分类号：_____ 学校代号：11845

UDC：_____ 密 级：_____ 学 号：2111105055

广东工业大学硕士学位论文

（工学硕士）

多组合分类器在局部区域气温预测中的研究与应用

李 俊 磊

指导教师姓名、职称：滕少华 教授

学科(专业)或领域名称：计算机应用技术

学 生 所 属 学 院：计算机学院

论 文 答 辩 日 期：二〇一四年五月

A Dissertation Submitted to Guangdong University of Technology

for the Degree of Master

(Master of Engineering Science)

Research and Application of multi-ensemble classifier in the local area's temperature forecast

Candidate: Li Junlei

Supervisor: Prof. Teng Shaohua

May 2014

School of Computers

Guangdong University of Technology

Guangzhou, Guangdong, P. R. China, 510006

摘 要

随着科技的进步、信息化的发展、气象研究技术的提高，气象领域积累的数据量与日俱增。如何从海量的气象数据中发现有价值的信息是气象科研人员的一项重要任务。气象信息与人民的生活息息相关，人民的生活和生产制造直接受天气的影响。如果能将数据挖掘应用到气象数据中充分挖掘出可用的信息，不仅能提高天气预报准确率和灾害天气预警能力，还能指导当地的工农业生产和提高人民的生活水平，造福人民。

在数据挖掘中，分类是一种非常重要的技术。现有的分类技术有决策树、贝叶斯、支持向量机、神经网络等，这些分类器都是单分类器。随着人们对分类器性能要求的提高，相关领域学者提出了集成学习的概念。所谓集成学习就是对同一个问题用多个单一的个体学习器进行组合学习，组合分类器就是将多个不同的分类器（基分类器)通过一定的方法组合起来构建而成的一个组合分类器。因此，组合分类器就是分类器的集成，同时也叫分类器的组合。实验证明，多个分类器组合在性能上超越于单分类器的性能。

本文在对气象数据的特点、气象数据挖掘现状和常用的气象数据挖掘方法进行了分析的基础上。利用数据挖掘中的决策树分类方法和集成学习思想构建组合分类器，并用来对广州某局部区域气象站搜集的历史气象数据进行分析和研究。

本文主要开展了下列研究：

1.设计并实现了基于决策树的并行组合分类器预测模型，将组合分类器和决策树分类方法用于局部区域的气温预测中，各基分类器分别对局部区域的气温进行预测，组合分类器综合各基分类器的结果，最后获得各基分类器的协同预测。

2.基于 C4.5 决策树算法，设计并实现了 Bagging、Adaboost 两种组合模型，同时基于 CART 决策树设计了随机森林模型。

3.针对局部区域气象数据，应用了 Bagging、Adaboost 和随机森林三种组合分类器，分别设计并实现了局部区域气温预测模型。

4.应用某局部区域的气象数据，验证了 Bagging、Adaboost 和随机森林三种气温组合预测模型的有效性，进而对三种模型的预测结果从准确率和性能上进行了细致的比较分析。

本文的研究成果，为当地的气象局提供了决策依据，给当地居民的社会生活和工业生产提供了指导性的效果。

关键词： 气象数据；组合分类器；Bagging；AdaBoost；随机森林

ABSTRACT

With the advancement of science and technology, the development of information, the promotion of Meteorological Research Technology, and the growing in the field of meteorology data with each passing day so fast. It is an important task for the meteorological research section to find the valuable information from the massive meteorological data. Weather information is closely related with people's lives. People's social lives and production are directly influenced by the weather. If data mining can be applied into the meteorological data, It will make full use of the available information, these information not only can improve the accuracy of weather forecast and the ability of disaster weather warning, but also can guide the local industrial and agricultural production and raise the living standards of the peoples.

In data mining, classification is a very important technology. There are decision tree, Bayesian networks, support vector machine and neural network in classification technology, these are all single classifiers. In order to improve the performance of classifier, some scholars proposed the concept of ensemble learning. Ensemble learning is a method that uses a number of basic classifiers to solve one problem, so a ensemble classifier is combining a number of different classifiers (basic classifiers) through a certain method using the related technology to fuse each base classifier eventually. Therefore, Ensemble learning classifier is also called classifier ensemble classifier. Experiments show that, the performance of ensemble classifiers is better than any one of single classifier significantly.

Based on analytical the characteristics of meteorological data, the present situation of meteorological data mining and common used methods for meteorological data mining. This paper used Decision tree classification and ensemble learning to construct ensemble classification then analyzed and studied the meteorological data of a local meteorological station.

This paper carried out the following research:

1. Designed and implemented the parallel ensemble classifier prediction model based on decision tree, the ensemble classifier and decision tree classification method were used to predict the temperature of local area, each base classifier was used to predict the temperature of the local area respectively, and the result of ensemble classifier was colligated each base

classifier, finally obtained each base classifier's collaborative forecasting.

2. Based on C4.5 decision tree algorithm, Bagging, AdaBoost two kinds of ensemble model were designed and implemented, the random forest model was designed based on CART decision tree at the same time.

3. According to local meteorological data, the application of three combination classifiers Bagging, AdaBoost and Random Forest, local temperature prediction model was designed and implemented respectively.

4. The application of a local meteorological data was verified the Bagging, AdaBoost and Random Forest three temperatures ensemble forecasting models was effective, and the prediction results of three models had carried on the detailed comparative analysis from the accuracy and performance.

The results of this research provided a basis decision making for the local weather bureau and also provided the guidance effect on the local residents of their social life and industrial production.

Key Words: meteorological data; ensemble classifiers; Bagging; AdaBoost; RandomForest

目 录

CONTENTS

第一章 绪 论

本文针对气象数据的特点，采用数据挖掘中的决策树方法和集成思想对气象数据进行挖掘分析，得出了有用的决策信息，这些信息对于气象数据的分类和预测用着非常重要的意义。下面就本文的研究背景和意义、主要研究内容、国内外研究现状以及论文的整体架构进行阐述。

1.1 研究背景及研究意义

随着科技的进步和国民经济的迅猛发展，信息化变得越来越普及，我国的气象卫星遥感技术远远超越了国际上的众多国家，已处于国际的领先地位。2006 年风云 D 星发射成功以后，我国实现了“双星观测、在轨备份”的业务格局。与此同时，我国的计算机发展也迅速，实现了高性能计算，这就促进了气象事业的蓬勃发展，从银河 I 号到银河 III 号，到 2010 年天河一号的成功研发。先进的气象数据收集设备的诞生导致气象数据量收集不断增大，然而数据处理技术的发展却相对滞后，对数据的处理方面绝大多数仅停留在增删查改等简单的操作阶段，缺少有效技术来对数据进行深层次的关系和规则挖掘。

数据挖掘是信息大爆炸时代背景下的产物，数据挖掘是从大量的数据中提取或挖掘出可能有价值的信息和知识，数据挖掘技术蕴含了传统的数据分析方法和用于处理海量数据的复杂算法[1]。目前，数据挖掘引起了信息业、科技界和商业界的极大关注和广泛的应用，在科技探索、市场营销、生产控制、金融机构、医疗诊断和商务管理等许多领域得到了应用。

近年来，随着气象业务数值预报系统的不断升级和互联网的迅速发展，人们对天气预报提出了更高的要求。例如飞机起飞、火箭的发射等严重依赖于当时的云、风、雷电、降雨量、气温等气象因素的影响，这就要求气象部门能够根据最近的历史气象数据进行精确的短时气象预测；为了科学指导农业生产、在节假日指导人们出行安排等日常活动，这就需要准确的中短期气象预测；国家大型项目的建设、自然灾害的预防、军事上的战略部署等更加迫切的需要高精度的中长期气温预测。尽管气象预测技术不断提升，气象预测水准与日俱增，预测的准确率不断提高，满足了人们的日常生活及社会活动的需求，然而在旱灾、洪灾、地震、等自然灾害的预测方面，目前仍没

