华北电力大学(北京)

博士学位论文

基于知识挖掘技术的智能协同电力负荷预测研究

姓名：王建军

申请学位级别：博士

专业：管理学；管理科学与工程

指导教师：牛东晓；李存斌

201106

摘 要

电力工业是国家的重大基础行业，对于我国经济建设、国家安全、社会稳定、居民生活质量具有至关重要的作用，精确的电力负荷预测对于制定发电计划、制定经济合理的电力调配计划、制定上网竞价计划、控制电网经济运营、降低旋转储备容量、进行电力市场需求分析、避免重大事故、有效化解风险、保障生产和生活用电方面具有十分重要的意义。然而电力负荷预测工作是十分复杂的，它除了包括负荷自身特性、经济人口等定量型因素的影响，同时也包含着不规则事件、日期类型、季节类型、描述性天气等非数字型的定性因素的影响，如果不考虑这些定性因素的影响，无论如何改进负荷预测模型，预测精度都很难有根本性的提高，负荷预测理论也难以有较大的突破。因此，本文提出了基于知识挖掘技术的智能协同电力负荷预测研究思想，旨在结合知识挖掘技术和智能电力负荷预测方法进行协同电力负荷预测，通过知识挖掘直接对数据库中的负荷变量属性及对应的各类影响因素变量属性进行分析处理，在预测时通过计算与预测目标各类知识特征的总体关联程度大小，自动提取具有高度相似性综合知识特征的同类历史数据，再结合智能算法和电力负荷预测方法建立具有针对性的自适应结构的智能预测模型对负荷进行预测，在遇到有少部分具有较大的预测误差点时利用知识挖掘形成的纠偏规则进行相应的后干预工作，能够进一步克服以前的预测方法的不足，使预测精度得到突破性的提高。本文进行的主要工作如下：

（1）提出了基于知识挖掘技术的负荷数据规范以及相应的预处理方法。在对影响负荷预测的属性变量进行分类的基础上，建立起相应的数据库结构规范，利用相应的数据预处理方法形成各种负荷预测需要的不同数据视图，可以方便对定性型因素和定量型因素的处理。

（2）建立了基于知识挖掘分类技术的自适应结构的日负荷曲线 BP 神经网络预测模型。在仅有纯负荷数据的情况下，首先通过计算负荷曲线的相似度对历史数据进行排序并进行初步的预测，然后再利用 BP 神经网络对误差进行纠偏工作来得到更加精确的预测结果；在具有较多气象数据可供分析时，首先利用聚类分析将日曲线负荷进行聚类分析，然后利用知识挖掘中的分类技术寻求气象数据和曲线负荷聚类之间的关系，形成相应的知识规则，并在形成分类规则时，利用粗糙集的属性约简技术剔除掉冗余属性来加快生成规则的速度，最后利用不同类别的数据训练出不同的 BP 神经网络模型；这样在进行负荷预测工作时，可以根据预先判断的气象数据找出相应的类别，并选取相应的 BP 神经网络进行预测来

提高预测的精度。在利用BP神经网络进行模型训练时，提出了一种简单的自适应BP神经网络对日负荷曲线进行预测，该自适应网络可以自动确定隐含层节点的个数，无需人为经验的干预。

（3）提出了利用微分进化算法调整参数的支持向量机中长期负荷预测模型。对于中长期负荷预测，由于其样本数据远少于短期负荷预测，因此适用于小样本数据量条件下的支持向量机智能预测方法，该方法可以有效地选取支持向量机所需求的相应参数，可以有效地提高中长期负荷预测的精度。

（4）提出了一种结合知识挖掘后干预纠偏技术、时间序列预测技术以及支持向量机预测方法的日最大负荷预测方法。由于日最大负荷预测不但需要考虑气象因素的影响，还需要考虑不同类型日期、不规则事件对日最大负荷的影响，本文提出的方法不但可以考虑负荷序列的趋势，而且可以考虑非线性因素的影响和不规则事件的影响，经过知识挖掘后干预纠偏后可以有效地提高负荷预测的精度。

（5）提出了基于负荷预测的预警监测指标并建立了基于知识挖掘的自然灾害预警方法。结合上文中的负荷预测方法和预测结果，对短期日负荷曲线进行偏离度的监测以及对中长期供需平衡进行监测，对于气象灾害的预防，给出了冰灾、沙尘暴以及台风三种典型气候的监测方法。在对短期预警监测中筛选监测行业以及形成实际区域冰灾气象条件监测时利用了知识挖掘中的决策树分类属性筛选技术，可以挑选出重要的监测对象以及冰灾预警的相应条件。

（6）对基于知识挖掘智能协同负荷预测技术的用电分析及预测系统进行了研究。该系统对上述知识挖掘和智能算法的研究成果进行了软件上的实现，有别于基于传统算法或是其他改进算法开发出的负荷预测系统。

关键词：知识挖掘；负荷预测；决策树；聚类；神经网络；支持向量机；预警

Abstract

Power industry is one of the most important basic industries in the energy field of our nation, it is the lifeline of national economy, and the economic development follows the electricity development. The power plays a very important role in China's economic construction, national security and social stability. Accurate load forecasting has an important significance of power generation plan, the reasonable development plan of the distribution system, reducing rotation storage capacity, avoid major accidents and prevent risks. However, load forecasting is a very complex problem because of the influenced factors includes such as economic factors, irregular event, date, season, weather and other non-numeric description factors. If we don't consider these factors, the load forecasting accuracy will not be improved further. In this paper, a collaborative model based on knowledge mining and intelligent load forecasting model is present. Through using the knowledge mining technology to deal with the factors, the relevancy of the history load data and the forecasting goal is discovered. The highly similar load pattern will be extracted in historical data, the intelligent load forecasting methods will be used to forecast the load, combined with adding the appropriate text mining interventions, large errors can be further overcomed. The model can make a breakthrough to improve prediction accuracy. The major work carried out is as follows:

(1) The criterion of data storage and the corresponding pre-data specification method is present. Based on the variables classification, the different data storage criterions are constructed and the corresponding data views are also constructed.

(2) A combined text mining classification techniques and BP neural network load forecasting model is established. When the data only contains load data, the similarity of load curve is calculated at first, and then the BP neural network error correction is used to get higher prediction accuracy. If the data contains weather data availables for analysis, the cluster technology is used to classify the load curve, and the desion trees technology is used to find the corresponding knowledge rules, in the classification process, the rough set can be used to reduce the attributes. Based on the classification data, the different BP neural network models are trained for load forecast. According to the rules, the appropriate data can choosed for prediction. When the BP neural network model is trained, a simple adaptive method is used, and the adaptive network can automatically determine the number of nodes in the hidden layer without human experience.

(3) An adaptative SVM long-term load forecasting model is proposed with the differential evolution algorithm. For long-term load forecasting, the number of

sample data is far less than the short-term load forecasting, SVM can effectively forecast in this situation. The experiment proves the differential evolution algorithm can select corresponding parameters, which can effectively improve the accuracy of long-term load forecasting.

(4) A novle model combined time series forecasting techniques, support vector machine methods and knowledge mining correcting technical of daily maximum load forecasting method is proposed. Daily maximum load forecasting not only need consider the impact of meteorological factors, but also need consider different types of dates and the effects of irregular events, the proposed method not only can treat the trend of time series, but also non-linear factors and irregular effects. The experimental results show that the method can effectively improve the accuracy of load forecasting.

(5) The early warning indicators for load monitoring and meteorological disaster are proposed. Based on above load forecasting methods, the short-term daily load curve deviation degree and long-term supply and demand ratio are present, then the three typical climate monitoring conditions including ice disaster, dust storms and typhoons are also given. In the ice disaster conditions presented, the decision tree classification technology is used to pick out the important monitoring industries of national economics.

(6) The collaborative knowledge mining technology and intelligence load forecasting method system is studied. Based on the above models, it is unlike the traditional methods and other improvements based on algorithms developed for load forecasting system.

Keywords: knowledge mining, load forecasting, decision tree, cluster, ANN, SVM, early warning

Contents

第1章 绪 论

1.1 课题背景以及研究的目的和意义

电力工业是国家在能源领域的重大基础行业，电力是国民经济的命脉，经济要发展，电力是先行，电力对于我国经济建设、国家安全、社会稳定、生活质量具有至关重要的作用。随着国际上电力市场的逐步建立，随着国内经济形势的快速发展，电力供求矛盾日益严峻，电力需求的影响因素逐渐增多，传统的电力工业负荷预测理论方法已经不再适用，适应新环境下的负荷预测理论方法研究迫在眉睫。准确及时地电力负荷预测对于制定经济优化的发电计划、制定经济合理的电力调配计划、制定上网竞价计划、在竞价上网中取得优势、最优制定电力现货和期货报价、控制电网经济运营、降低旋转储备容量、进行电力市场需求分析、搞好电力市场营销和电力客户关系管理、避免重大事故、有效化解风险、保障生产和生活用电等方面均具有十分重要的意义。

电力负荷预测工作作为电网管理部门的基础工作之一，能够为电网企业以及整个电力的发、输、配、送的电力工业链条直接产生重大的经济效益和社会效益。国外学者 Bunn 和 Farmer 早在上世纪 80 年代中期的研究成果中就已经表明，负荷误差提高 1%将会增加 10,000,000 英镑的电力经营成本。类似地，对于我国的一个中等规模的省级电网而言，按照常规假设其平均供电负荷为 4500MW，如果将系统日负荷预测精度提高 1%，就表示在系统发供电可靠率相同的条件下，电网发电出力富裕时可减少 50MW 的旋转备用容量，电网发电出力不足时可减少非计划限电 45MW，由此产生的主要效益为：因系统减少旋转备用容量产生的年经济效益 2000 万元，减少非计划限电增加售电量产生的年经济效益为 446 万元，共创电网年经济效益 2246 万元，假设每度电产值为 5.3 元、每度电的边际利润为 0.07 元、非计划限电电量损失率为 0.8 元、全年限电日 253 天考虑，、因系统减少非计划限电产生年社会上的经济效益将高达 3.38 亿元。此外，按照常规假设，我国火力发电所占比率为 75%，以我国 2009 年上半年的统计数据为依据，供电的煤耗率为 341 克/千瓦时，火电设备平均利用小时为 2934 小时，由此计算得的由于负荷精度提高 1%而产生的减少 50MW 备用容量将节约年用煤量 60000 余吨。由此可见，负荷预测精度的提高不但具有十分巨大的经济效益和社会效益，同时也对于我国的节能减排、资源利用以及可持续发展具有非常重大的意义。

然而电力负荷预测工作是十分复杂和困难的，以电力负荷预测的研究内容

为例：其包括负荷预测、供售电量预测以及电价预测等内容，按时间划分可分为中长期、短期、超短期负荷以及负荷曲线预测，其中又以短期日负荷预测为例，研究对象又可以包括日最大负荷预测、日 24/48/96 点负荷曲线预测、连续多日负荷预测、日扩展短期负荷预测、重大节日负荷预测等；按空间划分可以分为全国、地区、省或直辖市、市、县等负荷预测；按使用类型划分可以分为农业、工业、商业、公共事业等负荷预测，并且这些划分类别之间还可以相互交叉组合。由于电力的社会属性，预测工作受到例如气象、需求、现货、期货、经济、系统、市场、价格、竞争、政策、政治活动、背景、领域等大量复杂影响因素的多重干扰影响。而以往的电力负荷预测方法模型大多采用负荷自身的数据进行预测工作，忽略了这些影响因素对负荷的影响，因此在近期的研究中，越来越多的研究学者发现，如果想从本质上提高负荷预测的精度，就必须科学地、系统地、全面地考虑更多的影响因素，但目前少量考虑这些复杂影响因素的智能预测方法模型也只能考虑一些定量的影响因素，如 GDP、人口、温度以及湿度等，而没有办法处理那些用定性的、描述性知识表述的影响因素，例如：气象中的“晴”、“阴”、“多云转阴天”；“金融危机带来的不景气”；“北京于 2008 年 8 月 8 日 20：00 举办奥运会”等文本信息等，在这种情况下，缺少了这些非数字型的定性知识显性描述、隐性经验、推理规则、文本信息影响因素的考虑，就难以在模型中考虑这些定性因素的影响作用。在这种情况下，无论如何改进负荷预测模型，预测精度都很难有根本性的提高，负荷预测理论也难以有较大的突破。

正是基于上述原因，需要对传统的预测方法予以改进，引入并结合能够处理多种类型知识影响的新理论以及方法，而知识挖掘技术正是可以处理复杂多类型的知识因素影响的数学工具。因此，本课题提出的知识挖掘智能协同的电力负荷预测思想，旨将负荷、影响负荷的非线性因素以及其他因素综合起来利用知识挖掘技术和智能预测方法对负荷进行协同预测，利用预测结果进行相关的预警工作并结合研究成果开发出一套知识挖掘智能协同负荷预测及预警软件，通过软件进行推广使用以提高课题成果的实用性。这不但对于整个电力工业的经济效益和社会效益具有十分重大的意义，同时也对电力负荷预测理论而言也是十分有益的补充，此外，对于其它相关行业在复杂多类型知识因素影响下的预测工作也具有一定的借鉴和参考价值。

1.2 电力负荷预测理论及研究现状

1.2.1 电力负荷预测理论

电力负荷预测是指根据电力系统的运行特性、增容决策、自然条件与社会影响等诸多因素，在满足一定精度要求的条件下，确定未来某特定时刻的电力负荷数据[1]。其中的电力负荷指的是广义电力负荷，即除了电力用户常说的某一时刻的用电功率的狭义电力负荷外，还包括在电力工业整个环节中的电能生产的发电量、电能供给的供电量、电力企业销售给用户的售电量以及用户消耗的用电量等。由于电力负荷预测是根据电力负荷的历史数据和目前已知的或者是可预测的相关变量状态对未来的数值进行推测，因此需要利用预测工作的基本原理，用于指导负荷预测工作，其基本原理如下[2]：

（1）可知性原理

预测对象的发展规律未来的状况是可以为人们预知的，其所在的客观世界是可以被认识的，人们不但可以认识它的过去和现在，而且可以通过总结它的过去和现在推测其未来的发展趋势。这是人们进行预测活动的基本依据。

（2）可能性原理

因为事物的发展变化是在内因和外因共同作用下进行的。内因的变化及外因作用力大小不同，会使事物发展变化有多种可能性。因此，对某一具体指标的预测往往是按照其发展变化的多种可能性进行多方案预测的。

（3）连续性原理

预测对象的发展是一个连续统一的过程，其未来发展是这个过程的延续。它强调了预测对象总是从过去发展到现在，再从现在发展到未来。可以认为事物发展变化过程中会将某些原有的特征保持下来，延续下去。电力系统的发展变化同样存在着惯性，如某些负荷指标会以原有的趋势和变化率发展下去。这种惯性正是负荷预测的主要依据。因此，了解事物的过去和现在，并掌握其变化规律，就可以对其未来的发展情况利用连续性原理进行预测。

（4）相似性原理

尽管客观世界中各种事物的发展各不相同，但一些事物发展之间还是存在着相似之处，可以利用这种相似性进行预测。在很多情况下，作为预测对象的的现在发展过程和发展状况可能与另一事物过去一定阶段的发展过程和发展状况相类似，这时就根据后一事物的已知发展过程和状况，来预测所预测对象的未来发展过程和状况，这就是相似性原理，如传统预测技术中使用的类推法或历史类比法就是基于这个原理的预测方法。例如，当我们预测一个新的经济开发区的用电量时，由于其建成时期较短，没有很多历史数据可利用时，就难以

利用趋势外推、回归分析等方法建模预测。这种情况下，可以参考一个早已建成的、规模和条件具有可比性的其他经济开发区，用其发展时期相对应的用电量作为预测新经济开发区用电量的基础，从而可以作出相应的预测结果。

（5）反馈性原理

反馈是利用输出返回到输入端再调节输出结果的过程。预测的反馈性原理实际上是为了不断提高预测的准确性而进行的反馈调节。人们在预测活动实践中发现，当预测的结果和经过一段实践所得到的实际值存在着差距时，可利用这个差距，对远期预测值进行反馈调节，以提高预测的准确性。在进行反馈调节时，可以首先认真分析预测值和实际值之间的差距及产生差距的原因，然后根据已经查明的原因，适当改变输入数据，进行反馈来调节远期预测结果。反馈性预测实质上就是将预测的理论值与实际相结合，在实践中检验，然后进行修改、调整，使预测质量进一步提高。

（6）系统性原理

预测对象本身不但具有内在的系统，而且由于预测对象和外界事物的之间的联系同样也形成了它的外在系统。这些系统综合成一个完整的总系统，在预测时都要进行考虑。即预测对象的未来发展应该是系统整体的动态发展，并且整个系统的动态发展应该与它的各个组成部分和影响因素之间的相互作用和相互影响密切相关。系统性原理同时还强调系统整体最佳，只有系统整体最佳的预测，才是高质量的预测，才能为决策者提供最佳的预测方案。

在电力负荷预测的研究中，一般采用的是按时间期限进行分类的方法，本文也用的是这种分类方法，将电力负荷预测分为中长期、短期和超短期负荷预测[3]。其中中长期负荷预测一般指 1 年以上、10 年以下的以年为单位的预测。短期预测包括一年以内以月、周、日为单位的电力负荷预测，一般预测的是对未来一个月度、未来一周、未来一天的最高负荷、最低负荷或平均负荷等负荷特性指标。超短期负荷预测指的是预测一天以内的间隔 1 个小时、半个小时、15 分钟乃至更短的时间内的电力负荷预测。

由于电力负荷预测工作是一种对未来电力负荷数值的不确定估算工作，因此必然会和客观实际数值存在着一定的预测误差，目前的相关研究中常用的衡量电力误差指标有平均相对误差（MAPE）、绝对值相对误差（APE）以及相对误差三种，一般情况下，对于负荷逐点的评价指标采用的是绝对值相对误差，在对整体算法采用的误差评价指标是平均相对误差（MAPE），在需要统计正负误差工作时采用的误差评价指标是相对误差，在本文中同样采用这三个指标作为衡量误差的指标，其计算公式分别如下所示：

$$APE=\left|\frac{A(i)-F(i)}{A(i)}\right|\times 100\% \tag{1-1}$$

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{A(i)-F(i)}{A(i)}\right|\times 100\% \tag{1-2}$$

$$\frac{A(i)-F(i)}{A(i)}\times 100\% \tag{1-3}$$

其中 $A(i)$ 表示实际负荷，$F(i)$ 表示预测值，n 表示数值个数。

1.2.2 国内外负荷预测研究现状

自上世纪二十年代开始就有学者对电力负荷预测开始进行研究，但由于当时的电力系统规模小，变化较为平稳，因此电力负荷预测没有受到重视。而随着电力系统的市场化进程以及对能源的空前重视，使得负荷预测受到了更加广泛的重视。近二三十年来，国内外的很多专家学者对负荷预测的理论和方法进行了大量的研究工作，取得了很多成果。总的说来，国内外关于负荷预测的理论成果大多集中于短期负荷预测的研究，中期和长期的负荷预测研究偏少，超长期的负荷预测更少。从预测方法上来说，预测方法大致经历了四个阶段的发展。

（1）以线形回归方法为代表的传统统计预测方法阶段[3-14]；

（2）引入由 Box-Jenkins 提出著名的时间序列预测方法（包括 AR、MA 以及 ARMA）阶段[15-30]；

（3）灰色预测方法以及组合预测方法阶段[31-55]，其中组合预测方法国内重视程度较大，国外学者很少重视组合预测方法的研究；

（4）近期的以神经网络以及支持向量机为代表的智能预测方法阶段[56-104]。

在这些研究中，国外对负荷预测的研究起步早，研究也较为深入，国内对负荷预测研究虽然起步晚，但目前已经接近于国外的先进水平。对上述文献进行详细综述如下：

（1）线形回归预测方法

虽然传统的电力负荷预测方法包括许多方法，例如趋势外推法、均值预测、自适应指数平滑预测等方法，但是线性回归预测方法是其中最成熟、运用最广、影响力最大的传统预测方法。电力负荷回归模型预测技术就是根据历史负荷数据资料，依靠线性回归数学模型对未来的负荷进行预测，通过给定的一组或者多组自变量和因变量的历史数值，研究各自变量和因变量之间的关系，形成回归方程。常用的回归模型有一元线性回归、多元线性回归两种[3-4]。至今为止，线形回归模型由于方法简单、预测速度快的优点决定了其作为中长期负荷预测

研究必须引用的经典模型地位，也是软件实现的必须的基础算法之一。然而由于其模型的线性方法决定其必然无法描述复杂的非线性因素对其进行影响，并且其自变量所代表的主要因素需要依靠经验确定。

（2）时间序列预测方法

Box-Jenkins 提出的时间序列模型被认为是最经典、最系统、最被广泛采用的一类短期负荷预测方法，包括 AR、MA 以及 ARMA 三类方法。至今为止，大部分短期负荷预测研究文献中所用的比较基准都采用该算法，该模型将负荷数据看成是一个周期性变化的时间序列，然后根据给定的模型对未来的负荷进行预测[15-17]。由于时间序列预测方法所需要的数据本身就是历史负荷序列本身，因此其数据方便收集与获取，但同样也决定了时间序列方法具有难以考虑其余因素对负荷的影响这一缺点，该方法同样无法考虑非线性因素对负荷的影响。

（3）灰色预测方法及组合预测方法

值得一提的是，灰色预测方法是由国内学者邓聚龙首先提出的一种预测方法，因此，国内学者对灰色负荷预测方法的研究较国外学者更加关注，此外，由于国外学者更加关注于单一方法的优化改进研究，很少关注组合预测方法方面，因此，国内学者同样对组合预测方法较国外学者更加关注。

由于灰色预测法具有要求负荷数据少、不考虑变化趋势、运算方便、易于检验等特点，因此适合于中长期的负荷预测，尤其适合年度负荷预测，并取得了令人满意的效果，但灰色预测方法具有数据离散程度越大，预测精度越差的缺点，灰色预测方法适合具有指数增长趋势的负荷预测序列，对于具有其它扰动趋势的负荷序列预测精度难以提高。

组合预测方法是对传统统计预测方法的综合改进，其预测方法是选择多个预测函数模型进行加权平均组合，按照协方差最小确定权数形成最终的预测模型，组合预测方法虽然较单一的预测方法而言精度一般都有明显的提高，但是仍然从本质上难以适应短期内其余因素对负荷的影响，无论如何改变权重系数，都难以使得预测结果更加精确，组合预测方法均依赖于原始的多个预测方法的预测结果，这就意味着，组合预测方法不但具有这些方法的优点，同时也具有这这些方法的不足之处。

（4）智能预测方法

近些年来，越来越多的学者形成一种共识，即负荷预测精度的本质上的提高必须要考虑其他影响因素对负荷预测的影响，学者都开始对负荷预测方法研究的重点转向能够考虑其他影响因素，具有良好的非线性拟合能力的智能预测方法方向。因此，很多智能预测算法如神经网络[56-63]、支持向量机[64-66]、专家

系统、粒子群算法[93]、遗传算法开始大量地运用于负荷预测中，一时间，大量的智能负荷预测方法的研究成果开始在各种期刊上予以发表，在诸多智能算法中，属神经网络预测方法的应用最为广泛。例如 Henrique Steinherz Hippert 总结了1991年到1999年期间在电力相关杂志上所发表的40篇利用神经网络进行负荷预测论文后指出，神经网络在负荷预测上取得了巨大的成功，并且绝大多数研究在利用神经网络对负荷进行预测时不但利用日期变量对待预测负荷进行分类，而且考虑了温度因素对负荷的影响[94,96]，有少部分研究还考虑了湿度及其他因素对负荷的影响[97-98]。此外，Bahman Kermanshahi[99]在利用神经网络进行长期预测时，利用电价、油价、GDP、GNP 等 10 个非负荷变量对日本的长期负荷进行了预测。Che-Chiang Hsu 建立起了一个 3 个输入节点、2 个隐层节点、一个输出节点的神经网络对台湾的年度峰值负荷进行预测，其输入变量是 GDP、人口和最高气温[100]。随着 V.Vapnik 提出了基于结构风险最小化的支持向量机技术，支持向量机开始应用在负荷预测中，Ping-Feng Pai[101-102]分别利用遗传算法和模拟退火算法对支持向量机的三个参数 r，e，C 进行选取，然后利用支持向量机结合历史负荷数据对台湾的年负荷进行预测。此外，M.S. Kandil[103]利用专家系统对影响负荷的因素和规则进行选择，由此产生一系列的决策规则用以指导未来的负荷预测。而 Mohammed El-Telbany[104]利用粒子群算法对 Jordanian electrical 电力市场的日峰值负荷进行了预测，并将预测结果与 BP 神经网络的结果进行比较，其结果优于 BP 神经网络的结果。总的说来，上述智能算法的预测结果都优于传统的时间序列预测方法。但是在这些智能算法的预测过程中，都没有解决除气象数据和经济数据外的其他定性因素可预测因素对负荷预测的影响，以我国为例，2008 年 8 月份的北京奥运会召开造成北京市所需负荷增加，相应地政府管制措施同样也会影响负荷的变化。此外，上述研究虽然注重了其余因素对电力负荷预测的影响，但是却将这些因素和历史负荷数据放入同一个预测模型中对负荷进行预测，而无法考虑许多非参数因素造成的负荷变动影响，因此，如何利用这些因素进一步提高预测的精度是目前需要解决的一个难题。

以上的负荷预测研究工作和研究方法为搞好负荷预测工作奠定了基础，发挥了重要作用。但是总的说来，目前国内外的负荷预测研究工作仍然有下列不足：

（1）目前的智能预测方法虽然能够考虑更多的影响因素，例如 GDP、人口、气象条件中的温度以及湿度影响，但仍然缺乏对某些明显的经验型知识以及某些隐性知对负荷的扰动影响的考虑。

（2）目前的研究中并没有充分地考虑显性和隐性知识、经验和推理规则描

述的各类定性知识型复杂因素的影响，更没有考虑文本类型知识影响因素对预测建模的作用。

（3）负荷预测应用的智能预测方法不能客观筛选和确定主要影响因素、输入变量、模型结构，从而确定最优的负荷预测模型，这些在建模过程中都需要人为的靠经验进行确定，缺少一种能够自动确定模型结构的方法。

针对上述不足，近些年来，很多学者开始结合数据挖掘技术对负荷预测进行研究，有部分学者结合数据挖掘中的分类技术对负荷预测进行了研究。例如刘敦楠[105]利用数据挖掘的决策树技术建立知识库的方法事先判别负荷突变情况并做出适当处理，然后利用修正后的数据进行预测。朱六璋[106-107]基于数据挖掘决策树算法和通用的决策支持对象建模工具，结合区域电网气象负荷数据库建立起数据挖掘模型进行日负荷预测，此外，他还利用数据挖掘中的 C4.5 和 CART 算法挖掘影响负荷的气象因素差异与相应负荷变化率的规则，结合 BP 神经网络进行短期负荷预测。Wi Young-Min [108]通过数据挖掘技术将负荷类型分为工作日、周末及节假日三类，通过提取同一类别的数据对负荷进行预测。Mori Hiroyuki [109-110]利用数据挖掘技术对输入变量和输出变量进行分类规则的识别后进行短期负荷预测，此外他还利用数据挖掘技术生成回归树分类规则，结合神经网络进行日峰值负荷预测。

另有部分学者结合粗糙集理论对负荷预测进行了研究。如李秋丹[111]利用粗糙集和遗传算法选取与负荷相关的预测变量，再选取与预测日相似的训练模式，最后用神经网络对负荷进行预测。黎静华[112]采用粗糙集离散化算法对气温、湿度等影响负荷的属性进行离散化，通过计算形成不同层次上符合初定阈值的带粗糙集算子的网络规则集然后进行负荷预测，从而提高负荷预测的精度。熊浩[113]基于模糊粗糙集理论提出了循环采样方法，将负荷样本进行聚类，然后将得到的规则应用于短期、中期和长期空间负荷预测中。

此外，部分学者利用数据挖掘技术寻找与预测日相近的相似日进行预测。如赵磊[114]通过数据挖掘技术寻找与预测日同等气象类型的多个历史日负荷，由此进一步提取数据后构建优选组合预测模型。李邦云[115]提出了一种基于数据挖掘中时间序列相似性研究的电力负荷预测方法。牛东晓等人[116-117]通过多种挖掘技术寻找与预测日同等气象类型的多个历史日负荷，然后进一步提取相似的数据序列，通过构建人工神经网络预测模型和支持向量机模型来提高预测的精度。

还有部分学者利用模糊关联技术、聚类技术以及异常值判别技术等对负荷预测进行了研究。如沈海澜[118]采用模糊均值算法对负荷进行聚类，通过形成模糊关联规则进行匹配预测。崔昊[119]利用提出的改进聚类算法对不完备数据进行

补全，在此基础上进行预测。冯丽[120]通过离群数据挖掘算法和 Kohonen 网络提取相关负荷的特征曲线，将其用于不良数据的校正。Wang Jianzhou[121]利用数据挖掘技术剔除负荷中的背离点和识别负荷曲线的背离趋势，通过修正进行更加精确的负荷预测工作。

部分学者利用计算机技术对数据挖掘负荷预测系统进行了研究。李耀池[122]在采用数据挖掘技术对负荷变化规律分析的基础上提出了按日期类型分开建模的 96 点预测模型。开发了一套界面友好、功能完善的负荷预测软件。孙英云[123]提出了决策树技术和时间序列相结合的短期负荷预测算法,并在此基础上实现了基于 DotNet 架构的短期负荷预测系统。Lambert-Torres G[124]通过建立和模糊数据挖掘系统之间的接口形成了一套模糊挖掘负荷预测系统。

但是从上述检索的文献研究内容、目的和方法上来看，结合数据挖掘的负荷预测方法缺乏对定性因素进行挖掘的手段，并且在利用影响负荷的气象、经济、价格以及其他因素结合智能方法对负荷进行预测时一般直接将其作为输入变量直接将这些量化的数据代入到智能模型中进行预测，考虑的因素范围少。而目前国内将知识挖掘技术和预测相结合的相关研究仅见于汪寿阳教授的课题组所研究的课题中，其课题是基于文本处理和知识挖掘的金融市场的外汇预测问题，该团队对知识挖掘和预测理论的结合发展做出了很好的工作，为基于知识挖掘的预测研究奠定了基础[125]。电力负荷和金融的预测虽然都属于预测问题，但是电力负荷的特点和金融预测又有不同，例如对于影响电力负荷的预测因素而言，其对气象、市场、价格、需求、政治、网络系统故障等较多定性和定量的影响因素具有高度敏感性，是一类特殊的预测对象，而金融和外汇的预测的影响因素多和政策和经济发展形势相关。电力负荷预测在显现出周期性的同时也体现出含有其余因素对其的影响复杂性，这对于现有的电力负荷预测理论提出了挑战，因此要提出新的知识挖掘和智能协同负荷预测方法来应对这种挑战。

1.3 知识挖掘理论及研究现状

1.3.1 知识挖掘理论

随着计算机技术和互联网的发展，使得整个世界的范围变小，人们可以在互联网上无视空间和时间的限制攫取各种的资源，目前的企业基本都有自己的内部网和与外部互联网的接口以方便在网上进行数据和信息的协同工作，展现在人们面前的数据越来越多,人们通俗的给这种现象起了一个很形象的名称“数据过剩”或者是“信息爆炸”。数据在互联网上的传播从页面之间的文件传播转

向以利用数据库为主的数据库管理传播，渐渐的，数据库之中的数据越来越多，形成了“数据海洋”，但是与之相反的是，人们反而感到在如此多的数据中难以得到自己想要的知识，感觉自己“知识贫乏”，这就需要研究从数据海洋中提取出自己想要的知识，对数据进行去粗存精、去伪存真的技术。即需要利用机器学习的方法对数据库里的数据进行分析，在这种情况下，数据库中的知识发现KDD（Knowledge Discovery in Databases）学科产生了，这个学科涉及到机器学习、统计学、人工智能、数据可视化以及专家系统等领域内的交叉协同协作，是一门交叉性的学科。

KDD学科中最核心的技术是知识挖掘技术，所谓知识挖掘是指从数据集中自动识别出隐藏在数据中有用规则、概念、规律以及最终可理解的模式的过程，即从数据库中繁杂的数据中抽取隐含的、以前未知的、具有潜在应用价值的信息的过程。知识挖掘是所谓“数据挖掘”的一种推广，其定义几经变动，最新的定义是由Usama M.Fayyyad等给出的：“从知识资料集中识别有效的、新颖的、潜在有用的，以及最终可理解的模式的高级处理过程”。知识挖掘的目的是将大量的数据和知识融合成有序的、分层次的、易于理解的信息，并进一步转化成可用于干预预测和决策的知识，是一个智能化、自动化的过程。其知识资料集包含显性与隐性、定性与定量、经验与推理等类型知识，能对原有的数据、知识进行高度自动分析、归纳推理，从中挖掘出潜在的模式和规律，为预测决策、模式识别、故障诊断、生产过程优化等提供知识服务。

关于知识挖掘理论的研究在近几年才开始具备研究条件并得到重视，知识挖掘被认为是寻找内在规律的新方向，目前仍处于发展的早期，还有很多研究难题和面临的挑战，如知识数据的巨量性、动态性、噪声性、缺值和稀疏性，定性知识的转化和深层处理，发现模式的可理解性、价值性，系统的集成、智能交互、更新管理和复杂知识的处理等等。

1.3.2 智能负荷预测中可利用的知识挖掘技术

常用的知识挖掘技术包括序列分析、关联分析、分类、聚类、进化式程序算法以及预测分析等，结合电力负荷预测的特点，为解决自动获取具有高度相似性综合知识特征的同类历史数据进行历史负荷预测和确定智能算法中的参数问题，所需要的知识挖掘技术主要包含分类技术、聚类技术、进化式程序算法三种，其中分类技术主要用于提取一定的分类模式，通过分类模式进行判别和待预测时点负荷数据类似地历史数据序列，以此作为训练集建立预测模型，此外还可以通过分类技术进行属性选择的工作；聚类技术主要用于在已知输入变量的情况下，通过输入变量和历史负荷序列中的相同变量进行聚类分析，提取

和待预测日相似的序列进行智能预测模型的训练；进化式程序算法用于智能预测模型中参数的自适应调整，使得智能预测模型可以自动的调整自身所需要的相关参数，避免在确定相关参数时凭借人为主观性或者依靠经验进行反复的调整。下文对这三种较为成熟和常用的技术进行介绍。

（1）分类技术

分类的目的是学会一个分类函数或者分类模型（也常称为分类器）该模型能把数据库中的数据项根据其共同属性，映射到给定类别中的某一个。分类和回归都可以用于预测。预测的目的是利用历史数据记录自动推导出对给定数据的推广描述，从而能对未来数据进行预测。和回归方法不同，分类的输出是离散的类别值，而回归的输出则是连续值。要构造分类器，需要有一个训练样本数据也就是训练集作为输入。训练集由一组数据库记录或者元组数据构成，每个元组数据是一个关键字段（又称为属性或特征)值组成的特征向量，这些字段和大数据库（测试集）中的记录字段相同。另外，每个训练样本还有一个类标记。分类的过程一般分为以下几个主要步骤：

1）将现有的已知类别的数据划分为训练数据和测试数据两部分。

2）通过构造分类算法对训练数据进行学习，最终得到一个符合学习要求的分类模型，它可以以分类规则、决策树或数学公式等形式给出。

3）使用分类模型对测试数据进行检测，如果符合测试要求(如测试精度)，则进行第4步；否则，返回第2步。

4）应用得到的分类模型对未知类别的新数据进行分类。

其中，上述的第1步处理目前主要有两种划分方法：保持（holdout）方法和 k 倍交叉验证（k-fold cross validation）方法。保持方法将已知数据随机划分为训练数据和测试数据两部分，一般做法是三分之二数据作为训练数据，其余三分之一作为测试数据。使用训练数据导出分类模型，其在测试数据上的分类精度作为最终的分类精度。k 倍交叉验证将已知数据随机划分为 k 个互不相交的大致相等的数据子集 $S_1,S_2,\cdots S_k$。训练和测试迭代进行 k 次。在第 i 次迭代，S_i 作为测试数据，其余的子集用于训练分类法。最终的分类器的分类精度取 k 次测试数据上的分类精度的平均值。其中主要的分类方法有以下几种：

1）贝叶斯分类法

贝叶斯分类是统计分类方法，可预测类成员关系的可能性。贝叶斯分类的基础是贝叶斯定理。当类条件独立假设成立时，即假定一个属性值对给定类的影响独立与其他属性值，称为朴素贝叶斯分类算法。朴素贝叶斯分类算法性能可以和决策树与神经网络算法相媲美，在大型数据库的分类中具有高准确率和高速度的特点。贝叶斯网络是图形模型，它表示了属性子集间的依赖关系。

2）决策树分类法

决策树(Decsiionrtee)模型的基本原理是以一种递归方式来划分变量。其树状结构内部每一个节点表示在一个属性上的测试，每个分支代表一个测试输出，最终每个叶子节点表示了类或类分布。目前常用的决策树算法包括 ID3，AFCT，QUEST，CHAID，Magidson。其中 ID3 的增量版本包括 ID4，IDS，NIFERULE，KTE，SLIQ，BOAT 以及 Kmaber 等，其中以 ID3 和 CART4.5 的决策树分类技术最为常见，本文利用的就是 ID3 的决策树分类技术，其具体原理在下文中有详细的介绍。

3）神经网络分类法

神经网络是模拟人类大脑的结构和功能，由若干简单神经元按一定规则连接构成的网络系统。它能够采用某种学习算法从训练样本中学习，并将获取的知识存储在网络各单元之间的连接权重中。神经网络由于具有良好的非线性映射能力和对任意函数的准确逼近能力，用于分类问题往往能获得很高的分类精度，因而被公认为分类性能最好的分类方法之一。神经网络具有优良的鲁棒性，在噪声环境下也能很好的完成分类任务。另外，同粗集理论一样，神经网络也无需提供被分析数据之外的任何先验信息。从规则抽取的策略来看，现有的神经网络规则抽取方法可以分为两类：分解方法和教学方法。分解方法以神经网络中的每个神经元作为规则抽取的单位，首先为每个隐单元和输出单元生成各自的规则，然后依照网络的传导方向聚合这些规则得到整个网络的规则。教学方法也称黑箱方法，因为它在抽取规则时将神经网络看作一个“黑箱”，不考虑神经网络的内部结构和连接权重，直接从网络输出与网络输入之间的映射关系中来抽取相应的网络规则。

4）支持向量机（SVM）分类方法

支持向量机是统计学习理论的一个实现方法。统计学习理论是目前针对小样本统计估计和预测学习的最佳理论，它从理论上系统地研究了经验风险最小化原则成立的条件、有限样本下经验风险与期望风险的关系及如何利用这些理论找到新的学习原则和方法等问题。

在分类的过程中，经常遇到用于数据分析的数据可能包含数以百计的属性，其中大部分属性与挖掘任务不相关，是冗余的。尽管领域专家可以挑选出有用的属性，但这可能是一项困难而费时的任务，特别是当数据的行为不清楚的时候更是如此；此外，不相关或冗余的属性增加了数据量，不仅会减慢挖掘进程，而且会使发现的知识质量很差。而遗漏相关属性或留下不相关属性又是有害的，会导致所用的挖掘算法无所适从。因此在分类的过程中就衍生出了一个属性选择的研究方向。具体说来，属性选择就是一个从原有的属性集合中选择一个（相

对某种评价准则）最优属性子集的过程。这个属性子集应当保留原属性集合的全部或大部分分类信息。属性选择涉及两个关键问题：一是有效的属性子集的搜索策略；二是合适的评价标准。属性选择的思想是首先按某种搜索策略得到一个属性子集，然后使用选定的评价准则对此子集做出评价或比较，以决定是保留还是放弃。此过程反复进行，直到找到一个满意的属性子集。

原则上讲，只有穷举搜索才能找出属性的最佳子集，但是当属性个数和数据量急剧增加时，这种方法可能是不现实的。因此，通常采用压缩搜索空间的启发式搜索策略。这些策略通常是贪心算法，即在搜索属性空间时，总是做看上去是最佳的选择。它们的策略是做局部最优选择，期望由此导致全局最优解。在实践中，这种贪心算法是有效的，并可以逼近最优解。

属性选择的启发式搜索算法可分为以下几类：

1）包装(wrapper)算法：该算法开始时设定目标属性集合为空集，每一步依照选定的属性评价准则选择原属性集中最好的属性，并将它添加到目标集合中。在其后的每一次迭代，将原属性集剩下的属性中的最好的属性添加到目标集合中，直到得到满意的目标属性集合。

