油锅炉为最多，占 23.2%，其次为冬、夏两用的热泵型空调机组，占 18.9%，燃气锅炉占 15.8%，吸收式机组占 7.5%。此外，电锅炉和电加热也有一定的使用，分别占 13.7%和 6.3%。

通过以上的整理分析可以看出，电制冷还是上海市商用建筑夏季最主要的供冷方式，离心式制冷机组因其效率高、性能稳定而得到了广泛的应用；冬季热源仍然是以锅炉为主，各类能源形式的锅炉占到总数的 53.8%。

空气源热泵机组占地面积小、安装维护简便，因此也有相当的应用，但是空气源热泵机组制冷能效比不高，特别是在夏季高温季节有加剧用电高峰的缺点。空气源热泵机组可冬、夏两用，这样在主机的初投资上就有很大的优势，但是热泵机组的能效比不高，运行能耗相对较高，因此适用范围有一定的局限。

吸收式机组的应用取决于建筑的热源形式，当建筑物有比较稳定且代价低

廉的热源时，采用吸收式机组不失为一种节能、环保的能源利用方式。但是吸收式机组存在占地面积大、运行维护成本较高等缺点，限制了它的应用范围。

表 2.1 空调冷源形式汇总

能源形式	空调冷源形式	建筑数量	比例（%）
电	空气源热泵	17	17.9
	离心式制冷机	31	32.6
	螺杆式制冷机	1	1.1
	活塞式制冷机	1	1.1
	多种形式的组合	30	31.6
气	直燃型溴化锂吸收式机组	4	4.2
油	燃油型溴化锂吸收式机组	1	1.1
热网	热网驱动溴化锂吸收式机组	1	1.1
复合能源	电、气、油、热网、煤等能源的综合利用	5	5.3

表 2.2 空调热源形式汇总

能源形式	空调热源形式	建筑数量	比例（%）
电	空气源热泵	18	18.9
	电加热	6	6.3
	电锅炉	13	13.7
气	直燃型溴化锂吸收式机组	3	3.2
	燃气锅炉	15	15.8
油	燃油型溴化锂吸收式机组	3	3.2
	燃油锅炉	22	23.2
煤	燃煤锅炉	1	1.1
热网	热网驱动溴化锂吸收式机组	1	1.1
	热网直接供热	1	1.1
复合能源	电、气、油等能源的综合利用	3	3.2

特别值得一提的是，在所调查的商用建筑里，有 5.3%的建筑空调冷源采用了复合能源，3.2%的建筑空调热源采用了复合能源，这些建筑采取了离心式、热泵、吸收式、蒸汽锅炉等的组合机组方式。从建筑总的能源使用形式来看，有 54%的商用建筑使用了多种能源。这种灵活、多样的能源使用方式可以使建筑物不受单一的能源限制、充分利用各种能源政策和新技术、从容应对负荷变化和突发事件、降低运营成本等，是值得大力宣传和推广的。

另外，经过统计计算，所调查的 95 幢商用建筑的单位面积空调平均装机冷量为 119.97W/㎡，比文献[36]中 1997 年的调查结果降低了 5.5%，这是节能思想应用于空调设计阶段的一个直接体现。

2．空调系统形式

空调系统形式是影响室内舒适度和空调运行能耗的重要因素，95 幢商用建筑所采用的各种空调系统形式如表 2.3 所示：

表 2.3 商用建筑空调系统形式汇总

空调系统形式	建筑数量	比例（%）
风机盘管＋新风系统	31	32.6
定风量全空气系统	5	5.3
变风量全空气系统	2	2.1
风机盘管＋定风量系统	17	17.9
风机盘管＋变风量系统	30	31.6
定风量＋变风量系统	2	2.1
其他系统形式	8	8.4

从表 2.3 可以看出，风机盘管系统是上海市商用建筑采用最多的空调形式，这种系统的设计安装方便灵活、便于单独控制，特别适合于办公场合，但是也存在着室内温、湿度控制精度不高、新风的引入不方便、工作区域小和易发生“水患”等缺点，这在许多实际工程中都有典型的案例，因此应慎重选择、科学使用风机盘管系统。

随着对 VAV 系统的深入研究，变风量全空气系统在实际建筑中的应用也越来越多。但是这种控制精度高、动态追踪负荷、运行能耗低的空调系统要想充分发挥其优点，需要设计人员准确把握空调负荷特性、合理配置系统设备，并且需要有较完善的自控系统的支持，也需要物业管理人员具备一定的专业知识，科学地运行与维护。被调查的这些采用变风量全空气系统的建筑物有很多都存在各个区域冷热不均、新风量不足和实际运行能耗并不低的问题。

3．楼宇自控系统

上海市作为中国最大的国际化大都市，其建筑的智能化水平也是第一流的，拥有许多高档的智能建筑。在调查的商用建筑中，装有楼宇自控系统的商用建筑有 58 座，占总数的 61.1%，这个比例还是很高的。但是，我们也不得不面对这样一个现实：大多数商用建筑的楼宇自控系统都存在着这样或那样的问题，

不能充分发挥其功能；许多系统的软、硬件都比较落后，功能不完善，甚至处于失灵状态；很多楼宇自控系统的功能已被手动操作所代替。造成这一现象的原因是多方面的[50]，如功能完备的楼宇自控系统初投资很大、有些自控产品本身存在一定的质量问题、楼宇自控系统的运行维护费用很高、物业管理人员的专业水平不够等等。这个问题应引起各方面专业人士的高度重视，必须让楼宇自控系统摆脱目前如同“鸡肋”一般的尴尬处境，使这些商用建筑真正成为“智能化大楼”。

2.2.3 商用建筑能耗数据的统计与分析

上海市商用建筑信息数据库要求每幢商用建筑提供近几年全年各月的各种能源的使用情况，目前阶段我们只收集了这 95 幢商用建筑 2004 年全年逐月的能耗数据。图 2.4 为这些商用建筑的单位面积年一次能耗值的分布情况。这里的单位面积年一次能耗值是将全年各种能源的总能耗分别换算为一次能耗（单位为：MJ），加和后除以总的建筑面积得到该商用建筑单位面积全年一次能耗值。其中，1kWh 电能对应的一次能取为 9.47MJ。由图 2.2 中各数据点的分布情况我们可以看出，这些建筑的单位面积年一次能耗值主要集中在 1000～1600 MJ/m^2•y 之间，但由于建筑使用功能、系统运行方式的不同，这些商用建筑的单位面积一次能耗值有较大差异，个别建筑的单位面积能耗值比较高，最高值为 3342.14 MJ/m^2•y，最低值为 449.8 MJ/m^2•y，相差 7.43 倍。而能耗值最高的这座商用建筑宾馆客房的面积占总建筑面积的比例较高，达到 70.5%，如图 2.4 所示。

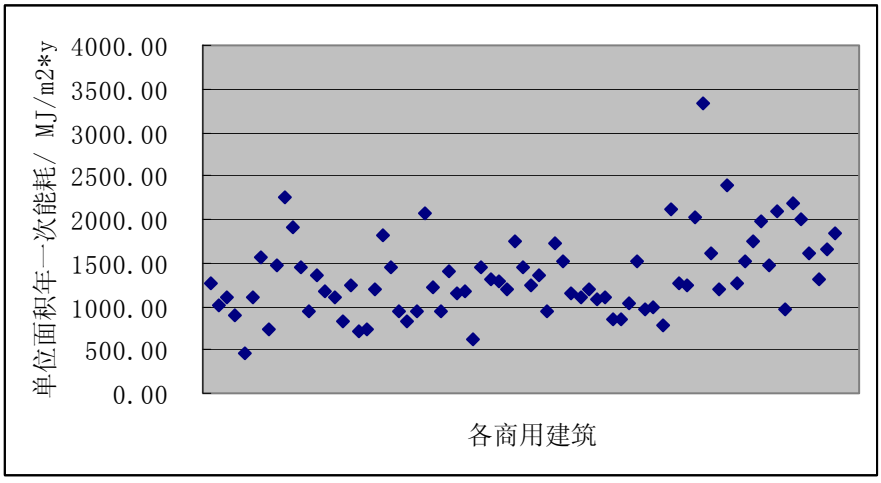

图 2.4 商用建筑单位面积年一次能耗分布情况

通过对这些数据的整理分析，我们得出了商用建筑的年平均一次能耗为 1427.4 MJ/m^2·y，比文献[36]中的调查结果（1800 MJ/m2·y）降低了 20.7%，如不考虑气候变化、调查对象不同等因素，可以认为近几年上海市商用建筑的能源消耗有所下降，人们的节能意识有所增强，新设备、新技术的应用也使建筑能源利用率提高很多。

同时，在调查的商用建筑中有 11 幢建筑无法准确提供各月详细的能源费用帐单，仅有全年总的能源费用。并且我们还了解到，大多数的商用建筑都没有按用途对能耗进行分项计量，而且物业管理部门也没有对各月的能源使用情况进行统计整理和科学、有效的分析。这些情况的存在使得物业管理部门只能被动地进行能源管理，甚至没有能源管理。不收集整理能源数据、也不进行全面的分析和比较就无法找到大楼的用能规律，也无法发现能耗过大的原因甚至出现问题的系统设备，建筑节能更是无从谈起。因此这些商用建筑的物业管理部门有必要建立起规范的能源管理制度，重视能源数据的采集和分析工作，切实有效地推进建筑节能工作的开展。

第 3 章 数据挖掘相关概念介绍

3.1 数据挖掘介绍

3.1.1 数据挖掘的产生背景

我们生活的社会已经处在一个高度发达的网络化时代，通信、计算机和网络技术正改变着整个人类社会。自动化的数据收集工具和成熟的数据库技术导致了大量的数据存储在数据库、数据仓库和其他信息媒介中，从一般的商场到银行、保险、证券等各个领域，数据量急剧增大。随着数据库技术的迅速发展以及数据库管理系统的广泛应用，人们积累的数据只会越来越多。大量的信息在给人们带来方便的同时也带来了一大堆问题[51]：一是信息过量，难以消化；二是信息真假难以辨识；三是信息安全难以保证；四是信息形式不一致，难以统一处理。在这些大量数据背后隐藏着许多重要信息，但由于目前所使用工具的局限性而无法将其挖掘出来，而这些重要信息可以很好地支持人们的决策。目前的数据库系统所能做到的只是对数据库中已有的数据进行存取和查询，人们通过这些数据所获得的信息量仅仅是整个数据库所包含的信息量的一部分，隐藏在这些数据之后的更重要的信息却很难掌握，比如关于这些数据的整体特征的描述及其发展趋势的预测等等，这些信息在决策生成的过程中具有非常重要的参考价值[52]。缺乏挖掘数据的手段，导致了“数据爆炸但知识贫乏”的现象，“人们被数据淹没，人们却饥饿于知识”。面对这一挑战，数据开采和知识发现技术应运而生，产生了数据挖掘技术。

3.1.2 数据挖掘的定义

知识发现是近年来新兴的研究领域，在 1989 年举行的第十一届国际联合人工智能学术会议上，为了强调“知识”是数据驱动发现的最终产物，首次用“数据库中的知识发现”(KDD：Knowledge Discovery in Databases)这个词来命名一种新的知识获取技术。第一个 KDD 实验室于 1989 年 8 月在美国底特律建立。

1996 年，Usama Fayyad 给出了 KDD 的经典定义：一个在数据中识别有效的、新颖的、潜在的、最终可理解的模型的非平凡过程[53]。它从大量的原始数据中挖掘出隐含的、有用的、尚未发现的信息和知识，被认为是目前解决“数据爆炸”和“数据丰富，信息贫乏（Data Rich and Information Poor)”的一种有效方法[54]。数据挖掘（DM：Data Mining）经常是作为 KDD 的同义词使用。但按照 Usama Fayyad 的观点，它们是有区别的，KDD 涉及从数据中发现有用知识的全过程，而数据挖掘则是这个过程中的一个特殊步骤，数据挖掘是指为了实现从数据中提取知识模型而对具体算法的应用。由于在产业界、媒体和数据库研究界，“数据挖掘”比“数据库中的知识发现”这个词更流行，因此现在多采用“数据挖掘”这个术语。

数据挖掘这个术语首先出现在 1989 年举行的第十一届国际联合人工智能学术会议上[55]，1991、1993 和 1994 年又连续举行了数据挖掘专题讨论会。随着参加会议人数的增多，从 1995 年开始每年都有知识发现和数据挖掘国际会议（KDD’95：Knowledge Discovery and Data Mining’95）召开。2002 年 7 月，在加拿大 Alberta 的 Edmonton 举办了第 8 届 KDD 国际会议[53]。另外，从 1997 年开始，数据挖掘也拥有了自己的专门杂志《Knowledge Discovery and Data Mining》

数据挖掘是指从大量数据中寻找潜在信息，如趋势(Trend)、特征(Pattern)及相关性(Relationship)的过程，也就是从数据中发掘信息或知识的过程，该过程也被称为数据考古(Data Archaeology)、数据模式分析(Data Pattern Analysis)或“功能相依分析”(Functional Dependency Analysis)[56]。更广义的说法是：数据挖掘是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[57]。

3.1.3 数据挖掘简介

数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。这个定义包括多层次的含义：数据源必须是真实的、大量的、含噪声的；发现的是用户感兴趣的知识；发现的知识要可接受、可理解、可运用；

发现的知识支持特定的待发现的问题[46]。数据挖掘是一个多学科交叉的新兴研究领域，它把人们对数据的应用从低层次的简单查询提升到从数据中挖掘知识，提供决策支持。在这个新兴领域中，汇集了来自机器学习、模式识别、数据库、统计学、人工智能以及管理信息系统等各学科的成果，多元化的投入使得这一学科得以蓬勃发展，而且已初具规模[58]。

数据挖掘是由机器学习发展而来的，其基本的思想是搜索出有用的描述。在数据库技术飞速发展的同时，人工智能领域的一个分支－－机器学习的研究也取得了很大进展。自 20 世纪 50 年代开始机器学习的研究以来，在不同时期的研究途径和目的也不尽相同，一般可大致分为几个阶段，其研究内容分别为：神经模型和决策理论、概念符号获取及知识加强等。根据人类学习的不同模式人们提出了很多机器学习方法，如：实例学习、观察和发现学习、神经网络和遗传算法等等。其中某些常用且比较成熟的算法已被人们用于实际的应用系统及智能计算机的设计和实现中。数据挖掘就是利用机器学习的方法从数据库中提取有价值知识的过程，是数据库技术和机器学习两个学科的交叉学科。数据库技术侧重于对数据存储处理的高效率方法的研究，而机器学习则侧重于设计新的方法从数据中提取知识。数据挖掘利用数据库技术对数据进行前端处理，而利用机器学习方法从处理后的数据中提取有用的知识[52]。目前，数据挖掘不仅被许多研究人员看作是数据库系统和机器学习方面一个重要的研究课题，而且被许多工商界人士看作是一个能带来巨大回报的重要领域。从数据库中发现出来的知识可以用在信息管理、查询响应、决策支持、过程控制等许多方面。

数据挖掘采用了各种分析方法和技术，包括分类、回归、聚类、关联建模和偏差检测等，数据挖掘技术给我们的能耗分析工作以新的启示：建立有关建筑物能耗的数据库，运用数据挖掘技术找出建筑能耗与各种影响因素的关系，并建立能耗模型，运用此模型可以简便、快捷地对建筑运行能耗进行分析和预测、对能源使用效率进行评定、并进行财务预算、节能措施评估等，并且此模型具有很好的通用性。

3.2 数据挖掘的基本任务和主要算法

3.2.1 数据挖掘的基本任务

数据挖掘的主要任务是对大型数据库中的海量业务数据进行抽取、转换、分析和模型化处理，从中提取辅助决策的关键性数据和隐藏的预测性信息。它能发掘数据间潜在的模式，找出人们可能忽视的信息，以便于理解和观察的形式反映给用户，并给出基于知识的决策分析意见和结论[59]。其任务一般分为描述（Descriptive）和预测（Predictive）两大类。描述性数据挖掘任务是对数据中存在的规则做一种描述，或者根据数据的相似性把数据分组，规划数据库中数据的一般性。预测性数据挖掘任务是在当前数据上推断，根据数据项的值精确预测某种结果，任务所使用的数据都是可以明确知道结果的。描述性数据挖掘任务不能直接用于预测[60]。由于数据挖掘所涉及的学科领域和方法很多，在各学科领域中数据挖掘均负有不同的发现任务，但以下几种发现任务是共同的，也是最重要的。

1．概念描述（Concept Description）[61]

数据库通常存放着大量的细节数据，然而用户通常希望以简洁的描述形式观察汇总的数据集。这种数据描述可以提供一类数据的概貌，或将它与对比类相区别。此外，用户希望方便、灵活地从不同的角度描述数据集。这种描述性数据挖掘称为概念描述。

2．汇总（Summarization）[62]

汇总的目的是对数据进行浓缩，给出它的总体综合描述。数据挖掘主要关心从数据泛化的角度来讨论数据总结，通过对数据的总结，数据挖掘能够将数据库中的有关数据从较低的个体层次抽象总结到较高的总体层次上，从而实现对原始基本数据的总体把握。

3．关联规则（Association Rule）

关联规则是指搜索事务数据库中的所有细节或事务，从中寻找重复出现概率很高的模式或规则。这类数据挖掘技术以大的数据库为对象，其中每个案例（case）都被定义为一系列相关数据项。这种查询要求用关联找出所有能把一组案例或数据项与另一套案例或数据项联系起来的规则。一个典型的例子就是：“在购买面包和黄油的顾客中，有 90%的人同时也买了牛奶”（面包+黄油＝牛

