数据挖掘技术在上海市商用建筑信息数据库中的应用

上海市房地产科学研究院 郑晓卫☆
同济大学 潘毅群 黄治钟
上海市节能监察中心 楼振飞

摘要 简要介绍了上海市商用建筑信息数据库和数据挖掘技术相关知识，运用回归归因法解决数据库中缺失数据问题，并对异常点进行检测；选择统计学方法进行数据挖掘，并分析评价由数据挖掘得到的商用建筑能耗预测模型。

关键词 建筑能耗 数据库 数据挖掘 数据处理 回归

Application of data mining technology to Shanghai information database of commercial buildings

By Zheng Xiaowei★, Pan Yiqun, Huang Zhizhong and Lou Zhenfei

Abstract Briefly presents the Shanghai information database of commercial buildings and some knowledge about data mining. Applies regression imputation method to solve the problem of missing data in this database, and checks the abnormal data. Selects statistic method to deal with data mining, analyses and assesses the regression model of building energy consumption.

Keywords building energy consumption, database, data mining, data processing, regression

★ Shanghai Real-Estate Science Research Institute, Shanghai, China

0 引言

建筑能耗是各国节能工作关注的重点之一。关于建筑能耗，国内外的专业人士进行了不懈的探索，如进行建筑能耗调查统计，开发建筑能耗模拟计算软件等。但目前建筑能耗统计调查中的信息没有被充分利用，本文将结合建筑能耗统计调查结果，建立上海市商用建筑信息数据库，运用数据挖掘技术找出数据库中隐含的信息，分析建筑能耗与建筑面积、建筑功能、冷热源等因素之间的内在联系，尝试建立商用建筑能耗预测模型。

1 数据库简介

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建筑。在我国，商用建筑能耗占总建筑能耗的 1/3 左右[1]。据统计，一般公共建筑单位建筑面积能耗大约是普通居住建筑的 5 倍，而大型高档商用建筑单位建筑面积能耗是普通居住建筑的 10～15 倍左右[2]。商用建筑堪称耗能大户，必须加以重视。在上海，商用建筑的数量比较多，是比较典型的公共建筑，因此笔者首先尝试建立上海市商用建筑信息数据库。

上海市商用建筑信息数据库主要包括建筑基本信息和建筑能耗两部分内容。建筑基本信息主要包括用户基本信息、建筑围护结构信息、空调系统信息及其他信息。建筑能耗数据则包括该建筑近几年来全年各月的电、气、油等能源的使用情况。由于上海市商用建筑信息数据库还处于完善阶段，用户还无法在线填写和查询信息，因此笔者通过现场统计调查方式收集上海市商用建筑信息。目前，笔者已经收集了 95 栋商用建筑的基本信息数据和建筑能耗数据。这些商用建筑是随机抽取的，涵盖了上海市各个城区、各种样式的商用建筑，有一定的代表性。随着将来现场调查的

☆ 郑晓卫，男，1980 年 10 月生，硕士研究生，工程师
200031 上海市复兴西路 193 号上海市房地产科学研究院工程所
(021) 64718289 - 209
E-mail: michael_zxw@163.com
收稿日期：2007-06-13
修回日期：2008-02-22

深入和用户在线填写工作的开展，数据库中数据量将越来越大，建立上海市商用建筑信息数据库的意义也将更加深远。

2 数据挖掘和数据处理介绍

数据挖掘是从大量的、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。数据挖掘是知识发现的一个处理过程，是知识发现的最核心部分[3]。知识发现与数据挖掘的关系如图 1 所示。在知识发现流程图中，数据处理（包括数据集成、预处理、数据转换等）对数据挖掘起着重要的作用。它是数据挖掘工作能顺利开展的前提。在实际的数据挖掘过程中，数据处理过程占用的时间大约为 80%[4]。目前上海市商用建筑信息数据库中建筑信息较少，存在数据缺失和异常点等问题，数据处理的作用更加突出。