有有效的预测方法。2009 年上半年，中国自然灾害受灾区域比较集中，涉及面积较大；邻近江河两岸的人民洪涝灾害严重；东北地区大面积受风雹灾害入侵；西南地区受地质灾害吞噬等。2009 年秋季，我国西南地区遭遇旱灾，局部地区遭遇百年一遇的大旱，持续六个月之久；2010 年 1 月 13 日，新疆塔城地区迎来了 60 年难得一遇的严寒暴雪天气，气温降至零下 33 度至零下 35 度；2010 年 5 月 5-6 日罕见暴风雨强势来袭，南方部分省市连连告急，受灾人数之多触目惊心：湖南 169 万，重庆 137 万，广东 87 万，江西赣南 26 万；2011 年初，华北地区旱灾致使大多数农作物绝产，几十万居民的饮水问题难以解决。这样的灾情在每年都数不胜数，夺走了无数宝贵的生命，给人民的生活带了极大的影响，同时对国民经济造成了严重的损失。因此，如何在海量的气象数据中挖掘出有用的信息，发现气象数据的内在规律是一个迫切需要解决的问题。

由于天气情况与人民的生活息息相关，关系着人民的衣食住行，所以气象预测一直是国内外研究的热点，数据挖掘的兴起，使得数据挖掘在气象领域中后来居上，成为了气象科研人员研究气象数据的核心技术之一，从海量的气象数据中挖掘隐藏的气象规律，并将数据挖掘的分类和预测方法应用到气象预报的业务中对天气进行有效的分类预测。数据挖掘技术的应用不仅提高预测的准确率和及时性，对气候灾害（如：洪灾、旱灾等)可以及时采取预防措施，减少受灾面积和受灾群众，同时也可以为特色作物的培育提供对应的决策[2]。最终满足人们的生活和社会活动的需求，避免一些不必要的损失。

综上所述，本文旨在从广州市某区气象站收集到的气象数据出发，利用组合分类器方法，对该区的局部气象数据进行分析与处理，预测当地的气温情况，以进一步指导气象预测，对于提高灾害天气预报的准确性做有益的探索。

1.2 课题的主要研究内容

本文运用数据挖掘中的组合分类器技术对广州市某区气象站收集到的气象数据进行分析和处理，采用决策树和集成思想建构 Bagging、 AdaBoost 和随机森林三种组合分类器模型，同时对当地的气温进行预测，通过分析和比较三种组合分类器之间的优劣得到了一些有用的信息，为气象部门提供了一定的参考信息。本课题的主要研究工作可分为以下几个部分：

(1)查阅相关文献。进一步了解当前已有的研究工作，为论文写作的研究工作提供基础支撑。

(2)研究数据挖掘中分类的相关算法，选择合适的算法对气象数据进行研究与处理。

(3)采用决策树的方法和集成思想构建组合分类器模型对气温等级进行预测，在合适的模型中加入协同并行思想。本文对气象数据主要考虑风速、相对湿度、气压、水汽压、露点温度、地面温度、时总蒸发量 7 个因素对气温的影响，建立气温预测的决策树组合分类器模型。

(4)利用建立的组合分类器模型对广州市某区的气象数据进行分析和处理。

(5)应用 Matlab7.0、Excel 等软件工具对所研究的气温预测模型进行实验。根据实验的结果验证了这些模型的可行性和有效性，同时提高了气温预测的准确率和及时性。

(6)对所研究的工作进行总结，提出下一步要做的研究工作。

1.3 国内外的研究现状

1.3.1 气象数据挖掘现状

数据挖掘(Data mining)是一门从大量资料或者海量数据中提取有用信息的学科，从资料中提取出隐含的潜在的有价值的信息[3-4]。数据库知识发现(KDD)是从数据库中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程，它是数据挖掘更为广义的一种说法，随着术语的变迁，两者通常被认为是同一个含义。

近年来，随着科学技术的飞速发展，网络中存在着大量的信息数据，而这些海量的数据中往往隐藏着潜在有价值的知识，希望能有一种技术能够从中提取出来，科学家们通过研究探讨便产生了数据挖掘理论这一门技术，因而数据挖掘的出现在信息产业界引起了极大关注，获得了广泛地使用，尤其是在气象研究领域。近年来，科技的进步，气象领域不断加大信息化程度，使得气象部门积累了气象数据每天在以好几倍的速度在增长，收集到的数据越来越大，形成了资料山和资料库，如何管理和使用好这些海量数据是气象研究工作者面临的主要问题，也是提高预报预测准确率和灾害天气预警能力的关键。目前，数据挖掘在气象研究领域的应用主要体现在以下几方面:

(1)气象预报

气象预报就是对未来的天气状况进行预测和报告的一种行为，一般天气预报是指短期的，其业务较多的是采用利用天气学基本原理去分析卫星或基站探测到的数据和统计预报的方法，如多元回归分析、灰色理论和回归分析建立组合模型、判别分析、均

生函数方法和时间序列分析等[5-10]，基于数据挖掘和统计的气象预报方法是目前国内外学者研究和探讨的热点，暂时未得到充分的应用。

杨淑群等[11]将支持向量机(SVM)分类方法首次对降水异常进行了分类预测，利用1958-2003年的气象资料建立了SVM推理模型，该模型用于预测降雨量和是否降雨具有较高的准确度，从已有的研究来看，SVM分类方法应用在降雨量的预测估计方面得到了较好的发展;由于气象数据具有很强的时空关联特性， Valliappa Lakshmanan[12]对陆地级的天气进行了实时拼图，对降雨估测、冰雹诊断等进行了实验；C.Piani 等人[13]对利用预测模型对欧洲的每日降雨进行了预报。

(2)气象预测

气象预测是指长期的气象预报，其主要内容是通过大量的历史气象数据进行分析然后对未来一定时间段内的气象状况如降雨量、气温、灾害天气等进行预测。分类预测分为连续值的预测和离散值预测两种，连续值预测是利用回归分析和神经网络对温度、降雨量等进行预测，离散值预测是利用决策树、支持向量机、神经网络、粗糙集等方法，对台风、暴雨、降霜等进行预测[14]。由于海量气象数据的存在，所以数据挖掘是最好的一种技术手段进行知识挖掘的特征。气候预测是气象数据挖掘研究的重点方向。

Siva Venkadesh 等[15]利用多年的历史气温数据进行分析，将人工神经网络和遗传算法相结合，建立了更为准确的人工神经网络模型，通过遗传算法确定气温数据集的每个输入数据的首选时间和分辨率，预测每一个水平等级上的空气温度;徐亮亮等[16]提出了一对一模糊支持向量机多分类方法的非线性夏季雨型预报模型，这种模型和传统的支持向量机多分类方法和线性物理统计方法相比具有明显的优势，它具有更好的预报能力和更强的抗干扰能力，克服了基于统计理论的相关性分析和回归方法在处理非线性问题上的弱点。Linli Jiang 等[17]将PSO和GA神经网络进行结合共同对每月的降水量进行分析建立预测模型，对每月的降水量进行预测。黎玉芳等[18]利用时间序列的季节指数建立气温预测模型，以桂林地区2001年至2010年各月份的平均气温为实验数据，对气温进行了预测；同时利用线性外推方法建立了降雨量预测模型，应用在1987年至2010年每年6月份的降水量的实测数据集中，对气温和降水进行预测，两个模型都得到了很好的预测效果。王定成等[19]利用支持向量机分类器的思想将灾害天气和正常天气作为分类器中的两个类别，以灾害天气为数据样本进行训练支持向量机建立灾害天气分类的模型，然后将构建的模型用于预测夏季的温度是否异常。滕少华等[20]提出了利用 Bagging 结合 KNN 算法利用协同思想对降雨情况进行了分类预测。姜文瑞

等[21]以陕北某县 30 年的气象历史数据为实验数据集，利用决策树中的 CART 方法针对一年四季的气候差异分别建立了四个季度的气温预测模型，对当地的各个季节的气温进行了预测。黄静华等[22]运用 K-means 聚类算法对气象数据进行了聚类研究，将聚类算法融合于气象分析领域；何伟[23]利用朴素贝叶斯方法建立的降雨量预测模型进行了相关的研究，结果表明朴素贝叶斯建立的模型具有较强的实用性和有效性。