奶)。关联规则发现的思路还可以用于序列模式的发现。用户在购买物品时，除了具有上述关联规律，还有时间或序列上的规律。目前，已经从单一概念层次关联规则的发现发展到多个概念层次的关联规则的发现。在概念层次上的不断深入，使得发现的关联规则所提供的信息越来越具体，实际上这是个逐步深化所发现知识的过程。

4．分类（Classification）和回归（Regression）

分类和回归代表了数据挖掘在当今应用的最大领域，它们同时都属于预测模型，通过创建模型来预测类成员（分类）或值（回归)。分类方法和回归方法最显著的区别在于预测的输出类型。分类方法用于预测变量的类属性，输出的结果是有类别的，被预测变量仅有几种可能的值。回归方法用于预测变量的某一特定的值，这些特定的值有无限多的可能取值，这样的被预测变量通常被称为连续的。事实上，分类方法和回归方法的联系非常紧密，这并不仅仅是因为分类方法和回归方法利用了许多相同的工具，更主要的是稍稍利用一些转换方法就能在它们之间相互转换。

5．聚类(Clustering)

聚类是数据挖掘领域的一个重要的研究课题。所谓聚类是把一组个体按照相似性归成若干类别，即“物以类聚”。由聚类所生成的簇是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其它簇中的对象相异。由于聚类算法不对数据做任何先验统计假设，故在模式识别和人工智能等领域，聚类算法也常常被称之为“无导师学习”或“自组织算法”。聚类问题与分类问题是截然相反的过程，分类是训练例子的分类属性值，而聚类则是在训练例子中找到这个分类属性值。当分析一个较大的、复杂的、连续的且有许多变量的数据库和完全未知的结构时，聚类是一个非常有用的工具。聚类常作为其它算法的预处理步骤，这些算法再在所生成的簇上进行处理[63]。

聚类可以帮助市场分析人员从他们的消费者数据库中区分出不同的消费群体来，并且概括出每一类消费者的消费模式或者说习惯；还可以从保险公司的数据库中发现汽车保险中具有较高索赔概率的群体；也可以用来从万维网上分类不同类型的文档等[64]。

6．偏差检验（Deviation Detection）

偏差检验技术用于抽取数据中的偏差和异常。那些令人关注的偏差包括：不适合于标准类的异常；与父类或兄弟类有很大差别；相邻两时间段内信息的

变动（如某一地区两个季度销售收入的变动）；处于模式边缘的冗余等。偏差分析有助于滤掉知识发现引擎所抽取的无关信息，也滤掉那些不适合的数据[45]。

除了上述介绍的几种基本任务之外还有特征规则、趋势分析、异常分析、模式分析等等[44]。

3.2.2 数据挖掘的主要算法

从不同的角度看，数据挖掘技术有多种分类方法，如根据发现的知识的种类、 挖掘的数据库类型、挖掘方法、挖掘的途径等进行分类。目前，常用的数据挖掘技术方法主要包括以下几种[57]：

1．决策树方法（Decision Tree）

决策树算法的起源是概念学习系统 CLS，然后发展到 ID3 算法而为高潮，最后又演化为能处理连续属性的 C4.5 和 C5.0。有名的决策树算法还有 CART,CHAID 和 Assistant。决策树可以以图形或文本形式的规则来描述和预测数据，但一般只能有一个依赖变量[65]。

决策树算法利用信息论中的互信息（信息增益）寻找数据库中具有最大信息量的字段，建立决策树的一个结点，再根据字段的不同取值建立树的分支。在每个分支集中重复建立树的下层结点和分支的过程，即可建立决策树，从而可从中生成分类规则，并利用规则和决策树生成复杂的知识结构。即用树形结构来表示决策集合，这些决策集合通过对数据集的分类产生规则。在决策树中每一个分支代表一个子类，树的每一层代表一个概念。从发展来看，易使用、易理解已成为一个潮流，一些新的技术导致原有的限制被突破，如 C5.0 利用 boosting 技术把多个决策树合并到一个分类器，而正在研究的 oblique tree 将产生在独立变量间的复合关系来分隔节点。采用决策树，可以将数据规则可视化，其输出结果也容易理解，决策树方法精确度比较高，同时系统也不需要长时间的构造过程。

2．神经网络方法（Neural Network）

神经网络最早是由心理学家和神经生物学家提出的，旨在寻求开发和测试神经的计算模拟。神经网络是一组连接的输入/输出单元，其中每个连接都与一个权相连。在学习阶段，通过训练数据逐步调整神经网络的权，使得能够预测样本的正确类标号来学习。此研究方法包括网络提取规则和灵敏度分析[66]。

神经网络方法是基于生物神经系统的结构和功能而建立起来的模拟人脑神经元方法。它从结构上模仿生物神经网络，以求达到模拟人类的形象直觉思维的目标。它是在生物神经网络研究的基础上，根据生物神经元和神经网络的特点，通过简化、归纳、提炼总结出来的一类并行处理网络。人工神经网络技术利用其非线性映射的思想和并行处理的方法，用神经网络本身结构可以表达输入与输出的关联知识。它通过不断学习、调整网络结构，最后以特定的网络结构来表达输入空间与输出空间的映射关系，是一种通过训练来学习的非线性预测模型。可以完成分类、聚类、特征挖掘等多种数据挖掘任务【59】。

神经网络方法以 MP 模型和 HEBB 学习规则为基础，可以建立三大类神经网络模型：（1）前馈式网络。它以感知机、反向传播模型、函数型网络为代表，可用于预测和模式识别等方面。（2）反馈式网络。它以 Hopfield 的离散模型和连续模型为代表，分别用于联想记忆和优化计算。（3）自组织网络。它以 ART 模型,Koholon 模型为代表，用于聚类。基于神经网络的数据分类通常具有较小的分类误差和对噪声较强的鲁棒性，其应用过程的关键性问题为：网络的构建和训练，根据属性的数目和类型的树目可以确定相应的网络输入输出模式，并形成合适网络结构；网络删减，在不影响分类误差的前提下，去除多余的网络节点和链接，经过删减的网络可以提供简洁和有意义的分类规则；规则提取，即从经过删减的网络中提取分类规则。

3．粗集方法（Rough Set）

粗集理论是近年来才兴起的一种研究不精确、不确定性知识的表达、学习、归纳等的新型数学理论,最初是由波兰数学家 Z.Pawlak 在 1982 年首先提出的。由于最初关于粗糙集理论的研究大部分是用波兰语发表的,因此当时没有引起国际计算机学术界和数学家们的重视,直到 20 世纪 80 年代末才逐步引起各国学者的注意。从 1992 年起,每年都举办粗糙集理论及其应用方面的国际学术会议【67】。

粗集理论模拟人类的抽象逻辑思维，以各种更接近人们对事物的描述方式的定性、定量或者混合信息为输输入、输入空间与输出空间的映射关系是通过简单的决策表简化得到的。它通过考察知识表达中不同属性的重要性，来确定哪些知识是冗余的，哪些知识是有用的。简化知识表达空间是基于不可分辨关系的思想和知识简化的方法来进行的，从数据中抽取推理逻辑规则作为知识系统的模型。它是基于一个机构（或一组机构）关于一些现实的大量数据信息，

以对观察和测量所得数据进行分类的能力为基础，从中发现、推理知识和分辨系统的某些特点、过程、对象等[59]。

在数据库中,将行元素看成对象，列元素看成属性（分为条件属性和决策属性)。等价关系 R 定义为不同对象在某几个（或几个）属性上取值相同，这些满足等价关系的对象组成的集合称为等价关系 R 的等价类。条件属性上的等价类 E 与决策属性上的等价类之间有三种情况：

（1）下近似：Y 包含 E。对下近似建立确定性规则。

（2）上近似：Y 和 E 的交集非空。对上近似建立不确定性规则（含可信度)。

（3）无关：Y 和 E 的交集为空。无关情况不存在规则。

该理论最大的特点是：不需要提供问题所需处理的数据集合之外的任何先验信息,如统计中要求的先验概率和模糊集中要求的隶属度,且算法简单、易于操作。该理论中提出的上下近似、核、约简等概念,为数据分析、决策分析提供了新的理论和方法。近年来,粗糙集理论方法日趋完善,并广泛应用于机器学习、决策分析、过程控制、模式识别与数据挖掘等领域,该理论还在医学、化学、材料学、地理学、管理科学和金融等其他学科取得了成功的应用[67]。

4．遗传算法（Genetic Algorithms）

遗传算法是一种较新的非线性优化技术。它基于生物进化理论中的基因重组、突变和自然选择等概念设计一系列的过程来达到优化的目的。这些过程包括基因组合、交叉、变异和自然选择。遗传算法作用于对某一特定问题的一组可能的解法，试图通过基因组合、交叉、变异过程来组合或“繁殖”现存的最好的解法来产生一个新的解集．然后利用基于“适者生存”的理论的自然选择方法来使较差的解法被抛弃，使繁殖的结果得到改善，从而产生更好的解集[59]。

遗传算法的基本思想是：根据自然选择原理以及自然遗传机制进行数据处理,从中寻找有用的信息。与其他算法比较，遗传算法的主要特点是：（1）规则的表达形式是以编码的形式给出的；（2）遗传算法中的规则不是固定的，而是进行不断的转换和进化，最初的规则是随机给出的，然后根据适者生存的原则，由最初的规则中选出最适用的规则，同时由最适用的规则产生它们的后代，这样不断地对规则群体进行进化，直到满足给定的条件为止；（3）遗传算法的处理过程中不需要其他任何辅助信息及附加的先决条件[67]。

遗传算法是模拟生物进化过程的一种算法，它由三个基本算子组成：（1）繁殖(选择)。它是从 1 个旧种群（父代）选出生命力强的个体，产生新种群（后

代） 的过程。(2）交叉(重组)。选择两个不同个体（染色体）的部分（基因）进行交换，形成新个体。(3）变异（突变)。对某些个体的某些基因进行变异（1 变 0,0 变 1)。这种遗传算法可以起到产生优良后代的作用。这些后代需满足适应度值,经过基于若干代的遗传，将得到满足要求的后代（问题的解)。遗传算法已在优化计算和分类机器学习方面显示了明显的优势。

为了应用遗传算法，我们需要把数据挖掘任务表达为一种搜索问题而发挥遗传算法的优化搜索能力。

5．聚类算法（Clustering）

聚类是一种重要的人类行为。早在孩提时代，一个人就通过不断地改进下意识中的聚类模式来学会如何区分猫和狗，或者动物和植物。通过聚类，人能够识别密集的和稀疏的区域，因而发现全局的分布模式，以及数据属性之间的有趣的相互关系[66]。二十世纪六十年代以来，聚类及其应用方面的论文大量地出现在模式识别等领域的文献中。聚类分析源于许多研究领域，统计学、生物学，以及机器学习等研究领域。聚类根据某种相似程度的度量，将数据对象分组成为多个类或簇（cluster)，在同一个簇中的对象之间具有较高的相似度，而不同簇中的对象差别较大。相异度是根据描述对象的属性来计算的，距离是经常采用的度量方式。

聚类是一门多元统计分类方法，它可避免传统分类法的主观性和任意性。在统计方法中，聚类称聚类分析，它是多元数据分析的三大方法之一（其它两种是回归分析和判别分析)。它主要研究基于几何距离的聚类，如欧式距离、明考斯基距离等。传统的统计聚类分析方法包括系统聚类法、分解法、加入法、动态聚类法、有序样品聚类、有重叠聚类和模糊聚类等。这种聚类方法是一种基于全局比较的聚类，它需要考察所有的个体才能决定类的划分，因此它要求所有的数据必须预先给定，而不能动态增加新的数据对象。聚类分析方法不具有线性的计算复杂度，难以适用于数据库非常大的情况[64]。

聚类算法是通过对变量的比较，把具有相似特征的数据归于一类。通过聚类以后，数据集就转化为类集，在类集中同一类中数据具有相似的变量值，不同类之间数据的变量值不具有相似性。区分不同的类是属于数据挖掘过程的一部分。应注意这些类不是事先定义好的，而是通过聚类算法采用全自动方式获得的。通常，聚类过程是数据挖掘过程的第一个阶段。它首先把数据区分于不同的类，以便于做进一步的分析。聚类法大至上可分为两种类型[59]：

（1）分层聚类（Hierarchical Clustering）。分层聚类是基于数学的标准，对数据进行细分或聚合，适用于数值数据。

（2）概念聚类（Conceptual Clustering）。概念聚类是基于数据的非数值属性，对数据进行细分或聚合，适用于非数值数据。

聚类分析已经广泛地用在许多应用中，包括模式识别、数据分析、图象处理、以及市场研究等。作为一个数据挖掘的功能，聚类分析能作为一个独立的工具来获得数据分布的情况，观察每一个簇的特点，集中对特定的某些簇做进一步的分析。

6. 模糊集算法（Fuzzy Set）

利用模糊集合理论对实际问题进行模糊判断、模糊决策、模糊模式识别、模糊簇聚分析等。系统的复杂性越高，精确能力就越低，模糊性就越强。

7. 规则归纳（Rule Induction）

规则归纳法是由一系列的 if... then... else... 类产生式规则来对数据进行归类。它通过统计方法，从海量数据中归纳、提取出有价值的 if... then... else... 产生式规则。由于这种方法所归纳的产生式规则可以直接应用于专家系统中，因而规则归纳技术在数据挖掘中得到了广泛使用，例如：关联规则的挖掘[59]。

8. 机器发现（Machine Discovery）

随着人工智能技术的发展，机器发现技术得到了长足的发展，即在工程和科学数据库中对若干数据项进行一定的数学运算，可以求得相应的数学公式。比较典型的系统有：科学定律发现系统 BACON 、数学概念发现系统 AM 等，它们都产生了巨大的影响[67]。

9. 可视化方法(Visualization Approach)

可视化技术基于“一幅图画胜过千言万语”这一事实，利用空间和非空间的属性（如：大小、颜色等），采用直观的图形方式将信息模式、数据关联或趋势等呈现给用户[59]。用户可以通过可视化技术交互地观察数据、分析数据关系，进而在一个相当高的层次上找出数据间可能的关系。可视化数据分析技术拓宽了传统的图表功能，使用户对数据的剖析更加清晰，它可以用于识别那些通过挖掘可能值得进一步观察的数据段[67]。例如，把数据库中的多维数据变成多种图形，这对充分揭示数据的内涵、内在本质及规律起了很大的作用。

3.3 数据挖掘的过程

知识发现是一个交互式、循环反复的整体过程，可以用一个等式非常形象地表示：KDD＝数据预处理＋DM（数据挖掘）＋解释评价，其中数据挖掘是其最核心部分，也是研究与实现的难点，正因为这个原因常常将数据挖掘等同于知识发现。具体来说，一个完整的数据挖掘过程包括以下几个步骤[67]：

3.3.1 定义问题

数据挖掘的过程就好像是从矿山中采矿或淘金一样，采矿必须首先确定金矿所在。同样地，从实际应用的角度出发，整个数据挖掘过程都必须建立在对挖掘对象（即所研究领域的大量数据）的深刻了解之上，对象不同所采用的挖掘技术也不同。因此，在数据挖掘之前就应熟悉相关对象的背景知识，明确数据挖掘的目的和任务，对目标进行可行性分析（是可操作的），证实数据可以满足实际使用需要。同时了解数据挖掘相关领域的情况，从而将数据挖掘技术和专业知识有机地结合在一起，对挖掘对象的了解将贯穿整个数据挖掘过程。这一步骤类似需求分析。

3.3.2 数据准备

数据准备是数据挖掘的第一个阶段，也是非常重要的一个阶段。数据准备的好坏将影响到数据挖掘的效率和准确度以及最终挖掘模式的有效性。数据准备阶段的工作包括以下四个方面的内容[46]：

1．数据的净化

数据净化就是根据用户要求，利用一些数据库操作对数据进行处理，从数据库中提取出需要挖掘的数据集合。清除数据源中不正确、不完整或其他方面不能达到数据挖掘质量要求的数据，对其中的噪声数据进行处理。例如推导计算缺值数据、消除重复记录等。进行数据净化可以提高数据的质量，从而得到更正确的数据挖掘结果。

2．数据的集成

数据的集成是在数据挖掘所应用的数据来自多个数据源的情况下，将多文件或多数据库运行环境中的数据进行合并处理，将数据进行统一的存储，解决

语义模糊性，处理数据中的遗漏和清洗脏数据等，并消除其中的不一致性。

3．数据的应用变换

数据的应用变换就是为了使数据适用于计算的需要而进行的一种数据转换，如进行离散值数据与连续值数据之间的相互转换、数据值的分组分类、数据项之间的计算组合等操作。这种变换可能是现有数据不满足分析需求而进行的，也可能是所应用的具体数据挖掘算法对数据提出的要求。