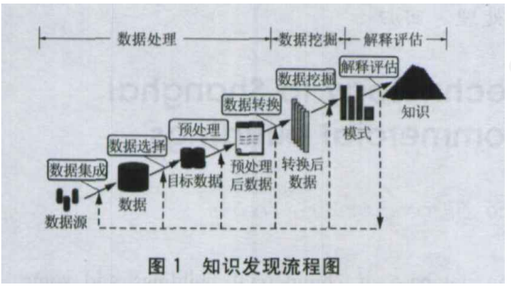

图 1 知识发现流程图

大部分数据挖掘方法都基于机器学习、模式识别、统计学等领域知识，主要有神经网络方法、粗集方法、决策树法、遗传算法、统计学方法等[5]。缺失数据处理方法有样本丢弃法、参数估计法、均值/众数归因法、回归归因法、多重归因法等[6]。异常点的检测方法主要有基于统计模型、基于距离、基于密度模型及基于偏离模型的异常点检测方法[7]。其他的数据处理过程如数据集成、数据变换、数据规约等在海量数据处理时作用更加明显，能起到减少数据量和维数、使数据挖掘工作更加有效的作用。

3 基于上海市商用建筑信息数据库的数据挖掘过程

基于上海市商用建筑信息数据库的数据挖掘过程将通过 SAS 软件的 Enterprise Miner（SAS/EM）模块实现。其流程按照 SAS 协会定义的数据挖掘方法——SEMMA 方法，即抽样（Sample）、探索（Explore）、修改（Modify）、建模（Model）、评估（Assess）紧密结合[8]，逐步建立完整的数据挖掘过程，如图 2 所示。

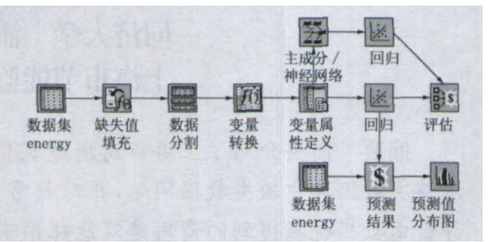

图 2 数据挖掘流程图

3.1 变量定义

在上海市商用建筑信息数据库的基础上，数据挖掘的目标是找出单位面积建筑年一次能耗与其他建筑参数之间的内在联系，并尝试建立商用建筑能耗模型。因此，对数据库中建筑信息数据进行初步整理，建筑年龄、建筑面积、办公比例、商业比例、宾馆比例、玻璃传热系数、玻璃镀膜情况、冷源装机容量、制冷一次能源利用率、热源装机容量、供热一次能源利用率、楼宇自控系统、空调运行时间、建筑年一次能耗 14 个变量进入数据挖掘过程。变量定义如表 1 所示。

在数据挖掘过程中，数据集分成训练样本集、验证样本集、测试数据集。但由于数据库中数据量不够充足，为了保证数据挖掘的效果，将全部数据作为训练样本集。

3.2 缺失数据处理

数据库中14个变量的数据缺失情况见表2。其中，玻

表 1 数据挖掘变量表

变量	变量名	变量类型	备注
建筑年龄	BUILTTIME	数值型，连续变量	按截止到 2004 年底算起
建筑面积	AREA	数值型，连续变量	
玻璃传热系数	WINU	分类变量	单层取 6.4 W/(m² · K)，双层取 3 W/(m² · K)
镀膜情况	FILM	0—1 型变量	0 表示无镀膜，1 表示有镀膜
办公比例	OFFICE	数值型，连续变量	皆为百分数，三者之和小于 1
商业比例	COMMERCIAL	数值型，连续变量	皆为百分数，三者之和小于 1
宾馆比例	HOTEL	数值型，连续变量	皆为百分数，三者之和小于 1
冷源装机容量	CCAPACITY	数值型，连续变量	各种类型的制冷机装机容量之和
制冷一次能源利用率	CPER	数值型，连续变量	统一为一次能源；*COP* 取值：活塞式为 3.75，螺杆式为 5.1，离心式为 5.13，热泵为 2.8，直燃型溴化锂式为 1.17
热源装机容量	HCAPACITY	数值型，连续变量	各种类型的热源装机容量之和
供热一次能源利用率	HPER	数值型，连续变量	统一为一次能源；热效率：燃油锅炉 90%，燃气锅炉 90%，燃煤锅炉 78%，直燃型溴化锂式 90%；热泵供热 *COP* 取 3.4
楼宇自控系统	BAS	0—1 型变量	0 表示无，1 表示有
空调运行时间	ACTIME	数值型，连续变量	按运行时间最长的空调系统取值
建筑年一次能耗	ENERGY	数值型，连续变量	建筑年能耗以一次能耗表示