2）过滤(filter)算法：该算法设定目标属性集合由整个属性集开始，每一步依照选定的属性评价准则删除属性集中的最坏属性，直到目标属性集合达到要求。

3）包装与过滤相结合算法：算法的每一步选择备选集合中的一个最好的属性加入目标集合，并从备选集合中删除一个最坏的属性。

一般说来，包装算法可以达到很高的分类精度，但是与过滤算法相比，它需要更多的计算，特别不适用于处理大规模数据。而过滤算法由于忽略了所选属性子集在分类算法上的性能表现会导致分类性能下降，而且不同的启发式算法往往偏好属性值较多的属性。使用组合方法在继承两者优点的同时抑制其缺点是一种常用的策略。

（2）聚类技术

聚类是把一组物理或者抽象对象按照相似性归为若干类，也称为“无指导分类”。其准则是使得同一类别中对象间的距离尽可能的小，而不同类别中对象间的距离尽可能的大。对于一个很大的多维数据集，在数据空间中数据点通常不会均匀分布。数据聚类方法可以找出稀疏和稠密的位置，进而发现数据集的整个分布模式。当要分析的数据缺乏描述信息，或者无法组织成任何分类模式时，利用聚类可以自动找到合适的分类。聚类方法包括统计方法，机器学习和神经网络方法等。概括的讲，聚类分析算法可以分层三种不同的类型：

（1）试图找到一个最优划分以把数据分成指定数量聚类的方法，简称为划

分聚类法；

（2）试图发现聚类机构的层次方法，也称为层次聚类；

（3）对潜在聚类建模的基于概率的模型方法，也成为概率聚类。

每种类型具体方法如下：

1）划分的方法

在基于划分的聚类中，聚类的任务就是把数据集划分成 K 个不相交的点集，使每个集合中的点尽可能同质，用数学语言描述就是给定 n 个数据点的集合，使每个点被分配到一个唯一的聚类 q。同质性是通过选取适当的评价函数使得每一点到它所属聚类中心的距离最小化。基于划分的聚类方法的重点就是评价函数。在聚类分析中，最大化(或最小化)评价函数通常是计算复杂度很高的搜索问题，因此经常使用递归的启发式搜索算法来优化评价函数。基于局部搜索的递归改善算法是聚类分析中最为流行的一种。其一般步骤为：从随机选取的聚类开始；然后重新分配点使评价函数最大程度的增长（或减少）；然后再重新计算更新后的聚类中心;再次重新分配点，如此继续直到评价函数没有变化或者聚类成员没有变化。这种方法的优点是简单而且保证至少得到评价函数的局部最大（或者最小），对应的缺点是无法知道是否收敛到的聚类 C 与最佳的可能聚类（所用评价函数的全局最优）相比的好坏程度。基于局部搜索的递归改善算法最著名的范例就是由 MacQuene 提出的 K-means 均值法。

2）层次的方法

层次的方法对给定数据对象集合进行层次的分解。和基于划分的聚类方法不同，层次法逐步融合点或者切分超聚类。根据这一基本思想，层次划分法可以分为两种：凝聚和分裂。前者是由单个对象开始，逐渐合并较小的簇，最终形成一个包含所有对象的簇或者达到某个终结条件；而后者先将所有对象置于一个簇中，然后逐步细分为越来越小的簇，直到最后每个对象都自成一簇或者到达终结条件。两者相比，凝聚法更重要而且应用更广。层次聚类法的一个特点就是很难把模型从评分函数和用来决定最佳聚类的搜索方法分离开。层次聚类虽然简单，但经常遇到合并或分裂点的选择困难。并且这种方法的可伸缩性比较差，因为合并或分裂的决定需要检查和估算大量的对象或簇。改进的一个方向是将层次聚类与其他的聚类技术集成，形成多阶段的聚类。层次聚类的主要算法有：凝聚层次聚类和分裂层次聚类法；CURE 算法以及 Chmaeleon 算法等等

3）基于密度的方法

基于密度的方法中的每一个聚类（分量）都对应于一个假定的概率模型。在这一框架中，假定数据来自于一个多元有限混和模型，建模的一般过程如下:

对于给定的数据集，首先决定想要用多少个聚类拟合数据（即确定 K），然后再为这 K 个聚类中的每一个选取参数模型（比如多元正态分布）；最后再用 EM 算法来根据数据确定分量的参数和分量的概率。通常使用数据的似然（对于给定的模型）作为评价函数。一旦找到了分解模型，便可以把数据分配到各个聚类中。

（3）进化式程序算法

知识挖掘中的进化式常用算法是指利用计算机技术来通过模拟生物进化的方法进行群体随机搜索学习来求出目标函数最优取值的过程，其基本思想是通过模拟由个体组成的一个种群的集体学习过程，在这个过程里，每个个体可以看成是给定问题求解可行域的一个可以随机移动搜索的解，通过利用群体间的个体进行随机的选择、交异和交叉过程来吸收别的个体的信息来在空间中向更好的可行解区域随机移动，达到一种自动优化的过程。

知识挖掘中的进化式算法以遗传算法为代表，最早意识到自然遗传算法可以转化为人工遗传算法的是 Holland 教授。1965 年，他首次提出了人工遗传操作的重要性，并将其应用于自然系统和人工系统中。1967 年，Bagley J.D 在他的论文中首次提出了遗传算法这一术语，并讨论了遗传算法在自动博弈中的应用。1975 年是遗传算法研究史上十分重要的一年。这一年，Holland 出版了他的著名专著《自然系统和人工系统的适应性》。该书系统地阐述了遗传算法的基本理论和方法，提出了对遗传算法的理论研究和发展极为重要的模式理论，该理论首次确认了结构重组遗传操作对于获得隐并行性的重要性。人们常把这一事件视作遗传算法得到正式承认的标志，Holland 也被视作是遗传算法的创始人。

遗传算法研究的真正兴起是在 80 年代末至 90 年代初，其不仅表现在理论研究方面，还表现在应用领域的研究方面。随着遗传算法研究和应用的不断深入,一系列以遗传算法为主题的国际会议也十分活跃：开始于 1985 年的 ICGA 每两年举办一次；在欧洲，开始于 1990 年的 PPSN 会议也每两年举行一次；还有如 FOGA 会议自 1990 年开始也是每两年举办一次。这些会议的举办体现了遗传算法正不断地引起学术界的重视，同时这些会议的论文也集中反映了遗传算法近年来的最新发展和动向。

遗传算法的基本思想可归为两点：第一，将物种进化的理论用于求问题的解，物种的进化又可分为遗传和变异两个方面；第二，只有最适合环境的物种才能保留下来，因而需经反复求解后才可以得到最佳的解。

遗传算法按照一定的规则生成经过基因编码的初始群体，然后从这些代表问题的可能潜在解的初始群体出发，挑选适应度强的个体进行交叉（或称交配、

交换）和变异，以便发现适应度更佳的个体，如此一代代地演化，得到一个最优个体，将其解码，该最佳个体编码则对应问题的最优解或近似最优解。

由于遗传算法是自然遗传学和计算机科学相互渗透而成的新的计算方法，所以遗传算法中经常使用一些有关自然进化的基本术语，了解这些术语对于学习和应用遗传算法以及其他进化式算法是十分必要的。

首先，生物遗传物质的主要载体是染色体，染色体由基因构成。基因是控制生物性状的功能单位。基因在染色体中的位置称作基因位，而基因所取的值叫做等位基因。基因和基因位决定了染色体的特征，也决定了生物个体的性状。

此外，染色体有两种相应的表现模式，即基因型和表现型。表现型是指生物个体所表现出来的性状，而基因型由与表现型密切相关的基因组成。同一种基因型的生物个体在不同的环境条件下可以有不同的表现型，因此表现型是基因型与环境条件相互作用的结果。

在遗传算法中，染色体对应的是数据或数组，而在标准遗传算法中，这通常是由一维的串结构数据来表示的。串上各个位置对应上述的基因位，而各位置上所取的值对应于上述的等位基因。遗传算法中处理的是染色体，或者叫基因型个体。一定数量的个体组成种群，而种群中个体的数目称为种群大小,每个个体对环境的适应程度叫做适应度。

执行遗传算法时包含两个必需的数据转换，一个是表现型向基因型的转换，另一个是基因型向表现型的转换。前者是把搜索空间中的参数或解转换为遗传空间的染色体或个体，此过程叫做编码操作；后者是把遗传空间的染色体或个体转换为搜索空间中的参数或解，此过程叫做解码操作。

遗传算法所代表的进化式算法具有很强的鲁棒性，这是因为它与传统的优化搜索方法相比，采用了许多独特的方法和技术，归纳起来，主要有以下几个特点：

（1）遗传算法的处理对象是问题参数的编码集，而不是参数本身，遗传算法将优化问题的参数编码成长度和代码集均有限的码。它是在求解问题的决定因素和控制参数上进行操作，从中找出高适应度的位串，而不是对函数和它们的控制参数进行直接操作，这样遗传算法就不受函数的限制条件，比如可导性、连续性、单峰性等的影响。

（2）遗传算法具有并行性

遗传算法的搜索是从问题的解位串集开始搜索，而不是从单个解开始。其并行性表现在两个方面：其一，遗传算法具有内在并行性，它本身适合大规模并行。遗传算法适合在目前所有的并行机或分布式系统上进行处理，而且对并行效率没有太大影响；其二，遗传算法具有隐含并行性，由于遗传算法采用种

群的方式组织搜索，因而可以同时搜索解空间内的多个区域，并相互交流信息。这样，使遗传算法具有极好的全局搜索性能，从而减少了陷入局部最优解的可能。

（3）遗传算法仅使用适应度函数指导搜索，而不需要其它辅助信息

传统搜索算法需要一些辅助信息，如梯度算法需要导数信息才能爬上当前的峰值点，这就要求目标函数可导，而遗传算法只需适应度函数和编码串即可。同时，遗传算法在增加收益和减小开销之间进行权衡，找到最优的策略。

（4）遗传算法采用概率转移规则，而不是确定性规则

遗传算法在搜索过程中以概率转移规则来指导它的搜索过程朝着更优化的解区域移动，这对于熟悉用确定性规则的人似乎不是很容易理解，但使用概率并不是说遗传算法只是简单的随机搜索，而是用随机工具去指导搜索，向着搜索空间中可能的最好区域进行。

（5）遗传算法易于介入已有模型，具有可扩展性，易于同其他技术结合

遗传算法是多学科结合与渗透的产物。遗传算法已经广泛应用于计算机科学、工程技术、管理科学、社会科学等越来越多的领域中，与其他技术结合使用，可以解决许多领域中用传统方法解决不了的问题。本文将进化式算法和智能负荷预测算法进行结合也是一个比较典型的应用实例。

遗传算法主要由六个部分组成：染色体编码、初始群体产生方法、个体适应度评价、遗传操作(选择算子，交叉算子，变异算子)、算法终止条件以及遗传参数的设置。要利用遗传算法成功地解决优化问题，每个部分的设计都非常关键。而后的进化式算法全部继承了这种思想，虽然在不同程度上有所改动，但是其基本程序仍然是这六个部分。

（1）染色体编码方法

遗传算法不能直接处理问题空间中的参数，而只能把它们转换成遗传空间中由基因按一定结构组成的染色体或个体，这一转换操作就叫做编码，它是实际问题与遗传算法建立联系的一种途径。

目前，大部分理论都是基于定长二进制进行编码，即将问题域参数空间中的一个点映射到个体的染色体上，二进制每一位即为染色体上的一个基因，字母表为{0,1}。模型参数的二进制编码是一种数学上的抽象，通过编码把具体的优化问题和生物演化过程联系起来，因为这时形成的编码字符串就相当于一组遗传基因的密码。在二进制编码中，按一定的顺序每几位二进制数（即一个基因链码）对应一个参数变量，从而通过二进制串表示了问题域信息。染色体长度取决于参数取值范围和模型分辨率。

但是，二进制编码可能出现不连续问题，即在欧氏空间中邻近点的二进制

编码的 Hamming 距离并不邻近。二进制编码通常不直接反映所求问题的特定知识，并且通常需要对问题的参数进行编码和解码，即需要实现基因型和表现型上的映射。

为了实现直接在解的表现上进行遗传操作，人们提出了浮点数编码方法.也即十进制数编码或实数编码方式。

实数编码方式是以实数位串进行编码的遗传策略。其优点在于：精度高，搜索空间大。与二进制编码相比，实数编码在变异操作上可以保持更好的群体多样性，但其搜索能力不如二进制编码强。大量试验表明：二进制编码和实数编码在解同一优化问题时，不存在显著的性能差异。Michalewicz 经过对比分析认为：二进制编码的进化层次是基因，而实数编码的进化层次是个体，二者的理论基础是不同的。本文应用的微分进化算法所基于的正是实数编码的进化式算法。

事实上，多种编码方法，如二进制编码方法、浮点数编码方法、符号编码方法、格雷码编码方法、多参数编码方法等等，各有优缺点，目前还缺乏一种理论来判断各种编码方法的好坏，并指导对它们进行设计。

（2）初始种群生成方法

初始种群的方法目前采用的方式大多是利用随机生成的方式，但是在某些问题的研究中由于初始种群的生产方式对于目标函数的求解过程或者是计算时间是非常重要的，因此也有一些对随机生成方式进行改变的初始种群生产方法，这些方法可以归结为两类：一类是使得初始种群的分布更加分散，另一类是根据某种原则或者是经验等先设置一些区域可行解内的较优点，然后在这些较优点附近进行随机生成。

（3）适应度函数的设计

在遗传算法中，利用适应度来衡量个体的优劣，采用适者生存的原则决定哪些个体进行繁殖，哪些个体要被淘汰。在具体应中，往往采用费用、盈利、方差等目标表达式，但为了统一起见，要做一些变换。在遗传算法的讨论中，一般希望适应度越大越好，且要求适应度非负。

（4）选择算子

选择算子又称复制算子、繁殖算子。选择是从种群中选择生命力强的染色体产生新种群的过程。选择的依据是每个染色体的适应度大小。适应度越大，被选中的概率就越大，其子孙在下一代产生的个数就越多。它在很大程度上决定着遗传算法的收敛速度和收敛效果。遗传算法是一种具有导向性的随机化搜索方法，选择算子正是遗传算法具有导向性的本质所在，也是遗传算法具有局部搜索能力的重要保证。选择算子的操作可以分为以下两步：首先，计算适应

度；然后，在适应度计算之后是实际的选择，按照适应度进行父代个体的选择。选择算子提高了群体的平均适应度，低适应度的个体趋向于被淘汰、高适应度的个体趋向被复制，但选择算子并没有产生新的个体。

（5）交叉算子

交叉算子又称重组、配对算子。当许多染色体相同或后代的染色体与上一代没有多大差别时，就利用交叉算子来产生新一代染色体，所以交叉算子可以产生新的个体。

染色体交叉分两个步骤进行：首先，在新复制的群体中随机选取两个染色体，每个染色体由多个位(基因)组成；然后，沿着这两个染色体的基因随机取一个位置，二者互换从该位置起至末尾的部分基因。根据交叉点的数目不同，可以将交叉分为单点交叉和两点交叉。在单点交叉中，交叉点之后的所有字符全部参加交换。在两点交叉中，只有两个交叉点之间的字符才参加交换。遗传算法的有效性主要来自选择和交叉操作，尤其是交叉操作，在遗传算法中起着核心作用。

（6）变异算子

选择和交叉算子基本上完成了遗传算法的大部分搜索功能，而变异则增加了遗传算法找到接近最优解的能力。变异就是以很小的概率，随机改变字符串某个位置上的值。在二进制编码中，就是将 0 变成 1，将 1 变成 0，它本身是一种随机搜索，但与选择、交叉算子结合在一起，在保证遗传算法的有效性和局部随机搜索能力的同时，使遗传算法保持种群的多样性，以防止出现未成熟而过早收敛。目前，比较常见的变异算法主要有反转变异、初始化变、步长变异和高斯变等等。

遗传算法是一种反复迭代的搜索算法，它通过多次进化逐渐逼近最优解，因此需要确定终止条件。最常用的终止方法是规定遗传（迭代）的次数。另外，如果目标函数是方差之类、有最优目标值的问题时，可以采用控制偏差的方法实现终止。第三种实现终止的方法是检查适应度的变化，一旦最优个体的适应度没有明显变化时，就终止算法。

1.4 本文的研究思路及研究内容

本文提出的基于知识挖掘技术的智能协同电力负荷预测研究思想，其研究思路是旨在结合知识挖掘、数据挖掘以及智能电力负荷预测方法进行协同电力负荷预测，通过知识挖掘对数据库中的历史负荷文本数据及对应的各类影响因素进行分析处理，从而得出一系列的知识、经验、推理规则等，放入知识库中。

预测时通过计算与预测目标的各类知识特征的总体关联程度大小，自动提取具有高度相似性综合知识特征的同类历史数据，结合智能算法和电力负荷预测方法建立具有针对性的预测模型对负荷进行预测以及和经验知识的后干预工作，用以进一步克服以前的预测方法的不足，使预测精度得到突破性的提高。以智能预测方法为主，知识挖掘方法为辅。目前利用智能方法对负荷进行预测的一般步骤可以分解为以下几步：

（1）收集数据，经过数据预处理后形成可规范的，可适用于智能预测模型进行预测的样本数据集；

（2）确定输入变量和输出变量，对于预测问题来说，输出变量一般比较固定，即待预测时间点的负荷；

（3）将样本数据集分成两个部分，分成训练集和测试集，其中训练集的作用是对智能预测模型的相关参数进行训练，测试集用来测试模型的预测能力，有时为了避免智能模型的过拟合能力，将训练集进一步分为训练集和验证集，利用验证集的数据对智能预测模型进行训练中的验证作用；

（4）调整预测模型的各种参数，根据设定的误差指标以达到较为理想的预测效果；

（5）利用训练好的模型进行实际应用中的预测。

从上述步骤来看，利用智能方法对负荷进行预测需要处理好以下三个主要问题：

1）对数据进行规范化；

2）确定输入变量；

3）调整智能预测模型的相关参数。

本文在引入知识挖掘技术与智能负荷预测方法进行协同预测时，也是从上述三个方面进行考虑并进行研究，利用知识挖掘进行协同智能预测首先需要进行相应数据的准备，第二章就是基于知识挖掘的相关知识对负荷相关数据的规范化方面进行了研究，主要是对影响负荷的属性变量进行分类，根据这些变量的特性进行数据库规范或者是文本规范的设计，根据数据规范对数据进行预处理方法的研究，形成可供挖掘的数据视图。

第三章利用知识挖掘的分类技术挖掘和预测日相似的分类或关联规则，通过对数据库中的数据进行挖掘和归类并序化排列，将具有高度相似影响因素特征的负荷历史数据提取出来，为负荷预测模型做好数据输入和建模的前期数据准备工作。通过自动提取与预测目标具有高度相似性综合知识特征的同类历史数据，可以强化负荷变化规律。该部分的研究内容分为两个部分，由于目前对于气象因素影响负荷预测的研究刚刚起步，因此在我国大部分的电力公司中还

没有开始对气象因素进行收集，或者只形成了少量的数据，并不具备利用气象数据进行知识挖掘的条件。因此本章分两种情况对基于知识挖掘分类技术对寻找和预测日具有高度相似性类型的日负荷历史数据进行了研究，第一种是在纯负荷数据情况下利用相似度排序技术按相似特征强弱提取负荷数据，由此组成规律强化、干扰弱化、具有高度综合相似特征的数据序列，另一种是在具有大量气象信息的基础上利用知识挖掘中的决策树分类技术、聚类技术和粗糙集约简技术对具有相似特征的历史数据序列进行提取，得到具有高度综合相似特征的数据序列，之后再通过建立自组织变结构优化神经网络负荷预测模型进行预测，经过处理后利用智能模型可以更准确地反映变化规律，使得预测精度就有明显地提高。

第四章利用知识挖掘中的进化算法对目前研究中较为流行的支持向量机方法的参数选取问题进行了研究，使得支持向量机方法能够更有效地进行中长期负荷预测。由于支持向量机方法利用结构风险最小化替代了传统神经网络利用的经验风险最小化的原则，通过引用核函数将输入空间中的非线性问题映射到高维特征空间中构造线性函数判别，由此转化为凸优化问题，可以保证算法的全局最优性，支持向量机能够有效地解决神经网络中的容易陷入局部最优值的问题，因此被很多学者看成是能够有效地替代神经网络的一种算法。由于支持向量机是以统计学理论为基础的，与传统统计学习理论不同的是它主要是针对小样本情况，因此支持向量机适合中长期负荷预测的情况，但是利用支持向量机进行预测需要确定三个参数，即学习参数 C，ε和利用高斯核函数的参数 σ，而利用知识挖掘中的微分进化算法结合支持向量机对这三个参数进行自适应调整可以使模型具有自适应参数优化的功能，可以避免根据人为主观意愿或者凭经验对参数进行调整的不足。

第五章利用知识挖掘对日最大负荷预测的预测结果进行后干预的调整，由于日最大负荷预测容易受到气温、湿度等各方面影响因素以及不规则事件产生的影响，因此需要利用知识挖掘的决策树分类技术对利用常规方法得到的预测结果进行进一步的调整纠偏工作，这部分的调整主要是纠正一些不规则事件对负荷的影响，这些影响带来的误差很难通过进一步调整智能预测模型的预测能力来进行适当的纠正，由此可以借助知识挖掘进一步挖掘不规则事件的影响模式，可以进一步提高负荷预测的精度。

第六章对基于知识挖掘技术的智能协同预测结果对电力负荷的预警工作进行研究，利用中长期负荷预测得到的结果对实际情况进行供需平衡方面的预警监测工作，利用短期负荷预测得到的结果对实际情况进行实时监测，以及利用知识挖掘的分类技术提取出相应的冰灾防范预警条件进行典型的极端天气进行

预警监测。通过研究可以使得电力部门的相关工作者能够提前准备相关工作，协助相关的决策使用。

第七章对上述的算法结合广东省江门市的实际情况进行了计算机软件上的实现，对知识挖掘智能协同负荷预测系统进行了研究，该系统不但实现了上述文章中的算法，而且对用电市场的指标进行分析，包括年度及月度供电量、售电量、分行业用电量以及负荷特性分析、日负荷特性分析等；在负荷预测部分还包括了传统的一元线性回归、灰色预测分析、二次多项式回归、自适应指数预测、指数预测、增长率预测法、非齐次指数预测、B.Compertz 模型、logistic 模型以及这些模型组合成的组合预测算法，在预测的基础上，系统还有相应的预警模块，可以对区域电网电力调度、规划、计划、用电等领域的相关指标进行良好的分析，从而辅助决策。

最后一章是本文的结论和展望，对本文的创新性进行了总结，并对未来的研究方向进行了展望。

第 2 章 基于知识挖掘技术的负荷数据规范设计

利用知识挖掘进行知识发现首先需要将数据进行规范的排列，按照不同的负荷预测实际要求形成相应的数据规范视图，本章首先对影响负荷预测的属性变量进行分析，然后对相应的数据规范进行设计，最后给出了形成相应数据视图的数据预处理方法。

2.1 影响负荷预测的属性变量分析

一般来讲，影响负荷预测属性变量按照是否是从负荷序列数据中得到的可以简单地分成负荷变量和非负荷变量两类，所谓负荷变量一般是指利用和待预测负荷类型相同的负荷序列中前几个时点的数据或者是前几个周期的同一时刻的数据作为自变量来对待预测点的负荷进行预测；而非负荷变量指的是利用和待预测负荷类型不同的因素对待预测点的负荷进行预测，如经济、气象数据、人口等；此外，当对电价进行预测时，如果用到电力负荷数据作为自变量，此时的电力负荷数据也归之为非负荷变量。

2.1.1 负荷变量

利用负荷序列中自身数据对待预测点的负荷进行预测是目前负荷预测研究中所选取的最常见的自变量。由于这些自变量已经是发生过确定性的数值，因此负荷变量的数据非常容易获取，并且。由于负荷带有连续性和周期性的特点，因此在选取负荷变量时一般选取前几个时点的数据和同一时点的前几个周期的数据，用数学语言描述即：假设待预测的负荷为 l_t，则选取的负荷变量可记为 $l_{t-1},l_{t-2},\ldots,l_{t-m},l_{t-c},l_{t-c*2},\ldots,l_{t-c*n}$，$c$ 表示一个周期的时点数（如月度负荷预测的周期是 12），一般来讲 $m<c$，其中表示 $l_{t-1},l_{t-2},\ldots,l_{t-m}$ 表示前 m 个时点的数据，这些变量用来预测外推的趋势性，而 $l_{t-c},l_{t-c*2},\ldots,l_{t-c*n}$ 表示取同一周期前 n 个时点的数据，这些变量用来预测周期性。

2.1.2 非负荷变量

非负荷变量大致可以分为日期型变量、气象类变量、经济类变量和其它变量四大类。各类变量的具体内容如下：

（1）日期型变量

时间变量：如上文所述，由于负荷具有连续性和周期性的特点，因此时间作为标记负荷数据的尺度同时也是一种影响负荷的因素，两条同周期的日负荷曲线在接近的时点会出现负荷峰值和负荷低谷，这是由于在不同的时间人们从事的活动不同而造成的。在知识挖掘中，时间变量有助于对负荷曲线进行聚类，以发现相同负荷模式来提高预测精度的作用。

工作日时间变量：对于日负荷曲线而言，由于生产和人们的活动在工作日和非工作日是有区别的，因此对于工作日和非工作日两者的日负荷曲线模式是不同的，而对于非工作日又可以分成平常周末和特殊节假日两种，这同样也是基于在非工作日中，对于平常的周末休息和特殊节假日的人类活动情况不同而造成的负荷曲线模式不同。其中平常周末指的是非节假日无串休情况的周六周日，而特殊节假日指的是元旦、除夕、春节以及连休日、五一劳动节、十一以及连休日，清明节，端午节等法定规定的节日。

（2）气象类变量

温度：由于空调、电暖器等电力调温设备的普及，使得人们在超出人体适宜温度范围的时候会采用相应的电器进行调温来使自己所在的环境舒适，在北方温度影响较大的是夏季的空调负荷，南方除了夏季的空调负荷外，还存在着在较低的温度下对负荷影响较大的冬季取暖负荷。温度对负荷产生的影响多是非线性影响，由于智能预测模型一般都具有比较好的非线性拟合能力，因此在目前的预测研究中开始被广泛的关注并采用，并且很多研究成果均证实在引入温度后能够提高负荷预测的精度。

湿度：同样温度下，不同的湿度给人们带来的感觉是不同的，以北京为例，夏季有名的“桑拿天”就是在温度较高的时候湿度也较高，使得人们在同等温度下感觉更加难受，因此带来空调负荷的增长，从而拉动负荷的需求。目前有少量的研究将湿度直接作为输入变量进行负荷预测。

人体舒适指数：人体舒适指数是反映人体对于环境的感觉程度，这个指数可以有效解释人体对于环境产生的反映，该变量有助于知识挖掘中对于负荷模式的分类和聚类研究，根据我国气象的相关规定，人体舒适指数的取值范围是1到11之间的整数。

天气描述：天气描述是指以语言描述的天气情况，例如：晴、阴、多云、小雨、大雨等，天气描述变量同样有助于知识挖掘对负荷模式的分类工作，例如预测日经气象预报得知将有“大雨”，而前 n 天的气象状况都为“晴天”，由于“雨天”和“晴天”的负荷变化差异极大，因此无论创建或使用多么先进的预测模型，根据一系列具有“晴天”特征的负荷历史数据建模，只能得出具有“晴天”特征的负荷预测值，而不能得出具有“雨天”特征的负荷预测值，可

能会导致较大的预测误差。此外，特殊的天气如台风，沙尘暴、暴雨、冰灾等天气描述不但会对负荷预测的模式分析起作用，同时也将会对预警分析起到相应的作用。

（3）经济类变量

经济类变量一般在中长期负荷预测或者是超长期预测中常常用到，例如在对电力工业做 5 年规划或者 5 年以上的长期规划所进行的相关负荷预测时，一般需要考虑经济发展的 GDP 值、人口数量等经济类变量，这类变量将会对长期负荷预测有很大的影响。

（4）其它变量

其他变量指的是除上述变量外对负荷模式起到影响的变量，这些变量一般都体现在描述性的语言中，例如 2008 年北京举办奥运会期间北京的负荷模式明显不同于以往同期的负荷模式，再如 2008 年底美国的金融危机导致许多地区尤其是广东地区的负荷明显呈现下降趋势，利用常规的负荷预测明显无法对这种情况下的负荷进行有效的预测，此外，相关的政策影响，如节能减排政策下的高耗能产业调整很有可能将会对未来的电力负荷产生较大的影响。

2.2 基于知识挖掘技术的负荷数据存储规范及数据视图设计

基于知识挖掘技术的智能协同电力负荷预测的负荷数据库结构如图 2-1 所示，其中，数据库构成可以分成三个部分，负荷数据库主要记载负荷序列数据和相应的时间节点数据。气象数据库主要储存气象类型的相关数据，其中气象数据既可以从气象预报的相关渠道给出的规定文本中提取，也可以直接从气象预报的相关网站上利用关键词搜索技术进行分析存储；而文本数据库存储一些描述性的文本内容，这些描述性的文本内容可以从 Internet 网上利用关键词搜索技术进行下载，经过处理后储存。数据抽取是从三个数据库中抽取出负荷预测要求的数据进入数据仓库，在数据抽取时对数据进行预处理，形成的视图可以直接供预测模型直接调用或者供知识挖掘技术进行模式分析调用。

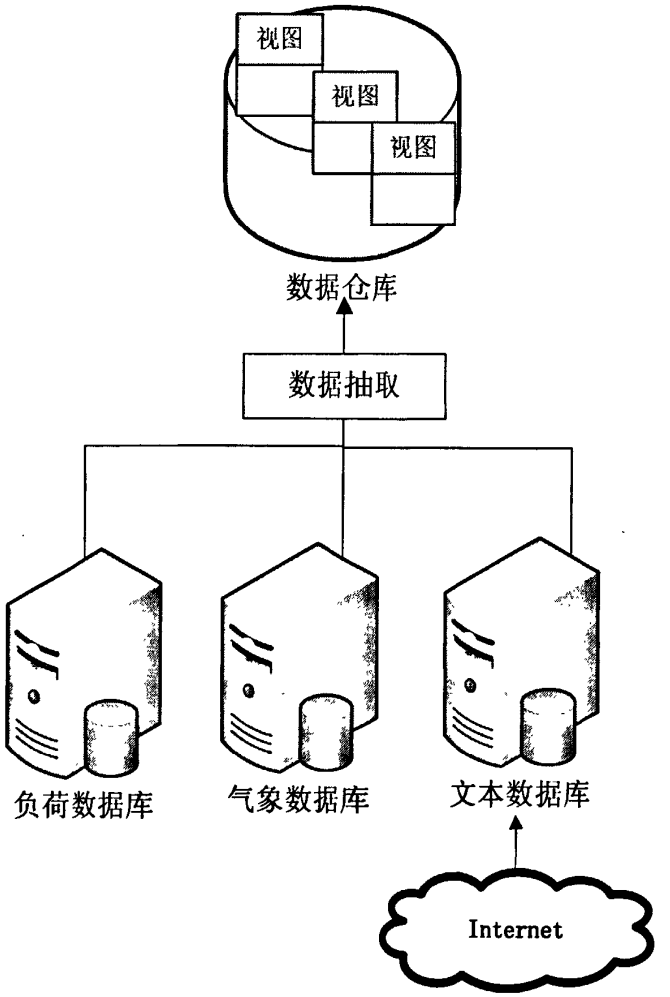


图 2-1 基于知识挖掘技术的智能协同电力负荷预测的负荷数据库结构图

Fig.2-1 Database structure of collaborative knowledge mining and intelligence load forecasting

2.2.1 负荷变量的存储规范

电网目前的负荷数据收集大多数来源于电网公司采用的 SCADA 系统，其中 SCADA 系统收集的数据是按照一定的时间间隔将实时的负荷数据传输到相应的数据库中，因此负荷数据库的数据规范可以设计成记载数据发生的日期变量和负荷变量两个方面以方便和 SCADA 系统之间进行相互通信，目前电网系统常用的负荷预测中的最高要求是 96 点负荷预测，即每 15 分钟预测一次，因此从 SCADA 系统数据库的接口提取出相关数据形成基于知识挖掘技术的智能协同电力负荷预测的负荷数据库规范如表 2-1 所示：

表 2-1 负荷变量的存储规范

Table 2-1 The storage criterion of load variables

字段名	字段类型	是否允许空值	含义
theTime	datetime	不允许	时间
point1	decimal(10,4),	允许	0：00 负荷值
point2	decimal(10,4),	允许	0：15 负荷值
point3	decimal(10,4),	允许	0：30 负荷值
……	……	……	……
point91	decimal(10,4),	允许	22：30 负荷值
point92	decimal(10,4),	允许	22：45 负荷值
point93	decimal(10,4),	允许	23：00 负荷值
point94	decimal(10,4),	允许	23：15 负荷值
point95	decimal(10,4),	允许	23：30 负荷值
point96	decimal(10,4),	允许	23：45 负荷值
BelongToCom	varchar(20),	允许	所属地区名称

从上述规范中可以方便的生成季度、月度、日、24 点以及 48 点的负荷预测视图供其余需求的负荷预测使用。

2.2.2 非负荷变量的存储规范

气象类型数据一般来源于当地的气象部门以及相关的气象发布网站上，此外对于一些描述型的天气数据可以从网上很容易的搜索到。类似地，关于气象类型的数据存储也需要带上日期信息以便和负荷数据库和文本数据库产生视图使用。气象类型数据的存储规范如表 2-2 所示。

描述性文本数据一般来源于负荷预测的相关管理人员的收集或者从 Internet 网上利用搜索功能进行获取，由于目前的搜索技术大多是基于较为成熟的关键词搜索技术，因此在对描述性文本进行存储时需要对关键词进行抽取。例如在对气象型的描述文本进行收集时，可以根据描述气象的相关关键词为“晴”、“多云”、“阴”、“小雨”、“中雨”、“大雨”、“暴雨”、“风”等等，再如对经济型的描述文本进行收集时，可以定义关键词为“平稳”、“迅速”、“增长”、“下降”等等，然后根据这些关键词对相关搜索的网站文本进行搜索形成相应的描述性文本，其数据存储规范如表 2-3 所示。

表 2-2 气象类型数据的存储规范

Table 2-2 The storage criterion of meteorological variables

字段名	字段类型	是否允许空值	含义
theYear	int	不允许,	年度
theMonth	int	不允许,	月度
theDay	int	不允许,	日期
DayWeather	varchar(200),	允许	当天天气描述
ForecastDayWeather	varchar(200),	允许	预计未来天气描述
MinTemperate	decimal(4,1),	允许	最小温度（摄氏度）
MaxTemperate	decimal(4,1),	允许	最大温度（摄氏度）
MaxHumidity	decimal(4,2),	允许	相对最大湿度（百分比）
MinHumidity	decimal(4,2),	允许	相对最小湿度（百分比）
UltravioletIncex	int,	允许	最高紫外线指数
HumanComfortable	int,	允许	人体舒适等级

表 2-3 描述性文本的存储规范

Table 2-3 The storage criterion of descriptive text

字段名	字段类型	是否允许空值	含义
theYear	int	不允许,	年度
theMonth	int	不允许,	月度
theDay	int	不允许,	日期
Keywords	varchar(200),	允许	关键词（用分号分隔）
Descriptions	varchar(2000),	允许	描述性文本
BelongToCom	varchar(20),	允许	所属地区名称

2.2.3 数据视图的规范

在定义了上述的基础数据表后，可以定义数据抽取技术对表格中的基础数据进行抽取成相应的数据视图作为预测的样本数据候选集，数据抽取相当于映射规则，通过三个数据库中的相关日期型数据和所属地区名称为主键对基础数据进行关联、抽取、交叉、组合等操作根据预测的实际需要生成各种数据视图。以按时间的负荷需求为例，下面给出 24 点负荷数据、日负荷数据的数据视图规范。

24 点负荷数据属于短期负荷预测，因此不但要采用负荷变量对未来预测点的负荷进行预测，也必须要考虑天气情况，不但需要全天的气象数据，而且需

要监测时点的数据，其数据视图的规范如表 2-4 所示。

表 2-4 日 24 点负荷数据视图的生成规范

Table 2-4 24-point load data view

字段名	字段类型	是否允许空值	含义
theYear	int	不允许,	年度
theMonth	int	不允许,	月度
theDay	int	不允许,	日期
BelongToCom	varchar(20),	不允许,	所属地区名称
Point0	decimal(10,4),	不允许,	0：00 负荷值
Point1	decimal(10,4),	不允许,	1：00 负荷值
Point2	decimal(10,4),	不允许,	2：00 负荷值
Point3	decimal(10,4),	不允许,	3：00 负荷值
Point4	decimal(10,4),	不允许,	4：00 负荷值
……	……	……	……
Point20	decimal(10,4),	不允许,	20：00 负荷值
Point21	decimal(10,4),	不允许,	21：00 负荷值
Point22	decimal(10,4),	不允许,	22：00 负荷值
Point23	decimal(10,4),	不允许,	23：00 负荷值
DayWeather	varchar(200),	不允许,	当天天气描述
MinTemperate	decimal(4,1),	不允许,	最小温度（摄氏度）
MaxTemperate	decimal(4,1),	不允许,	最大温度（摄氏度）
RealTemperate7	decimal(4,1),	不允许,	7 点实际温度（摄氏度）
RealTemperate16	decimal(4,1),	不允许,	16 点实际温度（摄氏度）
MaxHumidity	decimal(4,2),	不允许,	相对最大湿度（百分比）
MinHumidity	decimal(4,2),	不允许,	相对最小湿度（百分比）

虽然日最大负荷预测也同属于短期负荷预测范围，但是和上述短期负荷预测的日曲线负荷预测相比，日最大负荷预测的负荷数据的数量偏少，并且更容易受到其余因素的扰动影响，因此需要提供是否工作日的判别字段来对日期类别进行判断，此外，还需要和描述性文本关联来记载当天发生的特殊事件已供日最大负荷预测分析时使用。其规范见表 2-5 所示。

表 2-5 日负荷数据视图的生成规范

Table 2-5 Data view of daily load

字段名	字段类型	是否允许空值	含义
theYear	int	不允许,	年度
theMonth	int	不允许,	月度
theDay	int	不允许,	日期
BelongToCom	varchar(20),	不允许,	所属地区名称
Weekday	int	不允许,	判断是否周末，0 表示工作日，1 表示周末。
Special day	varchar	不允许,	记载何种类型的特殊日
maxloadvalue	decimal(10,4),	不允许,	日最大负荷值
DayWeather	varchar(200),	不允许,	当天天气描述
MinTemperate	decimal(4,1),	不允许,	最小温度（摄氏度）
MaxTemperate	decimal(4,1),	不允许,	最大温度（摄氏度）
MaxHumidity	decimal(4,2),	不允许,	相对最大湿度（百分比）
MinHumidity	decimal(4,2),	不允许,	相对最小湿度（百分比）
Keywords	varchar(200),	允许	关键词（用分号分隔）
Descriptions	varchar(2000),	允许	描述性文本

2.3 基于知识挖掘的数据预处理研究

在数据收集的过程中，难免会出现数据缺失、数据异常以及数据统计口径、单位等尺度不同的现象，因此在数据抽取时需要对原始数据进行某些相关的预处理工作对这些情况进行转换，利用映射规则生成数值反映到数据视图中，这样做既不会修改数据库中的原始数据，同样也不会使得数据视图中出现原始数据中出现的问题，使得预测工作更加简单。在进行知识挖掘技术的相关准备工作中，对知识挖掘所需要的数据质量有严格的要求，经相关资料统计，知识挖掘过程中的数据预处理工作是非常重要的，一般应该占整个数据挖掘准备工作的 80%以上的工作量[126]。数据预处理主要包括数据清理、数据集成和数据归约三类工作。其中数据清理工作指的是对数据中的遗漏值和异常值进行清理，而数据集成工作指的是将多个数据源中的数据进行整合的映射合并操作，数据归约指的是将数据集合中用于知识挖掘的部分进行识别，缩小数据处理的范围。本文在这三类所用的预处理方法如下：