4．数据的精简

数据精简是采用一定的方法对数据的数量进行缩减，或从初始特征中找出真正有用的特征来消减数据的维数，从而提高数据挖掘算法的效率与质量。

3.3.3 数据挖掘

数据挖掘阶段首先要确定挖掘的任务或目的是什么，如数据总结、分类、聚类、关联规则发现或序列模式发现等。确定了挖掘任务后，就要决定使用什么样的挖掘算法（包括选取合适的模型和参数）。同样的任务可以用不同的算法来实现，一般要考虑多方面的因素来确定具体的挖掘算法。例如：不同的数据有不同的特点，因此需要用与之相关的算法来挖掘；用户对数据挖掘有着不同的要求，有的用户可能希望获取描述型的、容易理解的知识，而有的用户或系统的目的是获取预测准确度尽可能高的预测型知识。最后将挖掘的结果用一定的方法表达成某种易于理解的形式。此过程往往需要反复替换挖掘算法。

3.3.4 模型解释与评价

数据挖掘阶段发现出来的模式需要进行专业的解释和分析，根据最终用户的决策目的对提取的知识进行分析，把最有价值的信息区分出来，提交给用户。经过用户或机器的评估，可能存在冗余或无关的模式，这时需要将其剔除。也有可能模式不满足用户要求，这时则需要返回前面处理中的某些步骤以反复提取，如重新选取数据、采用新的数据处理方法、设定新的数据挖掘参数值，甚至换一种挖掘算法等。

另外，数据库知识发现由于最终是面向用户的，数据挖掘的结果并不能直接被理解，有时还要对数据挖掘得到的结果分析其原因，这就需要应用一定的专业知识。值得注意的是，有时数据挖掘得来的规则只是表示某些数据的关系，

并不能代表某种具体的因果关系，所以需要根据实际情况分析原因。因此可能要对发现的模式进行可视化，或者把结果转换为用户易懂的表示[65]。

数据挖掘的完整流程如图 3.1 所示：

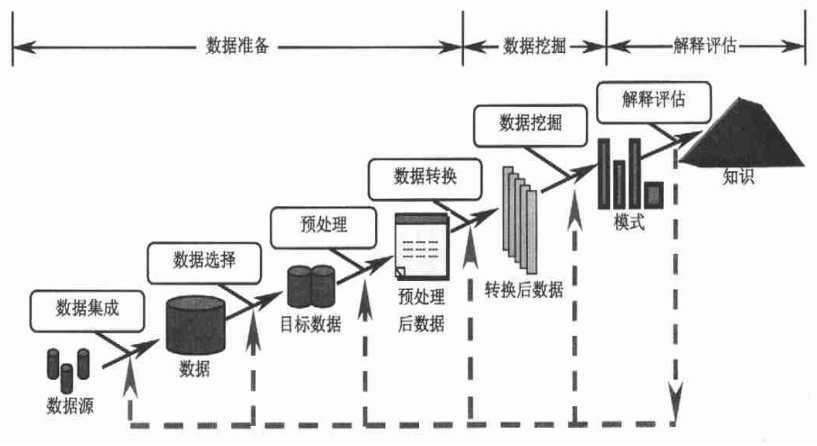

图 3.1 数据挖掘流程图

数据挖掘是一个复杂的处理过程，简单来说主要有以下几方面原因[52]：

（1）数据挖掘仅仅是整个知识发现过程中的一个步骤。数据挖掘质量的好坏有两个影响要素：一是所采用的数据挖掘技术的有效性，二是用于开采的数据的质量和数量（数据量的大小）。如果选择了错误的数据或不适当的属性，或对数据进行了不适当的转换，则挖掘的结果不会好。

（2）整个挖掘过程是一个不断反馈的过程。比如，用户在挖掘过程中发现选择的数据不太好，或使用的挖掘技术产生不了期望的结果。这时，用户需要重复先前的过程，甚至从头重新开始。

（3）可视化在数据挖掘的各个阶段都扮演着重要的作用。特别是在数据准备阶段，用户可能要使用散点图、直方图等统计可视化技术来显示有关数据，以期对数据有一个初步的理解，从而为更好地选取数据打下基础。在挖掘阶段，用户则要使用与领域问题有关的可视化工具。在表示结果阶段，则也会用到可视化技术。

3.4 数据挖掘与相关研究领域的关系[52]

数据挖掘是近些年来兴起的一门多学科综合的新技术，它能将大量数据中隐藏的复杂的规律揭示出来。数据挖掘是一个多学科相关技术融合生长的产物，

与数据挖掘技术有关的研究领域有很多，例如：机器学习、统计学、数据仓库、模式识别、机器发现、人工智能、知识获取、神经网络、数据可视化、定性推理、智能数据分析等等。数据挖掘来源于这些领域，并与它们交汇融合，但又有着显著的差别。

3.4.1 数据挖掘与机器学习

机器学习这门学科所关注的问题是：计算机程序如何随着经验积累自动提高性能。近年来，机器学习被成功地应用于很多领域，同时这个学科的基础理论和算法也有了重大的进展。机器学习算法在很多应用领域被证明有很大的实用价值，它在以下方面特别有用：（1）数据挖掘问题，即从大量资料中发现可能包含在其中的有价值的规律。（2）在某些困难的领域中，人们可能还不具有开发出高效的算法所需的知识。（3）计算机程序需动态地适应变化的领域[68]。

机器学习和数据挖掘都是研究如何从数据中抽取模式和模型的理论和算法。数据挖掘中的不少方法就源于机器学习。从算法角度看，数据挖掘（也就是前文所说的知识发现，KDD）与机器学习相比，数据挖掘主要研究真实世界中的大规模数据上的学习或发现算法。从系统模型角度看，数据挖掘是关于整个发现过程的，不仅仅是数据挖掘算法，还强调数据的准备和发现结果的解释和评估以及人机交互等等方面。数据挖掘和机器学习主要有如下区别[52]：

（1）数据挖掘是从现实世界中存在的一些具体数据中提取知识，这些数据在数据挖掘出现之前早已存在，而机器学习所使用的数据是专门为机器学习而特别准备的数据，这些数据在现实世界中也许毫无意义；

（2）由于数据挖掘使用的数据来自于实际的数据库，所要处理的数据量可能很大，数据的完整性、一致性和正确性都很难保证，因此数据挖掘中的学习算法的效率、可扩充性和数据处理就显的尤为重要；

（3）数据挖掘可以利用目前数据库技术所取得的研究成果来加快学习过程，提高学习的效率。

事实上，除了机器学习，数据挖掘和其他从数据中抽取模式或导出模型的方法之间的区别都比较类似，如模式识别和神经网络等。

3.4.2 数据挖掘与人工智能[69]

人工智能技术包括推理技术、搜索技术、知识表示与知识库技术、归纳技术、联想技术、分类技术、聚类技术等等，其中最基本的三种技术即知识表示、推理和搜索都在数据挖掘中得到了体现。

（1）知识表示

知识表示是指在计算机中对知识的一种描述，是一种计算机可以接受的用于描述知识的数据结构。由于目前对人类知识的结构及机制还没有完全搞清楚，因此关于知识表示的理论及规范尚未建立起来。尽管如此，人们在对智能技术系统的研究及建立过程中还是结合具体研究提出了一些知识表示方法：符号表示法和连接机制表示法。

（2）推理技术

推理技术从已知的事实出发，运用已掌握的知识找出其中蕴含的事实，或归纳出新的事实。推理可分为经典推理和非经典推理，前者包括自然演绎推理、归纳演绎推理、与/或形演绎推理等。后者主要包括多值逻辑推理、模态逻辑推理、非单调推理等。

一般而言，数据挖掘在处理过程中其基本思想是非经典的，而其依据的“剪枝”规则应该是经过经典推理严格证实的——有其严格的数学背景。比如，聚类处理时的基本思想是基于非经典推理，但为了提高效率而采取的“剪枝”技术必须保证完备性、正确性，经得起推理，否则便成了随意剪枝和删除信息，虽然提高了效率，但其正确性不能保证，就没有什么意义了。

（3）搜索技术

搜索是根据问题的实际情况不断寻找可利用的知识，从而构造一条代价较小的推理路线。搜索分为盲目搜索和启发式搜索，盲目搜索是按预定的控制策略进行搜索，在搜索过程中获得的中间信息不用来改进控制策略。启发式搜索是在搜索过程中加入与问题有关的启发性信息，用于指导搜索朝着最有希望的方向前进，加速问题的求解过程并找到最优解。

搜索机制在数据挖掘中得到了最详尽的体现。例如，在属性约简中，如果我们发现某一列属性的取值完全一样或区分能力不大，则可以提前删去；另外，在挖掘关联规则时，如果发现频繁 K 项集的任一（K-1）项候选集不存在，则终止搜索剩余的（K-1）项候选集，这就可以判断“频繁 K 项集是不存在的”等等。

搜索机制提高了数据挖掘的效率，这对解决人工智能中的 NP 难问题是一个积极的探索。

3.4.3 数据仓库与数据库

数据库为 KDD 提供数据操纵管理的理论和技术基础。这方面目前与 KDD 最相关的当属盛行的数据仓库技术，建立数据仓库的目的就是为了面向决策支持和联机分析。为了充分适应 KDD 技术的要求，数据库应该在更多的地方(不应仅仅是数据仓库技术)进行研究，这方面的工作还处于起步阶段。

若将数据仓库比喻作矿坑，数据挖掘就是深入矿坑采矿的工作。毕竟数据挖掘不是一种无中生有的魔术，也不是点石成金的炼金术，若没有足够丰富、完整的资料是很难期待数据挖掘能挖掘出什么有意义的信息的。从数据仓库挖掘有用的资料是数据挖掘的研究重点。换句话说，数据仓库应先行建立完成，数据挖掘才能有效率地进行，因为数据仓库本身所含资料是“干净”(不会有错误的资料参杂其中)、完整的，而且是经过整合的。因此两者的关系可说“数据挖掘是从巨大数据仓库找出有用信息的一种过程与技术”。为了更加清楚地了解数据仓库，这里给出数据仓库与传统数据库的比较：

表 3.1 数据仓库与数据库的比较

特性	数据仓库	传统数据库
主要目的	信息获取与分析	记录交易数据
数据模型	星型结构	表格
数据形式	分析性数据	交易性数据
数据储存状况	历史性、描述性数据	经常改变、及时性的数据
数据的时效性	经过处理的历史数据	当时的运算数据
数据库的结构设计	星型结构	个体－关系模式配合正规化
数据特性	大量重复储存,并预先加总	无重复储存
查询的范围	相当宽度	较狭窄
所包含的数据量	按 Gigabytes 算	按 Megabytes 算
内含数据的错误率	极少错误与缺失	可以允许错误与缺失
主题性	根据主题导向	根据功能导向区分数据库
暂存性	完整保留所有的历史数据	只保留目前最新的数据
适合建设的系统	多维数据库管理系统	关联式数据库管理系统

3.4.4 数据挖掘与统计学

1．数据挖掘与统计学的联系和相似点

统计学是“数据科学”，即收集、分析、展示以及解释数据的科学。数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。计算机技术、统计方法、各类算法的结合推动了数据挖掘技术的快速发展。

统计学和数据挖掘有着共同的目标：发现数据中的结构或模式。数据挖掘强调对大量观测到的数据的处理，它是涉及数据库管理、人工智能、机器学习、模式识别、以及数据可视化等学科的交叉学科。用统计的观点，它可以看成是通过计算机对大量的复杂数据集的自动探索性分析[62]。

统计学和数据挖掘有很多研究的共同之处。“数据挖掘”这个术语是 20 世纪 60 年代由于引入计算机进行数据分析而在统计学界传播开来的。数据分析是应用统计学的主要任务，特别是探索性数据分析，现在又出现了智能数据分析的概念，都是和数据挖掘密切相关的。统计方法既可在数据挖掘中的开采阶段提供建模手段，还可在数据预处理和数据转换阶段用于消除噪声和消减维数。

从概括数据、发现结构、建立模型、提取知识的角度看，统计学和数据挖掘有许多相似之处，表现在引导数据挖掘工具发现数据中潜在信息的基本算法常直接来自统计或来自统计中用到的相同的基本技术[52]。

2．数据挖掘与统计学的区别

尽管统计学是数据挖掘中的一个非常重要的手段和途径，与数据挖掘密不可分，但数据挖掘与统计分析是不同的，不能认为数据挖掘是统计学的分支。

数据挖掘同统计学的差异是，首先二者的应用对象不同，数据挖掘是面向最终用户，且用户无需掌握大量的统计学知识，而统计学则是面向统计学家。其次，数据挖掘所涉及到的问题及采用的方法有其自身的特点，表现在：数据挖掘所涉及到的数据集合远远大于统计分析牵涉到的数据对象，是非常大的数据库，达到 GB 甚至是 TB 数量级，这样的数据通常称为“海量数据”(huge data)，其计算工作量也是十分庞大的，而不是一般意义上的统计分析；数据挖掘与统计学建模的侧重点不同，数据挖掘的重点大多放在模型的学习上，分析的任务是找出特征、规律、联系，而不是验证，同时考虑模型的复杂性和需要的计算量，而较少放在对样本所属的总体的渐进推论上；数据挖掘必须多种技术相结

合，而不只是统计分析。

统计学有着较完善的理论基础和很强的数学背景：在采用一个方法之前先要证明，而不是像计算机科学和机器学习那样注重经验。数据挖掘的一个特定属性就是要处理的是一个大数据集。在统计学中由于可行性的原因，我们常常得到的只是一个或一些样本，但是需要描述样本取自的那个总体。而数据挖掘问题常常可以得到数据总体，例如关于一个公司的所有职工数据，数据库中的所有客户资料，去年的所有业务等。

统计学在对数据进行分析时，首先要建立统计模型，模型的好坏直接影响统计推断结果。相对于统计学而言，准则在数据挖掘中起着更为核心的作用，数据集的规模常常意味着传统的统计学准则不适合数据挖掘问题，不得不重新设计。部分地，当数据点被逐一应用以更新估计量，适应性和连续性的准则常常是必须的。尽管一些统计学的准则已经得到发展，但更多应用的是机器学习。

很多情况下，数据挖掘的本质是很偶然的发现非预期但很有价值的信息。这说明数据挖掘过程本质上是实验性的，只有那些可以依据过去经验形成的合理的解释的结构才会是有价值的。数据挖掘本质上假想数据已经被搜集好，关注的只是如何发现其中的秘密。统计学主要关注的是分析定量数据，数据挖掘的多来源意味着还需要处理其它形式的数据。统计学很少关注实时分析，然而数据挖掘问题常常需要这些[62]。

数据挖掘在许多情况下不能适用于古典统计学的框架，并有与统计学不相关的情况。即使数据挖掘采用了分类和回归这样典型的统计方法，它仍然具有某些特有的特性，表现在以下三个方面[52]：

（1）模型的复杂程度：统计学模型特别适用于一些相对简单的，且重要变量预知的模型。对于大型的数据集合，类与特性变量之间的关系是复杂的且难以理解的，对于此类问题数据挖掘较之统计学方法更为合适。例如神经网络和基于规则的分类器有能力对复杂的数据关系进行建模。

（2）大量的离散变量：统计学中的大多数多变量分析方法只适合于连续变量，但许多数据挖掘方法适合于离散变量，对于连续变量需将其离散化之后才能进行处理。

（3）交叉验证方法的广泛使用：数据挖掘方法在建立模型时通常将原始数据分成两部分，即训练集和验证集，通过对训练集的学习建立初始模型，验证集则用于对初始模型进行校正，防止出现拟合过度。

3.5 数据挖掘的应用

从已经出现的数据挖掘原型系统和应用系统来看，应用数据挖掘技术的领域都是信息丰富、环境多变、尚无模型、需要知识帮助进行管理和决策的领域[58]。国外，在大型商业、金融业、保险业、民航等大型企业都开始得到应用，国内目前总体上处于理论探讨、应用试验阶段。

使用数据库的有各个种类的部门，如政府、科研和商业或企业。各个部门在数据挖掘应用上既有相同之处，又有各自不同的独特地方。这里主要从科研和商业应用来总结数据挖掘的应用，因为它们分别代表了相当不同的应用领域，如商业中最主要和普遍的应用是分类预测。

（1）科研应用

从科学研究方法学的角度看，科学研究可分为三类：理论科学、实验科学和计算科学。计算科学是现代科学的一个重要标志。计算科学工作者主要和数据打交道，每天要分析各种大量的实验或观测数据。随着先进的科学数据收集工具的使用，如观测卫星、遥感器、DNA 分子技术等，数据量非常大，传统的数据分析工具无能为力，因此必须有强大的智能型自动数据分析工具才行。

在科学应用上一个非常有名的系统是加州理工学院喷气推进实验室与天文科学家合作开发的用于帮助天文学家发现遥远的类星体的一个工具 SKICAT。利用 SKICAT，天文学家已发现了 16 个新的极其遥远的类星体。SKICAT 使用了决策树方法构造分类器，结果使得能分辨的星体较以前的方法在亮度上要低一个数量级之多，而且新的方法比以往方法的效率要高 40 倍以上。

（2）企业市场营销[59]

在企业市场营销领域中，数据挖掘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。数据挖掘在行销业中的应用可分为两类：数据库行销(database marketing)和菜篮分析(basket analysis)。前者的任务是通过交互式查询、数据分割和模型预测等方法来选择

潜在的顾客以便向它们推销产品；后者的任务是分析市场销售数据（如 POS 数据库）以识别顾客的购买行为模式，从而帮助确定商店货架的布局排放以促销某些商品。

（3）风险评估领域

保险是一项风险业务，也是风险评估技术的最重要的应用，保险公司的一个重要工作就是对不同风险领域进行鉴定和分析，即风险评估。风险评估对保险公司的正常运作起着至关重要的作用，保费和保单的设计都需要比较详细的风险分析。通过数据挖掘技术，可以从过去的保单及其索赔信息出发，利用决策树的方法，寻找保单中风险较大的领域，从而得出一些实用的风险规则，对保险公司的工作起到指导作用，帮助保险公司规避风险。