表 2　数据缺失情况统计表

变量名称	建筑年龄	建筑面积	玻璃传热系数	玻璃镀膜情况	办公比例	商业比例	宾馆比例
原始数目	89	94	69	40	84	85	89
缺失情况	6	1	26	55	11	10	6
变量名称	冷源装机容量	制冷一次能源利用率	热源装机容量	供热一次能源利用率	楼宇自控系统	空调运行时间	建筑年一次能耗
原始数目	88	89	84	86	95	72	79
缺失情况	7	6	11	9	0	23	16

璃镀膜情况缺失最为严重，这是人为因素造成的，由于普遍认为采用镀膜玻璃，故没有在调查表中详细注明。

目前统计学领域比较流行的缺失数据处理方法有均值/众数归因法、回归归因法、多重归因法等。均值/众数归因法对连续性变量用样本均值填补，对于 0－1 变量或其他分类变量用样本中的高频值（即众数）替换缺失值。回归归因法是根据变量之间的线性关系，用其他自变量预测因变量，填充空缺值。多重归因法针对不同的数据缺失模式提供不同的方法填补缺失数据[9]。笔者分析发现，均值/众数归因法和多重归因法在数据量有限和缺失数据较多的情况下，填补的缺失数据误差大，对数据挖掘结果影响明显，而且在数据库中，冷、热源装机容量与数据库中其他因素（不包括建筑能耗）存在函数关系。因此笔者在继续调查和完善数据库中缺失数据的同时，选择回归归因法建立冷、热源装机容量回归模型，填补冷、热源装机容量的缺失值。

通过现场调查和查阅相关文献资料填补缺失数据后，热源装机容量、建筑年一次能耗还存在缺失数据。其中，热源装机容量有 6 个缺失数据，建筑年一次能耗有 15 个缺失数据。因此，这里只需用回归归因法建立热源装机容量回归模型。其中，建筑年一次能耗数据缺失的样本将不进入回归分析过程。

对含有 80 个样本的数据集进行回归分析，笔者认为因变量 HCAPACITY 与 AREA 等变量之间呈非线性关系。分别用逐步回归法、主成分分析、Mallow's Cp 统计量法[10]筛选变量，最终发现逐步回归法的回归模型拟合较好。逐步回归方法的 F 检验值和 T 检验值都小于 0.15，满足显著性水平。修正后的 R^2 值为 0.9947，接近于 1。逐步回归方法得到的热源回归方程为

$$LN(HCAPACITY) = 0.737LN(AREA) + 0.059(WINU) + 0.822(COMMERCIAL)$$

回归方程中的自变量有 AREA，WINU，COMMERCIAL。建筑面积越大，单位时间供热量将越大，因此建筑面积与热源正相关。玻璃传热系数影响建筑室内负荷，玻璃传热系数大，围护结构耗热量大，设计热源时必将选择更大或更多的热源设备。建筑中的商业比例影响建筑的峰值热负荷。变量 AREA，WINU，COMMERCIAL 进入回归方程是合理的。