（1）数据清理

数据清理的工作主要对数据中存在的缺失值进行补充，对数据中的异常值和孤立点进行识别、纠正等工作，其中数据清理的基本方法有以下几种：

1）缺失值的处理

对于缺失值处理较为常用的思路是利用最可能的取值来代替缺失值，目前常用的方法是利用回归、贝叶斯形式化方法工具或判定树归纳等进行缺失值的补充工作，此外利用全局的常量或者是中位数、平均值等替换进行缺失值的替换或者利用中位数、平均值等替换方法也是较为常见的缺失值处理方法，本文利用的缺失值处理方法是：对某一个负荷周期内只缺失少量数据的情况时，如果是序列中间的某一项资料空缺，利用缺失项两端数据的平均值近似代替，如果是序列开始的一项数据缺失，则利用公式 2-1 进行计算，并将结果近似代替。

$$l_t = l_{t-c} \times \frac{l_{t+1}}{l_{t+1-c}} \tag{2-1}$$

其中 l_{t-c} 表示周期为 c 的同一时点的上一周期值，l_{t+1} 表示下一时点值，l_{t+1-c} 上一周期的下一个时点值。

若一个负荷周期内缺少超过 80%以上的数据，或者是连续缺失数据超过 20%以上时，利用趋势外推的方法对缺失的数据进行预测，取预测结果作为替代值。

2）噪声数据的处理

噪声数据是指测量的变量中出现了随机错误和偏差的现象，其中较为常见的一种情况是出现了严重偏离期望的孤立点，一般对噪声处理采用的技术包括以下几种：

①分箱：分箱处理的思想是将出现噪声数据的局部区域数据封装成箱，然后利用箱平滑值、箱中值以及箱边界值进行平滑处理；

②回归：回归的思想是利用合适的两个变量来进行一元线性回归拟合，利用拟合的数值消除噪声，此外也可以利用多元线性回归的拟合数值进行噪声数据的处理；

③人工排查和计算机生成：将噪声数据隔离后交予人工进行审核，通过人工的干预，利用计算机的相关技术生成替代值对噪声数据进行替代。

④聚类：利用聚类技术对数据进行聚类，如果数据中出现了噪声孤立点，则这些噪声孤立点将会被划分成一类，对这些数据进行排查，将确实能够提供数据信息的点保留，剩下的点利用前三种手段进行修正。

3）异常值处理

异常值是指监测变量出现明显不符合常理情况下的取值，对异常值的判定可以用以下方法：规定负荷数据的理论最大值和理论最小值，当超出理论最大

值和理论最小值规定的范围并且是非缺失值时，可以视为监测数值出现异常；除此情况外，当序列中某一项值x_i超出本周期序列均值和前一周期序列均值的20%时，同样将x_i视为异常值。对异常值的处理方法是取上一周期序列的序列均值L_{t-1}，取本周期序列的序列均值L_t，求出上一周期和本周期的序列均值$\bar{x}_{t+1},\bar{x}_t$，按式2-2对异常值x_i进行修正，从而使历史数据序列趋于平稳。

$$x_i=\begin{cases}\max(\bar{x}_t\times(1+20\%),\bar{x}_{t+1}\times(1+20\%)), x_i>\max(\bar{x}_t\times(1+20\%),\bar{x}_{t+1}\times(1+20\%))\\ \min(\bar{x}_t\times(1-20\%),\bar{x}_{t+1}\times(1-20\%)), x_i<\min(\bar{x}_t\times(1-20\%),\bar{x}_{t+1}\times(1-20\%))\end{cases} \quad (2\text{-}2)$$

（2）数据集成

数据集成是指将多个数据源中的数据进行汇合，并将汇合后的数据进行合并存在同一个数据库、数据表或者是数据视图中，数据集成一般涉及以下三个问题：

1）模式集成

模式集成涉及的问题是实体识别，其目的是将多个数据源中的数据匹配进行模式的集成分析。

2）冗余数据剔除

冗余数据是指同一个属性或者是相同意义的数据在多个数据库中出现，在数据集成时需要保持属性的唯一性，剔除重复或者是多余的属性。

3）解决数据冲突

由于同一物体的同一属性由于在不同数据库中的编码不同，有可能造成语义上的歧义性，目前尚无方法解决该问题。

（3）数据归约

数据归约指的在保证在尽量保证数据完整性的前提下尽可能利用约简技术将多余的属性进行剔除，使得数据量进一步简化，数据归约能够有效地提高知识挖掘中相关算法的速度、节约算法运行的时间以及减少算法所用的计算机内存等资源，保证知识挖掘更加有效的进行，并可以产生几乎相同的知识规则结果。常用的数据规约方法有主成分分析法以及粗糙集属性约简方法，其中粗糙集约简技术的主要优点是该技术不需要任何预备的或者是额外的数据信息，并已经广泛地在知识挖掘中采用，在本文中选择的是粗糙集约简技术作为知识挖掘中的数据归约方法。

（4）数据的无量纲化处理

由于各属性的量纲尺度不同，因此在进行知识挖掘时一般需要将属性数据进行数据的无量纲化处理，也可以成为数据的标准化、规范化处理，它可以看

成是通过数学的变换对原始指标单位的量纲影响进行消除，尤其是在利用智能性的神经网络预测算法进行模型的训练和计算时，大多数的神经元激活函数的输入和输出均为[-1,1]之间的数据，这时需要将数据统一规范到[-1,1]的区间中。本文利用的标准化方法有两种，一种是将数据映射到[0,1]的区间中，利用的公式是：

$$l_{new}=\frac{l-l_{\min}}{l_{\max}-l_{\min}} \tag{2-3}$$

另一种方法是将数值映射到[-1,1]的区间中，利用的公式是：

$$l_{new}=2\times\frac{l-l_{\min}}{l_{\max}-l_{\min}}-1 \tag{2-4}$$

第一种归一化方法用于所有属性均无负值的情况，第二种情况用于有某些属性存在负值并且需要保留正负号进行进一步的知识挖掘时采用。

（5）数据离散化

由于电力负荷相关数据的数值属性在数据库中往往是连续存在的，因此在利用知识挖掘技术进行相关分析时，首先需要对连续属性的数据进行离散化处理，数据离散化是知识挖掘中的一个极其关键的问题。离散化问题的核心工作是将连续取值的属性值划分为少数的区间来减少连续属性值的取值范围。目前关于数据离散化的方法主要采用两种策略：划分方法和归并方法，其中划分方法的思路是将属性的整个取值区间作为一个离散化的属性值，然后将这个离散的属性值继续进行划分成更小的区间，一直到满足某种条件结束；而归并方法的思路正好相反，开始将每个不同的属性值看成是一个区间值，然后逐一进行合并成一个较大的区间，一直到满足某种条件结束。本文在离散化的方法选择上采用的是划分的策略，根据实际数据分布不同可以选取以下两种方法的一种。

（1）等宽度划分方法

等宽度划分方法比较适用于分布较为均匀的数据，该方法的优点是：在计算机实现时，只需对数据库进行一次遍历，算法的效率比较高，等宽度划分方法比较直观，易于理解。等宽度划分方法的算法如下：

1）扫描数据库，得到属性 A 的最大值 $max(A)$和 $min(A)$；

2）按照下式求出划分区间的宽度 w；

$$w=\frac{\max(A)-\min(A)}{n} \tag{2-5}$$

其中 n 表示要划分的区间数量。

3）按照区间宽度生成离散区间$[l_i,v_i]$，其中 $l_i=\min(A)+(i-1)\times w; v_i=l_i+w$。

（2）等距划分方法

当数据的分布不是很均匀时，需要利用基于距离的等距划分方法。等距划分方法类似从聚类的角度对数据语义进行分析，通过数据的聚类形成相应的属性区间。其原理依赖于同一类别的个体之间应该彼此距离较近，而不同类别的距离应该尽可能的无限大，等距划分方法的算法如下：

1）从数值属性 A 中选择 k 个不同的值作为数据初始簇的中心；

2）根据簇中 A 的评价值，将每个属性 A 的值赋给最类似的簇；

3）更新簇的评价值；

4）不断调整簇的中心进行计算，一直到满足相应的条件不再变化为止。

2.4 本章小结

本章对负荷预测的属性变量进行分析，将相关变量分成了负荷变量和非负荷变量两个部分，对于各个部分根据不同的需要提出了相应的存储规范，利用不同的存储规范给出了不同预测需求的数据视图规范。此外，针对数据库中可能出现的缺失值、异常点等问题给出了相应的数据预处理方法。

第 3 章 基于知识挖掘技术的 BP 神经网络协同日负荷曲线预测研究

神经网络是电力负荷预测中较为常见的一种人工智能预测方法，该方法具有可考虑数据型影响因素，无需识别变化规律，可以模拟任意非线性复杂映射、经学习训练能够得到最终模型的智能性优点。BP 神经网络（BPNN）是神经网络预测模型中采用较多的一种神经网络，该网络对非线性映射的逼近较为稳定，利用大样本的数据量进行训练后能够具有很好的非线性预测的能力，适用于短期的电力负荷预测。但利用 BPNN 进行电力负荷预测需要解决以下两个难点问题，一是采用何种神经网络结构对负荷进行预测，另一个是如何剔除不同变化规律的数据波动对神经网络训练的干扰。本章提出结合知识挖掘技术和神经网络的协同日负荷曲线预测模型。该模型的优点是利用知识挖掘的技术先对数据集进行分类研究，以自动提取确定与预测点同等类型的历史数据序列，形成周期性强化的神经网络训练数据，这样可以有效地提高预测的精度。此外，在采用 BP 神经网络进行预测时，提出了一种自适应结构的算法对神经网络的结构进行确定。

3.1 日负荷曲线预测及预测方法选择

日负荷曲线指的是负荷数值在一天中自 0 点到 24 点的随时间变化的负荷特性曲线。日负荷曲线预测一般指的是对未来下一日的日负荷特性曲线进行预测。日负荷曲线的预测值是电力系统进行日调度安排，决定下一日的开停机计划、在电力市场中安排下一日短期负荷交易分配以及安排相应的备用旋转容量的基础，其精确度直接影响电力系统运行的经济效益。日负荷曲线预测在目前的电力系统中根据具体要求的不同分为 24/48/96 点负荷预测，分别指以一小时、半小时和 15 分钟为间隔进行离散化预测。典型的日负荷曲线由于受到同一区域人们作息规律、季节影响以及社会因素的影响，一般都具有典型的周期性特征。从不同的时间角度进行观察，可以认为日负荷曲线具有周度、月度、季度以及年度变化的周期性，在预测时应该充分注意这种变化周期的特征。

目前对于日负荷曲线预测方法的研究有相似日负荷曲线预测法、典型负荷曲线叠加法、时间序列方法以及神经网络预测方法等。由于神经网络方法具有并行处理非线性拟合的能力，并且具有一定的自组织、自学习、联想记忆等优

良特性，非常适用于日负荷曲线预测的研究中，在众多的神经网络模型中，BP（Back Propagation）神经网络算法由于具有很好的函数逼近能力、存在反向记忆调节权重的功能以及可以方便考虑其余非线性定量因素的影响从而广泛地应用于日曲线负荷预测中，是近期受到学者青睐的一种预测方法。因此下文结合BP 神经网络的特点对知识挖掘协同的电力日负荷曲线预测进行研究。

3.2 BP 神经网络（BPNN）方法

BPNN 是一种正向前馈神经网络，利用最速梯度下降法的误差逆传播对网络权值和阈值进行不断的修正，一直到终止条件满足为止。BP 神经网络能学习和存贮大量的输入——输出模式映射关系，无需使用者具有描述这种映射关系的数学方程的相关知识，根据 Kolmogorov 的相关定理，可以由一个包括输入层（input）、隐层（hidden layer）和输出层（output layer）的三层 BP 神经网络对非线性映射进行任意精度的逼近。从数学意义上讲，若输入层的节点数为 n，输出层节点数为 l，BPNN 是从 R^n 到 R^l 的一个高度非线性映射，在所选网络的拓扑结构下，通过学习算法调整各神经元的阈值和连接权值使误差信号取值最小。图 4-1 所示的是一个典型的三层 BP 神经网络的网络拓扑结构。从图中可以看出 BP 神经网络的存储信息的结构可以分为以下两个部分：（1）网络的体系结构，即网络输入层、隐层和输出层神经元个数；（2）相邻节点之间的连接权值。

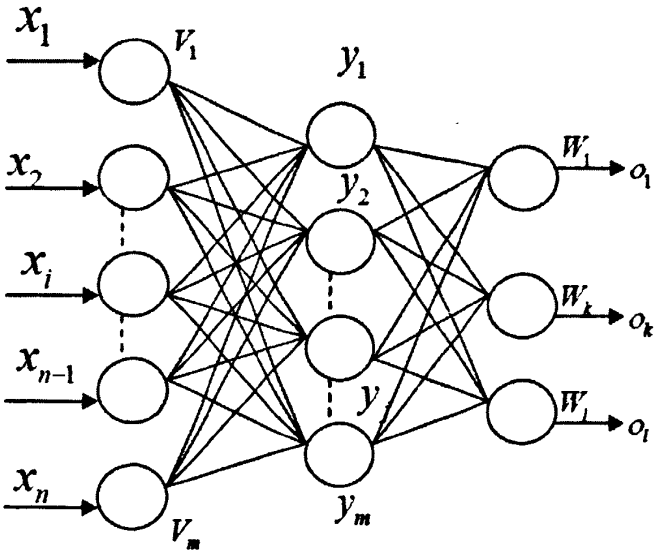


图 3-1 三层 BP 网络拓扑结构示意图

Fig. 3-1 Structure of three-layer BP network

整个 BP 神经网络算法学习过程由信息的正向传递和误差的反向传播两个部分组成。在正向传递中，输入信息从输入层经隐层逐层计算传向输出层，每层神经元的状态仅仅影响下一层神经元的状态，如果在输出层没有得到期望的输出值，则通过计算输出层的误差变化值转向反向传播，通过网络将误差信号沿原来的连接通路反向传播回去，用以修改各层神经元的权值，如此反复直到达到期望目标后终止运算。以图 4-1 中的神经网络结构为例，假设 BP 神经网络模型为三层网络结构，输入层单元数为 n 个；隐含层节点的个数为 m 个（一般是输入层节点个数的 2 倍以上，即 $m>2n$），输出层单元个数 l 个，E_P 为模式 p 下全网络的误差函数，t_{pj} 为模式 p 下节点 j 的希望输出值，O_{pj} 为模式 p 下节点 j 的实际输出值，w_{ij} 为节点 i 到 j 之间联系的权重，θ_j 为节点 j 的阈值。则 BP 神经网络的训练过程如下[127]：

（1）计算各节点希望输出值与实际输出值之差的平方和 E_p

$$\mathrm{E_P}=\frac{1}{2}\sum_{j=1}^{M}(t_{pj}-O_{pj})^2 \tag{3-1}$$

其中节点 j 的净输入 NET_{pj} 为

$$\mathrm{NET_{pj}}=\sum_{i=1}^{m}W_{ij}O_{pi}-\theta_j(j=1\ldots l) \tag{3-2}$$

其中，上式的求和是对所有节点 j 的输入量进行的，如果节点 j 处于输出层，则 O_{pj} 由下式加权及阈函数决定，即

$$\mathrm{O_{pj}}=\mathrm{f_j}(\sum_{i=1}^{L}W_{ij}O_{pi}-\theta_j)(j=1\ldots l) \tag{3-3}$$

其中阀值 θ_j 为常数，其取值范围为（-1,+1）。若使公式（3-1）的误差函数取最小值，需要使每次训练得到的 E_P 关于 W_{ij} 的导数为负值，由隐函数的求导法则可得：

$$\frac{\partial E_p}{\partial W_{ij}}=\frac{\partial E_P}{\partial NET_{pj}}\frac{\partial NET_{pj}}{\partial W_{ij}} \qquad (i=1,2,\ldots,m;j=1,2,\ldots,l_j) \tag{3-4}$$

将公式（3-2）代入式（3-4）中可得：

$$\frac{\partial NET_{pj}}{\partial W_{ij}}=\frac{\partial}{\partial W_{ij}}\sum_{k=1}^{L}W_{kj}O_{pk}=\sum_{k=1}^{L}\frac{\partial W_{kj}}{\partial W_{ij}}O_{pk}=O_{pk} \tag{3-5}$$

其中：

$$\frac{\partial W_{kj}}{\partial W_{ij}}=\begin{cases}0 & i\neq k\\ 1 & i=k\end{cases} \tag{3-6}$$

公式（3-4）右边第一个因子为

$$-\frac{\partial E_p}{\partial NET_{pj}}=\sigma_{pj} \tag{3-7}$$

将式（3-5）以及公式（3-7）代入公式（3-4），有

$$-\frac{\partial E_p}{\partial W_{ij}}=\sigma_{pj}O_{pk} \tag{3-8}$$

要减少误差函数 E_P 的值，就要使权重的变化量 ΔW_{ij} 正比于 $\sigma_{pj}O_{pk}$，即

$$\Delta W_{ij}=\eta\sigma_{pj}O_{pk} \tag{3-9}$$

其中 η 为比例系数。当求解出 σ_{pj} 时，可以通过公式（3-9）决定权重的改变量对权重进行调整，其中（3-8）中的 σ_{pj} 的表达式可以写成

$$\sigma_{pj}=-\frac{\partial E_P}{\partial NET_{pj}}=-\frac{\partial E_p}{\partial O_{pj}}\bullet\frac{\partial O_{pj}}{\partial NET_{pj}} \tag{3-10}$$

将公式（3-3）代入上式中的 $\frac{\partial O_{pj}}{\partial NET_{pj}}$ 可得

$$\frac{\partial O_{pj}}{\partial NET_{pj}}=f_j^{'}(NET_{pj}) \tag{3-11}$$

将公式（3-1）代入式（3-11）中可得

$$\frac{\partial E_P}{\partial O_{pj}}=-(t_{pj}-O_{pj}) \tag{3-12}$$

将式（3-12）代入式（3-11）得

$$\sigma_{pj}=f_j^{'}(NET_{pj})(t_{pj}-O_{pj}) \tag{3-13}$$

假设 j 是输出单元，可以将上述公式联立进行求解得到 σ_{pj} 和 ΔW_{ij}。若 j 不是输出单元而是隐含层单元，可以继续通过隐函数的求导方法得到下式：

$$\frac{\partial E_p}{\partial O_{pj}}=\frac{\partial E_p}{\partial O_{pj}}=\sum_{k=1}^{M}\frac{\partial E_P}{\partial NET_{pk}}\times\frac{\partial NET_{pk}}{\partial O_{pj}}=\sum_{k=1}^{M}\frac{\partial E_P}{\partial NET_{pk}}-\frac{\partial}{\partial O_{pj}}\times(\sum_{k=1}^{M}w_{ik}O_{pj}-\theta_j)$$

$$=-\sum_{k=1}^{M}\sigma_{pk}W_{jk}\quad(j=1,2,...,m) \tag{3-14}$$

其中 k 表示的是和隐含层节点 j 相连的输出节点，利用（3-14）得到计算出的数值后，可以得出中间层 σ_{pj} 的取值为：

$$\sigma_{pj}=f_j^{'}(NET_{pj}).\sum_{k=1}^{M}\sigma_{pk}W_{jk}\quad(j=1,2,...,m) \tag{3-15}$$

假设选取 S 型函数作为阈函数（S 的意义为修正系数），则有

$$f(x)=1/(1+e^{-sx}) \tag{3-16}$$

通过求导可得公式（3-17）

$$f'(x)=\frac{Se^{-sx}}{(1+e^{-sx})^2}=Sf(x)[1-f(x)]SO_{pj}(1-O_{pj}) \tag{3-17}$$

从上面的计算过程中可以看出 BP 神经网络的算法由以下三个计算过程组成：

（1）从输入层开始经隐层向输出层方向进行“模式顺序传播”过程；

（2）通过神经网络的的希望输出值和神经网络实际输出值算出误差信号，通过误差信号由输出层经隐含层向输入层逐层进行修正连接权的“误差逆向传播”过程；

（3）由“模式顺序传播”与“误差逆向传播”的反复交替过程进行的网络“学习记忆”训练过程，使网络趋向收敛的“学习收敛过程”。直至满足终止条件为止。

由于 BP 神经网络一般需要具有比较多的具有较强周期性的训练样本，因此适合与短期负荷预测研究，尤其适合电力负荷预测中的日曲线负荷预测研究，但是利用 BP 神经网络在进行日曲线负荷预测研究时具有无法客观上自动选择模型输入变量和自动优化确定神经网络结构的缺陷，因此需要利用知识挖掘的相关技术对此缺陷进行弥补，下文将分别对不同情况下的 BPNN 知识挖掘协同日曲线负荷预测问题进行研究。

3.3 仅含负荷数据下基于相似度的 BPNN 协同日负荷曲线预测

在进行短期负荷预测时，由于气象因素对于短期负荷的影响在近些年中才得到重视，因此在很多电力公司的数据库中还没有相应的气象数据，在这种情况下只能得到纯负荷历史序列，更甚者有时只能利用当前几个月的数据进行日负荷曲线预测，此时由于单纯负荷数据的限制难以进行如不同年份同时期或者是不同气象状态下的曲线模式复杂分析工作，这种情况下可以借助知识挖掘中简单的分类技术对相似负荷序列进行提取具有较强相似周期的训练数据来提高预测精度，其思路是首先计算负荷曲线的相似度，并通过确定相似度的阈值来筛选和预测日具有较强相似周期性的负荷数据来进行初步的负荷预测，然后利用神经网络进行预测负荷值的修正工作，通过修正工作从而进一步提高负荷预测的精度。

3.3.1 利用相似度进行初步预测

一般而言，由于日负荷曲线存在着连续性和周期性，因此在同一周期内的负荷序列在图形上具有一定的相似性，如图 3-2 是某城市临近两天的 24 点的日负荷曲线，可以看出，虽然这两天的负荷相差较大，但是均在 10 点和 18 点出现两个负荷高峰，在凌晨 3 点附近出现负荷低谷，虽然在细微方面存在着一定的偏离程度，但是在整体上具有着一定的相似性。基于此特点，可以建立负荷预测模型如公式 3-18 所示：

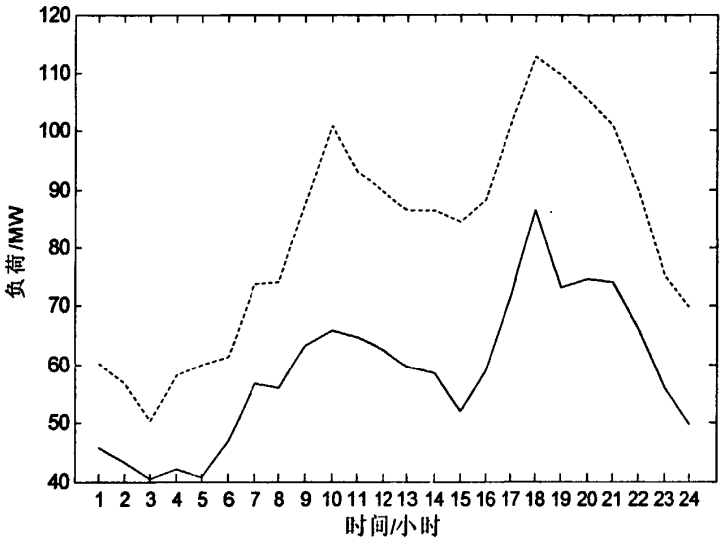


图 3-2 两条日负荷曲线

Fig. 3-2 two daily load curves

$$\hat{y} = \hat{y}_{sim} + \hat{y}_{\varepsilon} \tag{3-18}$$

其中 $\hat{y}_{sim}$ 表示根据相似度预测的部分，$\hat{y}_{\varepsilon}$ 表示偏离相似度的预测部分，而待预测负荷 $\hat{y}$ 由 $\hat{y}_{sim}$ 和 $\hat{y}_{\varepsilon}$ 两部分协同预测相加组成。其中 $\hat{y}_{sim}$ 可由相似度的相关计算预测得到，而 $\hat{y}_{\varepsilon}$ 可认为是由于某些诸如气温等非线性因素造成的偏离误差，这部分误差可以借助神经网络预测模型进行修正。

对于相似度预测的部分可以利用相似度的概念进行计算，关于相似度的概念为：在基于向量空间模型(VSM)的分类过程中，待预测特征向量与各类代表向量的夹角是决定该特征向量归属的重要依据之一，这些夹角的余弦被称作“相似度”，其广泛的用于数据挖掘的分类中。用数学语言可以表述如下：假设向量的特征空间为 $U=(u_1,u_2,\ldots,u_n)$，任取两个向量 $U_i=(u_1,w_{i1};u_2,w_{i2};\ldots,u_n,w_{in})$，$U_j=(u_1,w_{j1};u_2,w_{j2};\ldots,u_n,w_{jn})$；其中 w_{ik} 表示向量 U_i 在属性 $u_k(k=1\ldots n)$ 处的权重；

简记为 $U_i=(w_{i1},w_{i2},...,w_{in})$；其中 w_{jk} 表示向量 U_j 在属性 $u_k(k=1...n)$ 处的权重；简记为 $U_j=(w_{j1},w_{j2},...,w_{jn})$，则这两个向量的相似度可以用它们的夹角余弦值表示，其计算公式如 3-19 所示[128]：

$$Sim(U_i,U_j)=\cos\theta=\frac{\sum_{k=1}^{n}w_{ik}\times w_{jk}}{\sqrt{(\sum_{k=1}^{n}w_{ik}^2)(\sum_{k=1}^{n}w_{jk}^2)}} \tag{3-19}$$

$Sim(U_i,U_j)$ 越大，说明向量的相似度越高。由此启发，以日负荷曲线为例，若选取预测时刻前 n 个时点向量 $T_j=(t_1,t_2,...,t_n)$ 对 t_{n+1} 进行负荷预测（其中 t_i 表示在负荷在 i 时刻的负荷值），可以首先按照公式（3-20）对负荷的相似度进行排序，然后进一步对负荷进行预测。

$$Sim(T_i,T_j)=\cos\theta=\frac{\sum_{k=1}^{n}t_{ik}\times t_{jk}}{\sqrt{(\sum_{k=1}^{n}t_{ik}^2)(\sum_{k=1}^{n}t_{jk}^2)}} \tag{3-20}$$

其中：t_{ik} 表示历史负荷序列向量 T_i 在时点 $t_k(k=1...n)$ 处的取值；t_{jk} 表示待预测负荷序列向量 T_j 在时点 $t_k(k=1...n)$ 处的取值；$k=1...n$。

将负荷序列排序后，选取相似度高于一定阈值 α 的 m 组相关数据进行相似度部分的负荷预测，其中采用负荷预测的公式如下：

$$\hat{y}_{sim}=\mathbf{w}\cdot\mathbf{t}_n \tag{3-21}$$

其中：$\mathbf{t}_n=(t_1,t_2,...,t_l,...,t_m)$，$t_l$ 表示高于阈值 α 的 m 组相关数据中第 l 条数据在时点 $n+1$ 处的取值。$\mathbf{w}=(w_1,w_2,...,w_l,...,w_m)^T$，$w_l=\frac{sim(T_l,T_j)}{\sum_{l=1}^{m}sim(T_l,T_j)}$。$\hat{y}_{sim}$ 由向量 $\mathbf{w}$ 和 $\mathbf{t}_n$ 的内积表示。

3.3.2 利用自适应结构的神经网络进行误差纠正

由于电力负荷受到气象、经济、价格、政策、政治活动等因素的影响，使得电力负荷在保持与历史数据相似曲线运动的同时，总是或多或少的偏离其相似度轨迹，而构成这部分偏差的因素往往又与偏差呈非线性关系。BPNN 模型具有无需识别变化规律、可模拟任意非线性复杂映射、经学习训练可得到最终模型的智能性等优点。因此，可利用神经网络对这部分偏离进行预测。根据待解决问题的特点，建立的 BPNN 结构如下：

输入层：从提取出来的相似数据集中选择前 n 个时点和不同日的前 m 个同一时点的负荷值作为影响因素。

输出层：仅含一个单元 $\hat{y}_\varepsilon$，即预测时刻的偏离值。

隐含层：采用逐步确认的方式，即首先选取 2 个神经元，然后根据神经网络的 Kromogol 定理和设计经验确定神经元的最高上限以某种步长逐步训练神经网络，然后比较训练后的误差、收敛情况和网络性能来确定最佳的隐含层神经网络个数。

3.3.3 实例分析

根据本文提出的相似度与神经网络的协同预测模型，对某市电力公司 2009 年 1 月 1 日 0：00 至 2009 年 4 月 30 日 23：00 的每日 24 个时点数据进行仿真计算。其中选择 2009 年 4 月 29 日和 2009 年 4 月 30 日的 48 个时点的数据做为未知数据进行验证。对于每个预测点，选取该时点前 23 个小时的时点数据做相似度部分的预测，并在排序结果中选取前 10%的数据进行 $\hat{y}_{sim}$ 的预测。而由于历史数据的限制，选取前三个时点的误差值、日平均气温、日最高气温和日最低气温利用神经网络对偏离部分的 $\hat{y}_\varepsilon$ 进行预测。并与选取前三个时点的负荷值、日平均气温、日最高气温和日最低气温的 BPNN 负荷预测模型进行比较，比较误差分析采用平均相对误差 e_{MAPE} 作为比较的依据，通过 Matlab 编程计算所得结果如表 3-1 和图 3-3。

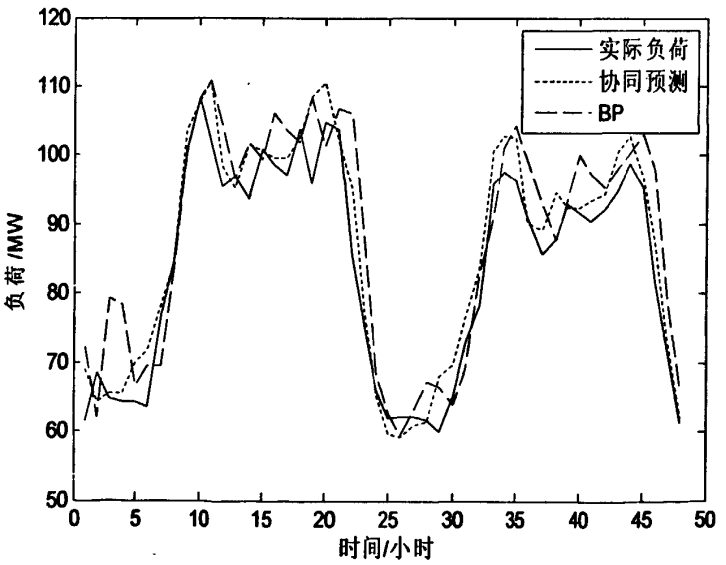


图 3-3 两种预测方法结果曲线图

Fig. 3-3 the results of the two forecasting methods

表 3-1 协同短期负荷日曲线预测结果

Table 3-1. the results of BPNN and collaborative knowledge mining and BPNN model

时点	实际负荷	协同预测	ape	BPNN	ape	时点	实际负荷	协同预测	ape	BPNN	ape
1	61.5	69.0438	0.1227	72.2297	0.1485	25	62	59.7208	0.0368	62.7868	0.0125
2	68.5	64.1092	0.0641	62.0436	0.1041	26	62.2	59.2778	0.047	59.4599	0.0461
3	64.8	65.5639	0.0118	79.4311	0.1842	27	62.2	61.0623	0.0183	63.2767	0.017
4	64.2	65.3926	0.0186	78.3482	0.1806	28	61.7	61.5194	0.0029	67.207	0.0819
5	64.2	69.9756	0.09	66.4961	0.0345	29	60	68.0051	0.1334	66.3818	0.0961
6	63.6	71.7306	0.1278	69.4441	0.0842	30	65.3	69.6135	0.0661	63.8665	0.0224
7	76.6	78.4178	0.0237	69.3271	0.1049	31	72.9	76.6144	0.051	69.0463	0.0558
8	84.8	84.3319	0.0055	83.535	0.0151	32	78.4	83.7898	0.0687	82.9184	0.0545
9	100.7	103.5305	0.0281	100.2538	0.0045	33	95.9	100.5979	0.049	91.0411	0.0534
10	108.3	107.8104	0.0045	108.3188	0.0002	34	97.5	102.9025	0.0554	101.1998	0.0366
11	102.1	110.8387	0.0856	110.6238	0.0771	35	96.4	102.6702	0.065	104.1881	0.0748
12	95.4	98.354	0.031	104.2989	0.0853	36	90.9	90.2496	0.0072	99.2012	0.0837
13	96.9	95.0803	0.0188	96.5438	0.0037	37	85.7	89.3017	0.042	92.8551	0.0771
14	93.6	101.6834	0.0864	101.8793	0.0813	38	88	94.567	0.0746	87.6829	0.0036
15	100.9	100.4703	0.0043	99.3584	0.0155	39	92.8	92.4476	0.0038	93.5478	0.008
16	98.5	99.6875	0.0121	105.964	0.0704	40	91.7	92.5378	0.0091	100.0029	0.083
17	97.2	99.4769	0.0234	103.4718	0.0606	41	90.4	93.4563	0.0338	97.2958	0.0709
18	103.8	102.4923	0.0126	101.7054	0.0206	42	92.2	94.4731	0.0247	95.3362	0.0329
19	95.9	108.5384	0.1318	108.2034	0.1137	43	94.8	100.6785	0.062	98.0158	0.0328
20	104.7	110.4231	0.0547	101.2132	0.0345	44	98.8	102.9175	0.0417	100.3238	0.0152
21	103.7	102.3217	0.0133	106.8572	0.0295	45	95.4	96.9041	0.0158	103.1185	0.0749
22	85.6	95.255	0.1128	105.9977	0.1924	46	82	87.6072	0.0684	97.9423	0.1628
23	75	77.0366	0.0272	86.8933	0.1369	47	71.2	72.5902	0.0195	78.6955	0.0952
24	66	65.2122	0.0119	68.5401	0.0371	48	61.2	61.9752	0.0127	65.8655	0.0708
MAPE									0.444		0.663

从结果上可以看出，本文提出的协同预测模型所得的预测结果精度整体上要高于 BP 神经网络的预测结果。

3.4 含天气数据时基于知识挖掘的 BPNN 协同日负荷曲线预测

在进行日负荷曲线预测时，如果具有较完备的含天气数据以及历史负荷数据供预测使用时，可以利用知识挖掘的分类技术和聚类技术协同对负荷预测进行进一步的分析来提取相似度高的数据序列集进行预测来提高预测的精度。首先可以利用聚类技术预先对历史负荷数据进行聚类分析来区分日负荷预测的模式，然后通过决策树的规则提取技术对条件属性的数据进行分类决策规则的分析，在预测时通过相关的分类分析来确定待预测日同等类型的历史数据序列，形成供 BP 神经网络使用的训练数据。由于知识挖掘的特点，在模式分析时可以对定性因素进行处理，通过知识挖掘的模式分析可以考虑定性因素对负荷预测点的影响。

3.4.1 基于知识挖掘技术分析的数据视图规范

在进行预测时，首先要形成供知识挖掘使用的数据视图，为可以方便的利用数据挖掘技术如粗糙集、决策树分类、聚类等技术进行分析，首先需要将天气数据与历史负荷数据组成如图 3-4 所示的记录结构。在数据记录结构前首先是天气数据类的定性因素，这些因素可以作为知识挖掘的条件属性，在数据记录结构的末尾是日负荷的曲线数据，即 24/48/96 点负荷作为知识挖掘的决策属性。

天气情况、湿度、风力……	24/96点负荷
定性因素数据	历史负荷数据

图 3-4 进行知识挖掘分析数据视图中的记录结构规范

Fig. 3-4 The record structure for data mining

当形成上述数据记录结构规范时，为利用这些历史数据进行知识挖掘的分类处理，需要对上述数据结构进行如下处理。

（1）利用聚类技术对历史负荷数据进行聚类分析；

（2）根据聚类结果对定性因素属性利用粗糙集进行属性约简；

（3）对约简后的定性因素进行决策树分类分析，寻找定性因素和负荷数据之间的联系。

3.4.2 日负荷曲线聚类分析

聚类分析是数据挖掘中的一个很活跃的研究领域，其算法可以大致分为划分方法、层次方法、基于密度方法、基于网格方法和基于模型方法五种。其中以划分方法中的 k-means 方法最为经典。k-means 算法以 k 为参数，把 n 个对象分为 k 个簇，以使簇内具有较高的相似度，而簇间的相似度较低。相似度的计算根据一个簇中对象的平均值(重心)来进行。

k-means 算法过程如下：首先从 n 个数据对象任意选择 k 个对象作为初始聚类中心，对于剩下的其它对象，根据它们与这些聚类中心的相似度（距离），分别将它们分配给与其最相似的（聚类中心所代表的）聚类；然后再计算每个所获新聚类的聚类中心（该聚类中所有对象的均值），不断重复这一过程直到标准测度函数开始收敛为止。一般采用均方差作为标准测度函数。如公式 3-22 所示：

$$E=\sum_{i=1}^{k}\sum_{p\in C_i}|p-m_i|^2 \qquad (3\text{-}22)$$

其中：E 是数据库中所有对象的平方误差的总和；p 是空间中的点，用来表示给定的数据对象；m_i 是簇 C_i 的平均值(p 和 m_i 都是多维的)。该准则可以使生成的结果簇尽可能的独立和紧凑。

对于电力负荷预测而言，能否将历史电力负荷准确的进行分类是能否进一步精确预测电力负荷的一个重要因素。因为分类可以避免由于电力负荷的类别不同而带来的数据扰动造成的神经网络预测训练精度缺失甚至神经网络预测不收敛等问题，此外，准确的分类也为进一步挖掘定性因素和历史负荷之间的关系奠定了良好的基础。而由于数据规范中所含的条件属性可能会很多，因此在聚类后进行分类之前需要对数据规范中的条件属性进行数据归约处理，本文利用知识挖掘技术中的粗糙集技术来对条件属性进行约简处理。

3.4.3 利用粗糙集进行属性约简

粗糙集理论将知识看作是关于论域的划分，认为知识是有粒度的，并引入代数学中的等价关系来讨论知识。该理论主要用于知识约简和知识相依性的分析，可以作为机器学习和复杂数据分析的工具。其基本理论如下：

设 U 是感兴趣的对象组成的有限集合，论域 R 是定义在 U 上的一个等价关系。则 U/R 表示 R 在 U 上导出的划分，$[x]_R$ 表示包含 x 的 R 的等价类，其中 $x\in U$。在粗糙集理论中，将序对（U,R）称为一个近似空间。任何子集 X 属于 U，称为一个概念。对每个概念 X 可定义下、上近似集如下：

$$\underline{R}X=\bigcup\{x\in U:[x]_R\subseteq X\}\quad \overline{R}X=\bigcup\{x\in U:[x]_R\cap X\neq\varphi\} \tag{3-23}$$

其中：下近似集表示由U中那些在现有知识R下肯定属于概念X的元素组成的集合，上近似集是可能属于概念X的元素组成的集合。对于U上的两个等价关系P，Q，Q的P-正区域定义为：

$$POS_p(Q)=\bigcup_{X\in U/Q}\underline{P}X \tag{3-24}$$

$POS_p(Q)$是U中所有那些通过知识P被肯定地分作U/Q的类的元素组成的集合。

设U是一个论域，P和Q是定义在U上的2个等价关系族。如果公式(3-25)成立。则称一个等价关系$R\in P$是Q-不必要的(或多余的)，否则，R在P中是Q-必要的。

$$POS_{IND(P)}(IND(Q))=POS_{IND(P-\{R\})}(IND(Q)) \tag{3-25}$$

其中：$IND(Q)=\cap P$（所有属于 p 的等价关系的交）也是一个等价关系，并且称为P上的一个不可区分关系。

P中所有Q-必要的等价关系组成的集合，称为P的Q-核，记作$CORD_Q(P)$。

基于上述理论，可以将粗糙集属性约简技术应用于电力负荷定性因素的约简分析中，若记P和Q分别表示影响电力负荷预测的定性因素属性和电力负荷的类别决策属性，若存在一个属性$R\in P$是Q-不必要的，则从P中去掉属性R不会改变对电力负荷类别的决策影响，而去掉P中那些属于Q-核中的属性将改变信息系统的决策。因此，可以利用粗糙集理论首先对影响电力负荷的定性因素进行约简处理，可以减少电力负荷预测的计算量，从而提高电力负荷预测的运算速度。

3.4.4 利用知识挖掘决策树分类算法进行分类规则提取

在进行完上述步骤后，可以利用知识挖掘中的分类技术进行分类规则的提取，分类方法是指在分析数据库中的一组对象的同时找出其共同属性，构造分类模型，然后利用分类模型对其它的数据对象进行分类。要构造分类模型，需要一个训练样本数据集作为输入，训练集由一组数据库记录或元组组成，每个元组包含一些字段值，又称“属性”或“特征”，这些字段和测试集中记录的字段相同，另外，每个训练样本记录有一个类别标识。分类目标是分析训练集中的数据，利用数据中能得到的特征，为每一类建立一个恰当的描述或模型，然后根据这些分类描述对测试数据进行分类或产生更恰当的描述。