(3)气象灾害预测

气象灾害预报主要是利用灾害天气动力学理论和定量遥感技术相结合的综合技术对灾害天气进行预报，气象灾害的发生严重影响到人们的生活和社会的安定，给人民带来巨大的损失。为了减少损失，必须提高灾害天气的预报能力，众多研究学者尝试将数据挖掘手段应用到灾害天气预测中[24]。Qin wang 等人[25]将粗糙集和人工免疫算法结合用于对数据的归约和提取来处理不平衡数据，对积雨云进行分类和预测研究；Tsegaye 等[26]采用双时间序列分析方法从众多的大气和海洋因子中确定了特定的海洋因子决定着干旱情况； Asanobu 等[27]将南北半球收集到的照片对台风进行了预报，通过聚类分析得出了台风云图模式。Cheng Tao 等[28]通过时空分析对森林火灾面积进行了有效的预测。

从目前的状况来看，数据挖掘技术的发展已经比较成熟，并且国内外众多的研究学者也对气象数据的挖掘进行了大量的研究。将数据挖掘技术应用到气象领域，针对待挖掘的应用领域和知识类型选择合适的挖掘算法等都是气象数据挖掘需要解决的问题。另外，目前气象数据的存储管理不适应数据挖掘，需要建立适合于数据挖掘的气象数据仓库，以便进行气象数据挖掘，提高数据预处理能力同时提高挖掘的高效性[29]。此外，气象数据由于收集、存储等多方面的原因导致气象数据含有的噪声较多，气象数据高维性等特点，以及气象业务领域的广泛性使得气象领域涉及到的挖掘算法众多，采用多种挖掘算法的组合方法对气象数据进行挖掘也是目前研究人员关注的重点。

1.3.2 组合分类器研究现状

分类器是机器学习、统计学习和数据挖掘等多个领域共同关注和研究的重要课题之一，分类器在医疗诊断、业务预测、入侵检测、安全防御等多个方面得到了应用。为了弥补单个分类器的缺陷或提高精度，经过学者的研究，将集成学习的思想应用于分类器中，提出了组合分类器这个概念。

在机器学习中，用集成思想构建组合分类模型的过程是：在使用数据集进行训练阶

段，先利用存在差异的多个训练集构建不同的基分类器，按照选定的集成组合方式将多个基分类器进行组合，然后在对待测样本进行分类阶段，分别用每一个基分类器对待测样本进行分类预测，通过对每个基分类器的分类预测结果进行投票（或加权投票）来确定最终的分类预测结果[30]。组合分类器模型的目的是为了提升弱分类器准确率，它以其较高的分类准确率，较强的稳定性和健壮性，以及对数据集的弱依赖性，被认为是分类方法中最有效的学习思想之一。自2000年以来以组合分类器为专题的多分类器系统（MCS）国际研讨会已举办了8届[31-32]。大量的理论和实验结果表明：多个基分类器集成的组合分类器不管在稳定性还是准确率或是复杂性上都有比单个分类模型更明显的优势[33]。目前存在的组合分类器构建方法有：(1)基分类器生成算法的不同，使用不同的分类算法(分别使用决策树[34]、Bayes[35]等)在相同的训练集上建立不同的基分类器；通过多个弱分类器组合成强分类器[36]。(2)相同的分类算法、各分类算法的参数不同(例如，对于神经网络[37]中的网络节点、动态误差、Sigmoid等参数的不同设定)；(3)相同的训练集，使用基分类器方法不同建立不同的基分类器；(4)通过从同一数据集中随机抽样产生不同的训练样本集、用同一分类方法构建不同的基分类器(常见的有装袋Bagging[38]、提升Boosting[39-40])。在这些方法中，被广泛研究和采用的是Bagging和Boosting，因为通过对训练数据集进行扰动建立的异构的基分类器，经过一定方式组合后形成的组合分类器的分类准确率很高。

通过运用Bagging和Boosting这两种策略的改进方案，组合分类模型已解决了如文本分类、数据流动态分类、入侵检测等多领域的分类问题，并取得了诸多研究结果。

Chih-Fong-Tsai 等[41]运用集成分类器对股票的投资收益进行预测，通过实验对比说明使用集成思想构建的多分类模型在投资收益预测的预测精度高于单个分类器的预测效果。Alexey Tsymbal 等[42]针对数据流的漂移问题提出运用动态的思想集成个体分类器的方式处理。郑春颖等人[43]将应用模糊积分技术加入集成分类模型中，提出了一种以集成分类模型为基础的动态模糊密度赋值方法，为了对比个体之间的差异性引入了影响因子这个因素，使得集成模型避免了过度学习问题，该文将模糊密度作为个体分类器集成的策略是它的最大亮点。Baoguo Yang 等人[44]针对传统文本数据流分类算法需要对数据进行既复杂又耗时手工人类的问题，提出了一种通过使用关键字对文本数据进行标记的分类方法，选用文本中的关键字对未标记的文本进行辅助标签，同时结合了集成分类的思想，这样便克服文本数据流中的概念漂移问题。含有噪声的数据在现实生活中已司空见惯了，针对噪声数据 Yong Wang 等人[45]提出了一种健壮的集成分

类器模型，这个模型克服了由噪声数据产生的虚概念漂移现象，能较好的处理含噪声的数据， 同时也验证了 RobustBooSting 算法能够很好的处理实概念漂移问题。上述研究成果均建立在集成分类模型上，由此可见集成分类模型的研究现在已经进入了高速发展阶段。

1.4 本论文的研究目标和组织结构

本论文的章节安排如下：

第一章简要介绍了相关背景知识和研究的意义和目的，以及国内外该领域的研究现状。

第二章介绍了气象数据挖掘的基本概念、数据挖掘理论以及相关的方法技术。以及介绍了集成学习的基本概念、算法。详细地分析了集成学习中的经典算法。

第三章基于 Bagging、AdaBoost 和 C4.5 结合的组合分类器以及随机森林组合分类器共三种分类器的模型的建立。

第四章主要对第三章建立的三个组合分类器模型应用于局部区域的气温预测中。对模型进行实验，并分析实验结果。

最后对论文进行了总结，并指出后续的研究工作。

第二章 研究领域与相关技术

本章主要介绍数据挖掘的基本概念和挖掘流程以及气象数据的特点等背景知识，然后对常用的数据挖掘技术在气象数据的应用进行了详细的阐述，最后给出本章小结。

2.1 数据挖掘简介

2.1.1 数据挖掘定义

数据挖掘的英文名为 Data Mining，但关于它至今为止还没有一个统一的定义。G.Baretsky-Shapiro 等人认为，数据挖掘是数据库知识发现的一个重要步骤，JayLouise Weldom 认为数据挖掘就是在一些历史数据的集合中寻找决策支持的过程。目前较为合理的一种定义是：数据挖掘的任务是从大量的、含有噪声的、随机的和模糊的实际应用数据中，提取出蕴藏在其中不易发现但又潜在有用的知识[1]。这定义包括几层含义：数据的来源必须是实际存在的历史数据，数据量要比较大，并且数据中含有噪声，它的形式既可以是结构化的也可以是非结构化的，甚至可以是异构型数据；发现的是用户感兴趣的知识，可以被用于商业决策、信息管理、风险评估等；发现的知识是可接受、可理解、可运用的。

2.1.2 数据挖掘过程

典型的数据挖掘过程可以概括为几个阶段：确定挖掘对象、数据定义和理解、数据质量改进、数据准备、数据挖掘算法模型开发、数据挖掘和知识评估[46]。简洁来说，就是数据确定、数据预处理、数据挖掘、知识评估。这几个阶段之间是不断反馈、循环往复的过程，通过评估和表示等过程最后得出用户满意的模式。数据挖掘的过程如图 2.1 所示：