用此回归方程预测数据库中 89 栋商用建筑的热源装机容量，得到预测值与实际值的相对误差，见图 3。由图 3 可知，该回归方程预测的热源装机容量值的相对误差多数

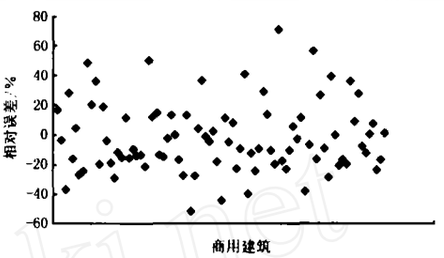

图 3　热源装机容量预测值相对误差

在±20%之间。因此在数据量有限的情况下，回归方程是可以接受的。用热源装机容量回归方程预测并填补 6 个缺失值，得到只含有建筑年一次能耗缺失值的数据集。

3.3　异常点检测

在 SAS 软件中，异常点检测的统计量有库克距离统计量 COOKD、删除学生化残差 SRE(i)等[11]。一般来说，当 COOKD>50%或 |SRE(i)| >3可以认为第 i 个观测值为异常点。笔者用基于统计模型的异常点检测方法查找数据集中的异常点，通过编程可得到每个建筑样本的 COOKD 和 SRE(i)值。经分析，正大广场和上海香港广场两个样本的|SRE(i)|值大于 3。正大广场建筑面积 247 000 m^2，冷源装机容量 45 708 kW，大约是建筑面积相近的其他商用建筑冷源装机容量的 2 倍。正大广场是纯商业建筑，而数据库中其他建筑多为办公建筑，故正大广场样本是异常点，应剔除。上海香港广场单位面积年一次能耗7 546.3 MJ/(m^2 · a)，远远大于其他商用建筑，因此上海香港广场样本是异常点，应从数据集中剔除。

3.4　数据挖掘

在数据挖掘方法中，决策树法、遗传算法主要用于分类模型；在数据库中建筑数量有限的情况下，神经网络方法的计算过程很难收敛，其结果的均方误差、平均误差等指标都不能满足要求；在小样本的情况下，统计学方法中的回归分析、主成分分析等理论能用于分析建筑能耗与其他因素的关系，可建立建筑能耗回归模型。结合数据库实际情况，笔者选择统计学方法进行数据挖掘工作。

填补缺失数据和剔除异常点后，93 个建筑样本将进入最终的数据挖掘过程。由于是以单位面积建筑年一次能耗为目标进行数据挖掘，因此在 SAS/EM 模块中，在变量转换节点创建新变量 EUI，并设定 EUI＝ENERGY/AREA。在变量属性定义节点把变量 EUI 设为因变量，而变量 AREA 和 ENERGY 将被拒绝进入数据挖掘过程，其他变

量均设为自变量。

在 SAS/EM 模块中，可用统计学方法中的向前选择法、向后选择法、逐步回归法、主成分分析等方法进行数据挖掘。经分析，逐步回归法的 F 检验值和 T 检验值都小于 0.15，满足显著性水平。修正后的 R^2 值为 0.908 1，接近于 1。进入模型的自变量有 ACTIME，CCAPACITY，OFFICE，HOTEL。由 T 统计量可知，自变量在模型中对因变量作用从大到小依次为 CCAPACITY，OFFICE，ACTIME，HOTEL。用统计学方法进行数据挖掘所得到的商用建筑能耗预测模型为

$$EUI = 36.704(ACTIME) + 0.040(CCAPACITY) + 710.7(OFFICE) + 1\,108.6(HOTEL)$$

3.5 模型评价

商用建筑能耗预测模型中，自变量 OFFICE 和 HOTEL 体现了建筑功能对建筑能耗的影响。由于宾馆要求高，使用时间长，所以 HOTEL 比 OFFICE 对建筑能耗的影响更加强烈。这从两者的回归系数可体现。在模型中，自变量 HOTEL，OFFICE 的回归系数分别为 1 108.6，710.7。变量 CCAPACITY 的回归系数为正，即与因变量为正相关。显然，冷源装机容量大的商用建筑的建筑能耗相对更高。建筑的空调运行时间越长，单位面积年一次能耗也将更大，变量 ACTIME 与因变量 EUI 正相关是合理的。

从节点预测值分布图可得到因变量 EUI 预测分布情况，见图 4。图 4 中，在横坐标上把总样本中单位面积年一次能耗值从小到大划分为多段，纵坐标表示单位面积年一次能耗值在某段里的样本数占总样本的比例。柱状体颜色越深，比例越高。93 个建筑样本中，单位面积年一次能耗处在1 210.2～1 290.4 MJ/(m² · a)之间的样本最多，所占比例约为 20.3%。