分类的过程是：分析输入数据，通过在训练集中的数据所表现出来的特性，经过有关算法，为每一个类找到一种准确的描述或者模型，并使用这种类的描述对未来的测试数据进行分类。决策树分类技术是以实例为基础的归纳学习算法。它着眼于从一组无次序、无规则的事例中推理出以决策树形式表示的分类规则。最早的决策树学习系统起源于概念学习系统(CLS 系统)，该系统第一次提出使用决策树进行概念学习，是许多决策树学习算法的基础。较早提出的 CART(Classification And Regression Trees）分类方法，是由 BreimanL，Friedman J H 和 Olshen R A 等人在 1984 年提出的一种决策树分类方法。这种方法选择具有最小 Gini 指数值的属性作为测试属性，最终生成二叉树，然后利用重采样技术进行误差估计和树剪枝(基于最小代价复杂性)，然后选择最优的树结构作为最终构建的决策树。这些算法均要求训练集全部或一部分在分类的过程中一直驻留在内存中。在决策树学习算法的各种算法当中，属 Quinlan 于 1979 年提出的以信息熵的下降速度作为选取测试属性标准的 ID3 算法最有影响。其算法如下：

设 S 是训练集，其中类别标识属性有 m 个独立的取值，即定义了 m 个类 C_i（$i=1\cdots m$），R_i 为数据集 S 中属于 C_i 类的子集，用 r_i 表示子集 R_i 中元组的数量。S 的期望信息量可以用以下公式计算：

$$\mathrm{I}(r_1,r_2\cdots r_m)=-\sum P_i\log_2 P_i \tag{3-26}$$

$$\text{其中 } P_i=r_i/|S|$$

设属性 A 有 v 个不同的取值$\{a_1,a_2\cdots a_v\}$，则通过属性 A 的取值可将 S 划分为 v 个子集，其中 S_j 表示在 S 中属性 A 的取值为 a_j 的子集，$j=1\cdots v$。如果用 S_{ij} 表示 S_j 子集中属于 C_i 类元组的数量，则属性 A 对于分类 C_i（$i=1\cdots m$）的期望信息量为：

$$E(A)=-\sum W_j I(s_{1j},s_{2j}\cdots s_{mj}) \tag{3-27}$$

$$\text{其中：}\ W_j=(s_{1j}+s_{2j}+\cdots+s_{mj})/|S|$$

$$I(s_{1j},s_{2j}\cdots s_{mj})=-\sum P_{ij}\log_2 P_{ij}$$

$$P_{ij}=s_{ij}/|S_j|$$

将 A 作为决策分类属性的信息增益为：

$$Gain(A)=I(r_1,r_2\cdots r_m)-E(A) \tag{3-28}$$

该算法需要计算每个决策属性的信息增益，具有最大信息增益的属性被选择为给定数据集 S 的决策属性节点，并通过该属性的每个取值建立由该节点引

出的分枝。在建立由该节点引出分枝的数据子集中继续计算除去已计算的决策属性节点后的各分类属性的信息增益，以此类推，一直计算至最后一个属性进而生成决策树。

3.4.5 基于决策树分类技术的自适应 BP 神经网络负荷预测

在进行完上述步骤后，可以得到一系列的分类规则，这时可以利用分类后的各类负荷分别进行神经网络的训练，其中 BP 自适应神经网络的结构确定方式和上节中的网络结构类似，其网络结构如下：

输入层：$L(t-24i), L(t-j)$，其中 $i=1,\ldots,p, j=1,\ldots,q$。分别表示预测时刻前几天同一时刻的历史数据以及前几个小时的历史数据。此外，输入单元还包括约简后的定性因素。

输出层：仅含一个单元 $L(t)$，即预测时刻的负荷值。

隐含层：由于根据 Kromogol 定理，仅含一个隐含层的神经网络模型就可以逼近任意一个非线性映射，因此，神经网络的隐含层仅选取一层。而隐含层的神经元输入采用逐步确认的方式，即首先选取 2 个神经元，然后根据神经网络的 Kromogol 定理和设计经验确定神经元的最高上限以某种步长逐步训练神经网络，然后比较训练后的误差、收敛情况和网络性能来确定最佳的隐含层神经网络个数。

由于对于每一类的神经网络利用的数据不同，因此各类的神经网络的隐含层节点数有很大可能不同，经训练后可以形成各类的 BP 神经网络，在进行日曲线预测时，首先利用待预测数据中的定性数据给待预测数据进行归类，然后选取相应的神经网络进行预测。

3.4.6 实例分析

基于上述算法流程，选取某区域 2009 年 5 月 1 日至 2009 年 6 月 30 日的 24 点负荷数据为例进行讨论，其中气象数据含有气压相关数据、气温相关数据、湿度相关数据、降水量、风速相关数据、日照时间等共 19 项属性相关数据，共组织形成 61 条记录形式。然后对此 61 条记录形式进行数据预处理。对于记录中的定性属性，首先对文字型的属性值进行数字型标识转化，如将一、二、三、四季度分别转化成 1，2，3，4 予以标识，而对于数值型的属性值利用专家经验或等距离方法将其离散化，从而得到初始分析记录集。另将 61 条记录中的前 60 条记录作为训练集，而最后一条记录作为测试使用。

按照上文中提到的流程对该记录集中的历史数据利用 Matlab 软件工具箱中的 K-means 算法进行聚类分析可将负荷分为 4 类，其聚类结果图如图 3-5 所

示。

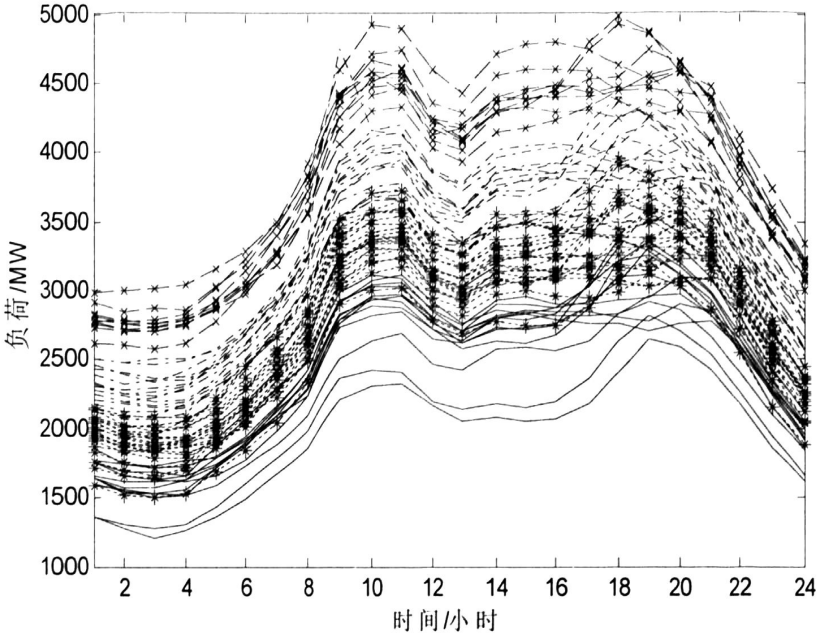


图 3-5 利用 K-means 算法对负荷数据进行聚类的结果图

Fig. 3-5 The Result of clustering the load with k-means

由图 3-5 中可以看出，利用 K-means 算法进行聚类后，61 条负荷记录按照 24 个点的数值特性进行了自动聚类，数值相对较近的点被归为一类。接下来将得到的聚类结果对记录数据进行标识，利用粗糙集理论对其定性因素进行属性约简，可以去除其中的 13 项因素，保留其中的 6 项因素，在对其中的 6 项因素利用 ID3 算法进行归类分析，根据此 6 项因素的负荷分类决策树规则供预测使用。

分别对四类数据开始进行训练，其中输入层选取 $L(t-24i), L(t-j)$，其中 $i=1,2,3, j=1,2$，神经网络采用 BP 神经网络，设置训练次数最高为 1000 次，精度为 10^{-5} 进行预测，根据逐步实验测试，当四类 BPNN 隐含层中的神经元节点分别选择 11，10，11，11 时精度最好。取测试集记录，对比决策树规则预先分析定性因素可将测试集记录分为第 4 类负荷，因此取第四类记录的负荷数据训练成的神经网络进行预测。另外，选取同结构未分类的神经网络以及自回归滑动平均模型 ARMA(1，1)（其中系数由 AIC 定阶准则得到）同样对上述数据进行预测，从而进行对比分析。误差分析采用平均相对误差 e_{MAPE} 作为比较的依据，见公式 1-1，其所得结果如表 3-2 所示。从表 3-2 中可以看出，本文提出方法得到的大部分预测值都比其它两种方法更精确。其平均误差值为 0.0216，而同结构 BP 神经网络的平均误差为 0.0267，根据 ARMA（1，1）得到的预测平均误

差为 0.0381。

表 3-2 三种方法预测结果比较

Tab. 3-2 three models forecasting result of three models

实际负荷值	本文方法		同结构 BP		ARMA(1,1)	
	预测值	e_{MAPE}	预测值	e_{MAPE}	预测值	e_{MAPE}
2810.3	2767.70	1.54	2893.10	2.86	2906.40	3.42
2763.2	2800.60	1.34	2600.20	6.27	2928.70	5.99
2764.6	2804.30	1.42	2816.90	1.86	2827.60	2.28
2800.8	2874.50	2.56	2822.30	0.76	2865.00	2.29
2910.4	2995.30	2.83	2894.90	0.54	2853.20	1.96
3044.4	3067.10	0.74	3068.50	0.79	3008.80	1.17
3278.3	3282.60	0.13	3242.60	1.10	3107.40	5.21
3727.9	3581.30	4.09	3601.10	3.52	3410.00	8.53
4407.2	4340.10	1.54	4085.30	7.88	3940.00	10.60
4709.8	4597.10	2.45	4591.10	2.59	4640.20	1.48
4735.6	4750.60	0.32	4687.90	1.02	4649.30	1.82
4359.7	4280.60	1.85	4500.00	3.12	4619.30	5.95
4288.7	4087.10	4.93	4214.00	1.77	4148.20	3.28
4548.6	4442.30	2.39	4427.50	2.74	4343.70	4.50
4595.4	4387.80	4.73	4578.80	0.36	4587.60	0.17
4603.7	4426.40	4.00	4498.50	2.34	4516.50	1.89
4581.8	4567.80	0.31	4528.30	1.18	4582.60	0.02
4445.7	4585.60	3.05	4518.60	1.61	4510.20	1.45
4460.2	4517.80	1.27	4347.60	2.59	4347.10	2.54
4423.6	4514.20	2.01	4560.00	2.99	4452.10	0.64
4392.8	4367.40	0.58	4412.30	0.44	4324.70	1.55
4041.9	4056.70	0.37	4301.20	6.03	4331.30	7.16
3598.1	3420.50	5.19	3865.00	6.91	3872.30	7.62
3219.9	3290.10	2.13	3316.60	2.91	3540.00	9.94
平均值		2.16		2.67		3.81

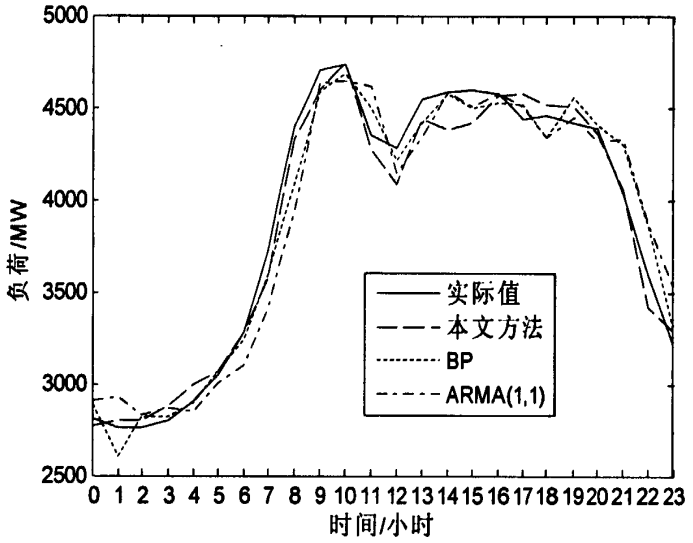


图 3-6 三种方法预测结果曲线图

Fig. 3-6 The Forecasting Result with Three Models

图 3-6 绘制出三种方法给出的预测值和实际负荷的曲线图，可以直观的看出，本章所提出的方法具有更高的拟合精度。从结果中可以说明本章提出的算法可以有效的提高负荷预测的精度。

3.5 本章小结

本章提出了适用于不同数据情况下的知识挖掘协同 BP 神经网络（BPNN）的日负荷曲线预测方法，首先提出的纯负荷数据量下的相似度协同 BPNN 的协同预测模型能够提取短期内的负荷相似序列进行日曲线负荷预测，实例说明预测精度较之未提取相似序列的 BPNN 预测模型有明显的改进。其次在具有完备天气情况数据情况下，利用知识挖掘 K-means 聚类技术、粗糙集、决策树分类技术与 BPNN 预测技术相结合，首先根据日负荷曲线的聚类结果对其相应的定性因素和文本因素进行分析约简，形成决策分类树，并生成相应的规则，然后利用 BPNN 模型进行日曲线负荷预测。由于对定性因素和文本因素进行了分析约简，可以剔除由于定性因素和文本因素的干扰对数据造成的波动，因此在进一步的 BPNN 预测上提高了精度。

在对 BPNN 进行训练时，输入层输入的数据是利用知识挖掘的分类技术后在数据库中利用相似度分类技术或者是决策树分类技术经过筛选后的相似度较高的数据，而隐含层神经元的选择采取了根据训练集自动选择精度结果最小的神经网络，实现了一种自动变结构的神经网络，因此神经网络的隐含层节点数

不需要人为经验的干预，能够有效的提高精度并且可以提高本章提出方法的实用性。

第 4 章 基于知识挖掘的自适应参数的支持向量机协同中长期负荷预测

支持向量机(support vector machine，SVM)是 Vapnik 于 20 世纪 90 年代中期基于统计学理论的 VC 维理论和结构风险最小化原则而提出的一种新的机器学习方法，起先应用于小样本、非线性的模式识别分类问题，随后 Vapnik 在 1998 年将其应用于非线性拟合中，并表现出良好的性能，近期成为学者最青睐的负荷预测方法之一。

由于支持向量机较神经网络对训练样本的数量要求少，并且支持向量机所解决的是一个凸优化的问题，利用同样的样本和参数所得到的训练模型给出的预测值是恒定的，理论上和实际应用中都要比神经网络模型性能更优，因此更加适用于数据量相对较少的中长期负荷预测中。此外，由于中长期的气象因素预报、以及经济情况等其余因素的预测问题也是一个非常困难的问题，因此在对中长期负荷预测进行影响因素的考虑时，尤其是年负荷预测时一般不考虑气象因素，常利用中长期负荷本身的数据进行预测。这些特点均有利于利用支持向量机模型进行中长期负荷预测的研究。然而支持向量机在应用时需要依靠经验选取一些学习参数，例如管道宽度ε和惩罚因子 C 等，这些参数的选取直接关系到支持向量机的预测精度，一般依靠经验选取参数，这非常不利于支持向量机模型的推广使用。本章利用知识挖掘技术中的进化算法对其所需参数进行选取，形成具有自适应参数功能的支持向量机协同中长期负荷预测模型。

4.1 中长期负荷预测及预测方法选择

中长期电力负荷预测是电力系统规划、运行与决策的基本工作，关系到整个电力系统发展的安全性和经济性。准确的进行中长期电力负荷预测有利于在保持电网安全稳定运行的基础上降低发电成本，提高经济效益。与短期负荷预测相比，以年度为代表的中长期负荷预测有着可利用的数据量少，考虑的影响因素不同，要求的预测精度较低的特点，因此中长期电力负荷预测方法的研究基本上是基于小样本数据量的统计预测方法，如线形回归，灰色预测，以及组合预测方法等，近些年，一些智能方法也被引入中长期负荷预测中，例如贝叶斯分类[129]、神经网络[130-131]、支持向量机[132-133]等。其中上文中提到的神经网络是倍受学者的青睐的研究热点之一，但是在小样本的情况下，神经网络容易

陷入局部极小点并且容易存在过拟合的问题。而支持向量机方法是近期提出的一种较新的智能算法[134-135]，其利用结构风险最小化原理和解决线性二次规划进行机器学习，与神经网络算法相比，具有要求确定参数少、在理论上有全局最优的唯一解的特点，被认为是可以替代神经网络的方法[136]，此外，由于支持向量机在小样本下具有良好的泛化能力，因此更加适合中长期负荷的预测。因此下文结合支持向量机的特点对知识挖掘协同的电力中长期负荷预测进行研究。

4.2 支持向量机回归（SVR）预测方法

假设有训练样本集 $G=\{(x_i,d_i)\},i=1\cdots N$，$x_i\in R^n$，$d_i\in R^1$。支持向量机回归的基本原理是寻找一个输入空间到输出空间的非线性映射 $\psi(x)$，通过映射将数据 x 映射到一个高维特征空间 F，并在特征空间中用下述估计函数进行线性回归[137-139]：

$$y=f(x)=w\psi(x)+b \tag{4-1}$$

其函数逼近问题等价于如下函数最小：

$$R(C)=(C/N)\sum_{i=1}^{N}L_\varepsilon(d_i,y_i)+\|w\|^2/2 \tag{4-2}$$

$$L_\varepsilon(d,y)=\begin{cases}0 & |d-y|\le\varepsilon\\ |d-y|-\varepsilon & \text{otherwise}\end{cases} \tag{4-3}$$

其中，$\|w\|^2/2$ 表示的是函数的平滑程度，$L_\varepsilon(d,y)$ 称为 ε-敏感损失函数。通过引入两个松弛变量 ζ，ζ^*，上述函数可以变成如下形式

$$\begin{aligned}&R(w,\zeta,\zeta^*)=\|w\|^2/2+C\sum_{i=1}^{N}(\zeta_i+\zeta_i^*)\\ &s.t.\\ &w\psi(x_i)+b_i-d_i\le\varepsilon+\zeta_i^*,i=1,2,\ldots,N\\ &d_i-w\psi(x_i)-b_i\le\varepsilon+\zeta_i,i=1,2,\ldots,N\\ &\zeta_i,\zeta_i^*\ge0,i=1,2,\ldots,N\end{aligned} \tag{4-4}$$

利用拉格朗日型，可将上式变成

$$
\begin{aligned}
&L(w,b,\zeta,\zeta^*,\alpha_i,\alpha_i^*,\beta_i,\beta_i^*)\\
&=\|w\|^2/2+C\sum_{i=1}^{N}(\zeta_i+\zeta_i^*)-\sum_{i=1}^{N}\beta_i[(w\psi(x_i)+b-d_i+\varepsilon+\zeta_i)] \qquad (4\text{-}5)\\
&-\sum_{i=1}^{N}\beta_i^*[(d_i-w\psi(x_i)-b+\varepsilon+\zeta_i^*)]-\sum_{i=1}^{N}(\alpha_i\zeta_i+\alpha_i^*\zeta_i^*)
\end{aligned}
$$

目标函数如果要达到极小值，其需要满足下面的条件：

$$
\begin{cases}
\dfrac{\partial L}{\partial w}=0\rightarrow w-\sum_{i=1}^{N}(\beta_i-\beta_i^*)\psi(x_i)=0\\
\dfrac{\partial L}{\partial b}=0\rightarrow \sum_{i=1}^{N}(\beta_i-\beta_i^*)=0\\
\dfrac{\partial L}{\partial \zeta}=0\rightarrow C-\zeta-\alpha_i=0\\
\dfrac{\partial L}{\partial \zeta^*}=0\rightarrow C-\zeta^*-\alpha_i^*=0
\end{cases} \qquad (4\text{-}6)
$$

利用 Karush-Kuhn-Tucker 条件并将公式 4-6 代入到公式 4-5 中，可以得到问题的对偶型为：

$$
\begin{aligned}
&\vartheta(\beta_i,\beta_i^*)=\sum_{i=1}^{N}d_i(\beta_i-\beta_i^*)-\varepsilon\sum_{i=1}^{N}(\beta_i-\beta_i^*)\\
&-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\beta_i-\beta_i^*)(\beta_j-\beta_j^*)K(x_i,x_j) \qquad (4\text{-}7)\\
&st:\sum_{i=1}^{N}(\beta_i-\beta_i^*)=0,\\
&0\le\beta_i\le C,0\le\beta_i^*\le C,i=1,2,\ldots,N.
\end{aligned}
$$

求解上述问题可得到支持向量机回归函数：

$$
f(x,\beta,\beta^*)=\sum_{i=1}^{N}(\beta_i-\beta_i^*)K(x,x_i)+b. \qquad (4\text{-}8)
$$

式中 $k(X_i,X)$称为核函数，需要满足 Mercer 条件，一般选取最常用的高斯核函数 $K(x_i,x_j)=\exp(-\|x_i-x_j\|^2/2\sigma^2)$。因此，利用支持向量机进行中长期负荷预测需要确定参数 ε、C 和 σ。而本文利用的 SVR 算法是 Scholkopf 提出的改进 SVR 算法，其引入一个参数 ν 来替代上述的参数 ε，将公式 4-4 等价变形为

$$
\begin{aligned}
&R(w,\zeta,\zeta^*,\varepsilon)=\|w\|^2/2+C(\nu\varepsilon+\frac{1}{N}\sum_{i=1}^{N}(\zeta_i+\zeta_i^*))\\
&s.t.\\
&w\psi(x_i)+b_i-d_i\le\varepsilon+\zeta_i^*,i=1,2,\ldots,N\\
&d_i-w\psi(x_i)-b_i\le\varepsilon+\zeta_i,i=1,2,\ldots,N\\
&\zeta_i,\zeta_i^*\ge 0,i=1,2,\ldots,N
\end{aligned}
\tag{4-9}
$$

其对偶型变成

$$
\begin{aligned}
&\vartheta(\beta_i,\beta_i^*)=\sum_{i=1}^{N}d_i(\beta_i-\beta_i^*)\\
&-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\beta_i-\beta_i^*)(\beta_j-\beta_j^*)K(x_i,x_j)\\
&s.t:\sum_{i=1}^{N}(\beta_i-\beta_i^*)=0,\\
&\sum_{i=1}^{N}(\beta_i-\beta_i^*)\le C\cdot\nu,\\
&0\le\beta_i\le\frac{C}{l},0\le\beta_i^*\le\frac{C}{l},i=1,2,\ldots,N.
\end{aligned}
\tag{4-10}
$$

改进后的支持向量机最终的回归函数和公式 4-8 是相同的，利用 Scholkopf 提出的改进 SVR 算法需要确定的三个参数就转换成了 C,ν 和 σ。目前在很多研究成果中，大部分研究成果均采用主观赋值和经验试验的方法对这三个参数进行调整，少量的参考文献中利用遗传算法（GA）和粒子群算法（PSO）对这三个参数进行调整，在本章中选用的方法是知识挖掘进化算法中的微分进化算法，这是因为该方法在诸多优化问题的求解过程中，其性能被证明要优于遗传算法、粒子群优化、模拟退火算法等其它算法[140]。

4.3 微分进化算法

微分进化算法（Differential Evolution，DE）是由 Rainer Storn 和 Kenneth Price 于 1995 年提出的一种新的进化算法，最初的设想是用于解决切比雪夫多项式问题，后来发现 DE 也是解决复杂优化问题的有效技术。该算法直接采用实数运算，不需要进行编码和解码操作，收敛速度快，稳定性好，对各种非线性函数适应性较强[141-142]。

微分进化算法和 1.3.2 节中介绍的知识挖掘技术中的遗传算法类似，也是一种基于群体进化的算法，具有记忆个体最优解和种群内信息共享的特点，即通过

种群内个体间的合作与竞争来实现对优化问题的求解，其本质是一种基于实数编码的具有保优思想的贪婪遗传算法。标准的 DE 算法如下[143-145]：

（1）初始化参数，首先需要设定种群中个体的数量 N，交叉率 R 以及变异率 F，另外，还需要设定进化的最大代数 g 作为判定算法终止使用。一般而言，交叉率 C 的取值在[0,1]之间。C 的取值越大将会使得收敛的速度越快，而 F 的选取范围一般在[0,1]之间。

（2）随机生成初始种群

算法开始时，首先将进化代数初始化，即设置 $g=0$。利用公式（4-11）生成一个 $N*D$ 的矩阵，其中 N 表示的是个体的数量，而 D 表示待优化参数解的维度，$high[j],low[j]$ 分别表示第 j 个待优化参数解的取值上界和取值下界。

$$X_{ij} = rand \cdot (high[j] - low[j]) + low[j] \tag{4-11}$$

（3）计算并记录每个个体值的适应度。

（4）变异操作

随机选取三个彼此不同的向量 X_a, X_b, X_c，按照公式（4-12）进行变异操作生成一个新的向量 $X_a^{'}$，其中 F 是上文中提到的变异率因子。

$$X_a^{'} = X_a + F(X_b - X_c) \tag{4-12}$$

（5）交叉计算

和遗传算法中的交叉计算类似，利用交叉计算来增加个体自建的相异程度，使得种群之间的个体分布更宽，避免出现收敛在局部最优的情况，其交叉计算的公式如 4-13 所示。

$$\begin{cases} X_b^{'}(j) = X_a^{'}(j) \ \ if \ \ rand(j) \le C \ or \ j = randn(j) \\ X_b^{'}(j) = X_a(j) \ otherwise \end{cases} \tag{4-13}$$

上式中 j 表示基因所在向量 X 的位置，$rand(j)$ 和 $randn(j)$ 均为随机生成的数字，不同的是 $randn(j)$ 的随机生成范围是 $[1,D]$，约束条件 $j = randn(j)$ 确保了每个染色体至少有一个元素进行了交叉操作。

（6）择优操作

通过择优操作可以选取出比较优秀的子代进行下一代的进化，择优操作的标准是依赖于适应度的计算，如果子代的适应度值高于父代，则选择子代，反之保留父代，用公式表示如下：

$$X_{i,G+1}=\begin{cases}U_{i,G}\ if\ f(U_{i,G})<f(X_{i,G})\\ X_{i,G}\ otherelse\end{cases} \tag{4-14}$$

其中$U_{i,G}$表示子代，$X_{i,G}$表示父代，$f(x)$表示 x 的适应度值。经过完择优操作后进入下一轮进化，一直到设定的最大进化代数 g_{max} 为止，最终的种群中获得最高适应度的染色体即使所要得到的最优解。

微分进化算法和遗传算法都采用了知识挖掘进化算法中的“优胜劣汰、适者生存”的自然进化法则，因此都归属于进化算法，而且两者都是群体搜索策略，因此有许多相似之处。在算法流程上，都要经过“初始种群的产生→种群评价→个体更新→新种群产生”这一循环过程，最终以较大概率获得问题的最优解。在功能上，两者本质上都固有并行性，在搜索中有摆脱局部极值的能力，而且都有与其它智能策略结合的固有优势。微分进化算法和遗传算法都是通过计算目标函数值对种群的个体进行适应度评价，对目标函数要求低，都可以求解非凸、多峰、非线性全局优化问题。但是两者之间也有区别，其主要区别在于：

（1）标准 GA 采用二进制编码，而 DE 采用实数编码。

（2）两种算法的变异操作区别很大。在标准 GA 中变异操作是一个概率非常低的操作，仅仅是改变基因串的一位。而在 DE 中变异是通过对种群中几个个体进行线性组合来产生中间个体。

（3）在标准 GA 中通过两个父代个体的交叉产生两个子个体，而在 DE 中通过第 i 个个体和三个随机选取的父代个体共同产生子个体，且在产生子个体的过程中用到了变异和交叉操作。

（4）在传统的 GA 中子个体以一定的概率取代其父代个体，而在 DE 中新产生的的子个体只有当它比种群中的目标个体优良时才对其进行替换。

（5）选择操作的含义不同。在标准 GA 中选择是从当前一代生成一组新的种群，通常使用与个体适应度值成正比的选择概率来选择作为父代的个体，最佳的个体最有可能被选中作为下次迭代的父代。DE 的选择方案是在完成变异、交叉之后由父代个体与新产生的候选个体一一对应地进行竞争，优胜劣汰，使得子代个体总是等于或优于父代个体。而且，DE 给予父代所有个体以平等的机会进入下一代，不歧视劣质个体。

4.4 利用微分进化算法自适应参数的 SVR 中长期负荷预测模型

结合微分进化算法的优化思想，将 SVR 方法中需要确定的三个参数作为优化变量，将微分进化算法和 SVR 进行协同预测的流程如图 4-1 所示。

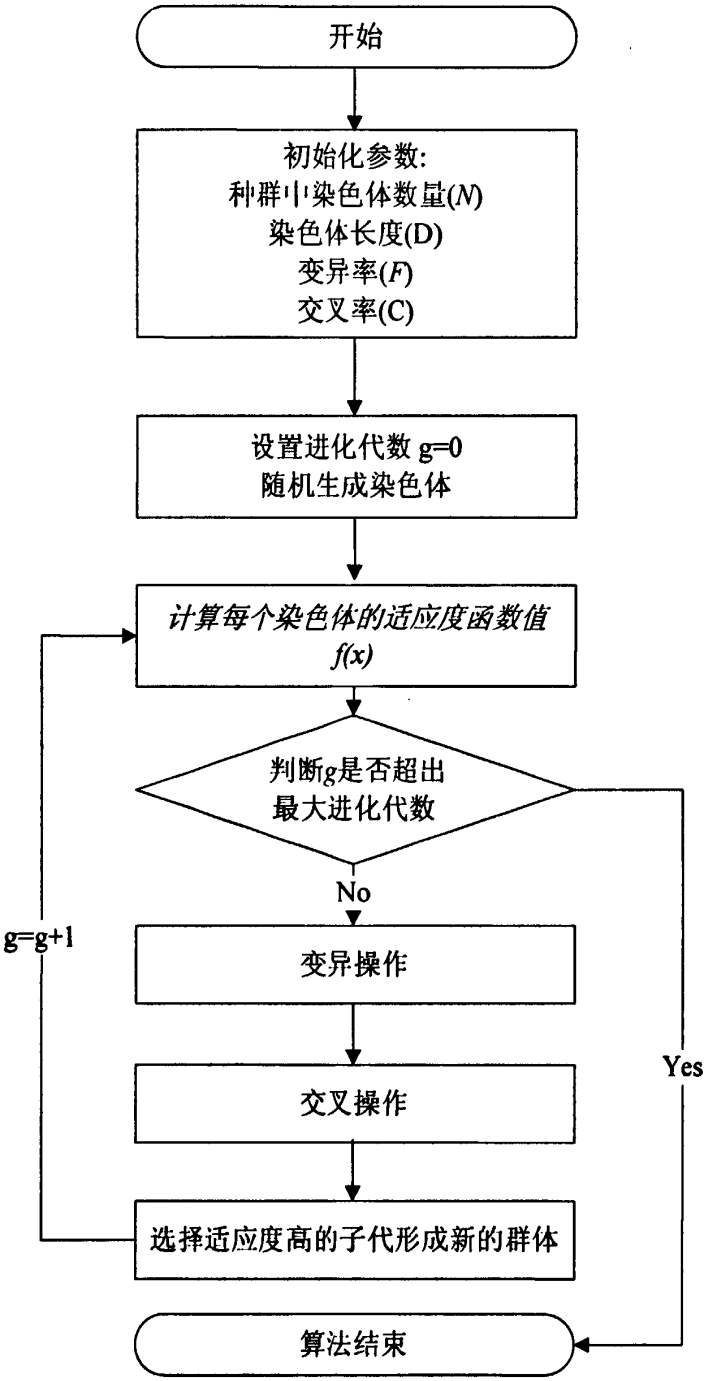


图 4-1 利用微分进化算法自适应参数的 SVR 方法流程图

Fig. 4-1 Adaptive differential evolution algorithm parameters of SVR method of flow chart

其中相应的步骤细节如下：

步骤 1：初始化参数，首先需要决定染色体成员总数 N，变异因子 F 和交叉率 R，每个染色体编码中包含 SVR 中的是所需要三个参数 C,ν 和 σ,因此，染色

体的长度是3，在本文中，这三个参数的取值是N=200，F=0.9，R=0.9。

步骤2：进化开始，设置g=0，然后按照公式4-11随机生成染色体。

步骤3：开始计算，将生成的染色体群输入到SVR模型中进行初步预测，根据预测结果，计算各个染色体的适应度并储存，其中适应度函数采用的常用的衡量预测精度的函数MAPE。

步骤4：子代生成，利用公式4-12、4-13、4-14生成子代，然后将子代输入到SVR中进行预测，并再次计算适应度函数值，设置$g=g+1$。

步骤5：循环计算直至满足停止条件。当g达到最大迭代次数时，算法停止并给出得到最优适应度的三个参数，否则从步骤3开始进行迭代计算。

4.5 实例分析

实例中采用北京市1978年至2008年的年用电量数据进行实证分析，其数据见表4-1和图4-2所示，上述的流程算法采用的环境是Matlab 2009a以及libsvm 2.8.8工具包，SVR的输入变量选取的是负荷前6个时点的数据，即$l_{t-6},l_{t-5},\ldots,l_{t-1}$，输出变量是待预测时点的数据$l_t$，按照4：1的训练集和测试集划分的标准，选取训练集的数据是1984年到2003年的数据，测试集采用的是2004年到2008年的数据。

在训练时采用的是滚动代入的方法，例如在预测1984年负荷的时候，将1978至1983年的实际数据作为输入变量代入，当预测1985年负荷的时候，采用的输入数据是1979年到1984年的实际数据，以此类推。在整个训练结束后，利用微分进化算法对3个参数进行进化选择，之后再进行模型训练，如此反复调整参数，直到最大进化代数满足条件为止。

为了方便比较本文提出算法的效果，本文将预测结果和利用默认参数的SVR模型、上文中提到的BP模型以及传统的回归算法进行对比，其中采用默认参数的SVR模型结构和DESVR的结构相同，而BPNN的输入变量和输出变量与SVR的输入变量和输出变量相同，中间层节点个数选择的方法采用的是上章中所提到的自适应方法，BPNN的终止精度设置为10^{-4}，最大进化代数为1000次，其中运行的环境均相同。

图4-2和表4-2列出了四种预测方法的结果，从图中可以看出，几种方法都预测出了负荷的增长趋势，并且智能预测方法的精度要比回归预测方法的精度好，其中DESVR的精度更好一些。这个结论可以从MAPE的精度比较中可以看出，DESVR无论是在测试集合还是训练集都表现的比较稳定，比默认参数下的SVR模型预测精度要高，虽然BPNN在训练集中的精度较高，但是在测试集中

却出现了较大的预测误差，在小样本的情况下，BPNN 出现了过拟合的情况，可以看出，利用微分进化算法进行自适应参数的选取能够有效地提高中长期负荷预测的精度。

表 4-2 DESVR，SVR，BPNN 和 Regression 对北京市年用电量的预测结果（单位：10^9kWh）

Table 4-1.Forecasting results of DESVR, SVR, BPNN and Regression (unit: 10^9kWh)

Year	Actual	DESVR		SVR		BPNN		Regression	
	Value	Result	Error(%)	Result	Error(%)	Result	Error(%)	Result	Error(%)
1984	10.294	10.410	1.127	10.834	5.246	10.291	-0.029	5.681	-44.813
1985	11.063	11.061	-0.018	11.448	3.480	11.054	-0.081	7.515	-32.071
1986	11.812	11.802	-0.085	12.112	2.540	11.825	0.110	9.350	-20.843
1987	12.850	12.485	-2.840	12.837	-0.101	12.857	0.054	11.185	-12.957
1988	13.786	13.424	-2.626	13.752	-0.247	13.760	-0.189	13.019	-5.564
1989	14.218	14.437	1.540	14.743	3.693	14.231	0.091	14.854	4.473
1990	15.048	15.270	1.475	15.698	4.320	15.032	-0.106	16.689	10.905
1991	16.140	16.271	0.812	16.682	3.358	16.155	0.093	18.523	14.765
1992	17.596	17.594	-0.011	17.764	0.955	17.581	-0.085	20.358	15.697
1993	19.250	19.082	-0.873	19.043	-1.075	19.252	0.010	22.193	15.288
1994	20.545	20.674	0.628	20.504	-0.200	20.542	-0.015	24.027	16.948
1995	22.259	22.158	-0.454	22.029	-1.033	22.257	-0.009	25.862	16.187
1996	24.437	24.019	-1.711	23.813	-2.554	24.433	-0.016	27.697	13.340
1997	26.361	26.474	0.429	25.906	-1.726	26.361	0.000	29.531	12.025
1998	27.621	28.957	4.837	28.160	1.951	27.618	-0.011	31.366	13.559
1999	29.726	31.043	4.430	30.302	1.938	29.725	-0.003	33.201	11.690
2000	38.443	33.571	-12.673	32.567	-15.285	38.441	-0.005	35.035	-8.865
2001	39.994	39.998	0.010	36.928	-7.666	39.994	0.000	36.870	-7.811
2002	43.996	43.995	-0.002	40.862	-7.123	43.996	0.000	38.705	-12.026
2003	46.761	46.774	0.028	45.082	-3.591	46.760	-0.002	40.539	-13.306
2004	51.318	51.483	0.322	49.125	-4.273	47.597	-7.251	42.374	-17.429
2005	57.054	56.637	-0.731	53.654	-5.959	62.396	9.363	44.209	-22.514
2006	61.157	62.393	2.021	58.629	-4.134	62.776	2.647	46.043	-24.713
2007	66.701	65.834	-1.300	62.678	-6.031	72.595	8.836	47.878	-28.220
2008	68.972	69.191	0.318	66.972	-2.900	74.909	8.608	49.713	-27.923

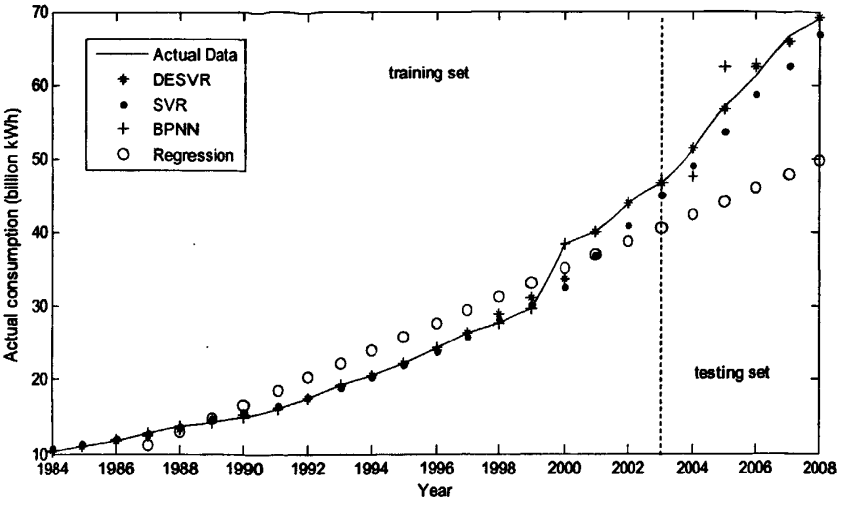


图 4-2 利用 DESVR，SVR，BPNN 和 Regression 对北京市年用电量的预测结果图

Fig.4-2. Forecasting results of DESVR, SVR, BPNN and Regression

表 4-3 DESVR，SVR，BPNN 和 Regression 四种方法的 MAPE 值（单位：%）

Table 4-3. MAPE values of DESVR, SVR, BPNN and Regression model (unit: %)

data set	DESVR	SVR	BPNN	Regression
training set	1.8	3.4	0.4	15.3
testing set	1.1	4.8	7.4	25.8
total	1.6	3.6	1.5	16.9

4.6 本章小结

SVR 预测模型的特点适合于电力负荷预测中的中长期负荷预测，但是需要对 SVR 模型相关参数进行确定，这不利于 SVR 的推广应用。针对这个问题，在本章中给出了利用知识挖掘技术中的微分进化算法协同 SVR 对长期负荷进行预测，通过利用北京市年用电量的实例分析中可以看出，将知识挖掘中的微分进化算法应用到 SVR 的参数选择中可以有效地提高预测的精度，并且能够在训练集和测试集两侧均保持较好的精度，不会出现像 BPNN 一样的过拟合现象的出现。

第 5 章 基于协同知识挖掘后干预纠偏技术的日最大负荷预测

本章提出了一种基于知识挖掘后干预纠偏技术的日最大负荷预测方法，该方法将传统负荷预测方法、智能方法以及知识挖掘三者结合进行电力负荷预测，具体来说是利用历史负荷数据利用传统电力负荷预测方法（如：ARMA）对未来负荷进行预测，对于影响负荷的非线性因素，采用智能方法对预测的未来负荷数据进行误差的修正，而对于非参数因素的影响，采用文本挖掘形成相应的知识和规则协同预测的方法，进而揭露其余因素对负荷的无规则影响，对结果进行后干预的纠偏工作。经实验证明，将负荷、影响负荷的非线性因素以及其他因素综合起来对负荷进行协同预测，以提高负荷预测的精度。