（4）金融投资

典型的金融分析领域有投资评估和股票交易市场预测，分析方法一般采用模型预测法（如神经网络或统计回归技术）。这方面的系统有 Fidelity Stock Selector，LBS Capital Management。前者的任务是使用神经网络模型选择投资，后者则使用了专家系统、神经网络和基因算法技术辅助管理多达 6 亿美元的有价证券。

（5）欺诈甄别

银行或商业中经常发生诈骗行为，如恶性透支等。这方面应用非常成功的系统有：FALCON 系统和 FAIS 系统。FALCON 是 HNC 公司开发的信用卡欺诈估测系统，它已被相当数量的零售银行用于探测可疑的信用卡交易。FAIS 是一个用于识别与洗钱有关的金融交易的系统，它使用的是一般的政府数据表单。

（6）生物工程

数据挖掘在生物学上的应用主要集中于分子生物学特别是基因工程的研究上。在著名的人类基因组计划中，通过用计算生物分子系列分析方法，尤其是基因数据库搜索技术已在基因研究上做出了很多重大发现，科学家们得以迅速的绘制出人类染色体基因图。目前，数据发掘技术正在用于对基因图进行解释从而发现各种蛋白质（有 10，000 多种不同功能的蛋白质）和 RNA 分子的结构和功能。

（7）工业制造领域与自动控制系统

随着现代技术越来越多地应用于产品制造业，制造业已不是人们想象中的手工劳动，而是集成了多种先进科技的流水作业。在产品的生产制造过程中常

常伴随有大量的数据，如产品的各种加工条件或控制参数（如时间、温度等控制参数）。这些数据反映了每个生产环节的状态，不仅为生产的顺利进行提供了保证，而且通过这些数据可以得到产品质量与这些参数之间的关系。这样通过数据挖掘对这些数据进行分析，可以对改进产品质量提出针对性很强的建议，而且有可能提出新的更高效节约的控制模式，从而为制造厂家带来极大的回报。这方面的系统有波音公司正在研制的 CASSIOPEE，主要用于诊断和预测在波音飞机的制造过程中可能出现的问题。

（8）体育界

众所周知，平时训练效果的好坏、赛场上的排兵布阵，决定了最后的比赛成绩。但运动员之间的个体差异是非常大的，如何根据运动员的实际情况，设计一套有较强针对性的训练方案和比赛策略则一直是体育界所面临的重要问题。在美国，IBM Watson 研究中心的研究员们对 NBA 的比赛数据进行了数据挖掘分析，取得了一定的成果，已经有不少教练将之运用到日常训练和临场指挥中。

3.6 数据挖掘工具

在数据挖掘技术日益发展的同时，出现了许多数据挖掘工具，如何选择满足需要的数据挖掘工具已成为一个问题。具体的评价标准应从以下几方面考虑：可产生的模式种类的数量、解决复杂问题的能力、多种模式和算法、验证方法、可视化、数据的选择和转换、可扩展性、可操作性、数据存取能力、与其他产品的接口等等[70]。

3.6.1 各种数据挖掘软件的介绍

目前，世界上已经有很多商业公司和研究机构开发出了各自的数据挖掘产品，而且功能和使用简易性也在日益提高。直接采用商业数据挖掘工具来帮助项目实施，是一个很好的选择。它既节省了大量的开发费用，又可以节约维护和升级的开销[71]。比较常用的数据挖掘软件有：

（1）SAS Enterprise Miner

SAS Enterprise Miner 是数据挖掘市场上非常杰出的工具，它利用 SAS 统

计模块的优势，同时增加一系列的数据挖掘算法，使它成为无论是初学者还是专业使用者都比较喜欢使用的数据挖掘工具之一，因而成为数据挖掘市场上的领导者，这个软件比较适用于企业发展数据挖掘及 CRM 决策支持。

（2）SPSS Clementine

Clementine 是 SPSS 发行的数据挖掘工具，它结合了多种图形接口分析技术，使用起来相当容易与直观，SPSS 公司也借此大幅提升了市场竞争力。

（3）IBM Intelligent Miner

IBM 的 Intelligent Miner 是数据挖掘领导地位强有力的竞争者，因为该工具不仅包含了最广泛的数据挖掘技术与算法，而且容纳数据的能力很大且计算能力强大，同时它包含丰富的 APIs 可供客户自定义数据挖掘应用软件，它通过精密的可视化技术及强大的基于 Java 的接口增加其可用性，支持 DB2 关系型数据库。整体上看来，Intelligent Miner（for Data）是市场上容量最大且功能强大的工具。

文献[71]对这三种数据挖掘工具进行了全面的评价，在数据存取、数据处理、模型算法、自动建模、可视化技术等各个方面进行了比较，并给出了评估得分，得分结果如表 3.2 所示：

表 3.2 三种数据挖掘软件的得分比较

功能和特征	总分			
	特征权值	软件		
		IBM Intelligent Miner	SAS Enterprise Miner	SPSS Clementine
数据存储	10%	75	90	80
数据处理	20%	93	100	98
模型算法	30%	91	96	91
自动建模	10%	92	100	86
可视化	15%	88	95	91
其它	15%	78	92	56
总分	100%	88	96	86

除以上三者，国际市场上众多公司针对不同的市场有很多软件公司推出了自己的数据挖掘产品，如统计软件公司推出的数据挖掘软件还包括 StatSoft 公司基于 STATISTICA 的 DataMiner 模块、S-PLUS 及 MATLAB、Mathematica 的数据挖掘模块等，另外像 SGI Mind Set、DataMind DataCruncher、INSPECT 等都

是比较知名的专用数据挖掘软件[56]。

3.6.2 SAS 软件介绍[72]

SAS（Statistical Analysis System）软件是数据处理和统计领域的国际标准软件，世界领先的数据分析和信息系统，本课题的研究就是采用了 SAS 软件的数据挖掘功能。

SAS 系统最早由美国北卡罗来纳大学的两位生物统计学研究生编制，并于 1976 年成立了 SAS 软件研究所——赛仕软件研究所（SAS Institute Inc.），正式推出了 SAS 软件。SAS 系统具有完备的数据访问、管理、分析、呈现及应用开发等功能，SAS 发展到今天已经成为一个由三十多个专用模块组成的大型集成式软件包，SAS 系统已被成功应用于 120 多个国家和地区的 31000 多个机构中，直接用户超过 3500000 人。30 年来，赛仕软件研究所一直致力于为金融、电信、交通、制造、政府以及科研教育等部门提供集成化的信息交付（Information Delivery）、数据仓库（Data Warehouse）和决策支持（Decision Support System）软件解决方案。作为全球十大独立软件开发厂商之一，赛仕在近 50 个国家和地区设有子公司或分支机构。

SAS 系统的功能特点是：模块式结构、把数据管理和数据分析溶为一体。它是一个综合的应用系统，它为计算机应用的数据访问（在分散的数据间建立联系）、数据管理（将数据置于可用状态）、数据分析（将数据转换成有用的信息）和数据呈现（按恰当方式表现所需的信息）这四大数据驱动任务提供了丰富的功能。

SAS 系统提供了 30 多个模块，各个模块之间既相互独立又相互交融补充，覆盖了信息处理和信息系统开发的各个环节，适当地组合 SAS 系统的模块，可用于：数据输入、数据检索、数据管理、数据分析、图形显示、图形分析、报表生成、统计计算、工程计算、质量控制、市场研究、调查分析、建立预测模型、管理信息系统、行政信息系统等方面。

SAS 软件一直被誉为数据处理和统计分析领域的标准软件，广泛应用于很多行业、不同领域中，发挥着重要的作用。尤其是其创业产品—统计分析系统部分，由于具有强大的数据分析能力而得到广泛应用。本文试图将 SAS 软件的应用引入到商用建筑能耗的分析与预测中，将 SAS 软件的功能和专业领域的研

究相结合，充分发挥 SAS 软件强大的数据处理和统计分析的能力，为数据挖掘技术在本专业领域内的应用提供参考。

第 4 章 数据挖掘过程

4.1 数据挖掘方法论

4.1.1 SAS/EM 简介

如前文所述，SAS 系统是一个模块化的软件系统，其数据挖掘功能由系统中的 Enterprise Miner 模块（SAS/EM）来完成[73]。

一个数据挖掘工程需要足够强的软件来完成分析工作，为了计划、实现和成功建立一个数据挖掘工程，需要一个集成了所有分析阶段的软件解决方案，包括从数据抽样到分析和建模，最后公布结果信息。大部分专业统计数据分析软件只能实现特定的数据挖掘技术，而 SAS Enterprise Miner 是一个集成的数据挖掘系统，允许使用和比较不同的技术，同时还集成了复杂的数据库管理软件。

SAS Enterprise Miner 是 SAS 中功能非常强大的数据挖掘环境，它集成了数据获取工具、数据抽样工具、数据筛选工具、数据转换工具、数据挖掘数据库、数据挖掘过程以及数据挖掘评价等多种工具，并且图形化的模块可以使整个挖掘过程组成一个处理的流程图【74】。

在 SAS Enterprise Miner 中，所有的分析工具都是以节点的形式出现的。在工具面板中除了数据挖掘工具外还提供了许多实用节点，这些实用节点使用户可以执行 SAS 程序说明、创建数据挖掘数据库、执行分组处理、创建子图等操作。用户可以根据分析的需要引入相应的节点，然后对该节点中的有关选项和参数进行设置，所用到的节点连接起来构成处理流程图。处理流程图是由一些节点按一定顺序连接起来的，它形象地展示了用户进行数据挖掘的整个过程。因此，在 SAS/EM 中进行数据挖掘的过程也就是创建处理流程图的过程【75】。

4.1.2 SEMMA 方法论

数据挖掘应用是一个解决实际问题的过程，在进行数据挖掘时，不仅要有

技术支持，还需要有先进的数据挖掘方法论的支持。为此，人们提出了一些数据挖掘过程的参考模型或标准，如 SPSS 提出的 5A，SAS 提出的 SEMMA，数据挖掘特别兴趣小组提出的“数据挖掘交叉行业标准过程”CRISP－DM 以及专业的数据挖掘技术咨询公司 Two Crows 提出的模型。但各种方法并不是迥然各异的，一般都要包括：问题的理解、数据的理解、收集和准备、建立数据挖掘模型、评价所建的模型、应用所建的模型等一系列任务[76]。

SAS Enterprise Miner 把统计分析系统和图形用户界面（GUI）集成在一起，并采用了 SAS 协会定义的数据挖掘方法—SEMMA 方法，即 Sample、Explore、Modify、Model、Assess 五个步骤，用户界面友好、直观、灵活、使用方便，即使是没有什么经验的用户也可以理解和使用。这一方法论的主要过程有[74][76]：

（1）Sample—数据采样

在进行数据挖掘时，应先从大量数据中选取有代表性、真实、完整和有效的，并且与问题相关的样本数据子集，而不是全部数据。通过数据样本的精选，不仅能减少数据处理量、节省系统资源，而且能使规律更加凸现出来。

包含的功能节点有：Input Data Source、Sampling、Data Partition

（2）Explore—数据分析和预处理

当拿到一个样本数据集之后，要先对它进行探索和分析：是否达到原来设想的要求；其中有无明显的规律和趋势；是否出现从未设想过的数据状态；各因素之间有无相关性；可以区分成怎样的一些类别等等。

数据预处理过程中的任务包括对数据表、记录属性的选择以及为了适合建模工具的要求对数据进行的转化和净化。

包含的功能节点有：Distribution Explorer、Multiplot、Insight、Association、Variable Selection、Link Analysis（Exp.）

（3）Modify—数据调整和技术选择

对原来模糊的问题进一步明确和量化，按照问题的具体要求来审视数据集，检验它是否适应问题的需要，对数据进行转换并做进一步的填充和剔除。

包含的功能节点有：Data Set Attributes、Transform Variables、Filter Outliers、Replacement、Clustering、SOM/Kohonen、Time Series（Exp.）

（4）Model—模型的建立和知识的发现

这一步是数据挖掘工作的中心环节，根据数据集的特征和要实现的目标，在各种技术手段（如数理统计、人工神经元网络、决策树等）中选择合适的技

术和算法，通过比较获得效果最好的模型。建立起最优模型之后，“挖掘”出隐藏的、有价值的、企业需要的信息，支持企业决策，并在友好的界面下描述出来以便于用户理解和利用。本阶段是整个数据挖掘的核心。

包含的功能节点有：Regression、Tree、Neural Network、User Defined Model、Princomp/Dmneural、Ensemble、Memory-Based Reasoning（Exp.）、Two Stage Model

（5）Assess—模型和知识的解释与评价

从上述过程中会得到一系列的分析结果、模式和模型，多数情况会得出对目标问题多侧面的描述，这时就要综合它们的规律性，提供合理的决策支持信息，确定是否有必要重新进行数据挖掘过程。评价的一种办法是直接使用原先建立模型的样本和样本数据来进行检验；另一种办法是另找一批数据并对其进行检验（已知这些数据能反映客观实践的规律性）；再一种办法是在实际运行的环境中取出新鲜数据进行检验。

包含的功能节点有：Assessment、Reporter

SAS/EM 系统的 SEMMA 方法论的流程图如图 4.1 所示：

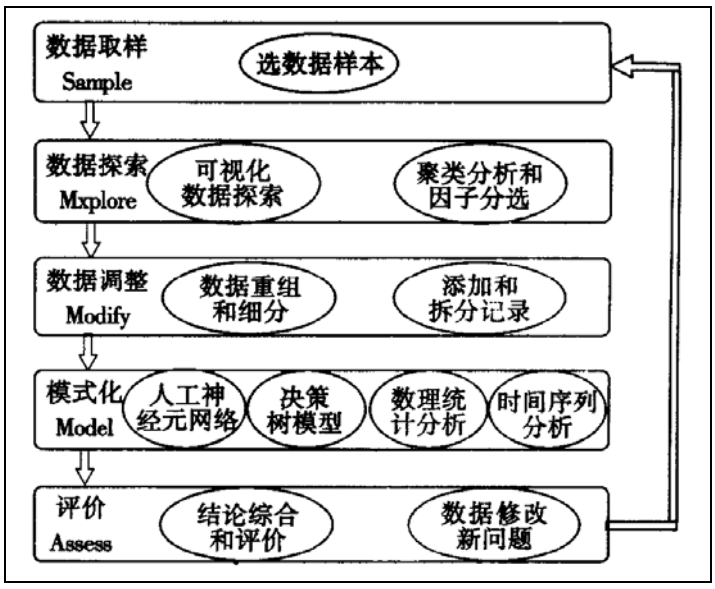

图 4.1 SEMMA 方法论流程图

作为智能型的数据挖掘集成工具，SAS/EM 的图形化界面、可视化操作可引导用户（即使是对统计分析的经验不太多的用户）按 SEMMA 原则成功地进行数据挖掘，用户只要将数据输入，经过 SAS/EM 运行，即可得到一些分析结果。有

经验的专家还可通过修改数据调整分析处理过程[77]。

本文使用 SAS 软件进行数据挖掘研究，按照该系统提出的 SEMMA 数据挖掘方法论原则，逐步建立了完整的数据挖掘过程，并最终得到了商用建筑能耗的预测模型，以下将逐一介绍这一数据挖掘过程。

4.2 定义问题

本课题以商用建筑的能耗为研究目标，以上海市商用建筑信息数据库为基础，使用 SAS/EM 模块的数据挖掘功能进行分析研究，得出各个影响因素与建筑能耗之间的关系，并建立商用建筑能耗的预测模型，验证评价之后用于预测商用建筑能耗。

整个数据挖掘过程在 SAS/EM 中由如图 4.2 所示的流程图完成。

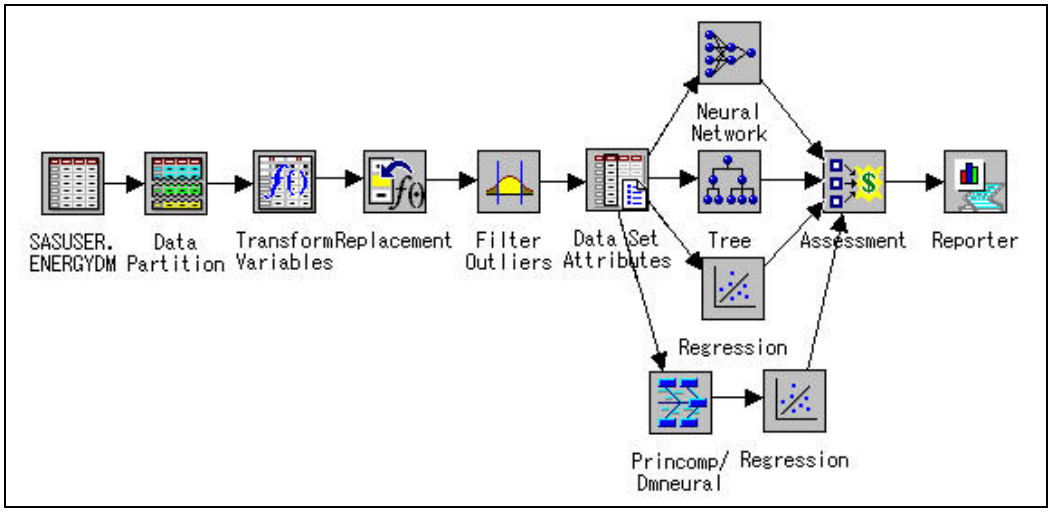

图 4.2 数据挖掘流程图

4.3 数据准备

本文的数据挖掘研究是以上海市商用建筑信息数据库为基础的，但是数据库中的数据由于变量类型、数据形式和含义、数据用途等方面的原因不能直接导入 SAS 系统进行分析，因此首先需要将数据进行处理和加工，为数据挖掘工作做好准备。