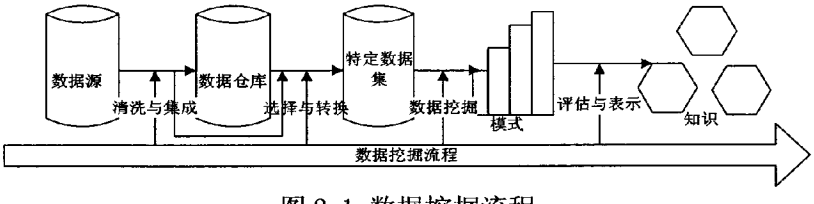

图 2-1 数据挖掘流程

Fig.2-1 Data Mining Process

（1）确定挖掘对象：确定数据挖掘的目的是进行挖掘的关键一步，如果连挖掘的目的都不清楚的话，那么就会很盲目，结果是不会成功的。

（2）数据清洗(data cleaning)：它的目的是将数据集的噪声数据、不一致数据或不完整数据进行清洗处理。使数据统一化，不含噪声，形式一致，便于挖掘。

（3）数据集成(data integration)：现实数据集往往来源不一、格式多样化，这时就需要把不同来源、格式、特点性质的数据在逻辑上或物理上有机集中整合到一个数据集中，形成一个完整的数据集。

（4）数据转换(data transformation)：将数据转换为易于进行数据挖掘的数据形式。

（5）数据挖掘(data mining)：选择相应的挖掘方法实施算法，如决策树、聚类、关联规则等，对经过转换的数据进行挖掘，挖掘出有用的模式形式或规律知识。

（6）模式评估(pattern evaluation)：对挖掘出的结果进行评价，按一定的评估标准提取出用户所需的模式知识。一般用的标准有：准确性、耗时性、泛化能力、可伸缩性、可解释性等。

（7）知识表示(knowledge presentation)：将挖掘的结果用可视化技术展现给用户。

2.2 气象数据的特点

随着科技不断引入气象研究中，近年来气象业务得到了突飞猛进的发展，气象局收集到的数据越来越大，气象数据作为大气科学数据，通过分布在各地的地面气象站收集得到的。为了对气象数据进行挖掘，了解气象数据的特性是第一步，气象数据的特性有以下特点：

（1）海量性

气象资料是我国记载最久远、保存最完善的信息资源之一，随着科技的进步，累计的数据量逐日庞大。据不完全统计中国大概有2610多个气象站台布及全国各地，每天至少收集300M以上各类气象资料。

（2）种类多样性

由于气象站遍及全国的各个地区，地区差异化、机器多样化，导致气象数据来源、表现形式和种类多种多样，如气象的种类包括高空气象数据、地面气象数据、海洋气象数据、日地物理数据等，他们所涉及的数据种类多样化，结构复杂。

（3）多维性

气象数据涉及的气象要素非常多，如地面气象站收集到的数据包括风向、风速、湿

球温度、相对温度、气压、地面温度、地温、时总日照、能见度、时总蒸发量、水汽压、露点温度等要素，每个属性构成了一个属性维，所以气象数据具有多维性。

（4）数据类型的复杂性

气象资料各要素的数据类型多样，它们可能是标称型、二元型、数值型、序数型等。如温度是数值类型，是否黄色预警是二元型。

（5）数据连续性

气象数据中的绝大部分数据类型是连续性的，如时总蒸发量、露点温度、能见度、温度、气压等要素。

（6）时空性

气象数据不仅具有较强的时间概念而且还具有空间概念，其中气象数据的精度是时间性的体现，数据的归属问题是空间性的表现。

2.3 气象数据挖掘的常用方法

数据挖掘中常用的算法是分类算法、聚类算法和关联规则算法等。应用在气象业务领域中的方法主要有分类分析、聚类分析、时间序列分析、关联规则、孤立点分析、主成分分析、回归分析、依赖关系分析等。有人采用空间聚类分析和时空关联规则挖掘了隐藏在海量气象数据中的聚类规则和时空关联规律，也有人利用孤立点分析技术挖掘和分析气象资料中的异常记录集，采用决策树模型建立了对降雨预测的模型和污染因子浓度值是否超标模型，同时也利用基于类轮廓的层次聚类方法对气象数据进行了聚类分析，都得到了很好的效果。

2.3.1 聚类分析

聚类分析(Clustering Analysis)[47]最早是由 Mac Queen 在 1967 年提出的，基本思想就是根据“物以类聚”的原理，对一组样本按照相似性归成若干个类别。简单地说，聚类就是将数据集划分为有若干个相似对象组成的多个组或簇的过程，使得同一组中对象间的相似度最大化，不同组中对象间的相似度最小化。或者说，一个簇就是由彼此相似的一组对象所构成的集合，不同簇中的对象通常不相似或相似度很低。与预测模型不同，聚类是一种无监督的机器学习方法，聚类中的簇不是预先定义好的，而是根据实际数据的特征，数据之间的相似度来定义的。

典型的聚类分析任务包括以下 5 步：

<1>模式表示（包括特征的提取和选择）。

<2>适合于数据领域的模式相似性定义。

<3>聚类或划分算法。

<4>数据摘要（如有必要）。

<5>输出结果的评估（如有必要）。

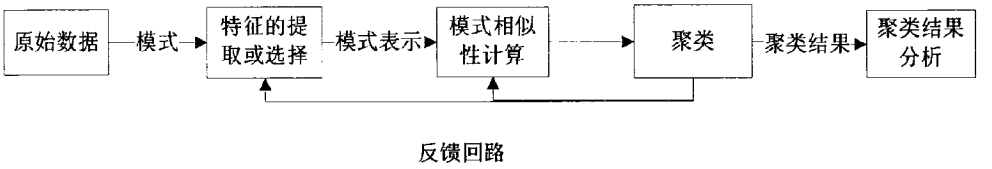

图 2-2 聚类分析的典型过程

Fig.2-2 The typical process of clustering analysis

现有的聚类算法大致可分为五种：划分聚类算法、层次聚类算法、基于密度的聚类算法、基于网格的聚类算法、基于模型的聚类算法。还有 DENCLUE 算法是结合了划分方法、层次方法和局部方法三者结合的一个综合聚类算法，STING 方法是基于网格方法和自上而下层次思想相结合的。

聚类算法是数据挖掘中应用最为广泛的技术之一，已被广泛应用于入侵检测、模式识别、反洗钱银行业务分析、商业财务数据分析、通讯等多个领域，同样在气象数据的分析中聚类分析也占着举足轻重的地位。学者们已经用聚类分析的方法解决了天气区域的划分、灾害天气评估、降水量分类等气象问题；聚类分析在多要素的分类问题中表现的尤为突出，经常会对天气图中的气象要素进行分类。

2.3.2 分类分析

分类分析[48]通过利用训练数据集构造分类器，然后使用该模型对未知类别样本进行分类确定样本的类别。数据分类过程主要包括建立模型(分类器)和使用模型进行分类，必须先建立模型才能进行未知样本的分类，构建分类器需要一个训练样本数据集作为输入数据，训练样本集是由大量的历史数据记录组成，每条记录都含有一些特征属性和一个类别标签，类标签是分类器输出。数据挖掘分类就是分析输入数据，根据训练集中的数据表现出来的特征为每个类别找到一种准确的描述或模型，然后，评估模型的分类准确度，如果准确率达到要求，就将生成的模型对未知类标记的对象进行分类。

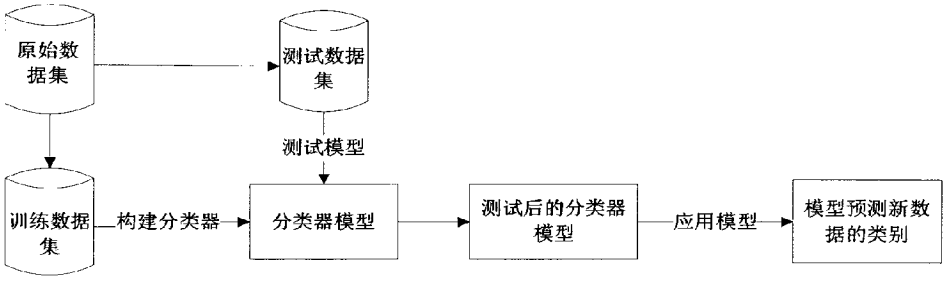

图 2-3 分类的模型建立图

Fig.2-3 The establishment of classification model

分类分析一直是数据挖掘的研究热点，也涌现出许多分类方法，构造分类器的方法主要有机器学习、统计方法和神经网络方法。如贝叶斯网、粗糙集、人工神经网络、支持向量机、决策树、K-近邻距离和基于关联规则的分类等，另外还有组合分类器算法，如装袋（Bagging）和提升（Boosting）等。