5.1 日最大负荷预测及预测方法选择

在负荷预测的研究中，日最大负荷预测是非常重要的一个预测指标，但是日最大负荷预测受到的影响因素也是最多的，和第三章中提到的日曲线负荷预测相比，虽然日最大负荷预测方法和日负荷曲线负荷预测均属于短期预测范围，但是两者的特点各不相同，首先，日负荷曲线由于存在着多点数据，因此可以对日负荷曲线整体进行聚类、分类方面的工作，这对于日最大负荷预测只有一个值的特点来说是很难做到的。其次重要的是，日最大负荷更容易受到不规则事件的影响。这就需要在日最大负荷预测中要利用组合的方法进行综合预测。和第四章的长期预测相比，日最大负荷预测必须考虑气象因素对负荷的影响，而隔日气象预报的结果是有一定的准确性的，可以作为影响因素参考。正是由于日最大负荷时需要考虑气象因素的影响，因此在对日最大负荷预测需要采用非线性的智能模型，上两章中提过的神经网络预测方法以及支持向量机预测方法均能胜任。在目前的研究中，支持向量机预测方法在 EUNITE 所组织的国际电力日最大负荷预测竞赛中脱颖而出[85]，取得了最佳的预测精度。因此本文选择支持向量机方法进行非线性因素的纠偏工作。但是利用综合方法仍然会存在少数因为不规则事件影响而出现的比较大的预测误差，因此需要利用知识挖掘的技术对预测结果进行后干预来纠正这些误差较大的点。

5.2 基于知识挖掘后干预技术的协同预测方法流程

基于知识挖掘后干预纠偏技术协同预测方法的框架图如图 5-1 所示，总体上可以分为三个模块，即：负荷预测模块，非线性因素误差纠正模块以及文本挖掘协同预测模块。其流程为首先利用数据库中的历史数据进行预测，得到初始预测结果 $\hat{L}_t$，然后对于预测的结果利用非线性因素和智能算法对预测产生的误差进行进一步的修正预测，得到 $\hat{E}_t$，最后，结合文本挖掘产生影响预测负荷的规则 $\hat{T}_t$ 对负荷预测结果进行进一步的集成预测从而得到最终的预测结果 $\hat{Y}_t$，其协同预测的公式可写成公式 5-1 的形式。

$$\hat{Y}_t = f(\hat{L}_t, \hat{E}_t, \hat{T}_t) \tag{5-1}$$

5.2.1 负荷预测模块

负荷预测模块主要由两个部分构成，分别是历史负荷数据库和负荷预测模型库。对于日最大负荷预测而言，在历史负荷数据库中提取出的数据视图中展现的应该是时间（包括年、月、日）和日最大负荷的数据，而负荷预测模型库中储存了大量的负荷预测方法，例如一元线性回归模型、ARMA 模型、神经网络模型等等。此外，还可以将这些基本模型进行组合预测。

在这些预测方法中，ARMA 是负荷预测中的一个重要的方法，目前在很多日最大负荷预测的研究文献中做为负荷预测中的基准方法用来比较其余算法的精度。本文选择该种方法作为日最大负荷预测的初步预测使用。其计算公式可以写成如下形式：

$$\phi(B)y_t = \theta(B)e_t \tag{5-2}$$

其中，y_t 表示 t 时刻时的负荷，e_t 表示 t 时刻时的随机误差，而 B 表示延迟算子，且 $B^n y_t = y_{t-n}$。$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$，$\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$，其中 p，q 是模型中的参数，而 e_t 是均值为 0，方差为 σ^2 的白噪声序列，即 $e_t \sim IID(0, \sigma^2)$。其中 p，q 参数估计可以利用自相关函数 ACF 和偏相关函数 PACF 计算得到或者由 AIC 准则得到。

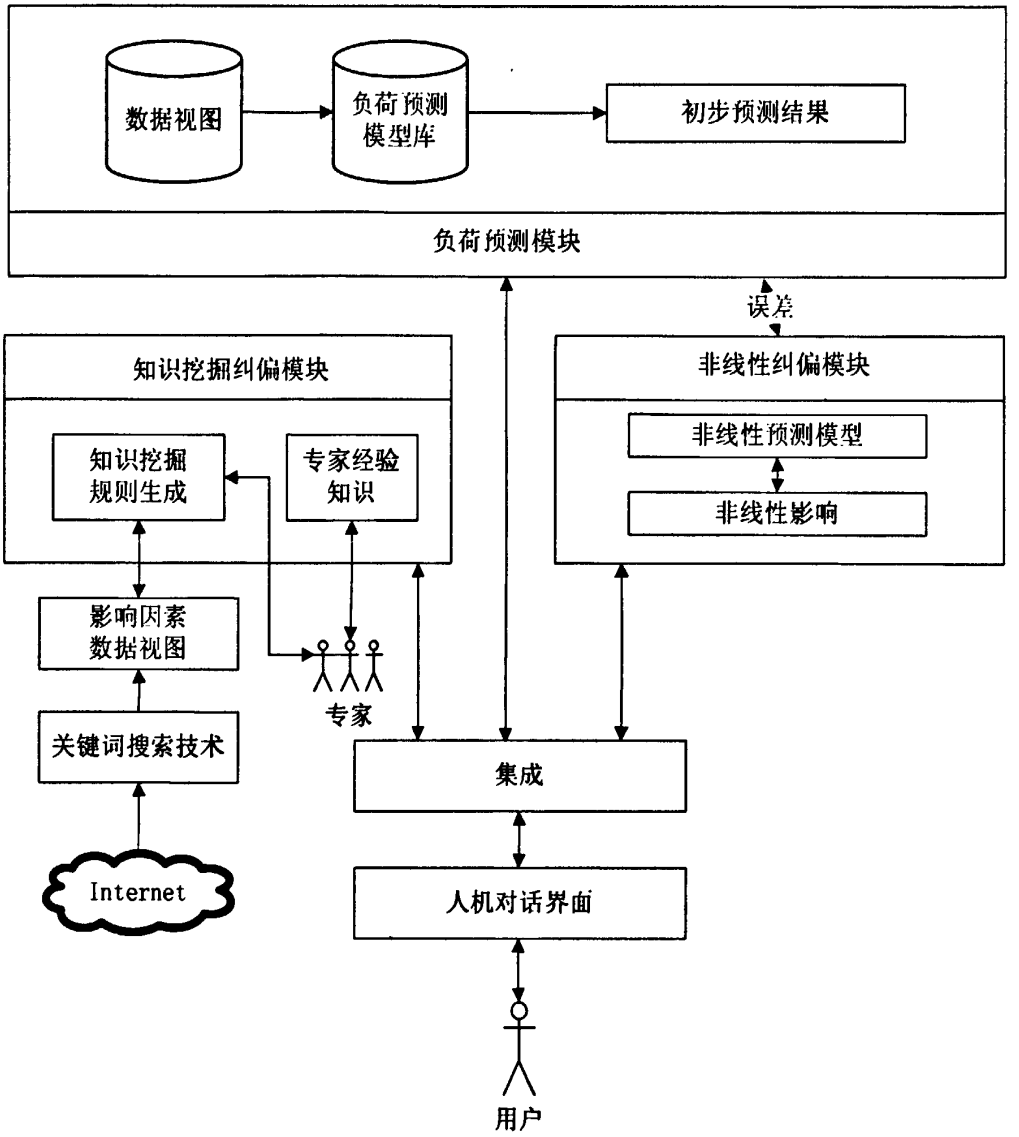


图 5-1 基于知识挖掘后干预纠偏技术协同预测方法的框架图

Fig. 5-1 The framework of combined knowledge mining correction techniques

利用 ARMA 对负荷进行日最大负荷预测大致分为以下几步：

（1）去掉相关的趋势（季节性周期趋势和线形趋势）使得负荷序列符合 ARMA 的预测条件；

其中去季节性周期趋势的方法如下：

1）收集至少三年以上各季度或者各月份的时间序列样本数据。

2）算出时间序列样本中历年所有季度或所有月份的算术平均值 $\bar{X}$。

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \qquad (5\text{-}3)$$

式中 n——时间列样本数据个数。

3）算出时间序列中同季度或同月份数据的算术平均值 $\bar{X}_i$。

4）算出季度或月份系数 β_i。

$$\beta_i = \frac{\bar{X}_i}{\bar{X}} \tag{5-4}$$

式中 i——季度或月份序号。

5）将所有原始数据乘以季度或月份系数可以得到去掉周期性趋势后的生产序列。

（2）确定 ARMA 中的参数 p，q；

（3）估计其余相关参数；

（4）进行负荷预测。

ARMA 具有良好的预测线性时间序列的能力，可以达到良好的精度，但是 ARMA 的缺点是其不能捕捉复杂负荷序列中的非线性因素影响以及其余因素对负荷的影响。因此，如果想要提高精度，就必须对预测结果进行进一步的修正。

5.2.2 非线性纠偏模块

如上文所述，利用 ARMA 时间序列模型虽然可以良好的预测负荷的趋势，但是很容易忽略非线性因素的影响，因此，非常有必要利用非线性的模型来拟合非线性因素的影响，上文中提到的 BPNN 和 SVR 模型都可以胜任此任务，本文采用的是第四章中的 SVR 模型对上述 ARMA 的预测结果进行非线性的纠偏工作。与第四章中预测不同的是这里采用的预测模型的输入输出数据基于的是前面的负荷预测模型和实际值之间的误差值。此外需要注意的是，对于 SVR 模型中所需要的三个参数，仍然可以利用知识挖掘中的微分进化算法来进行自适应参数的确定。

5.2.3 知识挖掘后干预纠偏模块

知识挖掘后干预模块是对预测结果进行进一步纠正的一个非常重要的模块，由于负荷预测模块和非线性因素智能预测纠正模块只能考虑因素对负荷的影响，而不能考虑其余可预测事件（如特殊节假日、电力系统故障等）对负荷的突然影响。因此，需要借助文本挖掘的手段和专家的参与生成一系列的知识进行对预测结果的后干预工作来进一步提高负荷预测的精度。

整个知识挖掘后干预预测模块由文本挖掘引擎和专家系统两部分构成，其结果是生成若干个 IF—THEN 规则供预测使用。文本挖掘首先从 Internet 上通过搜索引擎获取相关的数据，然后通过第二章介绍的文本挖掘的文本分类、聚类相关

模型对相关数据进行分析，之后形成一系列的预选规则后通过和专家进行交互从而提取出和负荷预测相关的规则。而专家系统则通过对历史数据的一系列推理，同专家进行交互同样得出相关的规则供预测使用。其预测部分用$\hat{T}_t$表示。最后，将从三个模块中得出的$\hat{L}_t,\hat{E}_t,\hat{T}_t$进行协同预测，一个非常简单的方法的是可以将三个结果直接相加进行协同预测，即：

$$\hat{Y}_t=\hat{L}_t+\hat{E}_t+\hat{T}_t \tag{5-5}$$

然而在实际中，由于知识挖掘协同预测模块是根据各种规则得出的预测值，因此预测得出的值$\hat{T}_t$在实际上往往并不是其余两者的补充部分，因此，三者协同预测的公式也可以进一步写为：

$$\hat{Y}_t=f(\hat{L}_t+\hat{E}_t,\hat{T}_t) \tag{5-6}$$

这样一来，最后一个重要的问题就是如何确定上式中的f，当然f同样也可以通过智能的算法进行训练后得到映射关系。其最优的准则可以采用均方差最小化，即：

$$\min(Y_t-\hat{Y}_t)^2 \tag{5-7}$$

5.3 实例分析

实例中采用的数据是中国广东省江门市 2005 年 1 月 1 日至 2007 年 12 月 31 日的日最大负荷数据，共计 1095 个数据点，整个序列用Y_t进行表示。如图 5-2 所示。

根据 5.1 节中介绍的基本流程，实例中实际操作步骤如下：

步骤 1：进行数据预处理的工作。可以很清楚的看到负荷序列存在着线性趋势和季节性趋势，因此首先提出掉周期为 365 的季节趋势，接下来利用最小二乘线性回归模型去除掉线性趋势。

步骤 2：ARMA 预测。通过 AIC 准则可以得到 ARMA 模型的p，q参数都是 1，利用 ARMA（1，1）的模型进行负荷时间序列的初步预测。

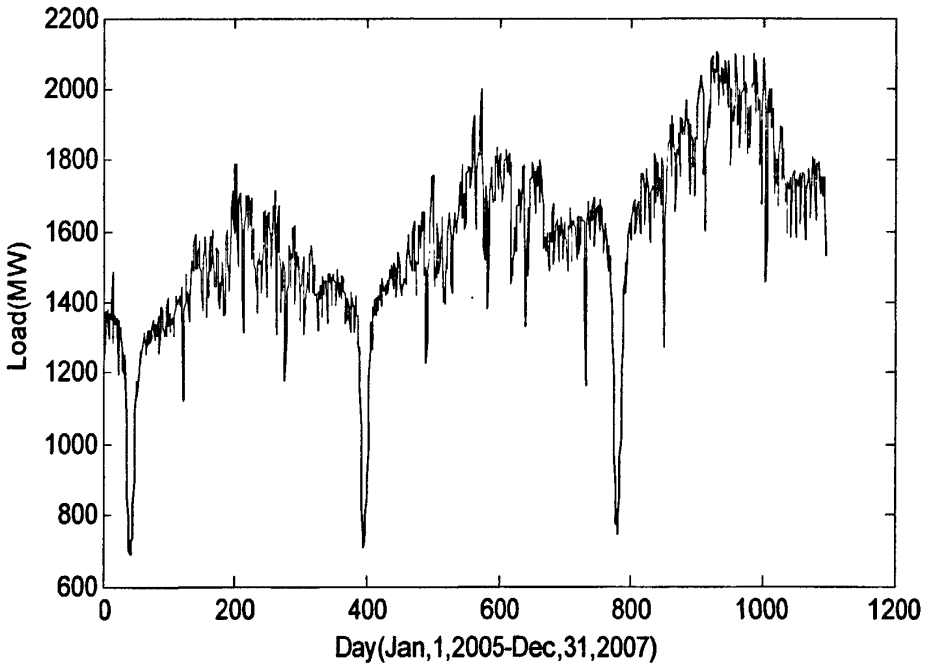


图 5-2 江门市 2005 年 1 月 1 日至 2007 年 12 月 31 日的日最大负荷数据

Fig. 5-2 Jiangmen City, 1 January 2005 to December 31, 2007 the daily maximum load data

步骤 3：利用 DESVR 模型进行非线性影响的纠偏，根据 ARMA 的预测结果 $\hat{Y}_t$，可以利用 $Y_t-\hat{Y}_t$ 获得误差序列 E_t，然后利用 DESVR 模型进行非线性影响的纠偏拟合。其中，DESVR 的输入因素是 $E_{t-1},E_{t-2},E_{t-3},E_{t-365}$，再将整个负荷数据序列分为两个部分，第一个部分是训练集，数据序列是从 2006 年 1 月 1 日至 2006 年 12 月 31 日，剩下的从 2007 年 1 月 1 日至 2007 年 12 月 31 日的数据作为测试集，从 DESVR 模型的结果中可以得到纠偏误差序列 $SVRE_t$。

步骤 4：对不规则影响进行知识规则分析。将 ARMA 的预测结果 $\hat{Y}_t$ 和 DESVR 的纠偏误差序列结果 $SVRE_t$ 相加可以计算出 Y_t'，然后利用公式 5-8 可以计算出误差偏离率 α_t。

$$\alpha_t=\frac{Y_t-Y_t'}{Y_t'} \tag{5-8}$$

根据前文中的知识影响规则模块的介绍，首先给出需要分析不规则情况的因素，具体见表 5-1。

表 5-1 知识挖掘中用的属性

Table 5-1 The properties of knowledge mining

因素		具体含义	知识挖掘中的取值或表示方法
日期型因素	周末	一周中的休息日	True/False
	特殊日期	元旦	NY
		和元旦连在一起的休息日	NYFree
		除夕	SFE
		春节	SF
		和春节连在一起的休息日	SFFree
		劳动节	ILD
		和劳动节连在一起的休息日	ILDFree
		国庆节	ND
		和国庆节连在一起的休息日	NDFree
	月份	一年中的月份值	1,2,3,...,12
气候因素	温度	最低温度，最高温度	mint,maxt
	湿度	最低，最高湿度	minh,maxh
	紫外线强度	紫外线指数	uvi
	人体舒适指数	人体舒适指数	hc

根据第二章提到的等距离散化方法，选取离散化的距离为 0.01，对 α_t 进行等距划分进行处理，其处理规则见表 5-2。

表 5-2 知识挖掘的预处理策略

Table 5-2 pre-knowledge mining strategy

SVREt 的范围	预处理策略	标记
[-1,-0.03)	保持原有值不变	原始值
[-0.03,-0.02)	调整为-0.02	-0.02
[-0.03,-0.02)	调整为-0.01	-0.01
[-0.01,0)	忽略	0
(0,0.01]	忽略	0
(0.01,0.02]	调整为 0.01	0.01
(0.02,0.03]	调整为 0.02	0.02
(0.03,1]	保持原有值不变	original value

在利用 Weka 生成决策树后，根据决策树可以抽取出 IF-THEN 规则，具体见表 5-3。

表 5-3 IF-THEN 纠偏规则（&表示“而且”）

Table 5-3 The corrective IF-THEN rules(& express and)

IF	Then
劳动节	-0.01
国庆节	-0.01
国庆节相关的串休日	-0.01
春节	-0.08
春节连续延长的放假日	-0.02
除夕	-0.02
元旦	0.02
元旦连续延长的假期	0.02
weekend = T & M >= 1.5 & maxt < 29.5 & M >= 4.5	-0.03
weekend = T & 1.5=<M<=4.5 & maxt >= 29.5 & mint < 24.5	0.02
weekend = T & 1.5=<M<=4.5 & 29.5=<maxt<= 31.5 & mint < 24.5	0.03
weekend=T & 1.5=<M<7.5 & 29.5=<maxt<33.8 & 24.5=<mint<=27.5& Minh<0.68	-0.02
weekend=T & 1.5=<M<7.5 & 33.8=<maxt<34.25 & mint>=24.5 & Minh<0.68	-0.04
weekend=T & M>=1.5 & maxt>=29.5 & mint>=24.5 & Minh>=0.68	0.01

步骤 5： 将 ARMA、DESVR 和 IF-THEN 规则进行集成后可以得到最终的预测结果。

$$FY_t = (1+\alpha)Y_t' = (1+\alpha)(\hat{Y}_t + SVRE_t) \tag{5-9}$$

利用 MAPE 指标对预测结果进行比较，其比较结果见表 5-4。

表 5-4. 三种方法的 MAPE 值

Table 5-4. MAPE values of three methods

method	MAPE		
	training set	test set	whole
ARMA	2.56%	2.82%	2.83%
ARMA+DESVR	1.82%	2.55%	2.18%
TIK	1.70%	2.30%	2.03%

从表 5-4 中可以很清楚的看出经过 DESVR 纠偏后的结果要比单纯利用 ARMA 的预测结果要好，最佳的预测结果是 TIK 方法得到的预测结果，TIK 方法得到的预测结果在训练集、测试集和整个样本集合中都含有最低的 MAPE 值。形成这种结果的主要原因是 DESVR 方法在 ARMA 预测结果中添加非线性因素的影响，经过纠偏之后有效地提高了预测的精度，而 IF-then 规则主要对不规则事件引起的较大预测偏差进行了后干预，使得预测结果的精度更进了一步。图 5-3 给出了最后 61 个时间点（数据集中最后两个月的最大日负荷值）的预测结果对比，从图中的对比可以进一步看出这一原因。

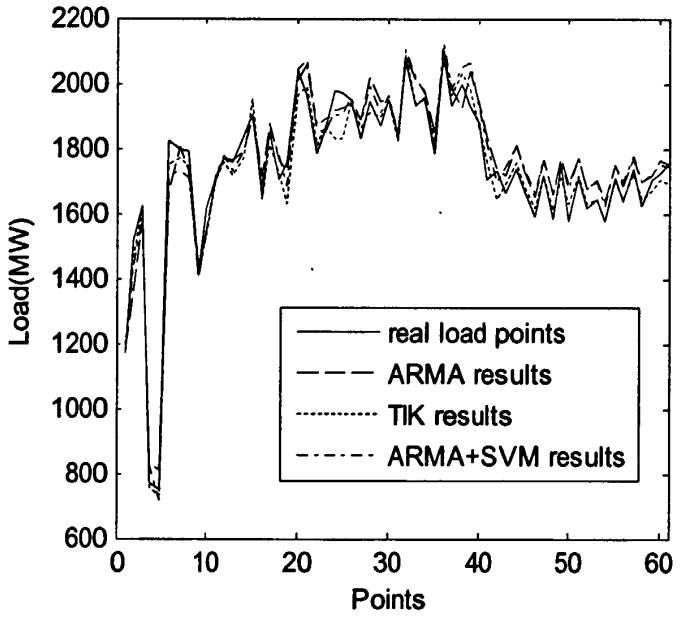


图 5-3. ARMA、ARMA 结合 DESVM 和 TIK 方法得出的预测结果

Fig 5-3. The results of ARMA results, ARMA combined DESVM method and the proposed TIK method

5.4 本章小结

由于日最大负荷较容易受到影响因素的干扰，因此在本章中提出基于知识挖掘后干预纠偏技术的日最大负荷预测模型，该模型集成了时间序列预测模型，非线性智能预测模型和知识挖掘方法，它可以同时考虑负荷预测趋势、非线性影响和不规则影响，首先 ARMA 方法对负荷序列进行初步的预测，接下来利用 DESVR 方法在 ARMA 预测结果中添加非线性因素的影响，对非线性因素产生的影响进行修正，经过纠偏之后可以有效地提高预测的精度，最后通过知识挖掘技术生成

IF-then 规则，通过这些规则对不规则事件引起的较大预测偏差进行了后干预，使得预测结果的精度更进了一步，这样可以有效地提高预测的效果。通过江门市的实际负荷数据进行实验，其结果证明了所提出模型的有效性，实验结果证明经过 SVR 纠偏可以显著地提高日最大负荷的预测精度，而且知识挖掘形成的规则可以进一步提高预测结果的精度，利用本章提出的方法可以得到最佳的预测精度。

第 6 章 基于协同知识挖掘技术智能预测结果的预警研究

由于电力的供给直接影响社会的生产和人民的正常生活，供电的意外中断非常可能打断正常的社会生产生活，如果引发了大面积的停电事故将会对社会造成难以估量的危害。再者，近年来在世界范围内的大面积停电事件发生的频率越来越高，影响的范围越来越大，造成的危害也越来越严重。大面积停电事件已引起了人们的重视。此外，由于恶劣自然条件引起的停电事故在近年也引起人们的重视，极端气候条件如台风、暴风雪、给电力设施带来极大损害的同时，对我国社会和经济的发展造成了极大的影响。本章在上述建立的基于知识挖掘技术的数据库的基础出发，以前文的日负荷曲线预测以及中长期预测结果为基础，对实际负荷预警监测以及供需预警监测和天气灾变预警三个方面进行研究。

6.1 基于短期负荷预测结果的负荷监测预警研究

由于电力的生产和销售过程涉及到发电、输电、配电、用电等多个环节，而电网作为连接发电和用电的整个电力运营中的一个纽带作用，将会受到来自其他各个方面的不同程度影响，例如简单的违反电力系统操作造成的突然断路、电力设备在长期的运行中受到有意无意破坏等原因都会都会给整个电网带来巨大冲击，影响电网正常运行。这时对短期负荷尤其是日曲线负荷进行监测预警工作可以能够及时发现不良的状态，进行相应方案的实施，从而保障电力的安全稳定运行。

6.1.1 短期负荷监测指标

（1）时点负荷偏离度

时点负荷偏离度是衡量电网的实际值和预测值之间出现偏差的程度，可以在一定程度上直观地反映电网由于突发事件影响所损失的负荷量，其计算方法为：

$$\eta_h = \frac{L_{loss}}{L_{forecast}} = \frac{L_{actual} - L_{forecast}}{L_{forecast}} \tag{6-1}$$

其中：η_h 为表示时点负荷偏离度，L_{loss} 为负荷损失值，$L_{forecast}$ 表示负荷预测值。

（2）重要用户时点负荷偏离度

由于电网所覆盖的区域中不同用户对负荷损失所造成的经济社会影响不同，因此在同样损失负荷的情况下，各区域的的警情也应该不同。根据我国《重要电力用户供电电源及自备应急电源配置技术规范》的相关规定，重要电力用户可以分成四类，分别是特级、一级、二级和临时性重要电力用户。同样需要按照上面的公式 6-1 来对时点偏离度进行监测。其中各类重要电力用户的划分标准如下：

1）特级重要电力用户

这类用户是指在管理国家事务中具有特别重要作用的用户，当中断供电时将有可能危害国家安全的电力用户，一般来讲，政府机关属于该类用户。

2）一级重要电力用户

这类用户是指在中断供电时可能产生直接引发人身伤亡事故、造成严重污染、发生中毒、爆炸或者火灾事件、造成重大政治影响、经济损失或者是有可能造成较大范围社会公共秩序严重混乱的用户。一般来讲，供电区域内的化工厂、工业用电用户以及移动通信用户等属于该类用户。

3）二级重要电力用户

这类用户指的是当中断供电时将可能引起较大环境污染、较大政治影响、较大经济损失或者造成一定范围社会公共秩序严重混乱的用户，一般来讲，供电区域内的学校属于这类用户。

4）临时性重要电力用户

这类用户指的是临时需要特殊提供供电保障的一类用户。

（3）累计损失电量比率

对于某些负荷中断现象，可能在一开始时由于负荷的中断量小很难利用负荷偏离度察觉，当累积到一定程度时才能被发现，这时需要利用累计损失电量指标进行衡量，从而达到预警的目的，累计损失电量比率的计算方法如下：

$$\eta=\frac{Q_{loss}}{Q_{forecast}}=\frac{\int_{i}^{t}l_{actual}-\int_{i}^{t}l_{forecast}}{\int_{i}^{t}l_{forecast}} \tag{6-2}$$

其中：η为表示累计损失电量比率，$\int_{i}^{t}l_{actual}$为从第 i 个时刻开始监测到第 t 个时点的累计电量值，$\int_{i}^{t}l_{forecast}$为从第 i 个时刻开始监测到第 t 个时点的预测累计电量值，当η超过一定的范围时将进行预警。

（4）停电人数比例

由于居民负荷所占的负荷比较小，而且也不属于重要用户的监测范围，因此，

利用负荷损失度或者是累计负荷损失比率不能完全衡量居民负荷的警情程度，这时，可以通过停电人数比例来进行衡量出现负荷中断警情所波及的程度，其计算方法如下：

$$\eta_p = \frac{P_{loss}}{P}$$

其中，η_p为停电区域人数占该供电区域覆盖总人口比例；P_{loss}为停电波及总人数；P为供电区域覆盖总人口。

6.1.2 短期负荷预测监测指标的警度设置

（1）时点负荷偏离度警度设置

根据短期负荷预测的精度判别标准，当时点预测偏离度在[-3%，+3%]的范围内时可以视为一般警情。参照电力行业的相关规定以及相关文献的研究，本文将有警情况下的警度设置为三级，及轻度警情，中度警情以及高度警情，分别用橙色、黄色、红色进行标识。时点负荷偏离度警度的具体设置如表 6-1 所示。

表 6-1 时点负荷偏离度警度设置

Table 6-1 The alarm degrees of load point deviation

警度	一般警情	轻度警情	中度警情	高度警情
取值范围	少于 3%	3%-5%	5%10%	超出 10%
预警颜色	蓝色	黄色	橙色	红色

（2）重要用户时点负荷偏离度警度设置

根据上述重要用户的设置原则以及预测精度的判别标准，同样地，当时点预测偏离度在[-3%，+3%]的范围内时可以视为一般警情。对于特级用户而言，当时点负荷偏离度超出[-3%，+3%]的范围时，直接给出红色预警信号；对于一级用户而言，当时点负荷偏离度在 3%至 5%的范围内给出黄色预警信号，当时点负荷偏离度超出 5%的范围时给出红色预警信号；对于二级用户而言，当时点负荷偏离度在 3%至 5%的范围内给出橙色预警信号，当时点负荷偏离度超出 5%但是在 10%的范围时给出黄色预警信号，当时点负荷偏离度超出 10%的范围时给出红色预警信号。

（3）累计损失电量比率警度设置

目前根据我国停电范围和事故严重程度的相关规定，将停电事件分成两个级别，分别是一级停电事件和二级停电事件两种。其中一级停电事件的状态是指发生下述停电事件之一：

1）区域电网减供负荷超过事故前总负荷的 30%。

2）重要政治、经济中心区域减供负荷超过事故前总负荷的 50%以上。

3）由于严重自然灾害引起的电力设施大范围破坏，省网大面积停电，减供负荷达到事故前总负荷的40%以上的，并且连带造成重要发电厂停电，电力设备严重受损，对区域电网以及跨区域电网安全稳定运行构成严重威胁的。

4）由于发电燃料供应短缺或者其他各类原因引起的电力供给严重危机造成区域60%以上容量机组非计划停机以及省网拉闸限电负荷达到正常值的50%情况的，并且对与区域电网和跨区域电网的正常电力供应构成严重威胁影响的。

5）由于重要发电厂、变电站以及输变电设备遭受到毁灭性的破坏或者打击的，造成了区域电网大面积停电，使得减供负荷达到事故前总负荷的20%以上，对区域电网、跨区电网安全稳定运行构成严重威胁的。

而二级停电事件是指发生下列情况之一时：

1）因电力生产发生重特大事故，造成区域电网减供负荷达到事故前总负荷的10%以上，30%以下；

2）因电力生产发生重特大事故，造成重要政治、经济中心城市减供负荷达到事故前总负荷的20%以上，50%以下；

3）因严重自然灾害引起电力设施大范围破坏，造成省电网减供负荷达到事故前总负荷的20%以上，40%以下；

4）因发电燃料供应短缺等各类原因引起电力供应危机，造成省电网40%以上，60%以下容量机组非计划停机。

上述规定属于国家层面的规定，以国家规定为参考，各省市区域电网根据自己的实际情况做出调整，分别制定了适合自己的停电事件等级状态，例如广东省的一级和二级停电事件规定如下：

一级停电事件：

（1）因电力生产发生重特大事故，引起连锁反应，造成南方电网大面积停电，减供电负荷达到事故前总负荷的30%以上；

（2）因电力生产发生重特大事故，引起连锁反应，造成广东省重要政治、经济中心城市减供负荷达到事故前总负荷的50%以上；

（3）因严重自然灾害引起电力设施大范围破坏，造成广东电网大面积停电，减供负荷达到事故前总负荷的40%以上，并且造成重要发电厂停电、重要输变电设备受损，对南方电网、跨区电网安全稳定运行构成严重威胁；

（4）因发电燃料供应短缺等各类原因引起电力供应严重危机，造成广东电网60%以上容量机组非计划停机，广东电网拉限电负荷达到正常值的50%以上，并且对南方电网、跨区电网正常电力供应构成严重影响；

（5）因重要发电厂、重要变电站、重要输变电设备遭受毁灭性破坏或打击，造成南方电网大面积停电，减供负荷达到事故前总负荷的20%以上，对南方电

网、跨区电网安全稳定运行构成严重威胁。

二级停电事件：

（1）因电力生产发生重特大事故，造成南方电网减供负荷达到事故前总负荷的10%以上，30%以下；

（2）因电力生产发生重特大事故，造成广东省重要政治、经济中心城市减供负荷达到事故前总负荷的20%以上，50%以下；

（3）因严重自然灾害引起电力设施大范围破坏，造成广东电网减供负荷达到事故前总负荷的20%以上，40%以下；

（4）因发电燃料供应短缺等各类原因引起电力供应危机，造成广东电网40%以上，60%以下容量机组非计划停机。

根据上述的停电事件级别判别依据，取上述规定的较小值作为警度设置界限可以得出，当累计损失电量比率在10%以内时，视为一般警情；当累计损失电量比率超过10%到20%时，进行轻度警情预警，预警颜色为橙色，当累计损失电量比率在10%到30%之间时，进行中度警情预警，预警颜色为黄色；当累计损失电量比率超过30%时，进行高度警情预警，预警颜色为红色。因此，具体设置方式见表6-2。

表6-2 累计损失电量比率警度设置

Table 6-2 The alarm degrees of total power loss ratio level

警度	一般警情	轻度警情	中度警情	高度警情
取值范围	少于10%	10%-20%	20%-30%	超出30%
预警颜色	蓝色	黄色	橙色	红色

（4）停电人数比例警度设置

关于居民停电人数造成的影响目前我国部分城市已经开始实施，以北京市为例，停电人数的影响级别设置为四个级别，分别是特别重大(Ⅰ级)、重大(Ⅱ级)、较大(Ⅲ级)和一般(Ⅳ级)四个级别。其中当北京市出现60万户以上居民停电事件时为特别重大电力突发事件(Ⅰ级)；当全市出现20万户以上60万户以下居民停电事件时为重大电力突发事件(Ⅱ级)；当城八区和亦庄地区发生20万户以下1万户以上居民停电事件，或远郊区县发生20万户以下2万户以上居民停电事件时为较大电力突发事件（Ⅲ级)；当城八区和亦庄地区发生1万户以下1000户以上居民停电两小时以上的事件，或远郊区县发生2万户以下2000户以上居民停电两小时以上的事件为一般电力突发事件(Ⅳ级)。

按照北京市统计局2009年公布的人口数字1755万人来计算，上述的四个级别对应的停电人数比例可以粗略的计算如下：当停电人数比例超过3.5%时，为特别重大事件，当停电人数比率在1.1%到3.5%的范围时属于重大事件，当停

电人数比率为0.1%到1.1%的范围时属于较大事件，当停电人口比例超过万分之一时属于一般电力突发事件。参照此标准，本文将停电人数比例警度设置如表6-3所示：

表6-3 停电人数比例警度设置

Table 6-3. The alarm degrees of power cut's people proportion

警度	一般警情	轻度警情	中度警情	高度警情
取值范围	少于0.1%	0.1%-1%	1%-3.5%	超出3.5%
预警颜色	蓝色	黄色	橙色	红色

6.2 基于中长期负荷预测结果的电力供需预警研究

由于中长期负荷预测和电网企业的规划紧密相关，而且由于电力能源的难以存储性，需要在同一时点达到瞬间的供需平衡，因此，电力的供应量和需求量之间在理想的状态下应该是一个平衡的状态。但是这种状态在预测中是几乎不可能实现的，在对未来长期负荷供电量和需求量进行预测时，即使剔除线损等因素也往往会出现一定的偏差，当这种偏差值超出一定的范围时，表示电力的供给和需求之间的平衡关系将要被打破。当供电量的预测超出需求量预测的情况下，表明电力的供给能力将要超出需求能力，这意味着大量的电力投资在空置浪费，如果需求的预测值高出供电量预测值一定范围时，表明电力投资不足，将会在一定程度上影响国民经济的发展。因此，有必要对电力供给和需求的平衡性进行预警方面的研究。

6.2.1 供需预警指标

（1）全社会电力供需比

全社会电力供需比是反映电力供应量和需求量之间平衡状态的最关键的指标，其计算方法为地区供电量和全社会用电量之间的比值，

（2）电力需求增长率

电力需求增长率反映的是电力系统需求增长速度的指标，由于我国城市化进程以及人民生活水平的逐步提高，加上家用电器的普及，使得电力需求必然加快增长，容易造成供需矛盾而引发的拉闸限电的情况发生。此外，2008年底发生的美国金融危机造成的影响使得我国的电力需求增长率呈现了下降的趋势，有些地方甚至出现了负增长。因此，电力需求的增长率也是供需预警需要监测的重要指标之一。其计算方法是电力需求量预测值/同时间上一期电力需求量-1。

（3）电力行业需求增长率

电力行业需求增长率指的是按国民经济行业的划分方法中各行业的需求增长速度，通过将主要行业的增长率进行分析来进行行业供需平衡之间的监测预警作用，其计算方法类似于电力需求增长率，所不同的是采用的是各行业的用电量。

（4）系统容量备用率

系统容量备用率反映的是系统能够承担的出于预测误差、平衡电力波动以及替代系统检修容量的能力，能够描述系统的潜在供给能力，其计算方法是 1-系统最大负荷/系统最大可调出力。其中系统的最大可调出力根据我国《电力系统安全稳定运行导则》的相关规定，可以按照地区的总装机容量×0.88 来计算。

6.2.2 供需预警指标的警度设置

本文对上文中供需预警指标的警度设置如下：

（1）全社会电力供需比的警度设置

本文将供需预警指标的警度划分成供给过剩、供给富裕、供给均衡，供给紧张以及供给严重不足五个类别，这五个警度区间的设置除了主要依靠国家和电力行业的相关安全规定和制度外，还需要依靠地区实际的情况。根据其余相关文献的研究，当供需差距在 5%范围内时，可以看成是供需基本均衡，这是因为可以将其中的差别看成是由于线损率以及中长期负荷的供电量预测和用电量预测两部分产生的误差造成的；当供给和需求差别在 5%到 8%时，可以看成是轻度警情，当供电量的预测值高出 5%到 8%时，可以看成是供给富裕；当用电量的预测值高出供电量的 5%到 10%时，可以看成是供给紧张，当供电量的预测值超出用电量预测值的 10%时，视为供给过剩，当用电量的预测值超出供电量预测值的 10%时，视为供给严重不足。其警度设置如表 6-4 所示。

表 6-4 全社会电力供需比的警度设置

Table 6-4 The alarm degrees of supply and demand ratio

警度	一般警情	轻度警情	中度警情	高度警情
取值范围	0.95%-1.05	[0.93,0.95)和(1.05,1.08]	[0.9,0.93)和(1.05,1.08]	低于 0.9,高于 1.1
预警信号	蓝色	黄色	橙色	红色

（2）电力需求增长率的警度设置

电力需求增长率应该和我国国民经济生产总值的发展呈线性关系，在理想状态下最好是保持一致的水平。目前我国国民经济生产总值的规划增长速度一般为 7%到 8%之间，本着电力先行的原则，电力需求的增长率可以稍高一些，因此当电力需求增长率在 10%以下时视为无警情，以后每上升 10%警情提高一度，即在 10%到 20%时为轻度警情；在 20%到 30%时为中度警情；超出 40%时为高度警情。当电力需求增长率呈现负值时需要引起高度的关注，因此当电力需求增长

率在[-5%，0）之间的范围时，则认为是轻度警情，在[-10%，-5%）之间的范围时，则认为是中度警情，当高于10%时认为是高度警情，其警度设置如表6-5所示。

表6-5 电力需求增长率的警度设置

Table 6-5 The alarm degrees of electricity demand growth

警度	一般警情	轻度警情	中度警情	高度警情
取值范围	[0,10%]	(10%,20%] [-5%，0）	(30%,40%] [-10%，-5%）	>40% <-10%
预警信号	蓝色	黄色	橙色	红色

此外，如果数据具有一定的数量时，如月度的电力需求增长率预警中也可以采用系统化的警度设置方法，所谓系统化的警度设置方法是通过对历史的电力需求增长率进行分析，按照系统化的多数原则进行划分，具体划分是将历史的电力需求增长率按照从小到大排列，按照从小到大排列的2/3作为警限，作为安全区域，剩下的作为预警区域，再继续从剩下的1/3中的从小到大排列的序列中选取2/3作为轻度预警警度，以此类推划分成中度预警警度，将剩下的部分作为重度预警警度。

（3）电力行业需求增长率设置

电力行业需求增长率的警度设置方法和上述的电力需求增长率的设置类似。其中相关的行业划分可以参照我国国民经济行业分类（GB/T 4754-2002）的标准，一共需要对95项大类和913项小类的指标进行监测。同时对这么多指标进行监测往往增加了预警的难度和复杂度，因此在对电力行业需求预警进行研究时，一般需要对所监测的国民经济行业指标项进行一定的筛选工作，选取出具有代表性的指标进行监测。其中我国国民经济行业分类（GB/T 4754-2002）的标准如表6-6所示：

（4）系统容量备用率

根据电力行业的经验，系统容量的备用率一般保持在8%～15%之间较为合适，其中如果以火电为主的电力系统的系统检修备用率一般保持在10%～15%之间，因此本文规定15%为警限值，当系统的容量备用率超过15%时，将进行预警信号的显示。