4.3.1 数据预处理

将上海市商用建筑信息数据库中的基本数据导出至 Excel 中进行预处理，根据数据库内的数据情况和数据挖掘的要求定义变量：

（1）ID

建筑编号，属性设为 id，仅作为观测的标记。

（2）BUILTTIME

建筑竣工年月，以数值形式表示，反映建筑设备的折旧损耗程度，竣工时间越早数值就越小，损耗程度也就越大。

（3）AREA

建筑总面积。

（4）WINTYPE

窗户类型，0 代表单层玻璃，1 代表双层玻璃。

（5）FILM

玻璃贴膜情况，0 代表不贴膜，1 代表玻璃贴膜。

（6）OFFICE、COMMERCIAL、HOTEL

分别表示办公、商场、宾馆等三种功能类型在商用建筑中的比例。

（7）CAPACITY

制冷机的装机容量。

（8）COOLTYPE、HEATTYPE

表示冷热源形式，1 代表电、2 代表煤气、3 代表燃油、4 代表煤、5 代表热网、6 代表复合能源。

（9）ACTYPE

空调系统形式，1 代表风机盘管系统、2 代表定风量系统、3 代表变风量系统、4 代表风机盘管＋定风量系统、5 代表风机盘管＋变风量系统、6 代表定风量＋变风量系统。

（10）BAS

楼宇自控系统，0 代表没有安装楼宇自控系统，1 代表装有楼宇自控系统。

（11）ACTIME

表示在空调季节里，空调系统每天平均的运行小时数。

（12）ENERGY

全年一次能耗值。

将处理好的 Excel 表格导入 SAS 系统中，以 SAS 系统的数据集格式保存在 SASUSER 目录里，以备数据挖掘使用。启动 SAS/EM 模块之后，首先使用“Input Data Source”节点将处理好的数据集引入数据挖掘的流程图内，作为数据挖掘的源数据。

4.3.2 数据集的划分

在挖掘时，一般把数据分成训练样本集和验证样本集。前者用于构建系统模型，后者用于验证系统的有效性。首先用数据挖掘技术作用于训练样本集，当系统模型稳定并且产生了一些有价值的结果（即知识）后，再用验证样本集作用于系统，这时应当产生相似的结果。在 SAS/EM 模块中，“Data Partition”节点可以实现此功能。

在本课题中，由于目前数据库刚刚建立，数据量还不够多，为了保证数据挖掘的效果将全部数据作为训练样本集，并以数据库之外的“新鲜”数据来验证模型。

4.3.3 数据转换

由于本课题的研究目标是预测商用建筑能耗，因此需要找出各个影响因素与建筑能耗之间的关系。考虑到建筑面积相对于其他因素对建筑能耗的影响太过明显，为了突出其它因素对建筑能耗的影响程度，本课题将商用建筑单位面积全年一次能耗值作为目标变量，在一些相关的建筑能耗研究文献中也以此作为研究对象【4】【36】【78】。

可以使用“Transform Variables”节点来实现这一转换。新建一个变量“ENERGY1”，将全年一次能耗值与建筑总面积的比值赋值给这个新变量。

4.3.4 数据填充及剔除

由于各种各样的原因，数据集中的数据很有可能不完整或者存在一些错误，这就需要对数据进行填充和剔除。

在本课题中，使用“Replacement”节点来对数据进行填充。对于连续型数

值一般采用均值填充法，对于离散型数值或分类变量一般以出现频率最大的值填充。同时，填充值的选取还要结合专业知识来考虑。

数据的剔除也是必不可少的步骤，使用“Filter Outliers”节点可以实现这一功能。为了保证数据挖掘的效果，本课题将缺少能耗数据、制冷机装机容量、空调运行时间等信息的建筑予以剔除。

4.3.5 数据属性定义

经过前面几步数据处理过程，数据集中的数据已基本符合数据挖掘的要求，下面将对各个变量的属性进行定义，确定因变量及自变量，以实施数据挖掘计算过程。

在“Data Set Attributes”节点中，首先将变量“ID”的属性设置为“id”，该变量不进入数据挖掘计算过程，仅仅为了便于研究。将变量“ENERGY1”的属性设为“target”，这是数据挖掘过程的目标变量，即因变量。同时将变量“AREA”和“ENERGY”设为“rejected”，表明这两个变量的信息已包含在变量“ENERGY1”中，不需要再进入计算过程，以免影响计算结果。由于变量“WINTYPE”的缺失数据太多，缺失率达到了 27%，无法反映真实情况，所以也将该变量的属性设为“rejected”。其它剩余变量的属性均自动设为“input”，作为模型的输入变量，即自变量。

至此，数据挖掘的数据准备阶段已全部完成，生成的数据集已满足数据挖掘计算的要求。数据集的详细情况见图 4.3、表 4.1。

表 4.1 数据挖掘源数据集变量表

变量名	变量含义	变量类型
ID	建筑编号	数值型，连续变量
BUILTTIME	建筑竣工年月	数值型，连续变量
FILM	玻璃贴膜情况	0、1 型变量
OFFICE	办公用面积占总面积的比例	数值型，连续变量
COMMERCIAL	商场用面积占总面积的比例	数值型，连续变量
HOTEL	宾馆用面积占总面积的比例	数值型，连续变量
CAPACITY	制冷机的装机容量	数值型，连续变量
COOLTYPE	冷源形式	名义型分类变量
HEATTYPE	热源形式	名义型分类变量
ACTYPE	空调系统形式	名义型分类变量

BAS	楼宇自控系统	0、1 型变量
ACTIME	空调系统运行小时数	数值型，连续变量
ENERGY1	单位面积全年一次能耗	数值型，连续变量

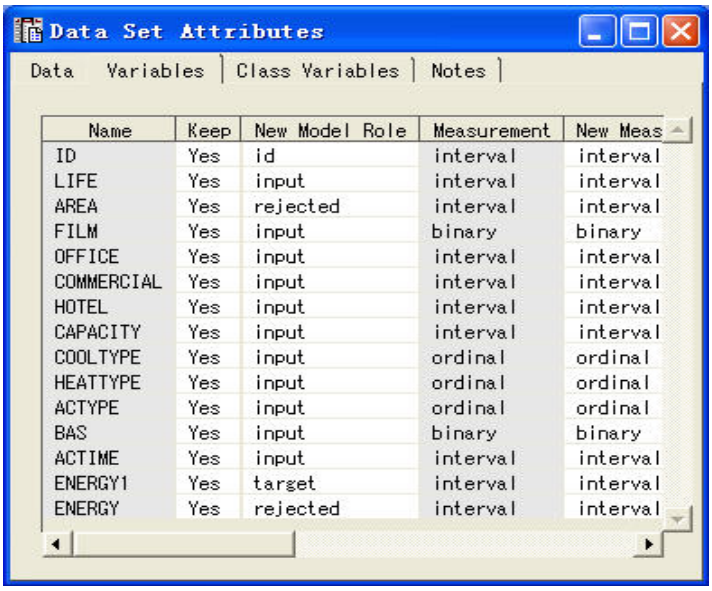

图 4.3 数据挖掘源数据集变量属性

4.4 神经网络模型与决策树模型

神经网络和决策树模型都是数据挖掘中非常有效的算法，它们可以在分类、预测等方面取得令人满意的效果，挖掘数据中隐含的深层次的信息和规律。神经网络模型的计算过程和结果输出都比较复杂，但是计算精度高；而决策树模型计算过程简便，输出结果直观、易懂，相应地计算精度稍差。

在本课题中，由于数据量和数据信息不够丰富，由 SAS/EM 模块得到的神经网络模型和决策树模型都不理想，计算过程很难收敛，其结果的均方误差、平均误差等指标都不能满足要求，因此本文将以回归模型作为数据挖掘的研究重点。

4.5 回归模型

“回归”（regression）这一名词，最初是由 19 世纪英国生物学家兼统计

学家 F.Galton（F·高尔顿）在一篇著名的遗传学论文中引入的。现代意义上的回归分析是研究一个变量（也称为因变量 dependent variable 或被解释变量 explained variable）对另一个或多个变量（也称为自变量 independent variable 或 explanatory variable）的依赖关系，其目的在于通过自变量的给定值，来预测因变量的平均值或某个特定值[79]。具体来说，回归分析需要解决以下问题：

（1）构建因变量与自变量之间的回归模型，并依据样本观测值对回归模型中的参数进行估计，给出回归方程。

（2）对回归方程中的参数和方程本身进行显著性检验。

（3）评价自变量对因变量的贡献。

（4）利用所求得的回归方程对因变量进行预测，对自变量进行控制。

本文使用 SAS/EM 模块的“Regression”节点来进行回归分析。将数据准备阶段完成的源数据集引入“Regression”节点，进行参数设置和条件设定之后就可计算出结果。

4.5.1 全回归模型

SAS 系统提供的回归模型有九种，分别是：全回归模型（NONE）、逐步引入法（FORWARD）、逐步剔除法（BACKWARD）、逐步回归法（STEPWISE）、最大 R^2 增量法（MAXR）、最小 R^2 增量法（MINR）、R^2 选择法（RSQUARE）、修正的 R^2 选择法（ADJRSQ）、Mallows 的 Cp 选择法（CP）。

对于回归结果的好坏一般以决定系数 R^2 作为评价标准。R^2 也叫确定系数，其定义为：回归平方和 $SS_{回}$ 占总离差平方和 $SS_{总}$ 的比例，它是因变量由模型中的自变量进行回归得到的拟合优度，它可以定量评价在 y 的总变异中，由 x 变量组建立的线性回归方程所能解释的比例。R^2 越接近于 1，则回归方程的效果越好。

本文首先采用全回归模型，建立以建筑能耗为因变量的所有自变量的多元线性回归模型。部分输出结果见图 4.4。

The REG Procedure
Model: MODEL1
Dependent Variable: ENERGY1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	1731920	157447	1.55	0.1526
Error	38	3847821	101258		
Corrected Total	49	5579741			

Root MSE	318.21134	R-Square	0.3104
Dependent Mean	1175.50266	Adj R-Sq	0.1108
Coeff Var	27.07024		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	Intercept	1	6552.13829	43676	0.15	0.8815	0
BUILTTIME	BUILTTIME	1	-2.28740	21.88037	-0.10	0.9173	-0.01839
FILM	FILM	1	-81.09062	116.14592	-0.70	0.4893	-0.12098
OFFICE	OFFICE	1	-968.41875	694.23467	-1.39	0.1711	-0.60081
COMMERCIAL	COMMERCIAL	1	-1814.39320	943.51403	-1.92	0.0620	-0.54061
HOTEL	HOTEL	1	-447.27809	794.62068	-0.56	0.5768	-0.22922
CAPACITY	CAPACITY	1	0.01192	0.00780	1.53	0.1344	0.23929
COOLTYPE	COOLTYPE	1	-61.81825	65.24849	-0.95	0.3494	-0.14025
HEATTYPE	HEATTYPE	1	24.18550	57.05465	0.42	0.6740	0.06150
ACTYPE	ACTYPE	1	16.32968	27.09249	0.60	0.5503	0.09090
BAS	BAS	1	100.62323	111.32923	0.90	0.3718	0.14621
ACTIME	ACTIME	1	-0.97400	24.43930	-0.04	0.9684	-0.00859

Correlation of Estimates

Variable	Label	Intercept	BUILTTIME	FILM	OFFICE
Intercept	Intercept	1.0000	-0.9999	-0.3827	0.0239
BUILTTIME	BUILTTIME	-0.9999	1.0000	0.3841	-0.0380
FILM	FILM	-0.3827	0.3841	1.0000	-0.0642
OFFICE	OFFICE	0.0239	-0.0380	-0.0642	1.0000
COMMERCIAL	COMMERCIAL	0.1138	-0.1265	-0.0311	0.8636
HOTEL	HOTEL	0.0184	-0.0284	0.0855	0.8491
CAPACITY	CAPACITY	0.2754	-0.2742	0.1752	-0.0917
COOLTYPE	COOLTYPE	0.2009	-0.2030	-0.1938	0.0967
HEATTYPE	HEATTYPE	-0.0016	-0.0003	-0.0057	0.0896
ACTYPE	ACTYPE	-0.0697	0.0691	-0.2779	-0.0583
BAS	BAS	0.4102	-0.4110	-0.2887	0.0354
ACTIME	ACTIME	-0.0690	0.0648	-0.2048	-0.0638

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation: Intercept	BUILTTIME	FILM	OFFICE
1	8.64387	1.00000	1.377307E-8	1.37497E-8	0.00227	0.00008180
2	1.01836	2.91343	3.21188E-10	3.23827E-10	0.00447	0.00033125
3	0.64081	3.67275	2.803859E-9	2.778144E-9	0.00229	0.00015216
4	0.56897	3.89772	2.460577E-9	2.579375E-9	0.23679	4.835925E-7
5	0.33226	5.10052	2.228894E-8	2.216093E-8	0.04447	0.00064498
6	0.25368	5.83724	6.458953E-8	6.348301E-8	0.02378	0.00096414
7	0.21511	6.33903	2.826758E-7	2.820612E-7	0.35159	0.00314
8	0.16109	7.32515	3.277743E-7	3.263969E-7	0.08972	0.00269
9	0.14211	7.79908	9.595128E-9	9.65786E-9	0.03966	0.00015410
10	0.01989	20.84425	0.00000264	0.00000265	0.05331	0.02139
11	0.00385	47.37399	0.00005643	0.00005614	0.00466	0.96949
12	5.303824E-7	4037.00710	0.99994	0.99994	0.14699	0.00095878

图 4.4 全回归模型输出结果

在全回归模型的输出结果中可以看出，模型的 R^2 值为 0.3104，修正后的 R^2 值为 0.1108，远小于 1；并且回归模型的 F 检验的概率值为 0.1526，大于显著性水平 0.05。另外，各个回归系数的 t 检验值的概率均比 0.05 大很多，没有显

著性。所以直接采用全回归模型的回归效果并不理想，变量间可能存在共线性，需要进行多重共线性判断和处理。

4.5.2 多重共线性分析

多重共线性是指自变量之间存在着线性关系或接近线性关系，即一个自变量可以被另一个或另几个自变量线性表示。多重共线性的存在会使回归方程的求解发生困难，同时也会影响到回归方程的效果[80]。因此在进行多元线性回归分析时，判断和处理自变量间多重共线性非常必要。

由输出结果－1 中各个自变量间的相关系数可以看出，“BUILTTIME”项与截距项的相关系数很大。在共线性诊断结果中，以最大特征根 8.64387 为 100%计算出各组的条件系数（Condition Index），比较条件系数最大的一行中方差比率（Var Prop）较大的几个自变量具有较大的共线性。显而易见，“BUILTTIME”项与截距项存在较大的共线性。

根据上述多重共线性的诊断，在全回归模型中去掉截距项，重新进行计算，得到了图 4.5 的回归结果。模型的 R^2 值达到 0.9484，修正后的 R^2 值为 0.9339，均非常接近于 1；并且回归模型的 F 检验的概率值<0.0001，此模型线性关系非常明显。此时，全回归模型已经能够很好地反映因变量与自变量之间的线性关系，且非常显著。

但是，上述分析只是对纳入回归方程的所有自变量对因变量所起的综合作用的效果评价，还应分别研究各个自变量与因变量之间的关系。在全回归模型的分析结果中，由各个自变量的回归系数的 t 检验概率值可以看到，各个回归系数并不都是显著的，有的自变量对因变量的影响很强，而有的则很弱。其中，“BUILTTIME”和“COMMERCIAL”两项的概率值较小（分别为 0.0049 和 0.0551），可以认为回归系数显著。因此，在本课题中并不是所有的自变量都对因变量有显著影响，不能将它们全都引入到模型中，全回归模型还不是“最优”模型，需要进行变量筛选重新建立回归模型。

The REG Procedure
Model: MODEL1
Dependent Variable: ENERGY1 ENERGY1

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	70819966	6438179	65.22	<.0001
Error	39	3850100	98721		
Uncorrected Total	50	74670066			

Root MSE	314.19821	R-Square	0.9484
Dependent Mean	1175.50266	Adj R-Sq	0.9339
Coeff Var	26.72884		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
BUILTTIME	BUILTTIME	1	0.99463	0.33373	2.98	0.0049	1.62609
FILM	FILM	1	-74.42300	105.95213	-0.70	0.4866	-0.04475
OFFICE	OFFICE	1	-970.90555	685.28388	-1.42	0.1645	-0.62034
COMMERCIAL	COMMERCIAL	1	-1830.49752	925.56530	-1.98	0.0551	-0.20755
HOTEL	HOTEL	1	-449.46591	784.46716	-0.57	0.5700	-0.06531
CAPACITY	CAPACITY	1	0.01160	0.00740	1.57	0.1250	0.09384
COOLTYPE	COOLTYPE	1	-63.78429	63.11267	-1.01	0.3184	-0.07232
HEATTYPE	HEATTYPE	1	24.19912	56.33503	0.43	0.6699	0.03799
ACTYPE	ACTYPE	1	16.61292	26.68577	0.62	0.5372	0.05173
BAS	BAS	1	93.77197	100.25011	0.94	0.3553	0.06042
ACTIME	ACTIME	1	-0.72099	24.07356	-0.03	0.9763	-0.00686