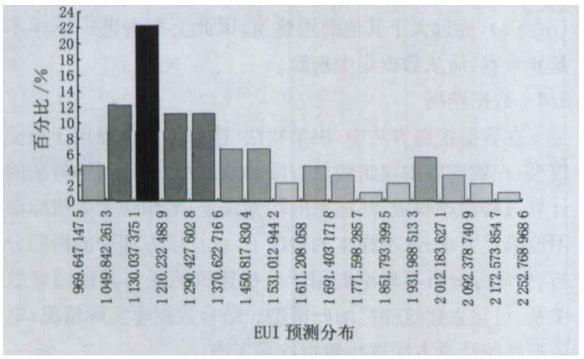

图 4 模型预测结果分布情况

在数据量有限的条件下，商用建筑的建筑能耗预测值的相对误差在－20%～20%之间是可以接受的。从图 5 可知，用商用建筑能耗预测模型预测数据库中的建筑能耗，大多数相对误差在可接受的范围内，只有 22%建筑的预测值相对误差绝对值大于 20%。因此在目前阶段，数据挖掘得到的商用建筑能耗预测模型是合理准确的。

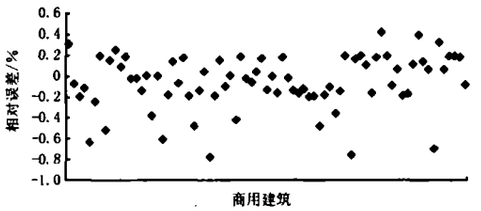

图 5 商用建筑单位面积年一次能耗预测模型预测值误差

4 结论

本文对上海市商用建筑信息数据库进行数据处理和数据挖掘分析，得到了商用建筑能耗预测模型，并提出了适合目前阶段的数据处理和数据挖掘方法。但由于主、客观原因还存在以下问题和不足，需在以后的工作中完善和改进：

1) 上海市商用建筑信息数据库中数据数量不足，数据质量也不尽如人意，影响数据处理和数据挖掘的效果；

2) 由于数据量的制约，一些数据处理和数据挖掘方法不能正常运用，无法充分体现数据挖掘的优势；

3) 数据处理方法掩盖了某些变量对建筑能耗的影响，造成这些变量最终不能进入模型。

通过对数据挖掘技术在上海市商用建筑信息数据库的研究应用，笔者认为，随着数据库在建筑能耗领域的广泛应用，数据挖掘技术将具有无可比拟的优越性，在建筑能耗领域应有非常广阔的应用前景。

参考文献：

[1] 涂逢祥，王庆一. 建筑节能——中国节能战略的必然选择(上)[J]. 节能与环保，2004(8)

[2] 郎四维.《公共建筑节能设计标准》要点[J]. 建设科技，2005(13)

[3] 张柏礼，孙志辉. 数据挖掘技术在能量管理系统中的应用[J]. 工业控制计算机，2002，15(12)

[4] Zhang S C, Zhang C, Yang Q. Data preparation for data mining[J]. Applied Artificial Intelligence, 2003, 17(5/6)

[5] 贺佳，陆健，曹阳，等. 医学统计学中的 SAS 统计分析[M]. 上海：第二军医大学出版社，2002

[6] 武建虎，加佳，贺宪民，等. 多变量缺失数据的不同处理方法及分析结果比较[J]. 第二军医大学学报，2004，25(9)

[7] 王宏鼎，童云海，谭少华，等. 异常点挖掘研究进展[J]. 智能系统学报，2006，1(1)

[8] 殷峻. 一个基于 SEMMA 的数据挖掘应用实例[J]. 冶金自动化，2003(3)

[9] 茅群霞，李晓松. 多重填补法与 Ad Hoc 法对模拟纵向数据集缺失值处理的比较[J]. 现代预防医学，2005，32(4)

[10] 金浩，高寨英. 最佳多元线性回归模型的选择[J]. 河北工业大学学报，2002，31(5)

[11] 岳朝龙，黄永兴，严忠. SAS 系统与经济统计分析[M]. 合肥：中国科学技术大学出版社，2003