表 6-6 我国国民经济行业分类的标准

Table 6-6 Standard Industrial Classification of National Economy

门类	大类	中类	小类
A 农、林、牧、渔业	5	18	38
B 采矿业	6	15	33
C 制造业	30	169	482
D 电力、燃气及水的生产和供应业	3	7	10
E 建筑业	4	7	11
F 交通运输、仓储和邮政业	9	24	37
G 信息传输、计算机服务和软件业	3	10	14
H 批发和零售业	2	18	93
I 住宿和餐饮业	2	7	7
J 金融业	4	16	16
K 房地产业	1	4	4
L 租赁和商务服务业	2	11	27
M 科学研究、技术服务和地质勘查业	4	19	23
N 水利、环境和公共设施管理业	3	8	18
O 居民服务和其他服务业	2	12	16
P 教育	1	5	13
Q 卫生、社会保障和社会福利业	3	11	17
R 文化、体育和娱乐业	5	22	29
S 公共管理和社会组织	5	12	24
T 国际组织	1	1	1
(合计) 20	95	396	913

6.2.3 基于知识挖掘分类技术的行业需求预警指标筛选研究

前文中提到，如果对全部的国民经济行业进行预警监测将会增加预警的复杂度和难度，因此需要对指标进行筛选，这可以利用前文中提到的知识挖掘中的决策树分类技术来进行处理，利用分类技术对电力行业需求预警指标进行筛选的步骤如下：

（1）首先在数据集内剔除掉区域电网内不涉及到的国民经济行业类型，即国民行业中的行业用电量值始终为 0 的类别。

（2）根据区域电网的实际情况在指标中设置必须需要监测的指标，即这些

行业属于必须进行预警监测的行业，一般以第一产业中的行业比较常见。

（3）将余下各个类别的行业用电量值进行升序排序，即按照行业用电量从小到大排序，计算累积行业用电量，去除掉累积电量不足整个行业用电量3%的类别，去除掉这些行业的原因一个是因为考虑到个别行业的用电量非常小，即使出现较大的预测误差对于整体误差并没有太大的影响，第二是这些用电量小的行业用电量的曲线模式一般都比较平稳，很难出现很大的预测误差。

（3)选取预测模型利用历史数据分别对电力需求和各行业的需求进行预测，得出各类需求的预测值，根据预测值和实际值的差异按照警度划分的原则进行警度划分，得出各行业的警度划分表以及整个电力需求的警度划分情况，将这些数值组成决策表，准备进行知识挖掘使用。

（4）利用决策树分类技术进行分析，得出相应的规则，选取规则中涉及到的行业需求预警指标作为候选行业。

（5）将候选行业和第二步中规定的必须监测行业结合起来，形成电力行业需求预警的指标集。

6.3 灾害气候预警研究

影响电网的危机的自然灾害是十分繁多的：例如近年来我国电网受到地震、水灾、风灾、雪灾、冰灾等的毁坏，这些灾害产生的影响可以直接造成电网企业的非正常经营，间接影响到国民经济的正常发展，因此对于自然灾害的预警监测是十分必要的。在造成电网危机的自然灾害中，有些灾害如地震灾害的预警目前仍然世界性的难题，但是有些灾害例如冰灾、台风等是可以通过一定的气候条件和季节规律进行预警，这些一般都是一些典型气候引起的自然灾害，这些自然气候灾害是有一定规律可循的。

6.3.1 威胁电力的典型灾害气候

（1）冰灾

冰灾尤其是夹杂着冻雨形成的冰灾对于电力系统的危害是巨大的，以国外为例，在1998年的一月份，长达80个小时的一系列冰冻雨覆盖范围包含安大略省东部，魁北克省南部，纽约州北部以及新英格兰地区北部四处。这次冰灾造成输电设施上的最大覆冰厚度达75mm，导致116条高压输电线路破坏和1300基输电塔倒塌。配电线路破坏350条，杆塔倒塌16000座(个)。冰灾造成100万用户停电，停电影响到的人口占加拿大人口总数的10%。号称拥有世界上最坚实的电塔魁北克在这次冰灾中，1000座电塔和35000个木制电线杆上被压倒，经济损

失就高达 20 亿美元。一些地区整整一个月都没有电。连接魁北克最大城市蒙特利尔电线都断了，停电数天。而蒙特利尔南部三市 Saint-Hyacinthe，Granby 和 Saint-Jean-sur-Richelieu 因为几个星期都不通电，被当地法语媒体称为“黑三角”。美国方面，缅因州 120 万居民中，有 70 万断电。在冰风暴三周后，仍有约 70 万人无电可用。再如 2005 年 1 月，瑞典遭受特大暴风雪的袭击，造成大量倒塔和断线，引起大面积停电，供电恢复时间长达 35 天。

以我国为例，从 1999 年 3 月 12 到 3 月 17 日，我国京津唐部分地区出现气温在 0 摄氏度左右的雨雪天气，由于绝缘子覆冰（雪）造成京津唐电网 10 条线路 47 条次的闪络，闪络线路涉及的电压等级包括 110kV、220kV 及 500kV。2005 年 2 月，我国重庆东南地区遭遇二十年一遇的特大风雪袭击，覆冰厚度 50～70mm。致使 220kV 黔秀西线 127 号、128 号两基铁塔因线路覆冰过重而倒塌，黔秀西线 97 号、129 号铁塔分别于 14 日和 15 日倒塌。2007 年 3 月 4 日，我国东北电网受到特大暴风雪袭击、造成 4 条 500kV 线路跳闸；多条线路通信光缆中断。大连地区电网与主网解列，且有 14 座 220kV 变电站、1 座牵引变、154 座 66kV 变电站停电。

我国电力系统遭受冰灾最严重、范围最广的是 2008 年 1 月份南方的大部分地区以及华中、华东的部分地区因为遭遇历史上罕见的持续大范围低温、雨雪和冰冻天气造成的灾害，其影响面积之大、损害之严重程度为建国以来最为罕见的一次。这次冰灾给黔、湘、桂、皖、赣、粤等 14 个省级电网（含自治区、直辖市）造成不同程度的损害，影响的居民过亿，近 570 个县的用户供电受到不同程度的影响，受灾害影响导致全站停电的 500kV 变电站高达 15 座，500kV 的电力线路共有 119 条，500kV 杆塔倒了 678 基、受损 295 基。从上述历史发生的事件中可以看出，冰灾是对电力危害最严重的灾害气候，因此对于形成冰灾形成的典型气候进行分析并进行预警是很有必要的。

（2）沙尘暴

沙尘暴是一种风中含沙的灾害性你的天气效应，由于其形成和过量砍伐森林、植被破坏、气候异常以及温室效应引起的多方效应相关，因此该灾害多发于气候干旱，植被稀疏的地区。沙尘暴对于电力的危害也是比较大的，例如 1990 年 4 月，由于遭受沙尘暴引起了埃及的首都开罗以及其余的几个主要城市的断电，引起了严重的混乱，在我国，在 2001 年新疆阿克苏地区的一次强沙尘暴袭击引起了大量的输电塔倒下，引起了较大面积的停电。

我国的沙尘暴天气一般分为三个路径，其一是西北路径，其沙尘起源一般来源于蒙古南部或者是我国内蒙西部，主要波及的地区为西北东部、华北北部和东北地区；其二是偏西路径，其沙尘来源于蒙古西南部，主要影响我国西北和华北

大部分地区；最后一个路径是偏北路径，沙尘来源一般来源于蒙古国乌兰巴托以南的广大地区，主要影响西北地区东部，华北地区大部和东北地区南部。

在沙尘暴灾害天气中，气流中的沙粒由于和地床面之间的电荷交换带上了电荷，使得空气中的介电常数增大，电阻率减小，从而形成较强的负极性电场，进一步改变输电线路电场的分布，威胁电力相关设备的外绝缘运行，影响到电力系统的可靠性。如果在沙尘暴过后如果遇到小雨的话，由于户外的变电站电力设备表面由于前期在沙尘暴天气中沉积大量的沙尘，加上下雨使得户外变电站短路跳闸事故的可能性加大。

对于沙尘暴灾害天气预警需要监测以下几个气象条件：一个是形成沙尘暴的动力因素风速，第二个是沙尘密度。此外，沙尘暴一般发生在春季的 3 月、4 月和 5 月三个月份，因此可以对这 3 个月份进行监测。还有一个排除条件是如果刚降过雨，在雨后的一段时间内是不会沙尘暴的。

（3）台风

台风是由于海洋上产生热带气旋而产生的，因此台风的防范一般仅限于沿海国家或地区，像日本和台湾地区以及我国沿海地区都是需要对台风进行预警的区域。台风给电力带来的危害也是巨大的，例如 1991 年到 1993 年登陆的 19 号台风对日本的高压输电塔和其余输变电设备造成了极大的损失，2002 年 10 月，日本 21 号台风造成茨城县 10 基高压输电塔连续倒塌的严重事故，一共造成 30 万用户停电共造成 30 万用户停电，289000kW 电力供给发生故障。

在我国，沿海地区每年平均有 7 个热带气旋（含台风）登陆，由这些热带气旋的登陆对我国电力安全构成严重威胁，例如 2004 年 8 月 12 日“云娜”台风在我国浙江登陆，损坏的输电线路达到 3342km。受这次台风影响，浙江电网 500kV 线路跳闸 10 次，全省共有 9 座 220kV 变电所失电；110kV 系统线路跳闸 68 次，主变压器跳闸 5 台次。2006 年 8 月 10 号，台风“桑美”登陆我国浙、闽沿海地区，中心风力最高 17 级，是建国 50 年以来登陆中国内地地区最大的超强台风(相当于 5 级飓风)。受“桑美”影响，我国华东电网 4 回 500kV 线路、6 回 220kV 线路、32 回 110kV 线路跳闸，浙、闽各有 1 座 220kV 变电站出于全部停电状态。

6.3.2 典型气候监测警度的设置

（1）冰灾警度设置

根据气象学的相关资料，冰状物的形成主要是和温度、风速和湿度三个因素密切相关，当气温在-5℃-0℃之间，风速为 1～10 m/s 时，空气湿度达到 85%以上时达到最佳的形成条件。当达到成冰条件后，由于大气环流移动中空气冷热交替的过程中会产生“过冷却”现象，使得过冷却水滴容易附着在固体表面形成雪

凇、雾凇、雨凇等冰状物，造成输电线设备覆冰厚度过厚从而危害输电安全性的后果。其中南方多以雨凇为主，北方多以雪凇和雾凇，相比较而言，雨凇造成的危害更大。从上述角度出发，对于冰灾进行预测的气象条件所监测的条件属性变量应该取为温度，风速和湿度三个因素，而用来判断冰灾危害与否的决策变量应取覆冰厚度。参照我国气象行业标准颁发的《电线结冰风险等级标准》中规定的警度设置，对冰灾的警度设置如表 6-7 所示：

表 6-7 冰灾警度设置

Table 6-7 The alarm degrees of ice disaster warning

警度	一般警情	轻度警情	中度警情	高度警情
覆冰厚度取值范围	0～10mm	10-20mm	20-30mm	超出 30mm
预警颜色	蓝色	黄色	橙色	红色

（2）沙尘暴警度设置

由于我国在发布的《沙尘暴天气等级》国家标准中规定：我国沙尘的天气等级分为浮尘、扬沙、沙尘暴、强沙尘暴和特强沙尘暴五个等级。具体设置如下：

1）浮尘：当天气条件为无风或平均风速≤3.0 米/秒时，尘沙浮游在空中，使水平能见度小于 10 千米的天气现象。

2）扬沙：风将地面尘沙吹起，使空气相当混浊，水平能见度在 1 千米—10 千米以内的天气现象。

3）沙尘暴：强风将地面尘沙吹起，使空气很混浊，水平能见度小于 1 千米的天气现象。

4）强沙尘暴：大风将地面尘沙吹起，使空气非常混浊，水平能见度小于 500 米的天气现象。

5）特强沙尘暴：狂风将地面尘沙吹起，使空气特别混浊，水平能见度小于 50 米的天气现象。

根据上述标准，将沙尘暴的警度设置如下：

表 6-8 沙尘暴警度设置

Table 6-8 The alarm degrees of dust storm warning

警度	一般警情	轻度警情	中度警情	高度警情
条件	浮尘	扬沙	沙尘暴	强沙尘暴及以上
预警颜色	蓝色	黄色	橙色	红色

（3）台风警度设置

根据我国气象局“关于实施热带气旋等级国家标准”GBT 19201-2006 的相关规定，我国台风按照热带气旋中心附近地面最大风速进行设置，划分为以下六个等级：

热带低压：底层中心附近最大平均风速 10.8-17.1 米/秒，也即风力为 6-7 级。

热带风暴：底层中心附近最大平均风速 17.2-24.4 米/秒，也即风力 8-9 级。

强热带风暴：底层中心附近最大平均风速 24.5-32.6 米/秒，也即风力 10-11 级。

台风：底层中心附近最大平均风速 32.7-41.4 米/秒，也即 12-13 级。

强台风：底层中心附近最大平均风速 41.5-50.9 米/秒，也即 14-15 级。

超强台风：底层中心附近最大平均风速高于 51.0 米/秒，也即 16 级或以上。

同时也规定我国的台风警度划分为蓝色预警、黄色预警、橙色预警、红色预警四级。各预警的相关含义如下；

蓝色预警信号表示 24 小时内可能或者已经受热带气旋影响,沿海或者陆地平均风力达 6 级以上，或者阵风 8 级以上并可能持续。

黄色预警信号表示 24 小时内可能或者已经受热带气旋影响,沿海或者陆地平均风力达 8 级以上，或者阵风 10 级以上并可能持续。

橙色预警信号表示 12 小时内可能或者已经受热带气旋影响,沿海或者陆地平均风力达 10 级以上，或者阵风 12 级以上并可能持续。

台风红色预警信号表示 6 小时内可能或者已经受热带气旋影响，沿海或者陆地平均风力达 12 级以上，或者阵风达 14 级以上并可能持续。

根据上述标准，将台风的警度划分如下：

表 6-9 台风暴警度划分

Table 6-9 The alarm degrees of typhoon warning

警度	一般警情	轻度警情	中度警情	高度警情
风力	24 小时内 6 级以上	24 小时以内 8 级以上	12 小时内 10 级以上	6 小时内 12 级以上
预警颜色	蓝色	黄色	橙色	红色

6.4 实例分析

实例中选取的数据是来自于广东省江门市 1999 年至 2009 年的数据，利用前面介绍过的基于知识挖掘技术智能协同预测结果对江门市的电力市场进行预警的监测分析。在分析中，以 1999 年到 2008 年的数据为测试数据集，以 2009 年的数据为验证数据集来进行验证。

6.4.1 短期负荷监测的实例分析

按照上文中的时点负荷偏离度的计算方法，通过 1999 年到 2008 年的数据对 2009 年日负荷曲线数据进行预测，并且以实际数据作为模拟，共计点数为 35040。

表 6-10 江门市 2009 年的日负荷监测结果图

Table 6-10. The monitor results of Jiangmen City's daily load in 2009

1/1	2/1	3/1	4/1	5/1	6/1	7/1	8/1	9/1	10/1	11/1	12/1
1/2	2/2	3/2	4/2	5/2	6/2	7/2	8/2	9/2	10/2	11/2	12/2
1/3	2/3	3/3	4/3	5/3	6/3	7/3	8/3	9/3	10/3	11/3	12/3
1/4	2/4	3/4	4/4	5/4	6/4	7/4	8/4	9/4	10/4	11/4	12/4
1/5	2/5	3/5	4/5	5/5	6/5	7/5	8/5	9/5	10/5	11/5	12/5
1/6	2/6	3/6	4/6	5/6	6/6	7/6	8/6	9/6	10/6	11/6	12/6
1/7	2/7	3/7	4/7	5/7	6/7	7/7	8/7	9/7	10/7	11/7	12/7
1/8	2/8	3/8	4/8	5/8	6/8	7/8	8/8	9/8	10/8	11/8	12/8
1/9	2/9	3/9	4/9	5/9	6/9	7/9	8/9	9/9	10/9	11/9	12/9
1/10	2/10	3/10	4/10	5/10	6/10	7/10	8/10	9/10	10/10	11/10	12/10
1/11	2/11	3/11	4/11	5/11	6/11	7/11	8/11	9/11	10/11	11/11	12/11
1/12	2/12	3/12	4/12	5/12	6/12	7/12	8/12	9/12	10/12	11/12	12/12
1/13	2/13	3/13	4/13	5/13	6/13	7/13	8/13	9/13	10/13	11/13	12/13
1/14	2/14	3/14	4/14	5/14	6/14	7/14	8/14	9/14	10/14	11/14	12/14
1/15	2/15	3/15	4/15	5/15	6/15	7/15	8/15	9/15	10/15	11/15	12/15
1/16	2/16	3/16	4/16	5/16	6/16	7/16	8/16	9/16	10/16	11/16	12/16
1/17	2/17	3/17	4/17	5/17	6/17	7/17	8/17	9/17	10/17	11/17	12/17
1/18	2/18	3/18	4/18	5/18	6/18	7/18	8/18	9/18	10/18	11/18	12/18
1/19	2/19	3/19	4/19	5/19	6/19	7/19	8/19	9/19	10/19	11/19	12/19
1/20	2/20	3/20	4/20	5/20	6/20	7/20	8/20	9/20	10/20	11/20	12/20
1/21	2/21	3/21	4/21	5/21	6/21	7/21	8/21	9/21	10/21	11/21	12/21
1/22	2/22	3/22	4/22	5/22	6/22	7/22	8/22	9/22	10/22	11/22	12/22
1/23	2/23	3/23	4/23	5/23	6/23	7/23	8/23	9/23	10/23	11/23	12/23
1/24	2/24	3/24	4/24	5/24	6/24	7/24	8/24	9/24	10/24	11/24	12/24
1/25	2/25	3/25	4/25	5/25	6/25	7/25	8/25	9/25	10/25	11/25	12/25
1/26	2/26	3/26	4/26	5/26	6/26	7/26	8/26	9/26	10/26	11/26	12/26
1/27	2/27	3/27	4/27	5/27	6/27	7/27	8/27	9/27	10/27	11/27	12/27
1/28	2/28	3/28	4/28	5/28	6/28	7/28	8/28	9/28	10/28	11/28	12/28
1/29		3/29	4/29	5/29	6/29	7/29	8/29	9/29	10/29	11/29	12/29
1/30		3/30	4/30	5/30	6/30	7/30	8/30	9/30	10/30	11/30	12/30
1/31		3/31		5/31		7/31	8/31		10/31		12/31

从结果表中可以看到，红色的高度警情大多出现在5-7月份的暑期时段，造成这种现象的的原因一是由于受到2008年经济危机影响使得江门市的出口相关工业部分停产、倒闭，当系统预测时利用了2008年同期的较高数据引起的同月份的预测值加大，二是由于夏季相应的气候炎热，引起空调负荷的增加，致使预测值在2-5点的时刻在某些点预测过低造成的。

在12月的中后期也出现了部分的中度警情，这同样也是两个原因，一个是由于2008年的同期数值较低，引起的负荷误差加大，另一个是出现低温现象使得利用电气设备加热造成的负荷增加造成的。从实际模拟监测中可以看出，时点负荷偏离度具有一定的预警作用。

6.4.2 电力供需预警监测的实例分析

（1）全社会电力供需比的实证分析

利用第4章中介绍过的中长期负荷预测方法，对江门市的全社会历年的年度供需比进行测算，得到结果如下：

表6-11 江门市2003-2009年的年度供需比

Table 6-11. the annual supply and demand ratio of Jiangmen city in 2003-2009

年度	供需比	警度
2003	1.403	红色
2004	1.357	红色
2005	0.979	蓝色
2006	0.983	蓝色
2007	0.983	蓝色
2008	0.986	蓝色
2009	1.080	橙色

从供需比中可以看出，2003和2004年均为红色预警状态，这是由于江门市的供电能力过剩所致，但从2005年到2008年均为蓝色预警状态，表示这几年处于供需平衡的状态，2009年出现橙色预警状态，造成这种状态的主要原因是因为在2008年底的时候受到美国金融危机的影响，使得2008年的供需电量出现了下降的趋势，使得预测模型在2009年预测的时候出现了一定的偏差，从而造成了供需比失衡，预警信号表明2009年将存在供给大于需求的情况，经实践证明，预警情况基本和实际相符。

（2）电力需求增长率的实证分析

利用上文中的电力需求增长率的计算方法和警度设置法对江门市的电力需求率指标进行计算和预警，其结果如表6-12所示。从结果中可以看出，2009年

给出的是红色预警状态，这是由于受到 2008 年的金融危机影响，造成 2008 年的实际需求增长率较低，而预测值给出的 2009 年的需求预测值没有很好的参照到 2008 年的电力需求出现下降趋势的情况，造成的预测值偏高的结果。2009 年给出的预警状态虽然是由于需求增长率偏大造成的，但是仔细分析原因可以得出实际上是由于 2008 年实际需求过低所致，因此给出该预警状态也是具有一定意义的。

表 6-12 江门市 2003-2009 年的电力需求增长率和预警结果

Table 6-12. The electricity demand growth early warning results of Jiangmen city in 2003-2009

年度	电力需求增长率	预警结果
2003	0.151	
2004	-0.207	
2005	0.439	
2006	0.102	
2007	0.211	
2008	0.010	
2009	0.527	红色

（3）电力行业需求增长率

按照江门市对实际行业监测数据如附表 1 所示，其中各大类和各小类共计 98 项。按照上文中提出的处理步骤，可以去掉重复类大项、无用电量的行业以及累计总和低于 3%的行业项共 31 项，将剩下的 67 项按照上文中提到的警度划分，划分所得的表格如表 6-14 所示。利用 weka3.6 软件进行属性的筛选可得最后的结果为：x_{12}，x_{15}，x_{23}，x_{28}，x_{30}，x_{46}，x_{65}；分别是工业、采矿业、食品、饮料和烟草制造业（轻）；印刷业和记录媒介的复制（轻）；石油加工、炼焦及核燃料加工业；水的生产和供应业以及教育、文化、体育和娱乐业。其中工业在江门市的用电量中占的比重是最大的，采矿业、、食品、饮料和烟草制造业（轻）和印刷业和记录媒介的复制（轻）都的用电量比例能占到整个工业的前 10 名以内；因此对筛选结果进行监测具有一定的实际应用意义。

6.4.3 典型气候预警监测的实例分析

由于江门市地处广东省广东省中南部，珠江三角洲西侧，全市总面积 9541 平方公里、人口 410 万，江门市属河流三角洲冲积平原，沿海有大小岛屿 96 个，海岸线总长 328.7 公里。江门属亚热带海洋气候，少霜无雪，温和多雨，阳光充足。因此在江门市需要监测的气候类型主要是冰灾以及台风两种类型，其中台风的预警监测可以从气象预报中及时获取，而对于冰灾的气候条件利用知识挖掘中

的决策树分类技术得出的规则如下：

表 6-15 江门市冰灾的预警条件

Table 6-15 Jiangmen city's ice disaster alarm condition

形成条件	天气描述	持续时间	警情
气温在-4℃-0℃	有雾或雨	1 天	轻度警情
气温在-4℃-0℃	有雨	1 天-3 天	中度警情
气温在-4℃-0℃	有雨	3 天以上	高度警情

6.5 本章小结

本章结合前面的负荷预测结果对相应的电网预警进行了研究，基于短期日负荷曲线预测的结果给出相应的短期负荷监测指标和警度设置，基于中长期负荷预测的结果给出相应的供需平衡监测指标和警度的设置，并对可预知自然气候灾害下的电网气候预警监测进行了警度的设置和研究。在对短期预警监测中筛选监测行业时利用了知识挖掘中的决策树分类属性筛选技术，可以挑选出重要的监测对象。

第 7 章 江门市供电局知识挖掘智能协同负荷预测系统研究

随着科学技术和社会经济的快速发展，高精度的电力负荷预测技术和应用系统研究所发挥的作用越来越重要，它对于经济优化地制定发电计划、制定经济合理的电力调配计划、制定上网竞价计划、在竞价上网中取得优势、最优制定电力现货和期货报价、控制电网经济运营、降低旋转储备容量、进行电力市场需求分析、搞好电力市场营销和电力客户关系管理、搞好电网规划、避免重大事故、有效化解风险、保障生产和生活用电等方面具有重要意义，有利于提高电力系统的经济效益和社会效益。

而随着计算机技术的发展，用电负荷预测系统的应用越来越广泛，不但大大简化了负荷预测的技术和分析过程，而且有效地提高了预测的精度。因此，将上述研究利用计算机技术进行系统化的实现可以进一步提高电力需求预测管理水平，增强相关理论技术的实用性。本章借助广东省江门市供电局的相关课题研究，建立起适合江门供电局的基于知识挖掘的智能协同负荷预测系统。

7.1 系统需求分析

通过对江门市供电局进行实际调研，总结其对系统的需求如下：

（1）用电市场分析及预警

1）分行业用电量分析及预警

需求：需要对分行业用电量进行分析，其细度划分至最低等级。其中最细等级可以通过用电量的排名进行筛选（需要采用一定的算法对用电量的指标进行筛选进行分析），其中可以定制必须分析和预警的用电量行业（例如：工业、或者是通过制定最低用电量值，例如超过 1000 万千瓦时以上的用电行业），挑选出相关行业进行分析。分行业用电量分析和预警时间最小单位为月，在此基础上完成年度的分行业用电量分析及预警工作。

其中用电分析经调研后其采取的指标包含以下几类：

①全社会用电量及其增长率；

②用电量景气度；

③年度最大全社会用电量；

④年度行业最大用电量（行业一级指标，如：一、二、三产业以及八大类）；

⑤各行业用电量及其增长率（行业一级指标，如：一、二、三产业以及八大类）；

⑥各行业用电量景气度（行业一级指标，如：一、二、三产业以及八大类）；

⑦各行业年度用电量所占比例（行业一级指标，如：一、二、三产业以及八大类）；

⑧各行业年度增长贡献率（行业一级指标，如：一、二、三产业以及八大类）；

⑨行业一级指标下的相关行业用电量及其增长率分析（一级指标下属类别，可选择排名或制定值进行分析）。

2）供电量分析及预警监测

需求：需要对供电量进行分析及预警，其时间最小单位为月，在此基础上完成年度的供电量分析。供电量分析部分选取的指标需要包含有月供电量及其增长率和年度最大供电量及其年最大增长率两项。

3）售电量分析

需求：需要对售电量进行分类别分析，其类别包含前六类（其原因是后六类业务在江门市没有用电需求，数值为零），分别是大工业用电、非工业普通工业用电、农业用电、商业用电、住宅用电、稻田排灌和脱粒用电六项。售电量分析的时间最小单位为月，在此基础上完成年度的分类别售电量分析。其采取的指标如下：

①各行业月售电量及其增长率（六类）；

②各行业年度售电量及其增长率（六类）；

③各行业年度售电量所占比例（六类）；

④各行业年度售电量增长贡献率（六类）。

4）负荷特性分析

需求：要求对负荷特性进行分析。其时间最小单位为日，在此基础上完成月度和年度的供电量分析。其采取的指标如下：

①日最大峰谷差及最大峰谷差率；

②月度、年度最大负荷；

③月度、年度不均衡系数；

（2）预测需求

1）中长期用电量预测

需求：对用电量分行业类别进行预测，其细度划分至第一等级（即：行业一级指标，如：一、二、三产业以及八大类）进行预测，其时间单位为月度和年度。其中需要考虑国民经济等因素。

2）中长期供电量预测

需求：对供电量进行预测，其时间单位为月度和年度。其中需要考虑国民经济等因素。

3）中长期售电量预测

需求：对售电量分类别进行预测，其类别包含前六类（理由同售电量分析）。其时间单位为月度和年度。其中需要考虑经济因素。

4）中长期以及短期负荷预测

需求：中长期预测包括月度和年度最大负荷预测，短期负荷预测值得是对每天按照15分钟的间隔对日进行96点负荷曲线预测，其中在短期负荷预测中需要考虑气象因素带来的影响。

（3）对典型的天气灾害进行预警的功能

需求：需要结合江门市的天气预报对影响电力的天气灾害具有一定的预警功能，包括台风、冰灾等天气灾害。

7.2 基于知识挖掘技术的智能协同电力负荷预测系统设计

根据对江门市供电局的系统需求分析，结合上文中提出的基于知识挖掘技术的智能协同电力负荷预测方法，给出基于知识挖掘技术的智能协同电力负荷预测系统的相关设计如下。

7.2.1 系统目标

运用用电分析的相关方法对广东省江门市供电局电力市场的相关指标进行分析，利用可以考虑定性因素关系的基于知识挖掘技术的智能协同电力负荷预测方法对江门市进行中长期、短期以及日曲线负荷进行预测，在预测和指标分析的基础上对江门市电力调度、规划、计划、用电等领域的相关指标进行分析，从而辅助决策。最终目的是有效地提高电力系统的经济效益和社会效益，减少损失，提高现代化管理水平。

7.2.2 系统架构设计

根据江门市供电局营销部以及其他各部门提出的需要建立一个以营销部局域网为环境，利用个人小型计算机和数据库服务器直接相连进行网络通信工作，其中仅服务器可以和江门市其余部门的相关系统（如SCADA系统）进行接口数据的传输的实际要求，系统架构设计为基于Client/Service（以下简称C/S）两层技术且需要客户端进行安装的系统架构。这种架构可以充分满足上述要求，并且简单易行，并且在局域网内运行，较B/S的网络访问速度快，安全性有保证。其

网络结构和系统逻辑结构图分别如图 7-1 和图 7-2 所示.

其中分布式系统中的各部分功能如下：

（1）客户端

客户端程序主要处理数据的展现、收集以及对数据的校验、对界面显示的控制等，直接与数据库服务器连接并且包含业务逻辑，所以需要安装客户端程序，但不需要安装数据库的客户端程序和数据库驱动程序，因此可以使客户端程序变得更小，更快且易安装。客户端程序属于 Win32 程序。

客户端程序可能直接通过 ADO 技术调用客户端中的数据模块层来实现与数据库服务器的交互，客户端程序中包含业务程序，客户端程序需要在 windows 所属的操作系统下运行。并需要保证在局域网中与数据库连通的网络环境。

（2）数据库

提供统一的中央数据库管理，需要与各客户端之间有连通的网络环境。

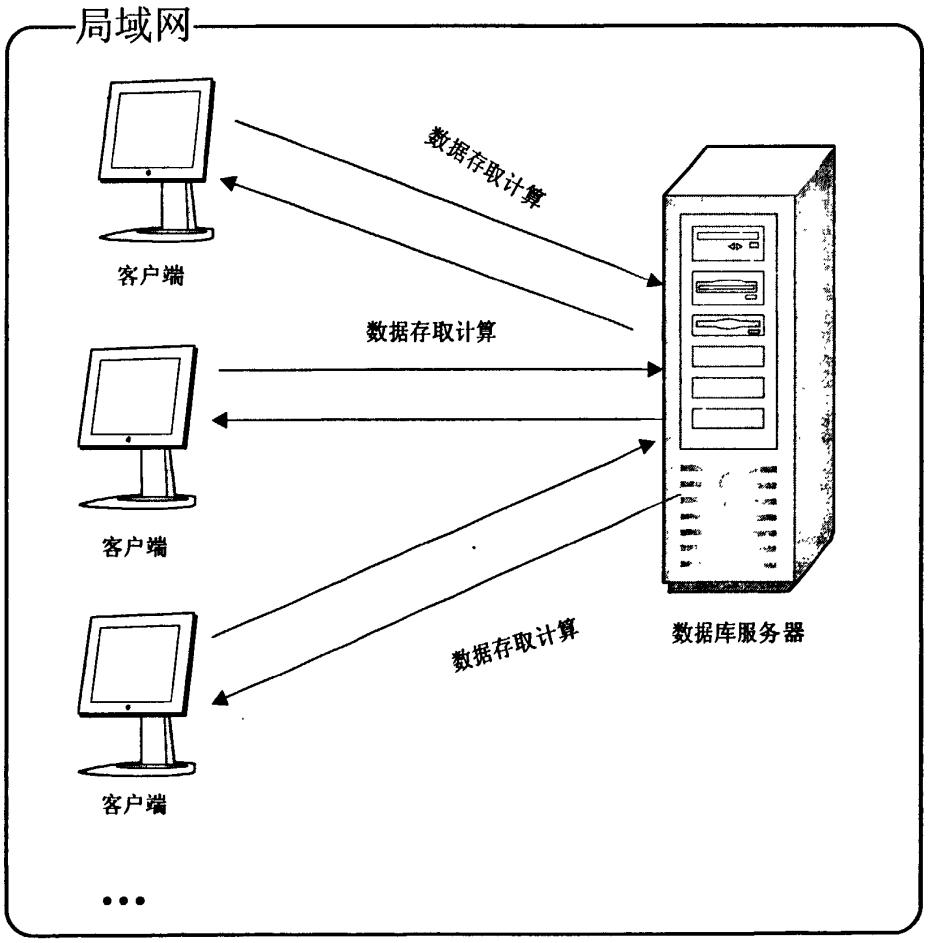


图 7-1 基于 C/S 的系统网络结构图

Fig. 7-1. The network structure of C/S system

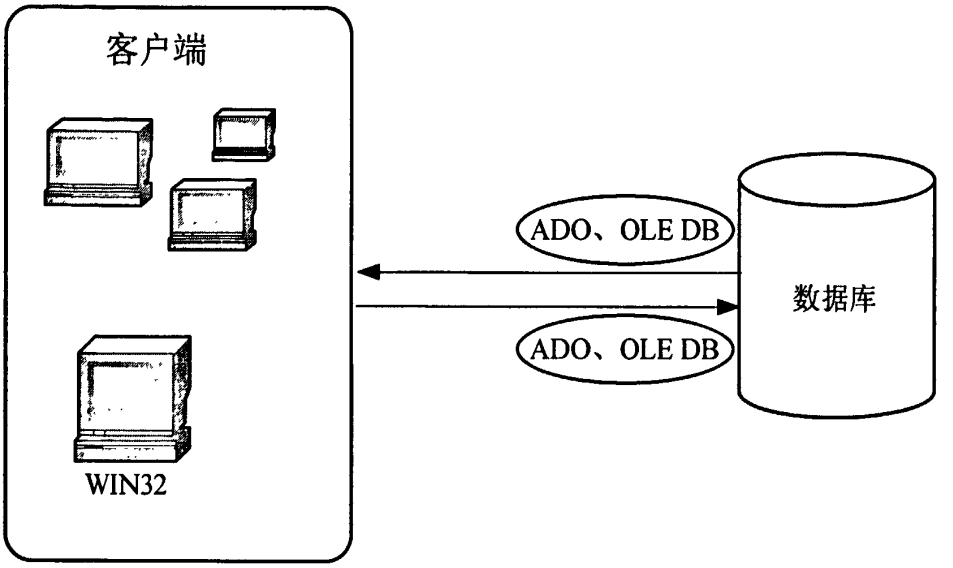


图 7-2 基于 C/S 的两层分布式系统逻辑结构图

Fig. 7-2. The two-tier distributed logical structure of system

7.2.3 系统运行环境

（1）网络与硬件环境

本系统是属于客户机/服务器（Client/Server）模式的具有两层结构的网络软件，本系统具有广阔的适用范围，可以在局域网网上安装使用，实现局域网内数据录入与查询；甚至在没有网络环境的单台计算机上也可方便地安装使用。若在局域网上安装使用本系统，为提高数据传输速度和网络安全性，建议构建 VPN 虚拟局域网。

后台数据库服务器的基本硬件要求：内存 256MB、CPU 为 Intel Pentium 800MHz 或以上，建议使用内存在 1GB 以上的高档次微机或专门服务器作为后台数据库服务器。

前台客户机的基本硬件要求：内存 128MB、CPU 为 Intel Pentium 500MHz 或以上，建议使用内存在 256 MB 以上、CPU 为 Intel Pentium 800MHz 以上的较高档次微机作为前台客户机。

需要注意的是：在没有网络环境的条件下，单用户使用的微机也可安装使用本系统，但要求微机具有一定档次，如内存在 128MB 以上、普通 586 档次 CPU。建议使用内存在 256MB 以上、Intel Pentium 800MHz 或以上的 IBM 系列微机。

（2）后台数据库服务器端软件环境

后台数据库服务器要求安装 Windows 2003 操作系统及其相应的补丁程序，

安装 SQL Server 2000 或更高版本的数据库服务器软件及其相应的补丁程序。

（3）前台客户端软件环境

前台客户机需要安装 Windows XP 或安装 Windows2000/2003 专业版或服务器版等操作系统，前台客户机还需要安装中文 Office 97/2000/XP/2003 中的 Word、Excel 等软件，作为文字编辑、报表输出的支持软件，有时还需要安装图片编辑等软件。

7.3 系统数据库设计

根据前文中的需求分析可知，由于中长期负荷预测需要结合国民经济的相关数据，因此对于国民经济相关情况的存储表格规范设计如表 7-1 所示，其中数据的获得可以从历年的统计年报和江门市相关的统计月报文本中直接进行摘取。

表 7-1 国民经济情况表(t_jj)

Table 7-1 National economy table

字段名	字段类型	允许空值	字段意义
theYear	int	不允许,	--年度
theMonth	int,	允许	--月度
GDP	decimal(12,3),	允许	--GDP 值
BelongToCom	varchar(20),	允许	--所属地区名称

短期负荷预测以及对于气象监测预警都需要利用气象信息中的相关信息，因此在数据库的基本信息存储表格中需要建立一个专门存储气象情况基本信息的表格，其中气象信息的来源来源于对江门市气象预报网站文本的摘取，在摘取时，首先将网站上的网页自动获取成文本格式，然后根据数据库中的关键词定位后进行数据的获取。其数据表格规范见表 7-2。

在需求中需要对行业用电量、供电量、售电量、中长期、日最大负荷以及日 96 点曲线负荷进行预测，因此需要建立存储这几类数据的基本表格，其中日最大负荷可以利用在对日 96 点的曲线负荷进行提取时得到，而中长期的最大负荷可以从日最大负荷中进行获取，因此中长期负荷可以利用视图技术进行解决，实际上需要建立的基本表格为行业用电量、供电量、售电量、日最大负荷以及日 96 点曲线负荷的基本信息表，其中日 96 点曲线负荷的基本信息表如前文中表 2-1 所示，其余的基本信息表设计规范如表 7-3 到表 7-6 所示：

表 7-2 气象基本信息表(t_qx)

Table 7-2 Meteorological information table

字段名	字段类型	允许空值	字段意义
theYear	int	不允许,	--年度
theMonth	int	不允许,	--月度
theDay	int	不允许,	--日期
ForecastDayWeather	varchar(200),	允许	--预计白天天气
MinTemperate	decimal(4,1),	允许	--最小温度（摄氏度）
MaxTemperate	decimal(4,1),	允许	--最大温度（摄氏度）
RealTemperate7	decimal(4,1),	允许	--7 点实际温度（摄氏度）
RealTemperate16	decimal(4,1),	允许	--16 点实际温度（摄氏度）
MaxHumidity	decimal(4,2),	允许	--相对最大湿度（百分比）
MinHumidity	decimal(4,2),	允许	--相对最小湿度（百分比）
UltravioletIncex	int,	允许	--最高紫外线指数
HumanComfortable	int,	允许	--人体舒适等级
HumanFeeling	varchar(10),	允许	--人体感觉
WholeRealLoad	decimal(10,4),	允许	--全负荷实测
MaxLoadTime	char(5),	允许	--最大负荷出现时间
WholeForecastLoad	decimal(10,4),	允许	--全负荷预测
ForecastRatio	decimal(6,4),	允许	--预测率
BelongToCom	varchar(20),	允许	--所属单位名称

表 7-3 售电量表(t_sdl)

Table 7-3 Table of electricity sales (t_sdl)

字段名	字段类型	允许空值	字段意义
theYear	int	不允许,	--年度
theMonth	int	不允许,	--月度
BigIndustry	decimal(10,3),	允许	--大工业用电量
NotGernalIndustry	decimal(10,3),	允许	--非普通工业用电量
Agriculture	decimal(10,3),	允许	--农业用电量
Business	decimal(10,4),	允许	--商业用电量
Houses	decimal(10,5),	允许	--住宅用电量
Irrigate	decimal(10,6),	允许	--稻田排灌用电量
BelongToCom	varchar(20),	允许	--所属单位名称

表 7-4 行业用电分类统计表(t_hyyd)

Table 7-4 Industry classification Tables (t_hyyd)

字段名	字段类型	允许空值	字段意义
theYear	int	不允许,	--年度
theMonth	int	不允许,	--月度
ClassifyId	decimal(8,5),	允许	--用电量项目编号
ClassifyName	varchar(100),	允许	--用电量项目名称
UsingLoad	int,	允许	--用电量
BelongToCom	varchar(20),	允许	--所属单位名称

表 7-5 供电量表(t_gdl)