Correlation of Estimates

Variable	Label	BUILTTIME	FILM	OFFICE	COMMERCIAL
BUILTTIME	BUILTTIME	1.0000	0.1029	-0.9166	-0.8285
FILM	FILM	0.1029	1.0000	-0.0596	0.0136
OFFICE	OFFICE	-0.9166	-0.0596	1.0000	0.8667
COMMERCIAL	COMMERCIAL	-0.8285	0.0136	0.8667	1.0000
HOTEL	HOTEL	-0.6504	0.1001	0.8491	0.7534
CAPACITY	CAPACITY	0.0759	0.3160	-0.1023	-0.0303
COOLTYPE	COOLTYPE	-0.1423	-0.1292	0.0939	0.0406
HEATTYPE	HEATTYPE	-0.1250	-0.0068	0.0897	0.0579
ACTYPE	ACTYPE	-0.0394	-0.3305	-0.0568	-0.0603
BAS	BAS	-0.0607	-0.1563	0.0281	0.0836
ACTIME	ACTIME	-0.2701	-0.2509	-0.0624	-0.0717

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation BUILTTIME	FILM	OFFICE	COMMERCIAL
1	7.67852	1.00000	0.00007239	0.00342	0.00010258	0.00099013
2	1.01801	2.74639	0.00000163	0.00550	0.00033651	0.00003469
3	0.63967	3.46466	0.00001280	0.00166	0.00014878	0.11871
4	0.56817	3.67619	0.00001211	0.27370	3.367095E-7	0.01585
5	0.33013	4.82276	0.00007914	0.04723	0.00056129	0.06910
6	0.25066	5.53470	0.00015752	0.05558	0.00061982	0.00044194
7	0.20267	6.15516	0.00125	0.46796	0.00349	0.00837
8	0.14935	7.17032	0.00306	0.04586	0.00645	0.00062901
9	0.14144	7.36801	0.00051246	0.03557	0.00133	0.00000993
10	0.01865	20.29006	0.01992	0.05583	0.04954	0.03844
11	0.00272	53.16666	0.97492	0.00769	0.93741	0.74743

图 4.5 去掉截距项的全回归模型输出结果

4.5.3 逐步回归模型

SAS/EM 模块提供了多种变量选元的方法，但在实践中最常用的是逐步回归法（STEPWISE）。逐步回归法综合了逐步引入法和逐步剔除法的优点，在向前引

入每一个新的自变量之后，都重新对已选入的自变量进行检验，以确定其有无继续保留在方程中的价值，引入和剔除交替进行，直到全部自变量根据给定的显著性水平没有一个再能被选入或剔除出回归模型为止。

用逐步回归模型替换全回归模型，显著性水平取为 0.05，重新进行计算，部分输出结果如图 4.6 所示。此时，模型的 R^2 值为 0.9386，修正后的 R^2 值为 0.9347，都接近于 1，且回归模型的 F 检验的概率值<0.0001。此回归模型的效果很好，虽然 R^2 值比全回归模型略有降低，但是修正后的 R^2 值有所提高。另外，回归模型经过变量筛选之后仅留下“BUILTTIME”、“HOTEL”、“CAPACITY”三个自变量，模型的规模减小很多，各个自变量对应的回归系数也都很显著。

The REG Procedure
Model: MODEL1
Dependent Variable: ENERGY1 ENERGY1

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	70084543	23361514	239.45	<.0001
Error	47	4585523	97564		
Uncorrected Total	50	74670066			

Root MSE	312.35288	R-Square	0.9386
Dependent Mean	1175.50266	Adj R-Sq	0.9347
Coeff Var	26.57186		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
BUILTTIME	BUILTTIME	1	0.51751	0.03320	15.59	<.0001	0.84606
HOTEL	HOTEL	1	533.01657	258.02261	2.07	0.0444	0.07746
CAPACITY	CAPACITY	1	0.01603	0.00659	2.43	0.0189	0.12964

图 4.6 显著性水平为 0.05 的逐步回归模型输出结果

为了得到更合适的回归模型，将显著性水平逐渐增大进行计算，得到了如图 4.7、图 4.8 所示的结果。从几次逐步回归计算的结果可以看出，随着显著性水平的提高，模型的 R^2 值由 0.9386 逐渐提高至 0.9421、0.9443，修正后的 R^2 值由 0.9347 逐渐提高至 0.9371、0.9394。回归模型的线性效果越来越好，选入模型的自变量为“BUILTTIME”、“OFFICE”、“COMMERCIAL”、“CAPACITY”，各自的回归系数也非常显著，此时显著性水平为 0.2。再增大显著性水平进行试算，得到的计算结果越来越差，而且也失去了回归分析的统计学意义。

The REG Procedure
Model: MODEL1
Dependent Variable: ENERGY1 ENERGY1

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	70348066	17587017	187.18	<.0001
Error	46	4322000	93957		
Uncorrected Total	50	74670066			

Root MSE	306.52327	R-Square	0.9421
Dependent Mean	1175.50266	Adj R-Sq	0.9371
Coeff Var	26.07593		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
BUILTTIME	BUILTTIME	1	0.55851	0.04075	13.71	<.0001	0.91308
COMMERCIAL	COMMERCIAL	1	-736.08538	439.52341	-1.67	0.1008	-0.08346
HOTEL	HOTEL	1	514.89056	253.43821	2.03	0.0480	0.07482
CAPACITY	CAPACITY	1	0.01464	0.00652	2.24	0.0297	0.11842

图 4.7 显著性水平为 0.15 的逐步回归模型输出结果

The REG Procedure
Model: MODEL1
Dependent Variable: ENERGY1 ENERGY1

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	70507863	17626966	194.81	<.0001
Error	46	4162203	90483		
Uncorrected Total	50	74670066			

Root MSE	300.80337	R-Square	0.9443
Dependent Mean	1175.50266	Adj R-Sq	0.9394
Coeff Var	25.58934		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
BUILTTIME	BUILTTIME	1	0.82911	0.11132	7.45	<.0001	1.35549
OFFICE	OFFICE	1	-600.79639	244.21806	-2.46	0.0177	-0.38387
COMMERCIAL	COMMERCIAL	1	-1444.54097	509.84720	-2.83	0.0068	-0.16379
CAPACITY	CAPACITY	1	0.01522	0.00641	2.38	0.0217	0.12313

图 4.8 显著性水平为 0.2 的逐步回归模型输出结果

至此，得到的回归模型经过各方面的检验均已达到要求，可以作为“最优”的回归模型，记为回归模型 I。得到的回归方程为：

$$\begin{aligned} ENERGY1 = {} & 0.82911(BUILTTIME) - 600.79639(OFFICE) \\ & -1444.54097(COMMERCIAL) + 0.01522(CAPACITY) \end{aligned} \qquad (4.1)$$

4.6 主成分分析

在实际的研究中，经常会遇到多变量或多指标问题，由于变量或指标较多，分析问题具有相当的复杂性，并且在多数情况下，这些不同的变量或指标间都存在一定的相关性。主成分分析方法就是设法将原来的变量或指标重新组合成一组新的、互不相关的几个综合变量或指标，同时根据实际需要从中选取几个较少的综合变量或指标来尽可能多地反映原变量或指标的信息。

主成分分析主要有两方面的应用：一是用于系统评估，将多指标问题转化为单一的综合指标的问题；二是用于筛选变量，从原始变量所构成的子集合中选择最佳变量构成最佳变量子集合。主成分分析往往不是研究的目的，而是达到目的的一种手段，可以应用到多元回归、聚类分析等过程中。

鉴于主成分分析方法的上述优点，本文将主成分分析引入回归模型的研究中，使用 SAS/EM 模块中的“Princomp/Dmneural”节点进行主成分分析，筛选变量之后再进行多元回归研究。

4.6.1 主成分分析过程

首先在“Princomp/Dmneural”节点中将“ENERGY1”之外的所有 11 个自变量选入主成分分析过程，计算后得到如图 4.9 所示结果。

确定主成分的个数有两个准则：一是主成分的累计贡献率在 70%～85%；二是主成分所对应的特征值大于或等于 1。这两个准则一般要结合起来考虑。由本例的计算结果来看，前 5 个主成分的累计贡献率达到了 76.12%，并且第 5 个主成分对应的特征值为 1.00919584，所以取前 5 个主成分已基本能够代表原先 11 个自变量的信息。

因本文以主成分分析的筛选变量功能为主要研究任务，所以对于这 5 个主成分的具体分析不再展开叙述。

The PRINCOMP Procedure

Observations 50
Variables 11

Simple Statistics

	BUILTTIME	FILM	OFFICE	COMMERCIAL	HOTEL	CAPACITY
Mean	1997.381800	0.5400000000	0.7528000000	0.0964000000	0.0472000000	7264.004000
StD	2.713030	0.5034574339	0.2093531339	0.1005446393	0.1729343393	6772.244180

Simple Statistics

	COOLTYPE	HEATTYPE	ACTYPE	BAS	ACTIME
Mean	1.160000000	1.720000000	3.320000000	0.6200000000	11.24000000
StD	0.765586343	0.858094710	1.878340562	0.4903143515	2.97479205

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.48840621	0.71204227	0.2262	0.2262
2	1.77636394	0.05181691	0.1615	0.3877
3	1.72454702	0.34991400	0.1568	0.5445
4	1.37463302	0.36543718	0.1250	0.6695
5	1.00919584	0.19387395	0.0917	0.7612
6	0.81532189	0.23128974	0.0741	0.8353
7	0.58403214	0.01585905	0.0531	0.8884
8	0.56817309	0.21550275	0.0517	0.9401
9	0.35267034	0.09274498	0.0321	0.9721
10	0.25992537	0.21319423	0.0236	0.9958
11	0.04673114		0.0042	1.0000

Eigenvectors

		Prin1	Prin2	Prin3	Prin4	Prin5
BUILTTIME	BUILTTIME	-.080414	0.590369	0.227854	0.145929	0.129381
FILM	FILM	-.012065	-.571303	0.213924	0.202880	0.197604
OFFICE	OFFICE	-.571872	-.149764	0.151255	-.063235	-.204944
COMMERCIAL	COMMERCIAL	0.187681	0.216571	-.354832	0.358728	0.590194
HOTEL	HOTEL	0.566942	0.023098	0.100776	-.225995	-.156567
CAPACITY	CAPACITY	-.059357	0.395062	0.423178	-.109899	-.091926
COOLTYPE	COOLTYPE	0.033143	-.014222	0.027639	0.668756	-.285963
HEATTYPE	HEATTYPE	0.129573	-.003425	0.200544	0.528924	-.344814
ACTYPE	ACTYPE	0.125093	-.297295	0.390683	0.023994	0.423301
BAS	BAS	-.081607	0.084066	0.568353	0.035735	0.315703
ACTIME	ACTIME	0.515737	-.058927	0.218311	-.113405	-.203420

Eigenvectors

	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11
BUILTTIME	0.199518	0.078306	-.256610	0.660378	-.080371	-.017407
FILM	0.182750	-.169300	0.428335	0.443553	-.324386	0.000755
OFFICE	-.034796	0.035759	-.010937	0.127553	0.343642	0.666847
COMMERCIAL	-.134174	-.108549	0.265578	-.048282	0.228139	0.400321
HOTEL	0.121408	0.057796	-.131947	-.034591	-.413868	0.619944
CAPACITY	-.327937	0.203328	0.669977	-.146319	-.146204	-.010438
COOLTYPE	0.404911	0.493516	0.065451	-.236327	0.019285	0.025579
HEATTYPE	-.544454	-.446642	-.209611	0.030271	-.054739	0.023566
ACTYPE	-.393176	0.535997	-.331519	-.003381	0.105772	-.024292
BAS	0.387555	-.415335	-.139471	-.465118	0.059596	0.018839
ACTIME	0.141880	-.067959	0.190190	0.228375	0.712325	-.090334

图 4.9 主成分分析部分计算结果

4.6.2 变量的筛选

主成分分析筛选变量的具体步骤是：

（1）从结果中找出主成分对应的最小特征值，一般是最后一个。本例为第

11 个主成分，特征值为 0.04673114，接近于 0；

（2）从特征向量表中找出最小特征值所对应的特征向量。本例为 Prin11 一列；

（3）在最小特征值所对应的特征向量中，将系数绝对值最大者所对应的变量删去。本例为自变量“OFFICE”，其系数为 0.666847；

（4）对剩余变量再进行主成分分析，按照上述过程删除变量，反复进行，直到最小特征值不是很小（没有具体的标准）为止。

经过上述步骤的反复进行，本例共删除了“OFFICE”、“HOTEL”、“BUILTTIME”3 个自变量，计算结果见附录。用剩余的 8 个自变量再做主成分分析，得到的最小特征值为 0.51131097，不必再删除变量。

4.6.3 回归分析

将经过主成分分析筛选出来的 8 个自变量导入“Regression”节点，再次进行回归分析，计算过程与 4.5 节相似，不再赘述。

与初次回归分析一样，全回归模型中去掉截距项后回归效果明显提高，部分输出结果见图 4.10。全回归模型中的 R^2 值为 0.9186，修正后的 R^2 值为 0.9030，回归模型的 F 检验的概率值<0.0001。各个自变量的回归系数的 t 检验概率值都较大，回归系数不显著，仅“ACTIME”项<0.0001。

经过逐步回归计算之后，回归模型的质量明显提高，部分输出结果见图 4.11。模型的 R^2 值达到 0.9153，修正后的 R^2 值为 0.9079，回归模型的 F 检验的概率值<0.0001；各个自变量的回归系数的 t 检验概率值都较小，回归系数比较显著，此时显著性水平取为 0.25。

至此，回归模型已满足各种检验指标，可作为“最优”回归模型，记为回归模型 II，得到的回归方程为：

$$\begin{aligned} ENERGY1 &= 0.01752(CAPACITY) + 36.62713(ACTYPE) \\ &+ 147.52038(BAS) + 70.50595(ACTIME) \end{aligned} \quad (4.2)$$

The REG Procedure
Model: MODEL1
Dependent Variable: ENERGY1 ENERGY1

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	68588968	8573621	59.21	<.0001
Error	42	6081098	144788		
Uncorrected Total	50	74670066			

Root MSE	380.51023	R-Square	0.9186
Dependent Mean	1175.50266	Adj R-Sq	0.9030
Coeff Var	32.37000		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
FILM	FILM	1	-89.23413	122.46038	-0.73	0.4702	-0.05366
COMMERCIAL	COMMERCIAL	1	-194.35747	537.78534	-0.36	0.7196	-0.02204
CAPACITY	CAPACITY	1	0.01455	0.00871	1.67	0.1022	0.11768
COOLTYPE	COOLTYPE	1	1.15532	74.08522	0.02	0.9876	0.00131
HEATTYPE	HEATTYPE	1	70.15159	66.87709	1.05	0.3002	0.11012
ACTYPE	ACTYPE	1	39.26145	31.40224	1.25	0.2181	0.12225
BAS	BAS	1	158.72441	119.67112	1.33	0.1919	0.10227
ACTIME	ACTIME	1	66.30049	13.45847	4.93	<.0001	0.63039

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation: FILM	COMMERCIAL	CAPACITY	COOLTYPE
1	5.78723	1.00000	0.00676	0.00786	0.00692	0.00620
2	0.62907	3.03309	0.00247	0.50127	0.11377	0.01692
3	0.56972	3.18716	0.30914	0.07558	0.27029	0.00295
4	0.32228	4.23759	0.03535	0.30478	0.00005385	0.39207
5	0.25159	4.79608	0.05477	0.01089	0.13464	0.14351
6	0.20208	5.35142	0.52999	0.06687	0.42742	0.04052
7	0.14260	6.37054	0.05637	0.00017213	0.01396	0.38046
8	0.09542	7.78798	0.00516	0.03258	0.03294	0.01736

图 4.10 去掉截距项的全回归模型输出结果

The REG Procedure
Model: MODEL1
Dependent Variable: ENERGY1 ENERGY1

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	68344664	17086166	124.26	<.0001
Error	46	6325401	137509		
Uncorrected Total	50	74670066			

Root MSE	370.82169	R-Square	0.9153
Dependent Mean	1175.50266	Adj R-Sq	0.9079
Coeff Var	31.54580		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
CAPACITY	CAPACITY	1	0.01752	0.00805	2.18	0.0347	0.14175
ACTYPE	ACTYPE	1	36.62713	28.53595	1.28	0.2057	0.11405
BAS	BAS	1	147.52038	113.40495	1.30	0.1998	0.09505
ACTIME	ACTIME	1	70.50595	10.09257	6.99	<.0001	0.67038

图 4.11 显著性水平为 0.25 的逐步回归模型输出结果

第 5 章 模型解释与评价

经过第 4 章的完整的数据挖掘过程，得到了有关商用建筑能耗的两个预测模型，回归模型 I、回归模型 II。为了验证所进行的数据挖掘计算过程是否达到了预期的目的，结果是否符合要求，需要对模型进行解释和评价，分析其合理性和可用性；另外，两个模型虽然都是回归模型，但模型 II 是借助主成分分析方法筛选变量之后进行的回归分析，而且两个回归方程的形式是不同的，因此有必要对两个模型进行比较和验证，分析其准确性和可靠性，以确定最终的商用建筑能耗预测模型。