2.3.3 时间序列分析

时间序列分析[49]是指按时间顺序对一些数据进行观察分析取得的一系列观测值，这里的“时间”具有广义坐标轴的含义，既可以指按时间的先后顺序排列的数据，也可以是按空间的前后顺序排列的随机数据。例如某地区的逐月降水量，其实际记录结果按月份先后排列就是一个时间序列。时间序列中包含着很多有用的信息，对其进行分析具有重要的价值。序列分析和时间序列在气象数据中的分析时，认为气象数据是时间序列数据，它的要素在任一时刻的量值都与其前一段时间要素变化有关。根据这种关系可以建立对应的模型来描述这些气象要素之间的变化规律，然后利用所建立的模型对要素在未来时刻的值进行预测。

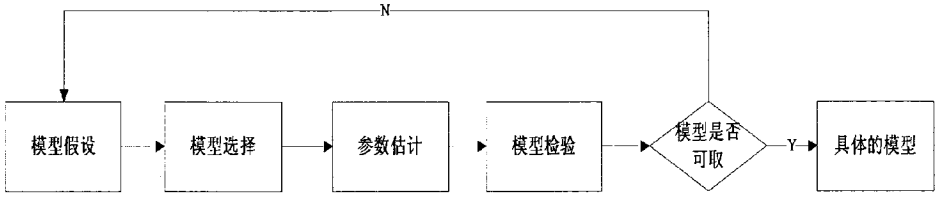

图 2-4 时间序列分析模型

Fig.2-4 The analysis model of time series

时间序列分析是统计方法体系中的一个重要分支，它直接以一个时间序列的变化规律为研究对象，通过分析时间序列数据的特征来揭示事物的变化规律。经典的时间序列分析方法有指数平滑法、趋势预测法和时间序列分解等，一些经典的时间序列分析

模型包括AR(自回归模型)、ARMA (自回归滑动平均模型)、ARCH (自回归条件异方差模型)等已被广泛应用于自然和社会科学领域。

2.3.4 关联规则

关联规则(Association Rule)[50]是一个事物与其它事物相互依存和关联关系的一种描述，两个或两个以上变量之间存在的某种规律称为关联。数据关联是数据库中一类重要的、可被发现的知识，关联规则挖掘研究首先要解决关联关系的问题，然后，如何提高关联规则挖掘的效率，如何从海量数据中进行关联规则挖掘。

关联分为简单关联、时序关联和因果关联。关联分析是指如果两个或多个事物之间存在一定的关系，那么其中一个事物就能通过其他事物进行预测，目的是挖掘隐藏在数据间的相互关系。关联规则是寻找同一个事件中出现不同项的相关性，比如在一次购买活动中不同商品之间的相关性，即关系分析就是利用关联规则进行数据挖掘。关联规则在干旱指标分析对干旱天气进行逐日预测；利用规则分析技术对灾害天气预测和降雨量预测以及地震、凝聚霜等预测取得了不少的研究成果。

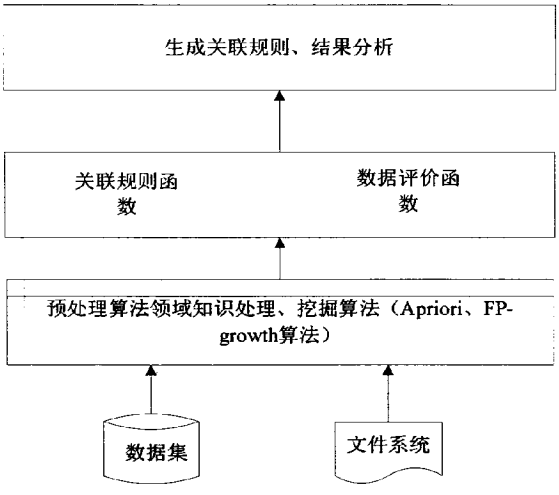

图 2-5 关联规则的数据挖掘系统体系结构示意图

Fig. 2-5 The architecture diagram of association rules in data mining

2.3.5 孤立点分析

孤立点分析[51]最早是由Hawkins提出的:孤立点是数据集中那些小模式数据，这些数据并非随机孤立点，而是由不同机制产生的。孤立点分析通常又被称为孤立点检测或异常点挖掘。孤立点分析的基本方法就是寻找观察结果与参照数据之间有差别的样本，主要工作就是在给定的数据集中定义什么样的数据是不一致的，另外使用合适的

方法来挖掘孤立点。孤立点数据通常被人们忽视于海量数据中，或者被忽略了其存在的特殊意义。孤立点数据通常被认为是噪声数据，所以在对数据预处理的过程中会使用孤立点分析剔除对数据挖掘结果有一定影响的孤立点数据（噪声数据），但不是所有的孤立点都是没有用处的，有时孤立点反而是隐含重要信息的数据。在气象数据中，孤立点数据往往可能是某种灾害天气的数据，因此，通常对气象数据中的孤立点进行分析很有可能发现气象灾害等。

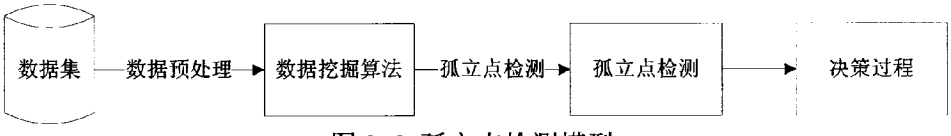

图 2-6 孤立点检测模型

Fig. 2-6 Outlier detection model

目前已有的传统孤立点分析方法主要有基于统计的方法、基于距离的方法、基于密度的方法、基于关联的方法和基于聚类的方法，另外还有基于属性的方法和基于时序、空间和时空的孤立点分析等。孤立点分析基本上已不再是单一的分析方法，而是与经典的数据挖掘算法相结合使用，这样可以更有针对性地发现孤立点。气象资料分析之前一般要经过多道预处理过程，过程中产生的一些疏漏使得部分数据发生异常，对于这些异常数据的处理是相对容易的，而对于事实存在的异常气象数据是很难处理的。

2.4 本章小结

数据挖掘是近年来伴随着信息技术的快速发展和数据量成倍增涨所激发的新技术和自动工具需求的成果。数据挖掘以智能的方式将大量的数据提取成有用的信息和知识。本章简要的介绍了数据挖掘的概念、数据挖掘的过程、数据挖掘的方法以及数据挖掘的应用等，同时也介绍了数据挖掘技术在气象业务中应用到的算法，对常用的算法进行了阐述。

第三章 基于决策树的组合分类器建模

分类器模型一直是数据挖掘中研究的重点领域之一，其中决策树算法是众多分类算法中较为常用的算法之一，因为它直观、明了能够给出清晰的树状模型，而且是不稳定的算法，属于弱分类器。组合分类器能够有效的提高分类准确率，尤其在处理复杂多样的大量数据中优势明显。本文在分析了气象数据的基础上，对气象数据进行预处理，并针对气象数据建立了三个组合分类器模型。

3.1 气象数据集描述

本文的实验数据来源于广州市某地区局部区域气象站的实际地面观测数据，气象数据是从2011年1月份到12月份全年的气象数据，气象站收集的是当地全年每天0-24小时每小时的即时数据共8760条，如果数据项是缺失值，则其值为空，任何一条数据的数据项包括风向、风速、时总雨量、最高气温、湿球温度、水汽压、本站气压、最高本地气压、最低本地气压、草面温度、5cm草面温度、10cm草面温度等共54项。

3.2 气象数据预处理

据目前统计，数据挖掘技术并没有完全运用到气象实用业务系统中，由于气象数据的复杂性导致对气象进行预测和规律的发现等方面还处于研究阶段。气象数据的质量是影响数据挖掘知识精度的关键因素，然而气象真实数据往往含有许多噪声、不完整、不一致的数据，对气象数据进行预处理是非常必要的，数据预处理的好坏直接影响到数据挖掘的结果，对气象数据进行预处理工作能够将数据统一，使得数据挖掘过程更高效、更容易。