Table 7-5 Power supply table (t_gdl)

字段名	字段类型	允许空值	字段意义
theYear	int	不允许,	--年度
theMonth	int	不允许,	--月度
SupplyLoad	int	不允许,	--供电量
BelongToCom	varchar(20),	允许	--所属单位名称

7.4 系统的主要功能

根据需求分析对本系统的主要功能结构主要划分成历史数据管理、历年用电情况分析、中长期负荷预测和预警四个部分，其功能结构图如下图 7-3 所示。

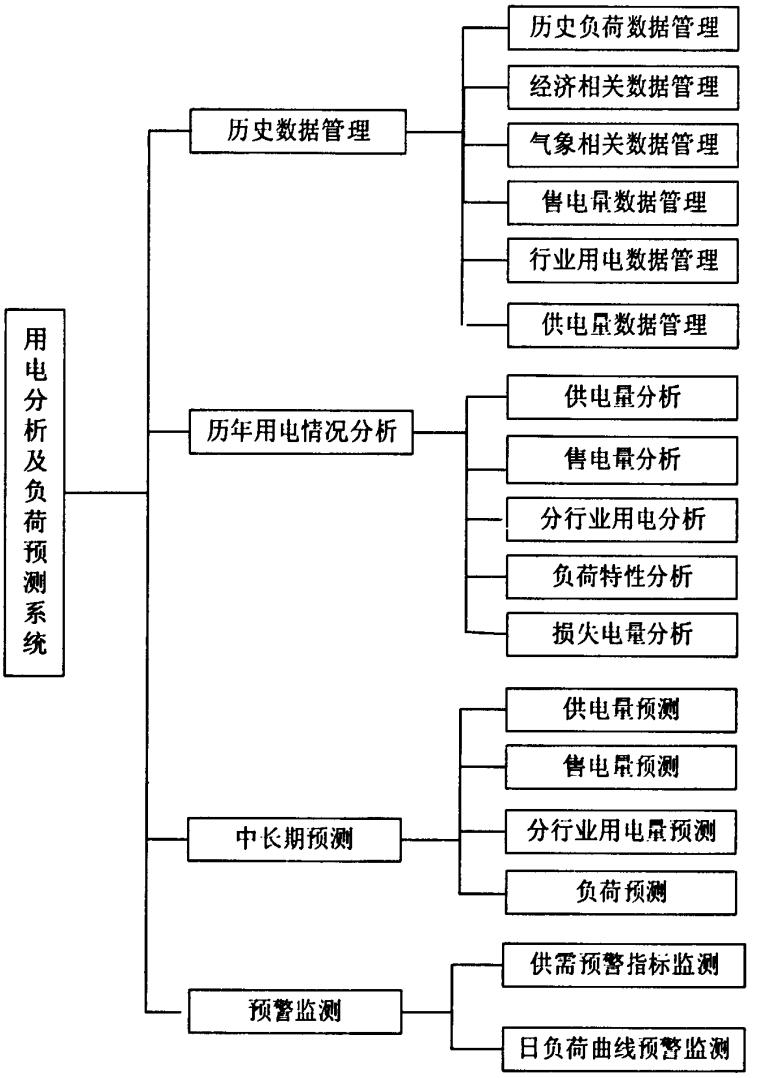


图 7-3. 系统功能结构框图

Fig. 7-3. System functions structure

7.4.1 历史数据管理

历史数据管理包括六个操作子功能：历史负荷数据管理、经济相关数据管理、气象相关数据管理、售电量数据管理、行业用电数据管理以及供电量数据管理子功能。以下以“售电量数据管理”为例对该方面的操作进行说明，其余六个子功能操作情况类似，故不再赘述。

点击目录树或者相应菜单中的“售电量数据管理”选项，可以打开图 7-4 所示的售电量数据管理窗口。该窗口的主要功能是对报装容量数据进行管理。其在主菜单下的操作界面如图 7-4 所示。该窗口的初始状态为浏览状态，即信息不可

编辑。窗口中提供了六个按钮对数据进行管理，分别是“导入”、“添加”、“删除”、“修改”、“保存”和“刷新”。下面依次对各按钮的功能予以说明。

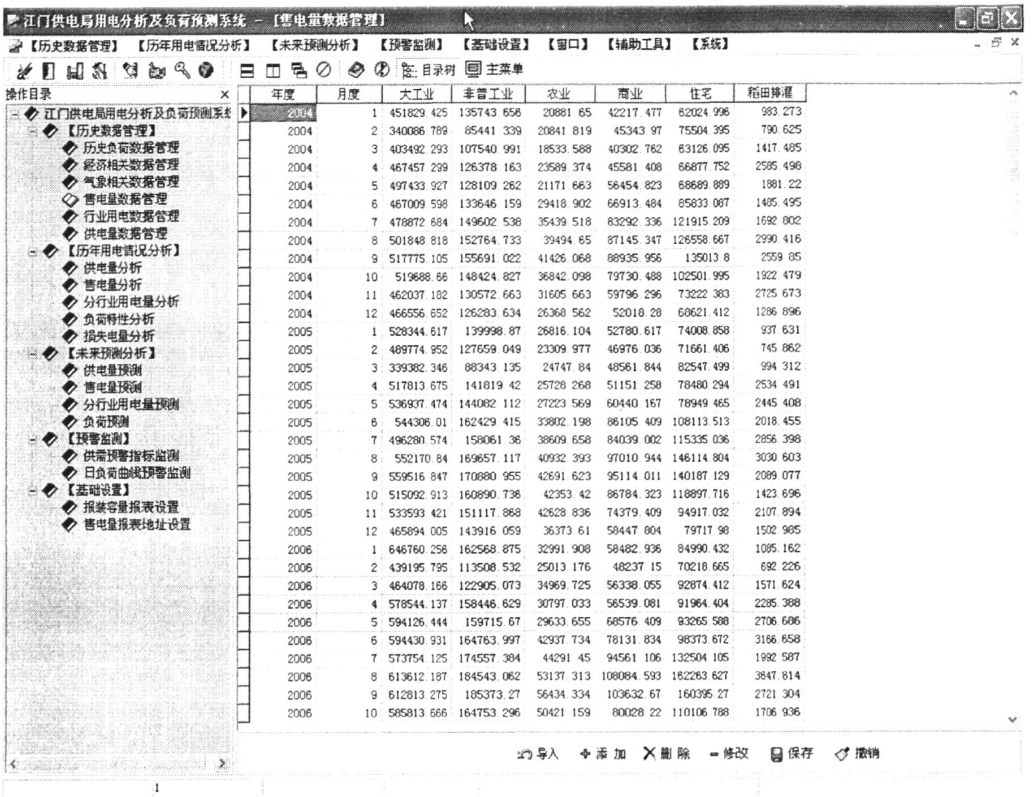


图 7-4 售电量数据管理界面

Fig. 7-4 Data Management of electricity sales

在需要对数据进行增加的操作时，可以选择“导入”或者是“添加”按钮进行操作，单击“导入”按钮后，将会弹出用户熟悉的 windows 的打开文件选择窗口，如图 7-5 所示，在该窗口中可以选择从和单位系统接口中自动形成的文本文件，点击确定后，系统将会自动导入相应的数据。

当接口自动生成的文本出现问题时，可以手动逐条录入，可以单击“添加”按钮，这时光标所在位置会出现一条空白行，如图 7-6 所示。用户可以按照相应的空白所对应的关系填入相关数据。在添加完成后需要点击“保存”按钮以确定输入完成。

点击删除按钮后，系统将会弹出如图“7-7”的提示信息窗口，单击“是”则删除用户所选记录，点击“否”将不做删除操作。需要注意的是点击删除按钮后，删除的记录不能恢复。

单击“修改”按钮后，用户可以对所选的记录进行修改校正。

单击“刷新”按钮后，可以将最新的数据库中的信息读取并显示，这可以避

免在其余客户端上输入信息但是由于网络延迟并没有在本机上显示的情况。

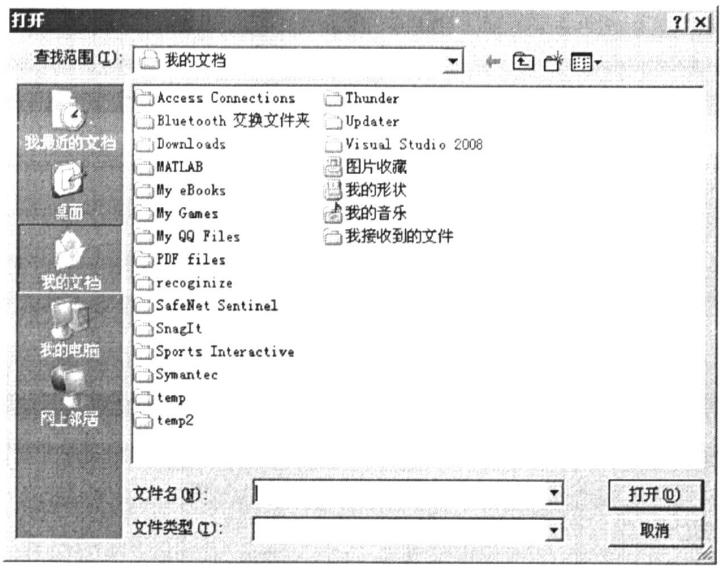


图 7-5 点击“导入”按钮后的文件选择窗口

Fig. 7-5 The file selection window after click the "Import" button

年度	月度	大工业	非普工业	农业	商业	住宅	稻田排灌
2007	1	772044.366	187230.339	39355.333	66695.074	98981.9	1102.824
2007	2	629986.33	159611.945	32550.435	56560.964	87508.343	967.601
2007	3	409856.81	103593.988	39807.064	61264.752	99829.506	1013.363
2007	4	628174.514	168966.067	40304.517	65583.836	98928.968	2355.511
2007	5	654786.114	171518.565	35939.944	65093.82	93439.626	2411.633
2007	6	662188.048	190087.705	49378.303	92304.373	114684.144	1966.996
2007	7	666012.13	206173.383	59548.829	107005.674	154942.733	2217.023
2007	8	667826.127	207866.016	63153.784	114673.368	182350.853	2891.367
2007	9	672287.711	204509.139	66057.701	111036.182	175987.059	2624.206
2007	10	662879.303	195668.167	59770.024	98221.504	138472.698	1879.064
2007	11	650135.896	190697.504	58384.914	91424.515	111513.248	2259.263
2007	12	570152.221	163220.884	40509.35	64687.69	95671.389	1934.462
2008	1	746578.467	195856.285	42436.747	70924.389	102675.497	1457.423
2008	2	611121.567	165555.835	37269.965	63349.342	97157.307	1012.508
2008	3	432636.153	116002.312	39096.64	68593.523	126216.606	1244.362
2008	4	692620.295	194835.593	43424.058	69480.474	117835.077	2310.35
2008	5	691765.112	194282.595	45287.369	80175.653	113672.914	2434.514
2008	6	713836.425	206037.418	47126.422	95600.598	113573.1	1634.731
2008	7	681713.492	206811.862	51323.028	101239.598	142170.655	3859.572
2008	8	714879.291	224807.01	52310.89	122245.129	158037.545	2554.977
2008	9	682188.714	219021.807	56712.753	122677.137	199737.658	2461.044
2008	10	611849.242	202445.281	55394.99	112184.037	165950.077	1984.169
2008	11	540432.5	183693.928	50183.647	94985.918	150876.398	1709.958
2008	12	402899.515	153598.845	31477.939	70157.292	106453.946	1302.207
2009	1	537324.075	199636.793	44522.163	77268.787	109470.337	1190.92
2009	2	592464.541	129736.527	36027.998	65681.111	112867.437	857.658
2009	3	491885.657	161756.211	50156.472	79855.355	140454.075	2111.615
2009	5	643866.702	179596.426	41445.494	78712.542	120245.421	1948.729
2009	6	646782.23	199581.117	50587.943	99102.61	122118.146	3035.088
2009	7	636935.115	215171.599	57543.507	111657.582	162126.036	2064.331
2009	8	692473.797	231036.388	64437.555	130582.614	187994.878	2334.518
2009	9	693266.247	239425.75	69900.288	135126.72	231830.835	2991.855
2009	10	704903.712	226615.155	65301.949	120555.19	210036.042	2202.74
*							

图 7-6 点击“添加”按钮后出现空白行

Fig. 7-6 Click the "Add"button then present the blank line

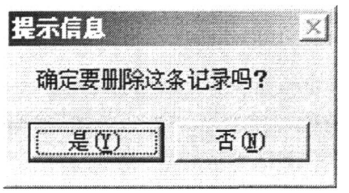


图 7-7 删除提示信息

Fig. 7-7 Delete messages

7.4.2 历年用电情况分析

该部分的功能包含供电量分析、售电量分析、分行业用电量分析以及负荷特性分析以及损失电量分析五个功能，其中前四个功能分别是对需求中提出的相关指标进行分析，最后一个损失电量分析是对已经发生过的日负荷曲线进行对比从而进行损失电量上的比对分析，其中相关的参照日是当选择分析日时，系统将自动按照第三章中介绍的相似日选取方法选取的除分析日外相似度最高的曲线，同时也可以进行人工选择。在选择后，可以选择计算区间参数选择来选择计算的起始点和截止点计算区间内的损失电量值。其计算结果显示在结果框中，供相关人员使用，其界面如图 7-8 所示。

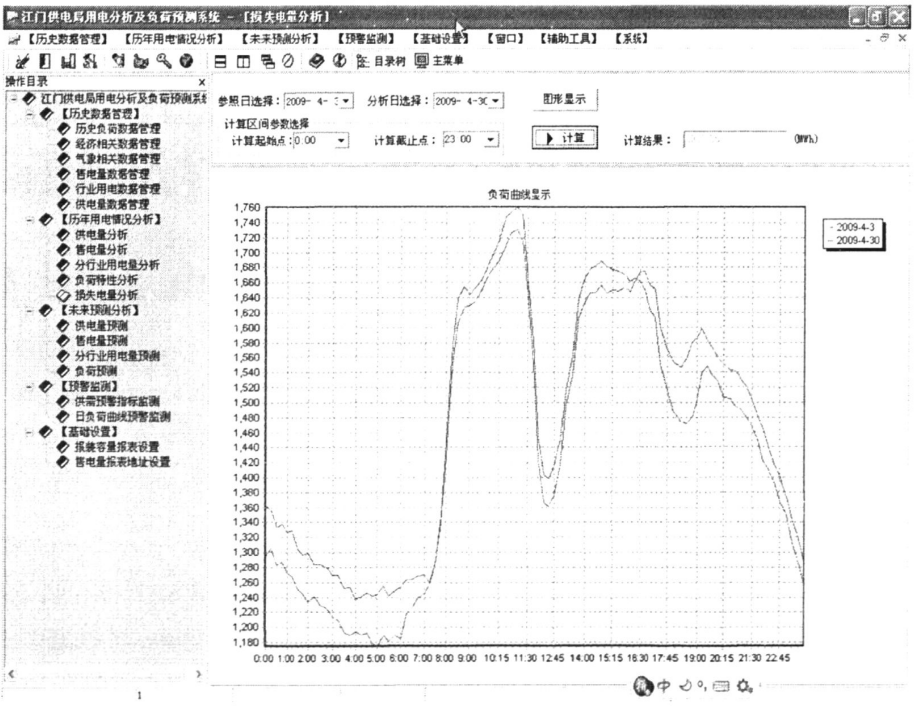


图 7-8. 损失电量计算界面

Fig. 7-8. The interface of power loss calculation

7.4.3 中长期负荷预测

中长期负荷预测部分是对需求中的年度和月度的行业用电量、供电量、售电量以及负荷情况进行相应的预测，在系统的预测模型库中既包含传统的一元线性回归、灰色预测分析、二次多项式回归、自适应指数预测、指数预测、增长率预测法、非齐次指数预测、B.Compertz 模型和 logistic 模型，同时也包括本文中提到的 BP 神经网络模型以及自适应参数的支持向量机模型和基于知识挖掘技术的文本后干预模型。

供电量预测在主菜单下的操作界面如图 7-9 所示。该窗口有两个选项卡，其中一个是初始值选取，在预测区间选择中选择预测区间类型，即属于“年”、“月”、还是“日”预测（其中只有负荷预测可以选择日预测代表日最大负荷预测，其余没有）。然后再预测时间选择区域选择相应的预测时间，点击“设定”按钮，系统将会在右侧给出相应时间的历史序列。

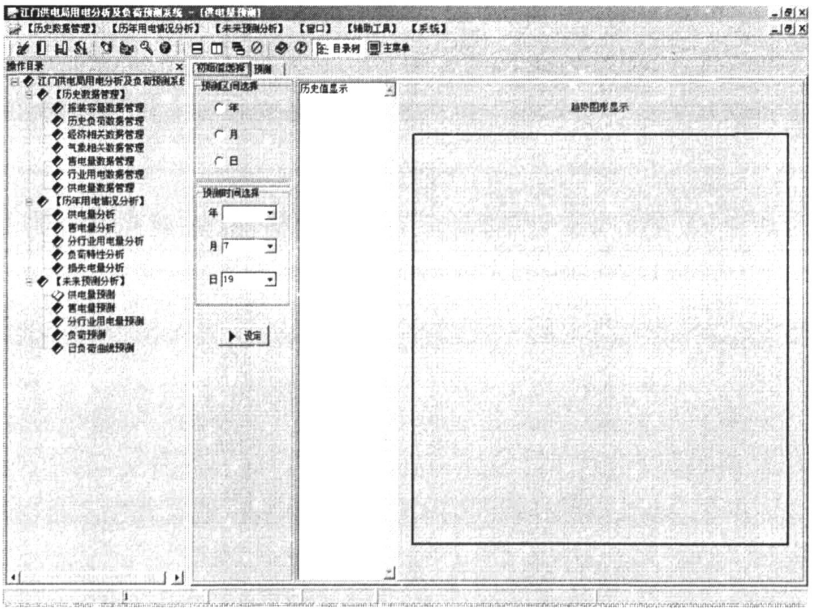


图 7-9. 供电量预测窗口

Fig. 7-9. power supply forecasting window

设定好数据后，切换至另一个“预测”选项卡，在左侧有预测方法选择区域，可以点击选取所要进行预测的预测方法（可多选）。点击预测后，在右侧将给出所选方法的预测结果、误差以及图形显示。如图 7-10 所示。

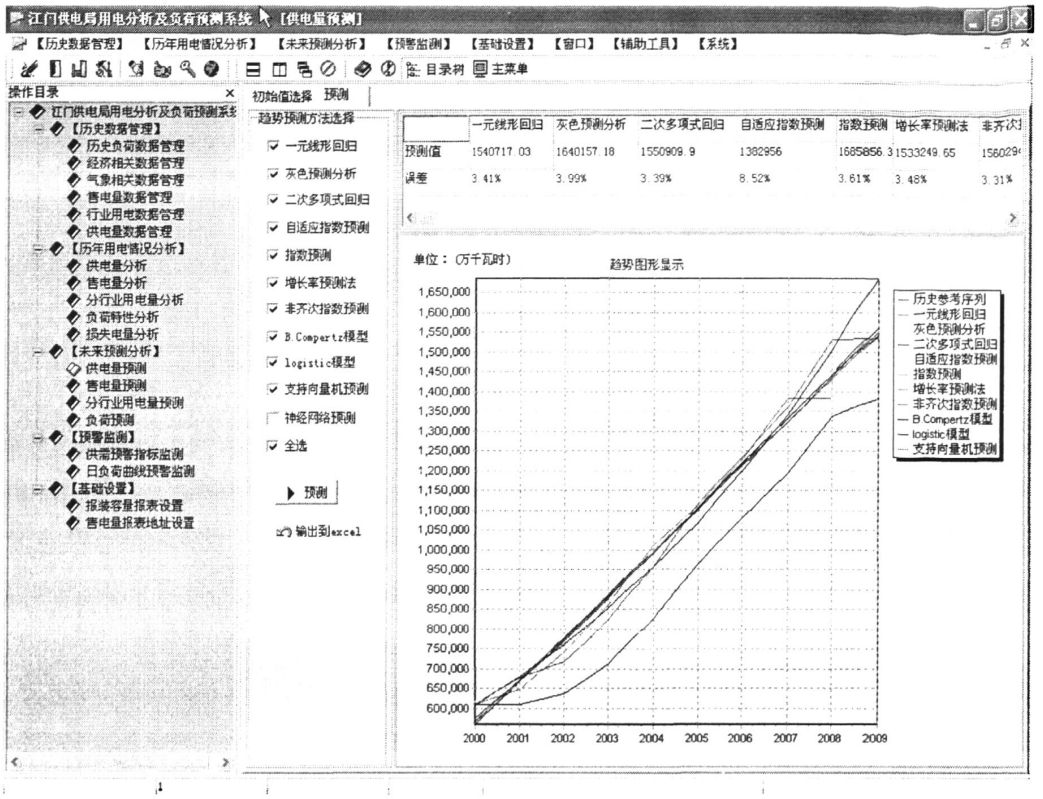
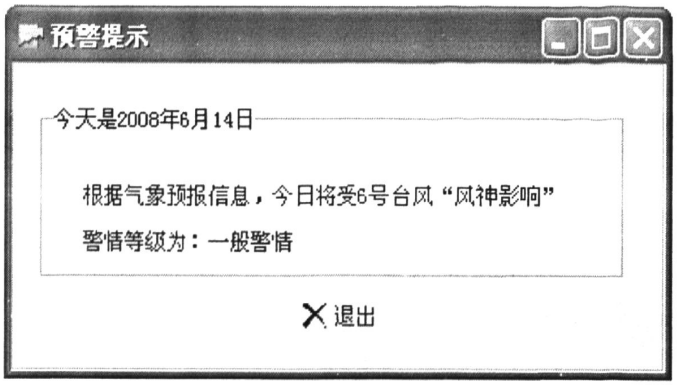


图 7-10 预测方法选择及结果显示窗口

Fig. 7-10 Prediction Methods and results display window

7.4.4 预警部分

预警部分主要包括供需预警指标监测和日曲线预警指标监测两个部分，其中由于典型气候预警的监测数据一般直接来源于气象信息数据表，因此在一进入系统的时候，如果根据天气预报信息具备预警条件，则在的首页面直接提示预警提示窗口。此外，如果在系统运行时遇到气象预报的突发性预警情况时，系统会弹出预警提示窗口进行自动预警，如图 7-11 所示的是江门市 2008 年 6 月 14 日的台风预警提示。



7-11. 台风预警提示窗口

Fig. 7-11. Prediction display window of typhoon

供需预警指标的监测根据江门市的实际情况选取了全社会电力供需比、电力需求增长率以及经过上章中描述的利用决策树分类技术的筛选行业电力需求增长率三个指标，图 7-12 所示的是全社会电力供需比指标的监测情况。

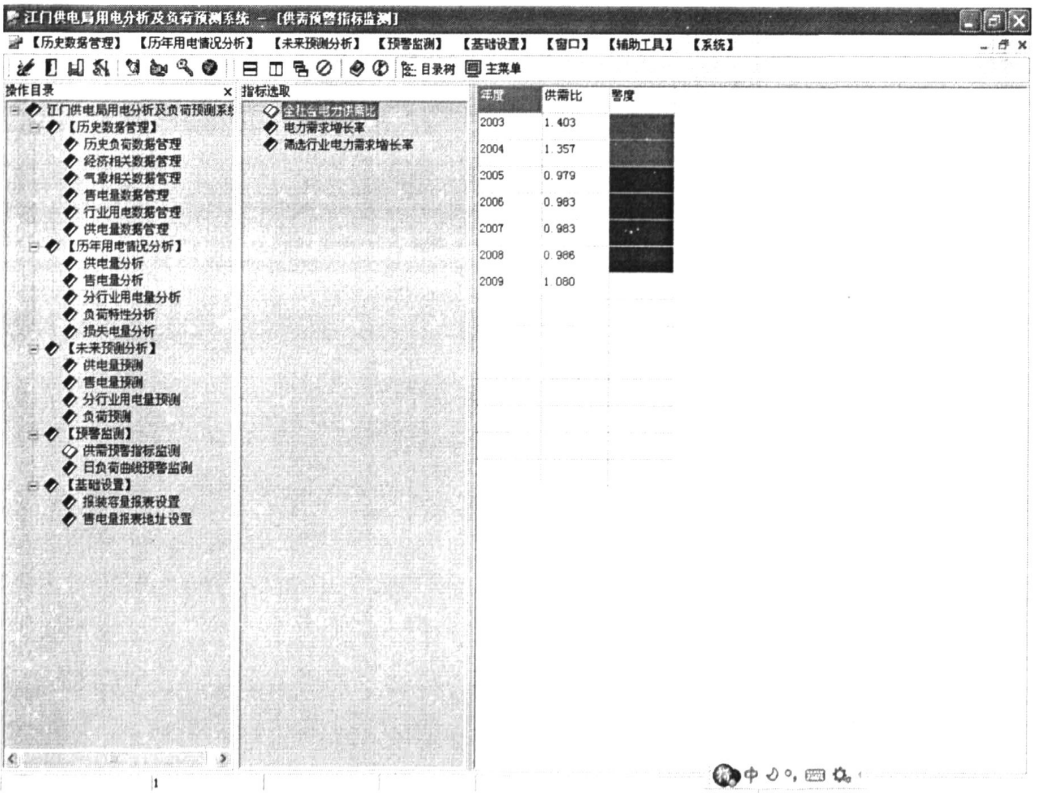


图 7-12. 供需预警指标检测界面

Fig. 7-12. Early warning window of supply and demand ratio indicators

图 7-13 给出的是日曲线预警监测界面，在这个界面中是按照每 15 分钟刷新一次的预测值和实际值的结果对比，在文本框中分别给出当前点的预测误差和检

测过的损失电量比率，如果点击界面上方的日期按钮将给出全天的日负荷曲线预测情况。

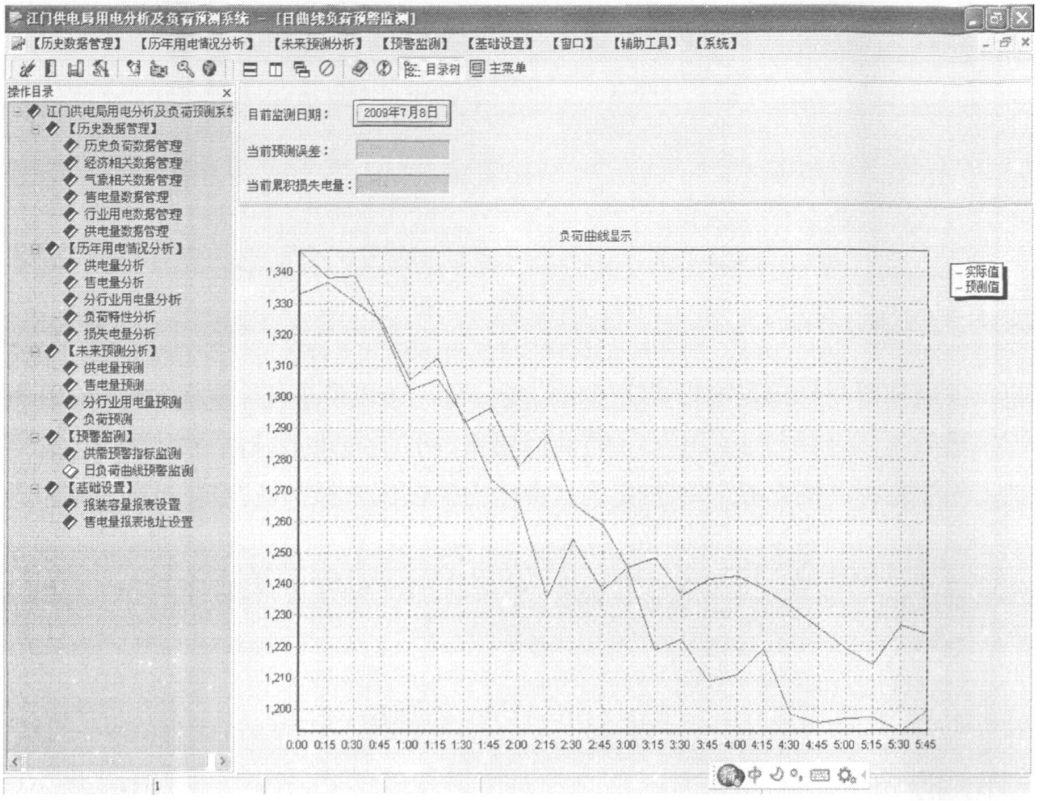


图 7-13. 日曲线负荷预警监测界面

Fig. 7-13.. Early warning window of the daily load curve

7.5 本章小结

根据广东省江门市的实际情况对上文中提到过的基于知识挖掘技术的智能协同电力负荷预测方法进行了计算机软件上的实现，结合预测方法和预警警度的设置完成了相应的预警监测功能。经过实际证明，本系统的应用有效提高了负荷预测的准确率，可以使得供电企业能够经济合理制定电力调配计划、做好电网规划和经营决策，做到合理投资建设，经济运行调度，减少公司财务费用，增加企业效益，能够为企业预先发现相应的警情。

第 8 章 结论与展望

8.1 结论及主要创新点

本文结合知识挖掘中的分类技术、聚类技术以及进化算法三类技术对智能预测技术进行了改进；首先通过分类和聚类技术建立起相应的规则来提取历史数据库中的相似数据，通过高相似性的历史数据训练神经网络模型能够提高神经网络模型在日曲线负荷预测中的预测精度，利用知识挖掘技术在提取相似性序列时可以直接对定性因素进行处理。在利用神经网络模型进行预测时，提出了一种自适应的网络结构优化方法，能够在预测时无需人为主观或经验上的干预开展预测工作；其次在针对利用支持向量机进行中长期预测时难以确定相应参数的问题时，引入了知识挖掘中的微分进化算法对相关的参数进行自适应的调整优化选择，经实例证明通过自适应调整优化参数后的支持向量机预测功能明显优于利用默认参数进行预测效果；此外，在对较容易受其它因素影响的日最大负荷预测中提出了一种结合传统时间序列预测方法、智能纠偏技术以及知识挖掘后干预技术的协同算法，该算法可以综合地考虑线性、非线性因素以及不规则事件对负荷的影响，能够进一步地提高日最大负荷预测的精度。在上述负荷预测的基础上开展了负荷监测预警的相关研究，通过设置相应的预警指标和警度对负荷进行监测并实施预警工作可以能够及时的发现相应的问题并及时的开展相应的应急调查和预案准备，最后利用计算机技术对上述研究根据广东省江门市的实际情况进行了系统上的实现。

本文的创新点主要有以下几点：

（1）建立了基于知识挖掘分类技术的自适应结构的日负荷曲线 BP 神经网络预测模型。在仅有负荷数据的情况下，首先通过计算负荷曲线的相似度对历史数据进行排序并进行初步的预测，然后再利用 BP 神经网络对误差进行纠偏工作来得到更加精确的预测精度；在具有较多气象数据可供分析时，首先利用聚类分析将日曲线负荷进行分类分析，然后利用知识挖掘中的分类技术寻求气象数据和曲线负荷分类之间的关系，形成相应的知识规则，在形成分类规则时，可以利用粗糙集的属性约简技术剔除掉冗余属性，最后利用不同类别的数据训练出不同的 BP 神经网络模型；这样在进行负荷预测工作时，根据预先判断的气象数据找出相应的类别选取相应的 BP 神经网络进行预测。在利用 BP 神经网络进行模型训练时，提出了一种简单的自适应 BP 神经网络对日负荷曲线进行预测，该自适应

网络可以自动确定隐含层节点的个数，无需人为经验的干预。

（2）提出了利用微分进化算法调整参数的支持向量机中长期负荷预测模型。对于中长期负荷预测，由于其样本数据远少于短期负荷预测，因此适用于小样本数据量的支持向量机智能预测方法，该方法可以有效地选取支持向量机所需求的相应参数，可以有效地提高中长期负荷预测的精度。

（3）提出了一种结合知识挖掘后干预纠偏技术、时间序列预测技术以及支持向量机预测方法的日最大负荷预测方法。由于日最大负荷预测不但需要考虑气象因素的影响，还需要考虑不同类型日期、不规则事件对其的影响，本文提出的方法不但可以考虑时间序列的趋势，而且可以考虑非线性因素的影响和不规则影响，经实验结果证明，该方法可以有效地提高负荷预测的精度。

（4）提出了基于负荷预测的预警监测指标并建立了基于知识挖掘的自然灾害预警方法。结合上文中的负荷预测方法和预测结果，对短期日负荷曲线进行偏离度的监测以及中长期供需平衡进行监测，对于气象灾害的预防，给出了冰灾、沙尘暴以及台风三种典型气候的监测方法。在对短期预警监测中筛选监测行业以及形成实际区域冰灾气象条件监测时利用了知识挖掘中的决策树分类属性筛选技术，可以挑选出重要的监测对象以及相应的形成条件。

（5）开发了基于知识挖掘智能协同负荷预测技术的用电分析及预测系统。该系统是基于上述知识挖掘和智能算法的研究成果，有别于基于传统算法或是其他改进算法开发出的负荷预测系统。

8.2 展望

伴随着知识挖掘以及负荷预测技术的进一步研究发展，结合知识挖掘技术和负荷预测的研究将会成为一个更加丰富多彩的研究领域，此外，新能源的介入同时也将会扩展负荷预测的应用领域，因此本文的研究仍有拓展的空间，对以下几个问题还有待于继续研究和探讨。

（1）将知识挖掘中的改进分类算法、改进聚类算法以及改进的进化式算法引入到负荷预测的改进模型或者是新算法中以取得更高的预测精度是一个将来所需要的研究方向。

（2）我国目前由于低碳经济理念的提出引起了很多的新能源以及分布式发电的应用，相比较传统的发电模式而言，这些能源的稳定性，可调节性较差，并且这些能源发电具有明显的不同于传统电力负荷预测的特征，如何建立这些方面的知识挖掘规范以及合适的预测模型也是以后的一个研究方向。

参考文献

[1]王志勇. 数据挖掘方法在短期负荷预测中的应用研究[博士论文]. 浙江：浙江大学，2007.
[2]牛东晓，曹树华，赵磊，张文文.电力负荷预测技术及其应用[M].北京：中国电力出版社，1998.
[3]程其云. 基于数据挖掘的电力短期负荷预测模型及方法的研究[博士论文]. 重庆：重庆大学，2004.
[4]赵宏伟，任震，黄雯莹. 基于周期自回归模型的短期负荷预测[J]. 中国电机工程学报，1997，17（5）：348-351.
[5]Kyung-Bin Song,Young-Sik Baek,Dug Hun Hong and Gilsoo Jang. Short-term load forecasting for the holidays using fuzzy liner regression method[J]. IEEE Transactions Power Systems，2005,20(1)：96-101.
[6]Tomonobu Senjuy, Paras Mandal, Katsumi Uezato,and Toshihisa Funabasi. Next day load curve forecasting using hybrid correction method[J]. IEEE Transactions Power Systems，2005，20（1）：102-109.
[7]王勇，黄国兴，彭道刚. 带反馈的多元线性回归法在电力负荷预测中的应用[J]. 计算机应用与软件，2008，1：82-84.
[8]吴曾，张庆丰. 基于稳健回归的电力负荷预测[J]. 电力科学与工程，2009，4（1）：25-27.
[9]唐杰明，刘俊勇，刘友波. 基于最优 FCM 聚类和最小二乘支持向量回归的短期电力负荷预测[J]. 现代电力，2008，25(2)：76-81.
[10]毛李帆，江岳春，龙瑞华，李妮，黄慧，黄珊. 基于偏最小二乘回归分析的中长期电力负荷预测[J]. 电网技术，2009，32(19):71-77.
[11]沈秀汶，吴耀武，熊信银. 基于有偏最小最大概率回归的短期负荷预测[J]. 电力系统及其自动化学报，2007，19（4）：46-49.
[12]李钜，李敏，刘涤尘. 基于改进回归法的电力负荷预测[J]. 电网技术，2004，28（17）：1-11.
[13]贺静，韦钢，熊玲玲. 负荷预测线性回归分析法的模糊改进[J]. 华东电力，2003，30（1）：99-104.
[14]赵宏伟，任震，黄雯莹. 基于周期自回归模型的短期负荷预测[J]. 中国电机工程学报，1997，17（5）：349-352.
[15] 李玲玲，朱博. 基于混沌时间序列的短期电力负荷预测[J]. 信息技术，2009，3：44-46.
[16] 张思远，何光宇，梅生伟，王伟，张王俊. 基于相似时间序列检索的超短期负荷预测[J]. 电网技术，2008，32（12）：56-59.
[17] 朱陶业，李应求，张颖，张学庄，何朝阳. 提高时间序列气象适应性的短期电力负荷预测算法[J]. 中国电机工程学报，2006，26（23）：14-19.
[18]张林，罗晓初，徐瑞林，赵理. 基于时间序列的电力负荷预测新算法研究[J]. 电网技术，2006，30（S2）：595-599.

[19] 雷绍兰，孙才新，周湶，邓群，刘凡．一种多变量时间序列的短期负荷预测方法研究[J]．电工技术学报，2005，20（4）：62-67.
[20]杨正瓴，张广涛，林孔元．时间序列法短期负荷预测准确度上限估计[J]．电力系统及其自动化学报，2004，16（2）：36-39.
[21]焦建林，芦晶晶．基于改进时间序列法的配电网短期负荷预测模型[J]．电工技术杂志，2002，5：25-28.
[22]王晛，张少华．一种应用时间序列技术的短期电力负荷预测模型[J]．上海大学学报(自然科学版)，2002，8（2）：133-136.
[23]王秋梅．时间序列法负荷预测的原理和应用[J]．华东电力，1993，4：37-39.
[24] J. Nowicka-Zagrajeka, R. Weron. Modeling electricity loads in California: ARMA models with hyperbolic noise[J]. Signal Processing, 2002, 82(12): 1903-1915.
[25]S. Sp. Pappas a, L. Ekonomou b,D. Ch. Karamousantas c, G. E. Chatzarakis b, S. K. Katsikas d, P. Liatsis. Electricity demand loads modeling using Auto Regressive Moving Average (ARMA) models[J]. Energy, 2008, 33(9): 1353-1360.
[26]Bonnie K. Ray. Long-range forecasting of IBM product revenues using a seasonal fractionally differenced ARMA model[J] . International Journal of Forecasting, 1993, 9 (2): 255-269.
[27]Jiann-Fuh Chen, Wei-Ming Wang, Chao-Ming Huang. Analysis of an adaptive time-series autoregressive moving-average (ARMA) model for short-term load forecasting[J]. Electric Power Systems Research,1995, 34 (3): 187-196.
[28]Celal Aksu, Jack Y. Narayan. Forecasting with vector ARMA and state space methods[J]. International Journal of Forecasting, 1991, 7 (1): 17-30.
Paolo Burlando, Renzo Rosso, Luis G. Cadavid, Jose D. Salas. Forecasting of short-term rainfall using ARMA models[J]. Journal of Hydrology,1993, 144 (1): 193-211.
[29]Kourosh Mohammadi, H. R, Eslami, Rene Kahawita. Parameter estimation of an ARMA model for river flow forecasting using goal programming[J]. Journal of Hydrology,2006, 331 (1-2): 293-299.
[30]Fong-Lin Chu . Forecasting tourism demand with ARMA-based methods[J]. Tourism Management,2009, 30 (5): 740-751.
[31]牛东晓，李春祥，孟明．基于灰色和偏最小二乘方法的年度负荷预测[J]．华东电力，2009，37(6)：899-902.
[32]卢建昌，韩红领．基于灰色神经网络组合模型的日最高负荷预测[J]．华东电力，2008，36(2)：60-63.
[33]高明，李芳竹，梁杰．基于灰色理论的中长期负荷预测[J]．吉林电力，2008，36(4)：22-24.
[34]徐剑．基于粗糙集和灰色系统模型的短期负荷预测方法研究[D]．北京：华北电力大学，2008.
[35]牛东晓，赵磊，张博，王海峰．粒子群优化灰色模型在负荷预测中的应用[J]．中国管理科学，2007，15(1)：69-73.
[36]李小燕．基于灰色理论的电力负荷预测[J]．华中科技大学，2007.