5.1 回归模型 I、II 的解释与说明

5.1.1 回归模型 I 的解释与说明

各个自变量的回归系数的符号表示了自变量与因变量的相关性，这可以定性地反映出各个因素对建筑能耗的影响规律。经过多次计算得到的回归方程剔除了对因变量影响不显著的自变量，最终选入四个自变量构成关于商用建筑能耗的回归模型 I。

自变量“BUILTTIME”的回归系数为正，表明自变量“BUILTTIME”与因变量正相关，即商用建筑竣工时间越晚，建筑能耗越大。这是由于目前引入数据库的商用建筑中，随着社会经济水平的发展，近几年建成的商用建筑大多功能齐全、舒适性高，能耗相对较高。而较早建成的商用建筑由于对舒适性的要求不是很高，各种用能设备不是很齐全，因此能耗相对较低。

自变量“OFFICE”和“COMMERCIAL”与因变量负相关，由于本课题将商用建筑中的可出租区域分为了办公、商场、宾馆这三种类型，所以它们之和近似为 1，因此负的回归系数反映了宾馆面积比例高的商用建筑能耗较高。自变量“CAPACITY”也与因变量为正相关，显而易见，制冷机装机容量大的商用建筑能耗相对较高。

由于一般回归方程中的回归系数都有单位，所以不能用回归系数的绝对值大小来比较自变量对因变量的作用大小。在图 4.8 的输出结果中给出了各个自

变量的标准化偏回归系数值，这一指标反映了自变量对因变量的贡献大小。由计算结果可知，各个自变量对因变量的贡献（即各个因素对建筑能耗的影响程度）由大到小依次为：“BUILTTIME”、“OFFICE”、“COMMERCIAL”、“CAPACITY”。

5.1.2 回归模型 II 的解释与说明

由图 4.11 回归模型 II 的输出结果可以看出，自变量“CAPACITY”的回归系数为正，表明自变量“CAPACITY”与因变量正相关，即制冷机装机容量大的商用建筑能耗较高。

自变量“ACTYPE”的回归系数为正，表明采用混合型空调系统的商用建筑比采用单一空调系统的建筑能耗要大，但这并不是指混合型的空调系统比单一的空调系统能耗大。由于本数据库里还未包含有关空调系统具体情况的信息，还不能得出各种空调系统形式对建筑能耗的影响规律，本课题的数据挖掘过程只能根据数据库中现有的数据进行分析计算。在本数据库中的商用建筑里，采用多种空调系统的商用建筑一般都是近几年建成的大规模、档次较高、功能齐全的楼宇，它们的全年能耗相对较高。

自变量“BAS”的回归系数也为正，它与因变量正相关。也是由于前面提到的原因，在本数据库中的商用建筑里，采用楼宇自控系统的建筑一般都是近几年新建成的大规模的楼宇，它们的运行能耗相对较大。并且如果运行使用不当楼宇自控系统不一定能实现节能的目标。自变量“ACTIME”也与因变量为正相关，每天的空调平均运行时间越长，建筑能耗相应也越大。

需要说明的是，数据挖掘是以数据库中的数据为基础进行的分析计算，目前数据库中的数据量还不是很多，得到的规律和结果还只适用于一定的范围。随着数据库的不断完善，数据量不断扩大，数据挖掘的结果将会更符合客观规律。

同回归模型 I 一样，通过计算结果可以得出各个自变量的标准化偏回归系数值，如图 4.11 所示。比较这一数值可以得到各个自变量对因变量的贡献由大到小依次为：“ACTIME”、“CAPACITY”、“ACTYPE”、“BAS”。

5.2 回归模型 I、II 的比较

从回归分析的结果来看，回归模型 I 的各项指标均好于回归模型 II，具体比较见表 5.1 所示。另外，从图 4.8 和图 4.11 的回归结果中也可以看出，回归模型 I 中的各个自变量的回归系数比回归模型 II 的更为显著。

表 5.1 回归模型各项指标的比较

评价指标	R^2	修正后的 R^2	误差的均方根	显著性水平
回归模型 I	0.9443	0.9394	300.8	0.2
回归模型 II	0.9153	0.9079	370.8	0.25

5.3 回归模型 I、II 的验证

由于本文所做的数据挖掘研究是以上海市商用建筑信息数据库的基本数据为基础的，得到的回归方程也是以这些数据为依据，因此要验证这两个回归模型就必须用数据库以外的“新鲜”数据，以保证结果的可信度。

首先以上海市的一幢综合性的写字楼作为案例来验证两个回归模型。笔者曾为该大楼做过详细的能耗分析和计算机模拟，因此掌握该建筑各方面的信息。按照两个回归模型的自变量构成，该商用建筑的基本信息为：“BUILTTIME”项为 1996,“OFFICE”项为 0.83,“COMMERCIAL”项为 0.1,“CAPACITY”项为 7735.2,“ACTIME”项为 11,“ACTYPE”项为 3,“BAS”项为 0，该商用建筑单位面积全年一次能耗值为 1477.47。将自变量取值代入两个回归方程分别进行计算，与实际能耗值和计算机模拟的能耗进行比较就可验证回归方程的可靠性。具体计算结果如表 5.2 所示：

表 5.2 回归模型的验证比较

比较项目	建筑实际值	回归模型 I	回归模型 II	计算机模拟
一次能耗值（MJ/m²·y）	1477.47	1129.52	1020.97	1656.68
相对误差（%）	0	-23.55%	-30.9%	12.13%

由表 5.2 的计算结果可以看出，两个回归模型的预测结果与该建筑实际的建筑能耗有较大误差，但本课题的研究重点是对数据挖掘这一技术方法的研究，

在目前的研究阶段这个结果是可以接受的。随着数据量的不断增多，模型的预测效果将会有所提高。

从表 5.2 的比较结果中还可以看出，回归模型 I 要略好于回归模型 II，它的预测结果与实际的能耗数据误差更小。另外，与采用数据挖掘方法得到的两个回归模型相比，计算机模拟得到的建筑能耗值更接近实际情况，充分体现了计算机模拟技术的优势。但该模拟结果是经过了多次模型校正后得到的，并且计算机模拟的建模过程复杂、通用性差，对专业技能要求较高，其应用有一定的局限性。这也从另一个角度验证了数据挖掘技术在商用建筑能耗预测中的可用性。

为了进一步验证和比较两个回归模型的可靠性，再选择两座上海市的商用建筑来做验证。具体的计算结果比较如表 5.3 所示：

表 5.3 两个实例的验证比较

	建筑实际一次能耗值（MJ/㎡•y）	回归模型 I	相对误差（%）	回归模型 II	相对误差（%）
实例一	2392.55	1775.25	-25.8%	1444.26	-39.64%
实例二	958.59	1171.54	22.22%	1282.11	33.75%

通过这两个实例的结果对比，两个回归模型的预测效果比较稳定，回归模型 I 的预测效果比回归模型 II 要好。

5.4 模型评价

由前述几节对模型的解释、说明、比较和验证可以看出，这两个回归模型都具有一定的可靠性，但无论是从回归分析的评价指标还是模型用于预测建筑能耗的准确性等方面，回归模型 I 都比回归模型 II 效果好，可以认为其更“优”，因此将回归模型 I 作为本研究阶段的最终的商用建筑能耗预测模型。

通过数据挖掘方法得到的回归模型 I，经过验证基本能够反映出建筑能耗的情况，能够实现预测的功能。图 5.1 所示为用回归模型 I 预测的能耗值与建筑实际能耗值的相对误差，由该图可以看出，由该模型预测的能耗值相对误差大多在±20%之间，还是可以接受的。

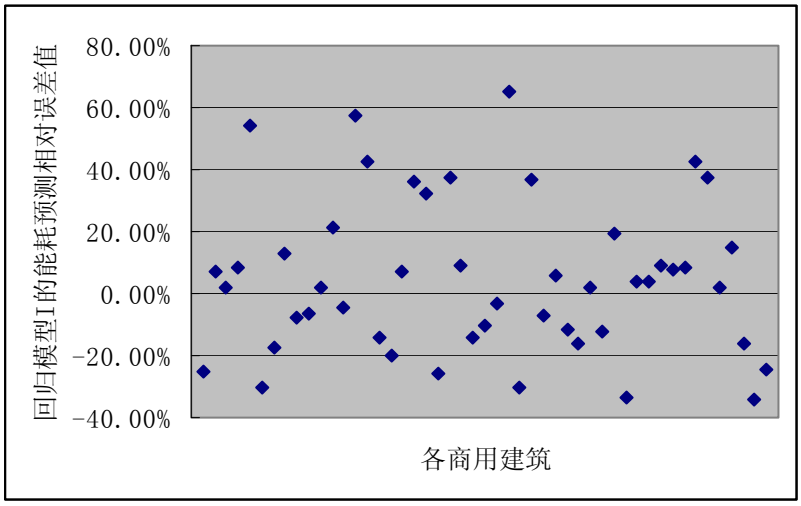

图 5.1 回归模型 I 能耗预测的相对误差

经过数据挖掘得到的回归方程形式简单，只需要四个自变量就可以基本反映出因变量的信息。但是，这并不是指没有引入回归方程的自变量就对建筑能耗没有影响。如气候、地域等因素，为了研究的方便，本课题在目前的研究阶段所建立的数据库仅仅是面对上海地区的商用建筑，所以暂时排除了这些因素对建筑能耗的影响。又如在回归模型的数据挖掘过程中剔除的自变量，都对建筑能耗有影响，但由于现阶段数据库中数据量并不足够多，经数据挖掘计算后这些自变量对因变量的影响并不显著，故将其排除出回归方程。

尽管回归模型 II 的准确性稍差，但并不能说明借助主成分分析方法建立回归模型不可行。在本例中，回归模型效果的好坏还受到数据数量和质量等因素的制约，并且主成分分析的方法能够有效地筛选变量，减少回归分析的计算量，提高数据挖掘的效率，因此有其继续探索和研究的价值。

本课题现阶段的研究重点是针对用数据挖掘技术来研究建筑能耗这样一种方法的探索和尝试，本文的研究为后续的研究工作提供指导策略和参考依据，因此目前得到的商用建筑能耗预测模型还只是阶段性成果。并且，由于数据量尚不够充足，目前的数据挖掘研究还不能准确地得到影响建筑能耗的规律。相信经过后续研究的努力，完善数据库、充实数据量，定会取得进一步的研究成果。

第 6 章 结论与展望

节能是我国社会经济发展需要解决的当务之急的重任，由于建筑能耗在我国能源消费中的比重越来越大，建筑节能成了节能工作中的重中之重。经过几年来全社会各个方面的共同努力，节能省能的意识已经深入人心了，并且也取得了可喜的成果。本文将数据挖掘技术应用到商用建筑能耗的分析和预测研究中，得到了一些很有价值的结论：

（1）本文首先以商用建筑为试点，建立了上海市商用建筑信息数据库。该数据库包含了基本信息和历史能耗数据两部分，通过调查问卷的方式收集了 95 幢商用建筑的信息作为数据库的基本数据。

通过对基本数据的整理和统计，得到了上海市商用建筑的一些基本状况，并着重分析了空调冷热源、空调系统、建筑能耗等方面的规律和特点。为改进和完善信息数据库提供了指导，并为数据挖掘研究提供了数据基础。

（2）数据挖掘是一个多学科交叉的新兴研究领域，它是信息技术发展到一定程度的必然产物，是利用积累数据的一个高级阶段。它把人们对数据的应用从低层次的简单查询提升到从数据中挖掘知识，提供决策支持。在这个新兴领域中，汇集了来自机器学习、模式识别、数据库、统计学、人工智能以及管理信息系统等各学科的成果，多元化的投入使得这一学科得以蓬勃发展，而且已初具规模。

本文简要介绍了数据挖掘技术的产生背景和发展历程，以及主要的技术方法，还介绍了数据挖掘技术与统计学、机器学习等相关研究领域的联系和区别，最后详细介绍了数据挖掘的过程和常用工具。通过全面的介绍和分析，明确了数据挖掘技术在暖通空调领域的可用性，可以将数据挖掘技术应用于建筑能耗的分析和预测研究中。

（3）使用 SAS 统计分析软件的 EM 模块（Enterprise Miner），按照 SEMMA 的数据挖掘方法论，完成了商用建筑能耗预测模型的数据挖掘流程。

经过数据预处理、数据转换等步骤形成了可供数据挖掘使用的数据集；使用多元线形回归分析方法进行数据挖掘计算，通过对计算结果的分析，校正模型、调整计算参数，得到了回归模型 I；采用主成分分析的方法筛选自变量，剔

除对因变量影响较小的变量后再进行回归分析，得到回归模型 II。

经过对两个回归模型的解释、说明、比较和验证，得出回归模型 I 是 “最优”模型，可以作为本研究阶段的商用建筑的能耗预测模型。

（4）通过本课题的研究，验证了数据挖掘技术在大量数据的处理和分析方面有其独特的优势，可以挖掘出隐含在数据背后潜在的、不易发现的、又具有很高价值的规律和信息。本文首次地将数据挖掘技术应用到商用建筑能耗分析和预测的研究领域，按照完整的数据挖掘过程建立了商用建筑能耗的预测模型，取得了很好的效果，得到了一个回归模型。通过本文的探索性研究，为数据挖掘技术在本专业领域，尤其是建筑能耗的分析和预测方面的应用提供了参考和借鉴。

本课题的研究取得了令人满意的效果，但是也由于主、客观方面的原因还存在一些问题和不足，需要后续研究来完善和改进：

（1）数据库的数据量和变量个数需要进一步扩充。通过本课题的研究可以看出，数据质量和数量是影响数据挖掘结果的关键因素。在后续研究中应增加数据库中的自变量个数，更全面地分析影响建筑能耗的因素；还应扩大数据库容量，以便更明显地反映数据中隐含的规律，样本数量越大客观规律就越显著。

（2）由于各方面条件不成熟，在目前的研究阶段还只能以商用建筑全年的一次能耗值为目标变量，这样就掩盖了建筑能耗逐月的变化规律，忽略了气候变化、空调系统运行策略等因素对建筑能耗的影响。因此需要在后续研究中考虑这一问题，收集相关的资料和数据，寻找合适的数据处理方法，以便更准确地预测建筑能耗。

（3）学习和研究更多的数据挖掘算法。数据挖掘技术是多种多样的，各有优、缺点，各有其适用场合。继续学习和尝试各种数据挖掘方法，寻找更适合商用建筑能耗分析和预测的数据挖掘模型。

（4）建立完整的商用建筑能耗分析和预测系统。将数据挖掘与数据库建立联系，直接利用数据库的信息进行数据挖掘，形成一个完整的基于数据挖掘技术的能耗预测系统，将数据挖掘的结果应用到实践中。

致谢

两年半的学习和科研工作汇集成这本论文文稿，给我的硕士研究生生活画上了句号。在毕业论文即将截稿之际，回顾在同济大学的这段读书生活不禁感慨万千，其中最大的收获便是自己各个方面的成熟和进步，以及对社会的了解和认识，同时也对今后的工作和学习信心十足。

在此我要特别地感谢我的导师潘毅群老师几年来对我的关心和培养，无论是在专业学习、课题研究还是日常生活等各个方面，潘老师都给了我无私的指导和帮助，并以其严谨、认真的治学态度和工作作风深深影响着我。同时，我也要感谢同济大学中德工程学院的黄治钟老师在我的整个论文研究工作中给我的建议和帮助，黄老师对待学术一丝不苟的精神和渊博的学识都是我学习的榜样。

最后，我还要感谢我的父母、家人、同学、朋友多年来对我的关心、理解和支持，这些都是我永远的财富。

2006 年 3 月 于同济

参考文献

[1] 谢浩，徐宇龙，张伦琳.建筑节能问题.住宅科技，2002，12:31-34
[2] 罗莹.建筑节能与设计.江西能源，1998（4）:32-35
[3] 2004 中国统计年鉴
[4] 曹叔维，初春玲.建筑能耗的综合性指标.节能，1999，12:10-12
[5] 涂逢祥，王庆一.我国建筑节能现状及发展.新型建筑材料，2004，7:40-42
[6] 胡平放，向才旺等.中国建筑能耗现状特征.武汉城市建设学院学报，1998，15(2):39-43
[7] 于涛，方修睦等.多层建筑能耗分析软件的开发与应用.暖通空调，2003，33（3）:87-89
[8] 陈华，涂光备，陈红兵.建筑能耗模拟的研究和进展.洁净与空调技术，2003，3:5-9
[9] 杨嘉，吴祥生，张锦松.建筑能耗模拟的应用与发展.云南建筑增刊，2002:112-115
[10] 方修睦，张道军等.采暖居住建筑物节能评价软件的开发.哈尔滨工业大学学报，2003，35（11）:1301-1306
[11] 寿炜炜，谭良才.HDY-SMAD空调负荷计算及分析软件的开发与应用特点.空调暖通技术，2002，4:9-13
[12] 王清勤.加拿大建筑能耗模拟软件和建筑节能规范实施软件.建筑科学，1997，4:59-60
[13] 何斌，王民.建筑能耗模拟计算软件ENER—WIN 9702简介及分析.建筑热能通风空调，2001，20（4）:38-39
[14] 周良，张国强等.建筑节能及暖通空调仿真软件的现状和发展.湖南大学学报，2000，27（6）:103-108
[15] 苏华，王靖.建筑能耗的计算机模拟技术.计算机应用，2003，23（12）:411-413
[16] 侯余波，付祥钊，郭勇.用DOE-2程序分析建筑能耗的可靠性研究.暖通空调，2003，33（3）:90-92
[17] 吴明，连之伟.变风量空调系统模拟及能耗研究.节能，2003，8:10-13
[18] 潘毅群，吴刚，Volker Hartkopf.建筑全能耗分析软件EnergyPlus及其应用.暖通空调，2004，34（9）:2-7
[19] 邓宇春，陈锋，江亿.建筑热环境设计模拟工具包DeST1.0介绍.全国暖通空调制冷2000年学术年会资料集，2000
[20] 石磊，李兆东，刘伟.冰蓄冷空调系统模拟优化技术探综.建筑热能通风空调，2003，2:37-38
[21] 龚波.自然通风的设计策略及模拟分析.建筑热能通风空调，2004，23（5）:30-34
[22] 朱光俊，张晓亮.住宅建筑采暖空调能耗模拟研究.住宅科技，2004，11:35-37
[23] 梁珍，赵加宁，路军.公共建筑能耗主要影响因素的分析.低温建筑技术，2001，85（3）:52-54
[24] 梁珍，程继梅，徐坚.商场建筑能耗主要影响因素及节能分析.节能技术，2001，19（3）:17-19