3.2.1 数据清洗

气象数据在收集的过程中会由于设备的限制或漏洞等因素，使得数据出现不完整、有噪声等问题。数据清理的目的包括消除原数据中的噪声数据、填补缺失值以及对脏数据的清洗。针对不同的数据，数据清理的方法是不一样的。

（1）遗漏数据处理

由于收集数据机器在收集数据的过程中出现各种小问题，导致遗漏一个或多个属性值的情况是常见的现象，而缺失属性会导致数据不完整。

在分类任务中，对于缺少类标号的数据，通常的做法是删除它们；气象因素属性值不是以随机数形式出现的，它们是时间序列数据，受季节、地形、地势等多重影响，缺失值的处理不能简单的以均值或指定值的形式进行替换，应该应用气象分析中的最优插值方式进行。为了尽量使用原始数据集中的有效信息，对于缺失值的处理采用了4种方式：

a）若某个属性缺失的值超过一定额度时，则在整个数据集中不考虑该属性列。

b）判断相邻的数据记录有没有相应属性的信息，如有则使用相邻数据项的加权值替换缺失值。如果相邻的项也缺失信息，则不用这种方法处理。

c）根据气象要素的实际具体情况进行判断，如大气压等信息，可以参考历史数据中同一天同一时间段的属性统计值，同统一时间段的平均值进行替换。

d）略缺失值的记录项，由于数据集数据量比较大，忽略一部分数据记录并不影响整体组合分类器模型的构建和准确性。

（2）偏值和极值处理

偏值和极值离了正常范围的值，即偏离均值数倍方差值的数据。偏值是偏离均值达3倍标准差的数据，极值表示偏离均值达5倍标准差的数据。偏值和极值表示的是单维属性在统计意义上的特性。但气象数据多维复杂的特点，导致我们不能简单的去判断他们，应该在多维数据上，通过查找奇异点的方法或孤立点技术，通过气象专家人机交互方法，判断是否是极端天气现象。若是潜在的数据错误，则进行相应的修正。

（3）噪声数据的平滑处理

气象基站的设备故障和数据录入的过程中人为的疏忽或者数据传输过程中的某些错误都会导致数据噪声的产生。对于含噪声的数据应该进行平滑，常用的方法主要有分箱（Bin）方法、聚类方法、人机结合检查方法、回归方法等四种。

3.2.2 特征归约

气象数据是时序数据，绝大多数时序数据都是多维数据，对于多维序列数据而言，如果只是考虑其中的一维序列情况来进行挖掘往往会引起较大的偏差。比如对某一地区月平均气温的预报，如果仅仅根据历年的月平均气温气象资料来建立一维的数据预测模型，预测结果往往不尽人意、相差甚远。因为气温是动态的，跟很多其他气象因素相关，并不是一个独立的因素，为此在进行数据挖掘时，往往会将与其联系的影响因子加以考虑，如气压、湿度、时总蒸发量、水汽压等，对主要的影响因子进行综合

分析和判断，以提高预测的准确率。

要想得到有效的预测效果，如何选择预测因子是一个非常关键的环节。预测因子的选择是数据归约的过程，要以当地气象局的预测经验和相关的气象业务领域内的知识为基础，从大量的气象资料中选取对于我们预测最有效的预测因子。在选取有效的预测因子时，要逐一地对所选的预测因子与预测目标进行单相关分析，同时从天气学、气象学等角度出发分析这种相关是否符合事实，是否具有实际意义。两者进行结合选出最有效的预报因子。同时也可以选择许多观测站的气象或各个高空站的各个高度上的气象要素作为预报因子。

本文主要是对局部区域的气温，鉴于原始数据集中的属性较多，首先使用主成分分析法对气象数据集进行特征约减和特征优选，获得和预测目标气温因子相关度高的影响因子。

主成分分析方法是一种利用降维思想，将多个指标化为少数几个不相关的综合指标（即主成分）的统计分析方法[52]。它常常用来分析多指标的数据之间关系和变化趋势。在多属性的研究过程中，往往由于多属性之间存在一定程度上的信息重叠现象，当变量较多时，在高维空间中研究样本的分布规律是很繁琐的事，利用主成分分析能够从众多的属性中找出一些综合因子代表，去除对研究目标无意义或不相关或重复的属性，减少维度，从而简化属性。

典型的主成分分析方法一般包括以下几个步骤[52]：

（1）将原始数据值进行标准化，构建样本阵，对样本阵元进行标准化变换；

（2）对标准化阵计算相关系数矩阵；

（3）将标准化后的指标变量转换为主成分；

（4）对主成分记性综合评价，计算各主成分的贡献率和他们的累积贡献率，最后通过贡献率确定主成分。

通过主成分分析法和结合气象部门的经验知识发现气温主要受风速、相对湿度、气压、水汽压、露点温度、地面温度、时总蒸发量 7 个因素的影响。因此，从某地区气象数据集中提取气温及对气温有影响的属性共 8 个属性，去掉一些异常或者属性缺失的气象数据，生成原始数据集。将风速、相对湿度、气压、水汽压、露点温度、地面温度、时总蒸发量作为分类属性，气温等级作为类属性。

表 3-1 华南地区要素正常范围

Table 3-1 Normal range of temperature in Southern China area

序号	1	2	3	4	5	6	7	8
要素	风速	相对湿度	气压	水汽压	露点温度	地面温度	时总蒸发量	气温
单位/精度	0.1m/s	1%	0.1hpa	0.1hPa	0.1	0.1	0.1mm	0.1
华南地区正常范围	0～102	10～100	9500～11000	30～403	小于等于当地最低温度	0～600	0～580	0～400

3.2.3 数据集成

数据挖掘经常需要对数据进行聚合，将两个或多个数据源中的数据，存放在一个一致的数据存储设备中。数据挖掘中的原始数据往往来源于不同的数据集，结合在一起并且形成一个统一的数据集合，为后期的数据过程提供良好的数据基础。在数据集成的过程中我们一般都要考虑三个问题：模式集成、属性是否冗余、数据值是否冲突。由于气象站分布不一，采集的信息不一，导致气象数据来源、表现形式和种类多种多样，这样对气象数据进行研究时往往要进行数据集成。

3.2.4 数据变换

数据变换是将数据转换成合适于挖掘的形式，将数据转换或归并已构成一个适合数据挖掘的描述形式。数据变换一般涉及平滑、聚集、数据泛化、规范化、数据离散型等处理。

（1）数据泛化：所谓数据泛化处理是用更抽象（更高层次）的概念来取代低层次或数据层的数据对象。对于数据值的属性，如温度属性，就可以映射到更高层次概念，如：寒冷、温暖和炎热等。

中国气象局将气温分为以下几个等级[53]：

表 3-2 气温等级划分标准

Table 3-2 The division standard of temperature scale

极寒	−40℃或低于此值	微温凉	12～13.9℃
奇寒	−35～−39.9℃	温和	14～15.9℃
酷寒	−30～−34.9℃	微温和	16～17.9℃
严寒	−20～−29.9℃	温暖	18～19.9℃
深寒	−15～−19.9℃	暖	20～21.9℃
大寒	−10～−14.9℃	热	22～24.9℃

续表 3-2

小寒	−5～−9.9℃	炎热	25～27.9℃
轻寒	−4.9～0℃	暑热	28～29.9℃
微寒	0～4.9℃	酷热	30～34.9℃
凉	5～9.9℃	奇热	35～39℃
温凉	10～11.9℃	极热	高于 40℃

结合实验数据集的情况和当地气候的分布情况可知，当地的气温比较平稳，地处亚热带季风气候，当地气温从未低于 0℃，而且高温也未高过 39℃，为了便于实验分类和预测，我们将本次实验数据集将温度分为以下 5 个等级：