[37]牛东晓，张彤彤，陈立荣，张博．基于关联分析的多因素电力负荷预测灰色模型群研究[J]．华北电力大学学报，2006，33(3)：90-92.
[38]蔡琼，陈萍．灰色 GM(1,1)模型及其在电力负荷预测中的应用[J]．自动化技术与应用，2006，25(3)：24-26.
[39]陈毛昌，穆钢，孙羽，彭茂君．基于 DFT 灰色预测理论在日电量负荷预测中的应用[J]．电力自动化设备，2005，25(9)：29-32.
[40]张俊芳，吴伊昂，吴军基．基于灰色理论负荷预测的应用研究[J]．电力自动化设备，2004，24(5)：24-27.
[41]朱芸，乐秀璠．可变参数无偏灰色模型的中长期负荷预测[J]．电力自动化设备，2003，23(4)：25-27.
[42]赵君有．基于灰色理论的中长期电力负荷的预测[D]．辽宁：沈阳工业大学，2007.
[43]牛东晓，陈志业，邢棉，谢宏．具有二重趋势性的季节型电力负荷预测组合优化灰色神经网络模型[J]．中国电机工程学报，2002，22（1）：29-32.
[44]Chin-Tsai Lin, Shih-Yu Yang. Forecast of the output value of Taiwan's opto-electronics industry using the Grey forecasting model[J]. Technological Forecasting and Social Change, 2003, 70（2）: 177-186.
[45]Ruey-Chyn Tsaur. The development of an interval grey regression model for limited time series forecasting[J]. Expert Systems with Applications, 2010, 37(2):1200-1206.
[46]P. Zhou, B. W. Ang, K. L. Poh. A trigonometric grey prediction approach to forecasting electricity demand[J]. Energy, 2006, 31（14）: 2839-2847.
[47]Yen-Tseng Hsu, Ming-Chung Liu, Jerome Yeh, Hui-Fen Hung. Forecasting the turning time of stock market based on Markov–Fourier grey model[J]. Expert Systems with Applications, 2009, 36（4）: 8597-8603.
[48]Yong-Huang Lin, Pin-Chan Lee . Novel high-precision grey forecasting model[J]. Automation in Construction, 2007, 16（6）: 771-777.
[49]Yong-Huang Lin, Pin-Chan Lee, Ta-Peng Chang. Adaptive and high-precision grey forecasting model[J]. Expert Systems with Applications, 2009, 36（6）: 9658-9662
[50]A X Wu, Y Xi, B H Yang, X S Chen, H C Jiang. Study on grey forecasting model of copper extraction rate with bioleaching of primary sulfide ore[J]. Acta Metallurgica Sinica(English Letters), 2007, 20（2）: 117-128.
[51]Diyar Akay, Mehmet Atak. Grey prediction with rolling mechanism for electricity demand forecasting of Turkey[J]. Energy, 2007, 32（9）: 1670-1675.
[52]Hsu C C, Chen C Y. Applications of improved grey prediction model for power demand forecasting[J]. Energy Conversion and Management, 2003, 44, (14): 2241–2249.
[53]Albert, W L Yao, S C Chi, J H Chen. An improved Grey-based approach for electricity demand forecasting[J]. Electric Power Systems Research, 2003, 67（3）: 217-224.
[54]Che Chiang Hsu, Chia Yon Chen. Applications of improved grey prediction

model for power demand forecasting[J]. Energy Conversion and Management, 2003, 44（14）: 2241-2249.
[55]康重庆，夏清，张伯明. 电力系统负荷预测研究综述与发展方向的探讨[J]. 电力系统自动化，2004，28（17）: 1-11.
[56]张亚军，刘志刚，张大波. 一种基于多神经网络的组合负荷预测模型[J]. 电网技术，2006，30（21）: 21-25.
[57]李春祥，牛东晓，孟丽敏. 基于层次分析法和径向基函数神经网络的中长期负荷预测综合模型[J]. 电网技术，2009，33（2）: 99-104.
[58]倪 明，高晓萍，单渊达. 证据理论在中期负荷预测中的应用[J]. 中国电机工程学报，1997，17（3）: 199-203.
[59]陶文斌，张粒子，潘弘，李振元，郑华. 基于双层贝叶斯分类的空间负荷预测[J]. 中国电机工程学报，2007，27（7）: 13-17.
[60]Syed M. Islam, Saleh M. A1-Alawi, Khaled A. Ellithy. Forecasting monthly electric load and energy for a fast growing utility using an artificial neural network[J]. Electric Power Systems Research, 1995,34(1):213–217.
[61]C C Hsu, C Y Chen. Regional load forecasting in Taiwan-application of artificial neural networks[J], Energy Conversion and Management, 2003,44 (12):1941-1949.
[62]P. F. Pai, W. C. Hong, Forecasting regional electric load based on recurrent support vector machines with genetic algorithms[J]. Electric Power Systems Research, 2005,74(3):417-425.
[63]Wei Chiang Hong. Electric load forecasting by support vector model[J]. Applied Mathematical Modelling,2009,33(5):2444-2454.
[64]牛东晓，谷志红，邢棉，王会青. 基于数据挖掘的 SVM 短期负荷预测方法研究[J]. 中国电机工程学报，2006，26（18）: 6-12.
[65]李元诚，方廷健，于尔铿. 短期负荷预测的支持向量机方法研究[J]. 中国电机工程学报，2003，23（6）: 55-59.
[66]谢宏，魏江平，刘鹤立. 短期负荷预测中支持向量机模型的参数选取和优化方法[J]. 中国电机工程学报，2006，26（22）: 17-22.
[67]庞松岭，刘岱. 基于经验模态分解与人工鱼群神经网络的短期负荷预测[J]. 东北电力大学学报(自然科学版)，2008，6: 10-16.
[68]魏俊，周步祥，林楠，邢义. 基于蚁群支持向量机的短期负荷预测[J]. 电力系统保护与控制，2009，37(4): 36-40.
[69]杨占刚. 中短期负荷预测系统设计与实现[D]. 天津: 天津大学，2007.
[70]王毅. 电力系统短期负荷预测技术的研究与实现[D]. 北京: 华北电力大学，2008.
[71]王波，徐泽柱，高松波. 混沌粒子群优化算法在短期负荷预测中的应用[J]. 水电能源科学，2009，27(2): 208-211.
[72]吴京秋，孙奇，杨伟，杨杰. 基于 D-S 证据理论的短期负荷预测模型融合[J]. 电力自动化设备，2009，4: 66-71.
[73]孟祥斌. 基于小波分析的短期负荷预测模型研究与实现[D]. 大连理工大学，2009.
[74]王德意，杨卓，杨国清. 基于负荷混沌特性和最小二乘支持向量机的短期负

荷预测[J]. 电网技术，2008，32(7)：66-70.
[75]徐冬生. 超短期负荷预测系统研究[D]. 浙江大学，2007.
[76]王鹏，郜能灵，王波，翟海青，叶剑，李磊，朱家栋，漆梁波. 针对气象因素的短期负荷预测修正方法[J]. 电力系统自动化，2008，32(13)：92-96.
[77]李广敏. 考虑光伏并网发电的短期负荷预测[D]. 河北：华北电力大学，2007.
[78]刘宝英，杨仁刚. 基于主成分分析的最小二乘支持向量机短期负荷预测模型[J]. 电力自动化设备，2008，28(11)：13-17.
[79]吴杰康，陈明华，陈国通. 基于PSO的模糊神经网络短期负荷预测[J]. 电力系统及其自动化学报，2007，19(1)：63-67.
[80]冷喜武. 支持向量回归在短期负荷预测中的应用研究[D]. 河北：华北电力大学，2008.
[81]畅广辉，刘涤尘，熊浩. 基于多分辨率SVM回归估计的短期负荷预测[J]. 电力系统自动化，2007，31(9)：37-41.
[82]Radwan E, Abdel-aal. Short-term hourly load forecasting using abductive network[J]. IEEE Transactions Power Systems，2004，19（1）：164-173.
[83]T. Senjyu, H. Takara,K. Uezato,and T. Funabashi. One-hour-ahead load forecasting using Neural network[J]. IEEE Transactions Power Systems，2002，17（1）：113-118.
[84]Apostolos kotsialos, Markos Papageorgiou, Antonios poulimenos. Long-term sales forecasting using holt-winters and neural network methods[J]. Journal of Forecasting，2005，24(5)：353-368.
[85] Bo-June Chen, Ming-Wei Chang, and Chih-Jen Lin. Load forecasting using support vector machines: a study on EUNITE Competition 2001[J]. IEEE Transactions Power Systems，2004，19（4）：1821-1830.
[86] R. E. Abdel-Aal. Improving electric load forecasts using network committees[J]. Electric Power Systems Research. 2005，74（3）：546-547.
[87] B. L. Zhang, Z. Y. Dong. An adaptive neural-wavelet model for short term load forecasting[J]. Electric Power System Research，2001. 59（2）：121-129.
[88] Dong-Xao Niu, Qiang Wang, Jin-Chao Li. Short Term Load Forecasting Model Based on Support Vector Machine[M]. Machine Learning and Cybernetics，Springer-Verlag Berlin，2006.
[89] Dong-Xiao Niu, Qiang Wang and Jin-Chao Li. Short Term Load Forecasting Model Using Support Vector Machine Based on Artificial Neural Network[C]. Proceedings of 2005 International Conference on Machine Learning and Cybernetics，2005，7(1)：4260-4265.
[90] Dong-Xiao Niu, Hui-Qing Wang, Zhi-Hong Gu. Short-Term Load Forecasting Using General Regression Neural Network[C]. Proceedings of 2005 International Conference on Machine Learning and Cybernetics, 2005, 7(1)：4076-4082.
[91] Dong-Xiao Niu, Yuan-Yuan Li. Integrated Optimum Grey Neural Network Model of Monthly Power Load Forecast[C]. ASME Power Conference,2005,4：324-327.
[92] Zhao Lei, Niu Dong-xiao, Ren Feng. Research on Synthetic Evaluation of Sustainable Development for Electric Power Industry Based on Improved BP Neural

Network Model[J] . Intelligent Information Management System and Technologies,2005,1(2)：780-786.
[93] Dong-Xiao Niu, Bo Zhang, Mian Xing. Application of Neural Network Based on Particle Swarm Optimization in Short-Term Lord Forecasting[J]. Third International Symposium on Neural Networks. 2006,5：1089-1092.
[94]Mohammed El-Telbany, F. E. K. Short-term forecasting of Jordanian electricity demand using particle swarm optimization[J]. Electric Power Systems Research, 2008,78(3):425-433.
[95]康重庆，夏清，张伯明. 电力系统负荷预测研究综述与发展方向的探讨[J]. 电力系统自动化，2004，28(17)：1-11.
[96]Kun-Long Ho, Y Y H，Chlsn-chuen Yang. short term load forecasting using a multilayer neural network with an adaptive learning algorithm[J]. Transactions on Power Systems，1992，7(1):141-149.
[97]Zhiling Lin, D Z, Liqun Gao, Zhi Kong.Using an adaptive self-tuning approach to forecast power loads[J]. Neuro computing,2008,71(4-6):559-563.
[98]K Kalaitzakis, G S S, E M Anagnostakis. Short-term load forecasting based on artificial neural networks parallel implementation[J]. Electric Power Systems Research, 2002，63，3：185-196.
[99]Bahman Kermanshahi, H I. Up to year 2020 load forecasting using neural nets[J]. Electric Power and Energy Systems, 2002,24(9):789-797.
[100]Che Chiang Hsu, C Y C. Regional load forecasting in Taiwan-applications of artificial neural networks[J] . Energy Conversion and Management,2003,44(12):1941-1949.
[101]Ping Feng Pai, W C H. Support vector machines with simulated annealing algorithms in electricity load forecasting[J]. Energy Conversion and Management, 2005,46(17): 2669-2688.
[102]Ping Feng Pai, W C H. Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms[J]. Electric Power Systems Research, 2005,74(3): 417-425.
[103]M. S. Kandil, S M ED, N E Hasanien. The implementation of long-term forecasting strategies using a knowledge-based expert system: part-II[J]. Electric Power Systems Research, 2001,58(1): 19-25.
[104]Mohammed El Telbany, F E K. Short-term forecasting of Jordanian electricity demand using particle swarm optimization[J]. Electric Power Systems Research, 2008,78(3):425-433.
[105]刘敦楠,何光宇,范旻,孙英云,陈雪青，周双喜.数据挖掘与非正常日的负荷预测[J].电力系统自动化，2004,28（3）：53-57.
[106]朱六璋，袁林，黄太贵.短期负荷预测的实用数据挖掘模型[J].电力系统自动化，2004, 28（3）：49-52.
[107]朱六璋.短期负荷预测的组合数据挖掘算法[J].电力系统自动化，2006,30（14）：82-86
[108]Wi Young-Min, Song Kyung-Bin, Joo Sung-Kwan. Data mining technique using the coefficient of determination in holiday load forecasting[J].Transactions of

the Korean Institute of Electrical Engineers, 2009,58(1)：18-22
[109]Mori Hiroyuki1, Kosemura, Noriyuki1, Kondo, Tom, Numa Kazuyuki1. Data mining for short-term load forecasting[C].Proceedings of the IEEE Power Engineering Society Transmission and Distribution Conference,2002：139-143.
[110]Mori Hiroyuki, Kosemura,Noriyuki. A data mining method for short-term load forecasting in power systems[J]. Electrical Engineering in Japan,2002,139(2)：12-22.
[111] 李秋丹，迟忠先，王大公.基于数据挖掘技术的负荷预测模型[J].大连理工大学学报,2003,43(6)：485-460
[112] 黎静华，栗然，牛东晓.基于粗糙集的默认规则挖掘算法在电力系统短期负荷预测中的应用[J].电网技术,2006,30(5)：18-23
[113]熊浩,李卫国,黄彦浩,张海峰,畅广辉.基于模糊粗糙集理论的综合数据挖掘方法在空间负荷预测中的应用[J].电网技术,2007,31(14): 36-40.
[114]赵磊,李媛媛,李金超,赵晓坤.基于数据挖掘技术的电力日负荷优选组合预测[J].华北电力大学学报,2005,32(3):19-22.
[115] 李邦云,丁晓群,程莉.基于数据挖掘的负荷预测[J].电力自动化设备,2003,23(8)：52-55
[116] 牛东晓，邢棉，孟明.基于联合数据挖掘技术的神经网络负荷预测模型研究[J].电工技术学报, 2004,19(9):62-68.
[117]牛东晓，谷志红，邢棉，王会青.基于数据挖掘的 SVM 短期负荷预测方法研究[J].中国电机工程学报，2006,26(18):8-14.
[118]沈海澜,王加阳,蒋外文,陈再良.模糊关联规则挖掘在电力负荷预测中的应用[J].计算机工程，2003,29(15):138-141
[119]崔旻,顾洁.基于数据挖掘的电力系统中长期负荷预测新方法[J].电力自动化设备，2004,24(6)：18-22
[120] 冯丽，邱家驹.离群数据挖掘及其在电力负荷预测中的应用[J].电力系统自动化，2004,28(22)：41-45
[121]Wang Jianzhou, Ma Zhixin,Li Lian. Detection, mining and forecasting of impact load in power load forecasting[J]. Applied Mathematics and Computation,2005,168(1)：29-39
[122]李耀池.数据挖掘技术在电力负荷预测系统中应用的研究[D].辽宁：辽宁工程技术大学,2006
[123]孙英云.基于数据挖掘的短期负荷预测研究[D].北京：清华大学，2004
[124]Lambert-Torres G，Marra W，Lage W.F.，De Moraes C.H.V，Costa C.I.A. Data mining in load forecasting: An approach using fuzzy techniques[D].2006 IEEE Power Engineering Society General Meeting, PES, 2006，1（1）：17-22
[125]余乐安，汪寿阳，黎建强.外汇汇率与国际原油价格波动预测--TEI@I 方法论[M]，湖南大学出版社，2006
[126] Jiawei Han and Micheline Kamber . Data Mining：Concepts and Techniques[M]. USA, Morgan Kaufmann Publishers. 2001. 70-95.
[127]段利东.火电厂建设项目运营初期风险评价管理研究[D].北京：华北电力大学，2010.
[128]李广原.属性论在文本相似度计算中的应用[J].广西师院学报(自然科学版),

2000，1（3）：199-203.
[129]陶文斌，张粒子，潘弘，李振元，郑华．基于双层贝叶斯分类的空间负荷预测[J]．中国电机工程学报，2007，27（7）：13-17.
[130]Syed M. Islam, Saleh M. Al-Alawi, Khaled A. Ellithy. Forecasting monthly electric load and energy for a fast growing utility using an artificial neural network[J]. Electric Power Systems Research, 1995,34(1):213–217.
[131]C.C. Hsu, C.Y. Chen, Regional load forecasting in Taiwan - application of artificial neural networks, Energy Conversion and Management, 2003,44 (12):1941-1949.
[132]P.F. Pai, W.C. Hong, Forecasting regional electric load based on recurrent support vector machines with genetic algorithms. Electric Power Systems Research, 2005,74(3):417–425.
[133]Wei-Chiang Hong. Electric load forecasting by support vector model. Applied Mathematical Modelling,2009,33(5):2444-2454.
[134]Vapnik, V. The Nature of Statistic Learning Theory[M]. New York: Springer-Verlag. 1995.
[135]Vapnik, V., Statistical Learning Theory. 1998, New York: Wiley.1995
[136] Niu, D., Y. Wang, and D.D. Wu, Power load forecasting using support vector machine and ant colony optimization. Expert Systems with Applications, 2010. 37(3): 2531-2539.
[137]Avci, E., Selecting of the optimal feature subset and kernel parameters in digital modulation classification by using hybrid genetic algorithm-support vector machines: HGASVM. Expert Systems with Applications, 2009. 36(2, Part 1): 1391-1402.
[138]Pai, P.-F. and W.-C. Hong, Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms. Electric Power Systems Research, 2005. 74(3): 417-425.
[139]Hong, W.-C., Chaotic particle swarm optimization algorithm in a support vector regression electric load forecasting model. Energy Conversion and Management, 2009. 50(1): 105-117.
[140]Vesterstrom, J. and R. Thomsen. A comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems[C]. in Proceedings of the 2004 Congress on Evolutionary Computation, 2004,(1):19-23
[141]Das, S. and A. Konar, Automatic image pixel clustering with an improved differential evolution. Applied Soft Computing, 2009. 9(1): 226-236.
[142]Maulik, U. and I. Saha, Modified differential evolution based fuzzy clustering for pixel classification in remote sensing imagery. Pattern Recognition, 2009. 42(9): 2135-2149.
[143]Al-Obeidat, F., et al., Differential Evolution for learning the classification method PROAFTN. Knowledge-Based Systems, 2010. 23(5): p. 418-426.
[144]Lu, Y., et al., An adaptive hybrid differential evolution algorithm for dynamic economic dispatch with valve-point effects. Expert Systems with Applications, 2010. 37(7): 4842-4849.
[145]Wu, L.H., et al., Environmental/economic power dispatch problem using multi-objective differential evolution algorithm. Electric Power Systems Research, 2010. 80(9): 1171-1181

附录

附表 1：江门市供电局监测的行业用电量的年增长量

行业名称	2002	2003	2004	2005	2006	2007	2008
全社会用电量总计	13.17	10.21	21.14	9.54	9.98	12.51	0.26
A、各行业用电量合计	15.18	12.04	24.06	9.22	10.04	12.5	-0.51
第一产业	9	5.6	32.75	46.06	19.2	15.05	-4.23
第二产业	15.56	11.53	25.73	10.7	14.23	13.11	-1.4
第三产业	15.8	17.02	13.21	-3.12	-11.32	8.02	6.23
B、城乡居民生活用电量合计	4.59	1.58	5.97	12.45	9.43	12.64	6.67
城镇居民	4.95	16.63	8.56	10.78	9.75	11.71	7.22
乡村居民	4.17	-16.05	1.75	15.35	8.9	14.2	5.76
各行业用电分类	0	0	0	0	0	0	0
一、农、林、牧、渔业	9	5.6	32.75	19.78	19.2	15.05	-4.23
1、农业	2.4	13.75	46.66	-45.86	3.91	9.83	0.49
2、林业	34	33.44	-17.39	-48.98	-20.12	-2.41	-10.34
3、畜牧业	3.77	16.43	-3.18	105.51	0.51	4.74	24.41
4、渔业	12.7	-1.8	0.36	1280.42	33.83	21.05	-10.65
5、农、林、牧、渔服务业	25.1	-20.21	21.83	-42.74	13.7	0.74	8.57
其中：排灌	-17.19	-46.76	30.89	3.55	3.43	-5.69	20.83
二、工业	15.71	12.01	25.85	18.29	14.75	13	-1.53
1、轻工业	21.17	8.72	25.8	2.2	-2.57	9.18	-0.56
2、重工业	13.96	32.1	8.1	32.26	24.86	14.75	-2.54
（一）采矿业	27.27	29.94	43.32	-6.72	-9.29	7.53	-3.99
1、煤炭开采和洗选业	-100	0	0	12900	-31.54	70.79	23.97
2、石油和天然气开采业	75	57.14	38.64	103.28	-58.87	-54.9	44.44
3、黑色金属矿采选业	0	0	-100	0	-62.5	-16.67	25
4、有色金属矿采选业	73.34	9.74	10.08	-85.37	312.6	-15.46	85.69
5、非金属矿采选业	23.22	32.06	48.92	-16.2	-22.7	13.97	-11.32
6、其他采矿业	48.57	89.1	-0.34	944.56	31.59	-5.07	-0.41

（二）制造业	15.6	11.7	25.57	19.48	18.39	12.89	3.44
1、食品、饮料和烟草制造业（轻）	20.67	-2.62	10.56	-49.04	37.02	23.49	14.63
其中：农副食品加工业	0	0	0	0	31.61	34.35	23.87
2、纺织业（轻）	19.59	7.79	27.98	3.53	-2.05	-3.26	-20.07
3、服装鞋帽、皮革羽绒及其制品业（轻）	0	0	0	24.74	57.16	3.24	2.95
4、木材加工及制品和家具制造业	0	0	0	36.69	108.48	57.67	12.89
其中：轻工业	0	0	0	0	38.31	39.16	13.21
5、造纸及纸制品业（轻）	0	0	0	0	20.03	10.74	9.67
6、印刷业和记录媒介的复制（轻）	0	0	0	-4.39	-2.32	29.26	3.64
7、文体用品制造业（轻）	0	0	0	1489.87	-22.29	21.31	7.12
8、石油加工、炼焦及核燃料加工业	-16.28	-6.94	-13.43	1234.48	-1.42	-5.37	41.4
9、化学原料及化学制品制造业	12.29	14.52	23.42	4.17	27.15	2.8	2.19
其中：轻工业	0	0	0	0	-25.21	31.98	9.21
其中：氯碱	0	0	0	0	12.92	1.04	-0.69
电石	0	0	0	0	0	0	0
黄磷	0	0	0	0	0	0	0
其中：肥料制造	0	0	0	0	13.78	-0.87	-9.21
10、医药制造业（轻）	21.9	7.61	27.91	-49.02	-8.74	19.43	6.06
11、化学纤维制造业（轻）	7.18	1.73	20.32	7.11	-29.14	16.79	5.33
12、橡胶和塑料制品业	19.18	26.2	22.42	82.05	28.45	11	3.24
其中：轻工业	25.05	7.57	5.51	59.53	-32.34	13.72	1.14
13、非金属矿物制品业	12.86	14.16	18.6	20.5	19.36	7.43	-0.9
其中：轻工业	6.92	138.06	-23.98	-42.42	-44.14	20.23	14.05
其中：水泥制造	0	0	0	0	19.66	9.94	-12.73
14、黑色金属冶炼及压延加工业	14.93	19.05	-1.5	187.79	91.34	23.78	6.27
其中：铁合金冶炼	0	0	0	0	383.31	27.93	10.36
15、有色金属冶炼及压延加工业	-19.6	24.65	109.46	178.49	24.37	11	17.86
其中：铝冶炼	0	0	0	0	-9.26	22.66	34.27
16、金属制品业	26.35	19.79	46.4	51.99	41.5	20.96	8.08
其中：轻工业	25.29	12.91	64.26	21.46	-5.77	8.8	3.41
17、通用及专用设备制造业	11.14	7.29	45.68	-40.69	14.39	38.86	14.32
其中：轻工业	10.02	16.94	147.07	-83.91	-53.98	55.06	10.34

18、交通运输、电气、电子设备制造业	13.51	22.24	20.9	71.91	14.45	14.26	2.38
其中：轻工业	36.88	21.8	23.38	-3.97	-55.17	8.5	-0.23
其中：交通运输设备制造业	0	0	0	0	3.6	5.26	-0.33
19、工艺品及其他制造业（轻）	0	0	0	-30.13	0.98	11.2	2.21
20、废弃资源和废旧材料回收加工业	0	0	0	0	-35.36	149.06	221.75
（三）电力、燃气及水的生产和供应业	7.53	1.81	10.38	16.82	-1.09	14.44	-36.77
1、电力、热力的生产和供应业	0	0	0	16.7	0.86	13.51	-41.79
其中：电厂生产全部耗用电量	0	0	0	0.43	-32.18	100.05	-19.28
线路损失电量	0	0	0	4.52	7.03	10.29	-49.99
抽水蓄能抽水耗用电量	0	0	0	0	0	-33.33	0
2、燃气生产和供应业	0	0	0	-75.67	20.08	176.9	-66.27
3、水的生产和供应业	7.53	1.81	10.38	58.11	-15.77	6.49	8.95
其中：轻工业	0	0	0	0	0.71	2.64	19.94
三、建筑业	12.86	2.39	23.09	-60.84	-32.12	30.25	15.36
四、交通运输、仓储、邮政业	20.62	4.26	-2.75	-33.87	-61.23	19.98	4.6
1、交通运输业	27.78	-3.02	8.94	-26.64	-50.25	26.07	5.67
其中：城市公共交通	31.68	-2.47	0.4	0.08	-78.05	20.79	21.28
管道运输业	27.27	-19.39	11.39	-6.82	114.63	156.82	5.63
电气化铁路	0	0	0	0	-100	0	0
2、仓储业	0	0	0	768.84	-75.73	25.09	8.81
3、邮政业	17.33	7.89	-8	-56.15	-58.45	5.85	-1.23
五、信息传输、计算机服务和软件业	-6.78	-2.83	23.22	236.24	25.14	15.23	15.78
1、电信和其他信息传输服务业	0	0	0	50.89	32.92	15.06	16.69
2、计算机服务和软件业	-6.78	-2.83	23.22	172.17	-14.12	16.55	8.74
六、商业、住宿和餐饮业	0	0	0	1.19	3.77	7.87	6.12
1、批发和零售业	0	0	0	-9.59	7.85	7.97	4.51
2、住宿和餐饮业	0	0	0	34.15	-4.64	7.66	9.92
七、金融、房地产、商务及居民服务业	16.82	24.02	12.08	-11.67	-30.46	0.02	1.48
1、金融业	16.49	27.49	10.06	-19.55	-40.25	-8.12	-6.9
2、房地产业	18.62	5.43	25.14	38.8	-22.64	13.04	11.37
3、租赁和商务服务、居民服务和其他服务业	0	0	0	226.2	18.38	17.42	16.25
八、公共事业及管理组织	7.93	5.92	47.72	139.11	-1.05	15.01	9.37

1、科学研究、技术服务和地质勘查业	0	0	0	-21.74	-6.66	-5.72	2.72
其中：地质勘查业	63.16	-61.29	-100	0	0	0	30
2、水利、环境和公共设施管理业	0	0	0	16.91	-3.75	14.05	15.49
其中：水利管理业	-3.04	-8.7	-48.55	-33.37	-49.48	-0.61	12.05
其中：公共照明	13.36	7.96	40.09	-22.58	11.5	31.58	-7.59
3、教育、文化、体育和娱乐业	-1.15	4.5	63.57	5.03	-2.61	20.91	9.95
其中：教育	0	0	0	0	-36.66	20.94	18.68
4、卫生、社会保障和社会福利业	0	0	0	-3.76	-15.3	11.68	2.34
5、公共管理和社会组织、国际组织	22.1	7.71	28.31	42.97	18.32	15.39	9.67

附表 2：江门市供电局初步筛选后的行业以及警度划分结果

Table 6-13. The Industrials after initial choosing and the warning degree

变量	名称	2002	2003	2004	2005	2006	2007	2008
x1	第三产业	2	2	2	1	2	1	1
x2	B、城乡居民生活用电量合计	1	1	1	2	1	2	1
x3	城镇居民	1	2	1	2	1	2	1
x4	乡村居民	1	2	1	2	1	2	1
x5	一、农、林、牧、渔业	1	1	4	2	2	2	1
x6	1、农业	1	2	4	4	1	1	1
x7	2、林业	4	4	2	4	3	1	2
x8	3、畜牧业	1	2	1	4	1	1	3
x9	4、渔业	2	1	1	4	4	3	2
x10	5、农、林、牧、渔服务业	3	3	3	4	2	1	1
x11	其中：排灌	2	4	4	1	1	1	3
x12	二、工业	2	2	3	2	2	2	1
x13	1、轻工业	3	1	3	1	1	1	1
x14	2、重工业	2	4	1	4	3	2	1
x15	（一）采矿业	3	3	4	1	1	1	1
x16	1、煤炭开采和洗选业	4	0	0	4	4	4	3
x17	2、石油和天然气开采业	4	4	4	4	4	4	4
x18	3、黑色金属矿采选业	0	0	4	0	4	2	3
x19	4、有色金属矿采选业	4	1	2	4	4	2	4
x20	5、非金属矿采选业	3	4	4	2	3	2	2
x21	6、其他采矿业	4	4	1	4	4	1	1
x22	（二）制造业	2	2	3	2	2	2	1
x23	1、食品、饮料和烟草制造业（轻）	3	1	2	4	4	3	2
x24	2、纺织业（轻）	2	1	3	1	1	1	3
x25	3、服装鞋帽、皮革羽绒及其制品业（轻）	0	0	0	3	4	1	1
x26	4、木材加工及制品和家具制造业	0	0	0	4	4	4	2
x27	5、造纸及纸制品业（轻）	0	0	0	0	3	2	1
x28	6、印刷业和记录媒介的复制（轻）	0	0	0	1	1	3	1

x29	7、文体用品制造业（轻）	0	0	0	4	3	3	1
x30	8、石油加工、炼焦及核燃料加工业	2	1	2	4	1	1	4
x31	9、化学原料及化学制品制造业	2	2	3	1	3	1	1
x32	10、医药制造业（轻）	3	1	3	4	1	2	1
x33	11、化学纤维制造业（轻）	1	1	3	1	3	2	1
x34	12、橡胶和塑料制品业	2	3	3	4	3	2	1
x35	13、非金属矿物制品业	2	2	2	3	2	1	1
x36	14、黑色金属冶炼及压延加工业	2	2	1	4	4	3	1
x37	15、有色金属冶炼及压延加工业	2	3	4	4	3	2	2
x38	16、金属制品业	3	2	4	4	4	3	1
x39	17、通用及专用设备制造业	2	1	4	4	2	4	2
x40	18、交通运输、电气、电子设备制造业	2	3	3	4	2	2	1
x41	19、工艺品及其他制造业（轻）	0	0	0	4	1	2	1
x42	20、废弃资源和废旧材料回收加工业	0	0	0	0	4	4	4
x43	（三）电力、燃气及水的生产和供应业	1	1	2	2	1	2	4
x44	1、电力、热力的生产和供应业	0	0	0	2	1	2	4
x45	2、燃气生产和供应业	0	0	0	4	3	4	4
x46	3、水的生产和供应业	1	1	2	4	2	1	1
x47	三、建筑业	2	1	3	4	4	4	2
x48	四、交通运输、仓储、邮政业	3	1	1	4	4	2	1
x49	1、交通运输业	3	1	1	3	4	3	1
x50	2、仓储业	0	0	0	4	4	3	1
x51	3、邮政业	2	1	1	4	4	1	1
x52	五、信息传输、计算机服务和软件业	1	1	3	4	3	2	2
x53	1、电信和其他信息传输服务业	0	0	0	4	4	2	2
x54	2、计算机服务和软件业	1	1	3	4	2	2	1
x55	六、商业、住宿和餐饮业	0	0	0	1	1	1	1
x56	1、批发和零售业	0	0	0	1	1	1	1
x57	2、住宿和餐饮业	0	0	0	4	1	1	1
x58	七、金融、房地产、商务及居民服务业	2	3	2	2	4	1	1
x59	1、金融业	2	3	2	2	4	1	1
x60	2、房地产业	2	1	3	4	3	2	2

x61	3、租赁和商务服务、居民服务和其他服务业	0	0	0	4	2	2	2
x62	八、公共事业及管理组织	1	1	4	4	1	2	1
x63	1、科学研究、技术服务和地质勘查业	0	0	0	3	1	1	1
x64	2、水利、环境和公共设施管理业	0	0	0	2	1	2	2
x65	3、教育、文化、体育和娱乐业	1	1	4	1	1	3	1
x66	4、卫生、社会保障和社会福利业	0	0	0	1	2	2	1
x67	5、公共管理和社会组织、国际组织	3	1	3	4	2	2	1
y	全社会用电量总计	2	2	3	1	1	2	1

攻读博士学位期间发表的论文及其它成果

（一）发表的学术论文

[1]DongXiao Niu（牛东晓）,JianJun Wang（王建军）,JinPeng Liu（刘金朋）. Knowledge Mining Collaborative SVM Correction Method in Short-term Load Forecasting. Journal of Central South University of Technology,2011.（SCI 期刊，已录用，2011 年 8 月份出刊）

[2]DongXiao Niu（牛东晓）, JianJun Wang(王建军). Combination of Text Mining and Corrective Neural Network in Short-term Load Forecasting, Journal of Computers, 2009, 4(12):1188-1194.(EI 国际期刊, EI 收录号: 20095012547812).

[3]Jianjun Wang（王建军）, DongXiao Niu（牛东晓）, Li Li（李莉）. Middle-Long Term Load Forecasting Based on Dynamic Architecture for Artificial Neural Network, Journal of Information And Computational Science, 2010,8(7):1711-1717 (EI 国际期刊，EI 收录号：20110313600261).

[4]Dongxiao Niu(牛东晓),Jianjun Wang(王建军),Li Li(李莉). Combined Model for Load Forecasting with Differential Evolution, Algorithm Advanced Materials Research,2011,181(1):594-598(EI 国际期刊，EI 收录号：20110713657648)

[5]牛东晓，王建军，李莉，谭忠富. 电力需求侧响应利益联动机制的系统动力学模拟分析[J]. 系统工程理论与实践,（已录用，EI 检索）

[6]牛东晓，王建军，李 莉，李存斌. 基于粗糙集和决策树的自适应神经网络短期负荷预测方法研究，电力自动化设备，2009.10(29):30-34（EI 期刊，EI 收录号：20095012547812）.

[7]WANG Jian-jun(王建军), NIU Dong-xiao, LI Li. An ARMA cooperate with Artificial Neural Network approach in Short-Term Load Forecasting, Proceedings of the 2009 Fifth International Conference on Natural Computation,2009,1(1):60-64（EI 检索，EI 收录号：20101512839960）.

[8]王建军，牛东晓，李莉. 基于相似度与神经网络的协同短期负荷预测模型[J]. 华东电力，2009,37(1):64-66.

[9].谭忠富，李莉，王建军，姜海洋，王成文.多智能体代理下电力双边谈判中的模糊贝叶斯学习模型[J]. 中国电机工程学报, 2009,29(7):106-113.

[10]李莉, 谭忠富, 王建军, 姜海洋, 候建英, 王成文. 可中断负荷参与备用市场下的可靠性风险电价计算模型[J]. 电网技术, 2009,33(4):81-87.

（二）申请及已获得的专利

2011 年 01 月，获得“区域电网用电市场分析与智能预测系统 V1.0”软件著作权一项，登记号 2011SRBJ0082。（排名 2/8）

（三）获得的科技奖励

2009 年 12 月获得中国商业联合会科学技术奖全国商业科技进步一等奖，“区域电网用电市场分析与智能预测系统研究”（省部级一等奖，排名 5/10）。

攻读博士学位期间参加的科研工作

[1]国家自然基金“基于协同知识挖掘的电力负荷预测理论研究”，2008.09-2009.12，参与人。

[2]国家自然基金“智能电网中适应不稳定大规模清洁能源发电的联合智能调度管理理论研究”，2011.01～2013.12，主研人之一。

[3]国家自然基金“信息化环境下企业项目链风险元传递理论模型研究”，2011.01～2013.12，参与人。

[4]国家自然基金“广义项目风险元传递理论模型及其应用”， 2011.01-2013.12，参与人。

[5]中央高校基本科研业务费专项资金资助项目 “智能电网环境下的自适互动智能电力预测研究”， 2010.07-2011.07，课题负责人。

[6]华北电力大学拔尖博士培育资助项目 “知识挖掘智能协同电力负荷预测研究”，2010.07-2011.07，课题负责人。

[7]广东省江门市供电局，用电市场分析及预测系统开发，2008.11-2009.12，负责人。

[8]广东省江门市供电局，江门供电局农电规范管理方案研究，2009.05 -2009.11，负责人。

[9]广东省江门市供电局，江门供电局开拓用电市场方案研究，2009.05-2009.11，负责人。

[10]吉林省电科院系统所 220KV 靖宇输变电工程后评价，2009.11-2010.08，负责人。

[11]大唐发电集团，大唐发电集团三个项目打包后评价研究，2009.01-2009.07，参与人。

[12]青海省电力公司，青海省电力公司十二五发展战略规划研究，2010.05-2010.09，负责人。

致 谢

值此论文完成之际，我衷心的感谢我的导师牛东晓教授，牛老师以他那渊博的知识、开阔的思路、严谨的治学态度和谆谆不悔的教导不断地精心指导着我的博士学习和研究，在过去的三年里，牛老师教会了我很多的东西，使我受益匪浅，他将我引入了电力负荷预测的研究领域里，结合我过去的知识结构给我不断的引导我进入了知识挖掘和电力负荷预测的交叉研究领域中，此外，牛老师对我的学习生活同样也是照顾有加，可以使我能够尽可能的投入到自己喜欢的研究中，身为经济与管理学院院长的他，为学生倾注的心血，难以言尽。他严谨的学风、正直的人品、高尚的师德将永远是我学习的楷模。本篇论文同样也倾注着牛老师的心血，从论文的开题到论文的最终稿都是在导师的精心指导下完成的。

此外，还要感谢我的博士生副导师和硕士生导师李存斌教授，李老师在我上硕士期间帮助我打下了信息管理系统开发以及数据挖掘相关方面的基础，并且引领我进行论文的初步写作和研究，尤其在硕士的三年中，李老师同样以他严谨的学风、正直的人品和高尚的师德作为我在学术上和实践中的指明灯。

感谢我的父母，使我能够心无旁骛的在浩瀚的学海中勇往直前，在我迷失方向的时候为我提供温暖的问候和无尽的动力，感谢他们无私的支持和鼓励，才能使我最求我喜欢的东西。

感谢我的爱人李莉，近十年的华电生活她一直在我身边默默的陪伴和支持着我，陪我度过硕士和博士期间的学习生活，使我在求学的道路上并不孤独。

感谢我的儿子王子轩，他给我的博士求学期间添加了无尽的乐趣。

还要感谢研究生院和经济与管理学院的领导、老师为我的论文完成所给予的大量帮助，这些都使得我的毕业论文得以顺利完成。

感谢华北电力大学技术经济预测与评价研究所的同门，他们是刘达、李金超、谷志红、王永利、刘金朋、房芳、王宁、林帅、张云云、褚烨、张烨、吕海涛、李莹莹、李建青、许聪、王汉梅、李欣、周萍、崔璐、马东、周昆、贾睿彪、田洁、嵇灵、魏亚楠、孙蕾等，和你们相处的日子将是一段中美好的回忆。

感谢华北电力大学信息管理与信息系统开发研究所的樊建平、郭晓鹏、董威、刘小亚、穆海、母德宝、刘天星、胡喆、马同涛、李伟、李贤、刘学艳、王恪诚、孙安黎、王丽娜、马伟、张亮等硕士阶段的同门表示感谢，与你们相处的日子是一段美好的回忆。

感谢我的同窗好友朱丽丽、谢维、白泉涌、唐慧、苏志雄、谢维、周景宏、

于超、张金良等同学在博士期间对于学习和论文期间给我无私支持和帮助。

感谢华北电力大学对我的培养和教育。还要感谢审阅本论文的所有专家和老师们。

本课题承蒙国家自然科学基金“基于协同知识挖掘的电力负荷预测理论研究”资助，特此致谢。

再回首博士三年的学习生活中，在收获知识的同时，也得到了很多磨炼，这将是我一生的财富，使我面对未来和面对挑战时更加充满信心！

王建军

二零一一年四月十二日

作者简介

本人与1981年8月25日出生于吉林省白山市。

2001年09月考入华北电力大学工商管理学院工商管理专业，2005年07月本科毕业并获得管理学学士学位。

2005年09月——2008年03月在华北电力大学工商管理学院管理科学与工程专业学习并获得管理学硕士学位。

2008年09月——2011年06月在华北电力大学经济与管理学院管理科学与工程专业攻读管理学博士学位。