[25] 周巧航，赵加宁，施雪华.深圳市某办公楼空调系统节能潜力分析.
[26] 魏玲，何嘉鹏，阎丽萍.中国东部窗户能耗模拟及节能分析.南京工业大学学报，2002，24（4）:70-73
[27] 左现广，唐鸣放.国内外建筑能耗调查与统计研究.重庆建筑，2003，2:16-18
[28] 涂逢祥等. 英、法、德三国建筑节能标准近期发展. 建筑节能，2002，37:131-138
[29] The Ove Partnership. Building Design for Energy Economy，by The Pitman Great Britain，1980:101-105
[30] Oktay Ural. Energy Resources and Conservation Related to Built Environment，Volume1. Miami beach，Florida，1980:365-369，598-611
[31] Dennis R. Landsberg，Ronald Steward. Improving Energy Efficiency In Buildings. State University of New York Press. Albany. 1980:51-56，290-321
[32] 俞英鹤.哈尔滨市民用建筑物能耗统计方案的编制与数据库开发:[硕士学位论文].哈尔滨：哈尔滨工业大学，2003
[33] 梁珍.城市民用建筑能耗模拟和统计方案设计:[硕士学位论文].哈尔滨：哈尔滨工业大学，2001
[34] 武建勋.空调建筑统计能耗模型探讨.暖通空调，1996，6:13-16
[35] 涂逢祥等.《建筑节能经济技术政策研究》. 北京：中国建筑工业出版社，1 991年5月：2-3
[36] 龙惟定，潘毅群，范存养等. 上海公共建筑能耗现状及节能潜力分析[J]. 暖通空调，1998，28（6）:13-16
[37] 武海斌等. 北京市城市居民家用空调器耗电量的调查研究. 暖通空调新技术，2000，2:52-56
[38] 翟超勤等. 全国家用电器年耗电量的估算. 暖通空调新技术，2000，2:52-56
[39] 建设部科技司发展促进中心、哈尔滨工业大学. 建筑物能耗统计方案及其数据库软件编制，2002年9月
[40] 鲍士雄，姚峻，蔡立群等.暖通制冷领域专家系统的国外发展概况.制冷，1996，2:28-32
[41] John Kelly Kissock, Jeff S.Haberl, David E.Claridge. Inverse Modeling Toolkit - Numerical Algorithms. ASHRAE Transactions, Vol.109, Part 2
[42] Robert C.Sonderegger. A Baselining Model for Utility Bill Analysis Using Both Weather and Non－Weather Related Variables. Baseline Calculations For Measurement And Verification Of Energy And Demand Savings, ASHRAE Summer Meeting, Toronto, Canada, June 18-25, 1998
[43] Satkartar Kinney, Mary Ann Piette. California Commercial Building Energy Benchmarking Final Project Report. California Energy Commission Public Interest Energy Research Program
[44] 赵丹群.数据挖掘原理方法及其应用.现代图书情报技术，2000，6:41-44
[45] 丁夷.数据挖掘—技术与应用综述.西安邮电学院学报，1999，4（3）:41-44
[46] 于之虹，郭志忠.数据挖掘与电力系统.电网技术，2001，25（8）:58-62
[47] Christoph Morbitzer, Paul Strachan, Catherine Simpson. Application of data

mining techniques for building simulation performance prediction analysis. Eighth International IBPSA Conference, August 11-14, 2003
[48] 倪德良.上海建筑节能现状与对策.新型建筑材料，1998，8:7-10
[49] 杨星虎，倪德良.建筑节能势在必行.上海建设科技，1998，2:40-41
[50] 谢仲华，龙惟定.建筑采暖空调能耗与节能潜力分析.能源技术，2000，3:158-163
[51] 赵志明，邓又明，王辉艳.浅谈数据挖掘技术原理及应用.吉林省经济管理干部学院学报，2004，18（4）:52-53
[52] 叶家成.数据挖掘若干方法的研究:[硕士学位论文].广州：华南理工大学，1999
[53] 伍星，陈进，李如强等.基于数据挖掘的设备状态监测和故障诊断.振动与冲击，2004，23（4）:70-74
[54] 陈林生，邱锦平.掘胜千里KDD.互联网周刊，2000，1:44
[55] Learner E E，Specification Searches. Ad hoc Inference with Nonex－perimental Data. Wily，New York，1978
[56] 戴稳胜，张阿兰，谢邦昌.数据挖掘的方法、流程及应用.中国统计，2004，7:53-54
[57] 王洪元，史国栋等.数据挖掘技术在故障诊断中的应用.江苏石油化工学院学报，2001，13（4）:42-44
[58] 张柏礼，孙志辉.数据挖掘技术在能量管理系统中的应用.工业控制计算机，2002，15（12）:1-2
[59] 肖攸安，李腊元.数据挖掘与知识发现的理论方法及技术分析.交通与计算机，2002，20（1）:57-61
[60] 林建勤，林筑英.数据挖掘与智能化信息处理研究.贵州大学学报，2003，20（3）:294-298
[61] 解正安.数据挖掘技术及应用研究:[硕士学位论文].南京：东南大学，2004
[62] 陈伟志,魏振军,王春迎.多元统计分析在数据挖掘中的作用.信息工程大学学报，2003，4（4）:22-25
[63] 行小帅,焦李成.数据挖掘的聚类方法.电路与系统学报.Vol.8 No.1 February 2003 P59-67
[64] 丁学钧，杨克俭等.数据挖掘中聚类算法的比较研究.河北建筑工程学院学报，2004，22（3）:125-127
[65] 邵华.数据挖掘技术及应用.软件工程师，2000，1:62-64
[66] 夏幼明,解敏,周雯.数据挖掘方法分析与评价.云南师范大学学报,2003,23(2):7-16
[67] 谭立云，高学东，武森.数据挖掘方法与应用.华北科技学院学报，2004，1(2):52-55
[68] 朱世武，崔嵬，张尧庭等.数据挖掘与其他技术的比较.统计研究，2003，7:58-61
[69] 钟智，尹云飞.数据挖掘与人工智能技术.河南科技大学学报，2004，25（3）:44-47
[70] 郑宏珍，柳明欣.数据挖掘及其工具的选择.计算机应用，1999，19（10）:109-110
[71] 刘世平，姚玉辉.数据挖掘工具的评判.数字财富，2003，6:74-76
[72] 凌云.SAS编程入门.SAS广州办事处
[73] Getting Started with SAS Enterprise Miner 4.3. SAS Institute Inc., Cary, NC, USA, 2004

[74] 贾琳，李明.基于数据挖掘的电信客户流失模型的建立与实现.计算机工程与应用，2004，4:185-187
[75] 候晓智.基于数据挖掘技术的上海市肝胆肿瘤病例住院费用研究:[硕士学位论文].上海：第二军医大学，2004
[76] 殷峻.一个基于SEMMA 的数据挖掘应用实例.冶金自动化，2003，3:5-7
[77] 吴修霆.SAS数据挖掘技术的实现.计算机世界，2000，14:44-45
[78] 许雷，范存养，龙惟定.上海高层建筑空调冷热源装置与能耗调查.制冷技术，1998，1:4-9
[79] 岳朝龙，黄永兴，严忠.SAS系统与经济统计分析.合肥：中国科学技术大学出版社，2003
[80] 金新政，胡彬等.SAS for Windows 统计系统教程. 武汉：华中科技大学出版社，2001

附录 部分数据挖掘计算结果

The PRINCOMP Procedure

Observations 50
Variables 10

Simple Statistics

	BUILTTIME	FILM	COMMERCIAL	HOTEL	CAPACITY
Mean	1997.881800	0.5400000000	0.0964000000	0.0472000000	7264.004000
StD	2.713030	0.5034574339	0.1005446393	0.1729343393	6772.244180

Simple Statistics

	COOLTYPE	HEATTYPE	ACTYPE	BAS	ACTIME
Mean	1.160000000	1.720000000	3.320000000	0.6200000000	11.24000000
StD	0.765586343	0.858094710	1.878340562	0.4903143515	2.97479205

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.94634918	0.19595469	0.1946	0.1946
2	1.75039449	0.23823234	0.1750	0.3697
3	1.51216215	0.15880797	0.1512	0.5209
4	1.35335418	0.43495809	0.1353	0.6562
5	0.91839608	0.10632188	0.0918	0.7481
6	0.81207420	0.22938049	0.0812	0.8293
7	0.58269371	0.01465089	0.0583	0.8875
8	0.56804282	0.22165643	0.0568	0.9443
9	0.34638639	0.13623959	0.0346	0.9790
10	0.21014680		0.0210	1.0000

Eigenvectors

		Prin1	Prin2	Prin3	Prin4	Prin5
BUILTTIME	BUILTTIME	-.289517	0.561860	-.032509	0.149886	0.153216
FILM	FILM	0.327664	-.264863	0.527886	0.037431	0.066683
COMMERCIAL	COMMERCIAL	-.137519	-.084806	-.234594	0.408853	0.791652
HOTEL	HOTEL	0.519176	0.167570	-.444017	-.046231	0.010616
CAPACITY	CAPACITY	-.077680	0.569839	0.050610	-.120432	-.103456
COOLTYPE	COOLTYPE	0.030371	0.007766	0.154763	0.673827	-.261239
HEATTYPE	HEATTYPE	0.186270	0.153340	0.117792	0.558333	-.221146
ACTYPE	ACTYPE	0.377157	0.072087	0.364357	-.096483	0.442735
BAS	BAS	0.071028	0.436054	0.448728	-.125188	0.145124
ACTIME	ACTIME	0.575688	0.187775	-.305931	0.039635	-.017949

Eigenvectors

	Prin6	Prin7	Prin8	Prin9	Prin10
BUILTTIME	0.226864	0.056597	-.253795	0.659063	-.007020
FILM	0.178577	-.154418	0.431782	0.492287	0.231203
COMMERCIAL	-.016497	-.143220	0.282003	-.149623	0.047869
HOTEL	0.145961	0.039143	-.128690	-.024559	0.680422
CAPACITY	-.349888	0.235653	0.659574	-.110428	0.127653
COOLTYPE	0.384425	0.497655	0.051022	-.233404	0.015897
HEATTYPE	-.571314	-.443649	-.199930	0.039645	0.050056
ACTYPE	-.341193	0.518652	-.343475	-.019249	-.103532
BAS	0.391580	-.416335	-.129150	-.472244	0.018947
ACTIME	0.169456	-.092492	0.201984	0.093118	-.671639

The PRINCOMP Procedure

Observations 50
Variables 9

Simple Statistics

	BUILTTIME	FILM	COMMERCIAL	CAPACITY	COOLTYPE
Mean	1997.881800	0.5400000000	0.0964000000	7264.004000	1.160000000
StD	2.713030	0.5034574339	0.1005446393	6772.244180	0.765586343

Simple Statistics

	HEATTYPE	ACTYPE	BAS	ACTIME
Mean	1.720000000	3.320000000	0.6200000000	11.24000000
StD	0.858094710	1.878340562	0.4903143515	2.97479205

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.79094701	0.10358608	0.1990	0.1990
2	1.68736093	0.33237964	0.1875	0.3865
3	1.35498129	0.34299005	0.1506	0.5370
4	1.01199124	0.09400410	0.1124	0.6495
5	0.91798713	0.15729936	0.1020	0.7515
6	0.76068778	0.17879435	0.0845	0.8360
7	0.58189343	0.03392184	0.0647	0.9006
8	0.54797159	0.20179198	0.0609	0.9615
9	0.34617960		0.0385	1.0000

Eigenvectors

		Prin1	Prin2	Prin3	Prin4
BUILTTIME	BUILTTIME	0.606108	0.188718	0.148564	-.161234
FILM	FILM	-.532621	0.289946	-.012208	-.332480
COMMERCIAL	COMMERCIAL	0.078203	-.245330	0.424430	0.017703
CAPACITY	CAPACITY	0.446815	0.364523	-.114690	0.240358
COOLTYPE	COOLTYPE	-.055916	0.114574	0.651179	-.323214
HEATTYPE	HEATTYPE	-.053089	0.290630	0.555667	0.215044
ACTYPE	ACTYPE	-.281789	0.450087	-.117203	0.034426
BAS	BAS	0.168736	0.554126	-.164779	-.363841
ACTIME	ACTIME	-.171365	0.282888	0.103906	0.721871

Eigenvectors

	Prin5	Prin6	Prin7	Prin8	Prin9
BUILTTIME	0.137332	0.151921	0.012389	-.275657	0.656407
FILM	0.045109	0.110607	-.079573	0.496435	0.504485
COMMERCIAL	0.794635	0.058285	-.089320	0.298454	-.146110
CAPACITY	-.079339	-.224779	0.345578	0.640337	-.101372
COOLTYPE	-.289872	0.255923	0.499632	-.040902	-.233274
HEATTYPE	-.193755	-.532219	-.481873	-.053592	0.042828
ACTYPE	0.453960	-.363145	0.444480	-.408592	-.025127
BAS	0.114927	0.283268	-.431719	-.054083	-.471364
ACTIME	0.022587	0.591518	-.018296	-.063573	0.062719

The PRINCOMP Procedure

Observations	50
Variables	8

Simple Statistics

	FILM	COMMERCIAL	CAPACITY	COOLTYPE
Mean	0.5400000000	0.0964000000	7264.004000	1.160000000
StD	0.5034574339	0.1005446393	6772.244180	0.765586343

Simple Statistics

	HEATTYPE	ACTYPE	BAS	ACTIME
Mean	1.720000000	3.320000000	0.6200000000	11.24000000
StD	0.858094710	1.878340562	0.4903143515	2.97479205

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.69781033	0.29283051	0.2122	0.2122
2	1.40497983	0.19451952	0.1756	0.3878
3	1.21046031	0.23400877	0.1513	0.5392
4	0.97645153	0.09429246	0.1221	0.6612
5	0.88215907	0.14718020	0.1103	0.7715
6	0.73497887	0.15312980	0.0919	0.8634
7	0.58184908	0.07053810	0.0727	0.9361
8	0.51131097		0.0639	1.0000

Eigenvectors

		Prin1	Prin2	Prin3	Prin4
FILM	FILM	0.464622	0.263670	-.475929	-.148292
COMMERCIAL	COMMERCIAL	-.247444	0.379900	0.138841	0.275676
CAPACITY	CAPACITY	0.172810	-.486880	0.565561	-.027008
COOLTYPE	COOLTYPE	0.139691	0.546122	0.295090	-.448947
HEATTYPE	HEATTYPE	0.296292	0.369581	0.512360	-.008570
ACTYPE	ACTYPE	0.521297	-.015744	-.228034	0.251927
BAS	BAS	0.456430	-.327034	0.062049	-.345946
ACTIME	ACTIME	0.325045	0.084042	0.170043	0.718686

Eigenvectors

	Prin5	Prin6	Prin7	Prin8
FILM	-.010736	0.080358	-.070048	0.674259
COMMERCIAL	0.773258	0.268063	-.077534	0.152891
CAPACITY	0.146369	-.021705	0.364888	0.507404
COOLTYPE	-.133209	0.299260	0.501730	-.185931
HEATTYPE	-.047079	-.524167	-.485353	0.022369
ACTYPE	0.376423	-.392510	0.431823	-.360979
BAS	0.281983	0.449841	-.426344	-.312333
ACTIME	-.373314	0.447756	-.019963	-.040142

个人简历 在读期间发表的学术论文与研究成果

个人简历：

陈晨，男，1980 年 3 月生。

1998 年 9 月至 2002 年 7 月，就读于青岛建筑工程学院（现青岛理工大学），供热、供燃气、通风与空调工程专业，获工学学士学位。

2003 年 9 月至今，就读于同济大学，供热、供燃气、通风与空调工程专业，硕士研究生。

已发表论文：

陈晨，潘毅群.无成本、低成本节能措施在商用建筑中的应用.制冷空调与电力机械,2006,1