表 3-3 温度等级表

Table 3-3 Temperature scale

等级	气候情况	温度范围	数据集中的数目
A	寒	0～4.9℃	50
B	凉	5.0～13.9℃	1613
C	温	14.0～17.9℃	1024
D	暖	18.0～21.9℃	1268
E	热	22.0～39.0℃	4789

为了便于实验对比，我们将气温的等级 A、B、C、D、E 分别用 1、2、3、4、5 标称数值来代替。

（2）数据规范化

为了缩小因属性之间的值大小不一造成挖掘结果的偏差，将气象数据属性按比例进行缩，实验数据首先要进行规范化处理，规范化最大好处就是消除量纲的影响，未规范化，数值取值较大的指标会削弱取值较小的那些指标对模型的影响。通常我们将每一个指标规范到[-1，1]或[0.1]区间。

常用的规范化方法有最大最小值法、Z-score 标准化、指数函数法等。

最大最小值标准化方法是对原始数据的线性变换，假如气象属性 A 的一个原始值为 x_i，它的最大值为 $x_{\max}$，最小值为 $x_{\min}$，通过最大最小标准化变换到[$x'_{\min}$， $x'_{\max}$]区间中，如果 $x'_{\min}$=0， $x'_{\max}$ =1，则表示数据标准化到[0，1]区间。标准化处理公式为：

$$x'_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}\left(x'_{\max} - x'_{\min}\right) + x'_{\min} \tag{3.1}$$

最小化—最大化标准化处理方法保留了原数据之间存在的关系，在转换的过程中并没有考虑到数据分布的正态性，会导致转化后数据均匀量化，很多值会趋于相同，特

别是偏值和极值对结果影响比较严重。

Z-score 标准化是基于原始数据的均值和标准差进行的数据标准化。其计算公式是：

$$x_i' = \frac{x_i - \bar{A}}{\sigma} \tag{3.2}$$

其中，$\bar{A}$ 是属性 X 的均值，σ 是它的方差，当属性 X 的最大值和最小值未知的情况下，这种方法是最有效的。

3.3 建模体系

对于一个组合器而言，必须先要选择好基分类器和集成方法，基分类器可以用分类的方法，数据挖掘中的分类方法有决策树、贝叶斯、基于关联规则、支持向量机等分类算法，经典的分类算法在不同的领域中广泛应用，决策树分类算法用于金融分析、医疗诊断、个体信用度等领域；支持向量机分类方法应用于模式识别、文本分类、基因分析、语言识别、回归分析等领域；神经网络对噪声具有很好承受能力，导致神经网络广泛应用于字符识别、生物学、人脸识别等多个领域。常见的组合分类器的构造方法有投票表决法(Voting)、加权多数法(Weight Majority)、堆积泛化法(Stacking)、Bagging(Bootstrap aggregating)、AdaBoost(Adaptive Boosting)等。

本文的构建组合分类器主要包括以下几个模块：

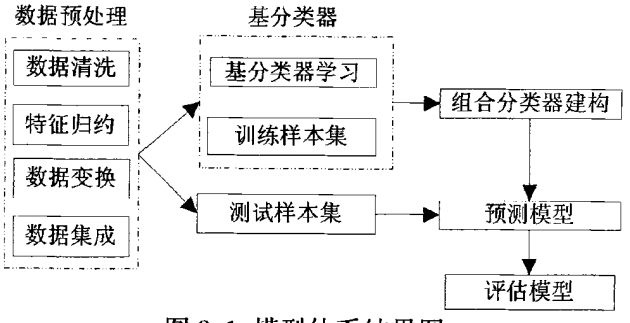

图 3-1 模型体系结果图

Fig.3-1 Model System structure Chart

（1）数据预处理：主要是对数据进行清洗、集成、特征归约和数据变换，处理数据中的噪声、不完整数据、不一致数据以及对缺失属性数据的处理，为后期的数据挖掘提供良好的数据基础。为了试验考虑，按一定的方法将数据集分为训练数据和测试数据。

（2）从数据集中选择出来的训练样本对决策树学习进行训练构建基分类器。

（3）组合一定的集成组合方法将各基分类器进行组合构造组合分类器，形成分类预测模型。

（4）将测试样本集应用到构建好的组合分类器中，对其进行测试，评估模型的准确率等方面的性能。

3.4 组合分类器的构建过程

分类器是用于分类的模型，它的目的是能够利用针对对于数据集建立好的分类器模型用于后期的分类。提高分类器的准确率一直是分类器研究的目标，分类器组合方法则是将多个不同的基分类器进行组合对同一个实例进行分类，最终利用多个基分类器的分类结果按一定的方式进行判断确定最终的结果，从而提高分类器的精度。分类器组合方法能够有效的提高分类器的性能，本文提出了一种以决策树分类器作为基分类器，并以 Bagging、AdaBoost 算法作为分类器集成的方法，最后构建组合分类器。

3.4.1 基分类器设计

决策树学习是一种以事例为基础的归纳学习算法，它是从一组无规则的事例中推理出一种用树状结构表示的分类规则，通常用来形成分类和预测模型。典型的决策树构建包括两个阶段：第一阶段是利用训练样本集建立决策树模型，这是一个机器学习过程。第二阶段是用已建好的决策树对新的数据集进行分类预测。

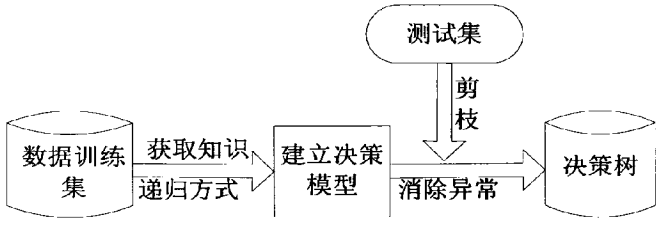

图 3-2 决策树的生成过程

Fig.3-2 The generation process of decision tree

在使用决策树模型对待分类的样本进行分类预测时，从根节点开始逐步对样本的属性值进行测试，对比样本的属性值是否在节点对应的值之内，并沿着对应的分支向下走，直到到达某个叶节点，此时落到的叶子节点代表的类别便是该样本的类别。所以决策树方法对分类问题而言其关键点在于根据训练数据构建决策树分类器，如果构建的决策树是良好的，那么预测的准确性也就高。

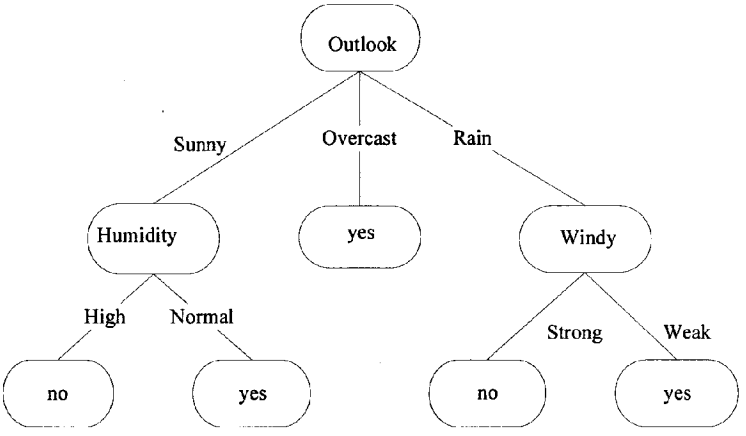

图 3-3 典型的天气数据决策树分类器模型

Fig.3-3 The decision tree classifier model of typical weather data

构造决策树的方法有多种算法，不同的决策树方法之间的主要差异在于节点之间的“差异”衡量不同。常见的决策树算法有 ID3、C4.5 、CART 和 SLIQ 等[1]。但 ID3 算法只能处理连续属性，C4.5 算法既能处理连续属性也可以处理离散属性，CART 算法主要强调树是二叉树的[54]。

ID3 是基于信息熵的决策树分类算法[1]。对于集合 S 的信息熵的计算公式：

$$x_i^{'} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}\left(x_{\max}^{'} - x_{\min}^{'}\right) + x_{\min}^{'} \tag{3.3}$$

那么对于训练集 T 按照一个属性检验 X 的输出而进行分成 n 个子集， 所需信息可通过这些相应的子集的熵的加权和求得， 计算公式：

$$Info_x(\mathrm{T}) = \sum_{i=1}^{k}((|\mathrm{T}_i|/|\mathrm{T}|)\times\log_2(\mathrm{T}_i)) \tag{3.4}$$

信息增益 Gain(x)的计算公式：

$$Gain(x) = Info(T) - Info_x(T) \tag{3.5}$$

C4.5 算法继承了 ID3 算法的优点并进行了改进：用信息增益率代替信息增益来选择属性，在树建造过程中进行剪枝，能够完成对连续属性的离散化处理，在精简决策树的同时，提高了分类的准确率能够对缺省数据进行处理等[55]。

训练集 T 按离散属性 x 的 n 个不同取值，划分为 T1、T2、…Tn 共 n 个子集， 则用 x 对 T 进行划分的信息增益率为：

$$Gain\_ratio(x) = Gain(x) / Split\_Info(x) \tag{3.6}$$

其中：

$$Split\_Info(x)=-\sum_{i=1}^{n}((|T_i|/|T|)\times Log_2(|T_i|/|T|)) \quad (3.7)$$

使用增益比率 Gain_ratio(x)建立决策树比使用 Gain(x)建立决策树要健壮，增益比率也是采用选择使给出的比率最大的检验属性。

针对于气温预测而言，决策树的构建过程，输入的将是与气温相关联的属性特征所组成的气象数据集和各属性特征，输出向量是预测项气温所属的级别。

决策树的建树流程图如图 3-4 所示，其中 S 表示气象训练样本集，A 表示气温等级分类样本集合，N 表示一个决策树中的节点。

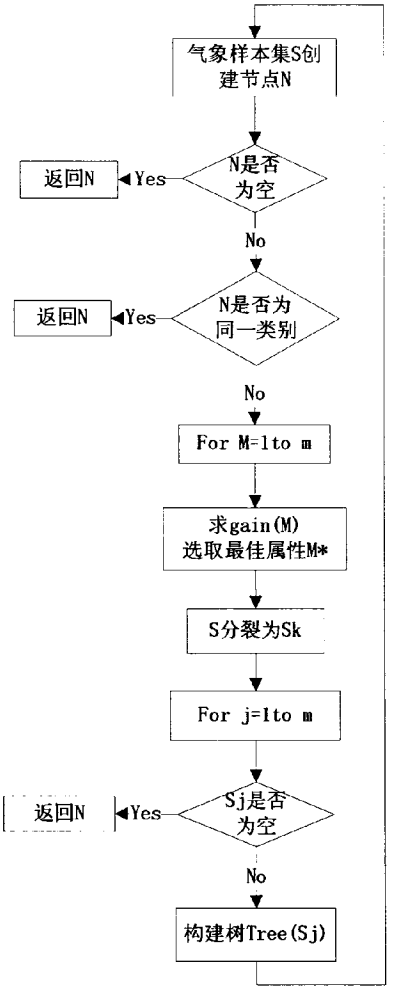

图 3-4 决策树的建树流程

Fig.3-4 The process model of decision tree

C4.5 算法构建气温决策树的构建过程：

输入：气象数据训练样本 samples；候选属性的集合 A={风速、相对湿度、气压、水汽压、露点温度、地面温度、时总蒸发量}。

输出：一棵气温等级决策树

(1) 创建根节点 N;

(2) IF T 都属于同一类 C，则返回 N 为叶节点，标记为类 C;

(3) IF A 为空 OR T 中所剩的样本数少于某给定值则返回 N 为叶节点，标记 N 为 T 中出现最多的类。

3.4.2 用于气温预测的 Bagging 方法

从大小为 n 的原始数据集 D 中，分别独立随机地采用有放回采样的方式进行抽样 N 个数据形成子样本集，抽样的过程是相互独立的，直到产生多个自助数据集，所以这里可以加入协同并行的思想进行数据采样。然后利用每个自助数据集作为训练集训练对应的“基分类器”。本论文利用 3.4.1 节所讲的决策树作为基分类器，在分类的过程中，所有的基分类器都会对测试样本进行分类，最后将各个基分类器的结果采用投票法来决定最终的分类结果。

设样本的总类别为 C；组合分类器模型有 m 个基分类器 Ci(i=1，2，3，…，m)，基分类器的预测函数为 $f_i(x)$，这里的 x 是输入向量。

（1）简单的多数投票

模型的最终预测分类结果可以表示为：

$$h_f(x)=\underset{c\in\{1,2,3,\ldots C\}}{\operatorname{argmax}}\sum_{i=1}^{m}1_{\{f_i(x)=c\}} \tag{3.8}$$

（2）加权投票法

加权法就是对每个基分类的结果进行加权处理，假如基分类器 C_i 的权重为 β_i，则模型的最终分类结果可以表示为：

$$h_f(x)=\underset{c\in\{1,2,3,\ldots C\}}{\operatorname{argmax}}\sum_{i=1}^{m}\beta_i * 1_{\{f_i(x)=c\}} \tag{3.9}$$

其中 β_i 反映出基分类 C_i 的重要性。

为了验证组合分类器采用投票法进行结果裁决能够提高模型的整体性能，那基分类器应该满足什么要求呢？如果组合分类器模型是由 m 个基分类器组合而成的，样本

的类别是2，各个基分类器的输出结果是相互独立，并且假设各个基分类器的分类器准确率均为P，那么进行简单的投票得到的分类准确率为：

$$P=\sum_{i=0}^{(m-1)/2} C_i^m p^{(m-i)}(1-p)^i \qquad (3.10)$$

由此可知，如果每个基分类器的分类准确率 $p<0.5$，则模型的整体分类准确率 P 将会随着 m 的增大反而减小，$P<p$。所以为了保住模型的整体准确率高于各个基分类器，各基分类器准确率 p 必须大于 0.5，这样才能体现组合模型的意义所在，而且模型的整体分类准确率 P 会随 m 的增加而增大。

表 3-4 给出了在各个基分类器相互独立的构成的组合分类器模型中，当 m=3，5，7，9 和 p=0.6、0.7、0.8、0.9 情况下，模型的整体分类准确率。

表 3-4 模型分类器的准确率

Table 3-4 the classifier accuracy of model

	m=3	m=5	m=7	m=9
p=0.6	0.6480	0.6826	0.7102	0.7334
p=0.7	0.7840	0.8369	0.8740	0.9012
p=0.8	0.8960	0.9421	0.9667	0.9804
p=0.9	0.9720	0.9914	0.9973	0.9991

本文将对基分类器的权重进行调整优化以提高模型的整体泛化能力。

Bagging 算法的基本原理是：通过给定一个学习算法来训练样本进行构建多个弱分类器，得到一个预测函数序列，最后通过投票法得出最好的分类结果。Bagging 算法实现了将多个弱分类器集成得到一个强分类器，同时具备很强的泛化能力和稳定性。用 Bagging 来构建一个组合分类器模型如图 3-5 所示。

由于基分类器的学习算法对训练数据越敏感，Bagging 的效果越好，因此对于决策树、支持向量机和人工神经网络这样的弱分类器算法，Bagging 的效果越好，因此对于决策树这样的弱分类算法 Bagging 相当有效。所以本文采用决策树作为基分类器。由此本文针对气象数据，利用决策树作为基分类器用 Bagging 集成思想进行组合形成 Bagging 气温预测模型，模型结构如图 3-6 所示：

模型中 D1、D2、D3、…、Dn 是在 D 数据集中进行 Boostrap 抽取选择的子样本集；Boostrap 的思想就是有放回的随机抽取得到的训练样本。

C1、C2、C3、…、Cn 是一组基分类器，对应的数据集中的训练数据用 C4.5 算法构

建而成的决策树分类器。各基分类器之间是相互独立的，所以可以采用协同并行的思想进行训练。

C+分类器融合各个基分类器的预测结果，每个分类器都可以得到一个结果，n 个分类器的结果构成了一个函数序列，将这个序列进行等权重投票，得票数最高的类别为最终组合模型的分类结果。

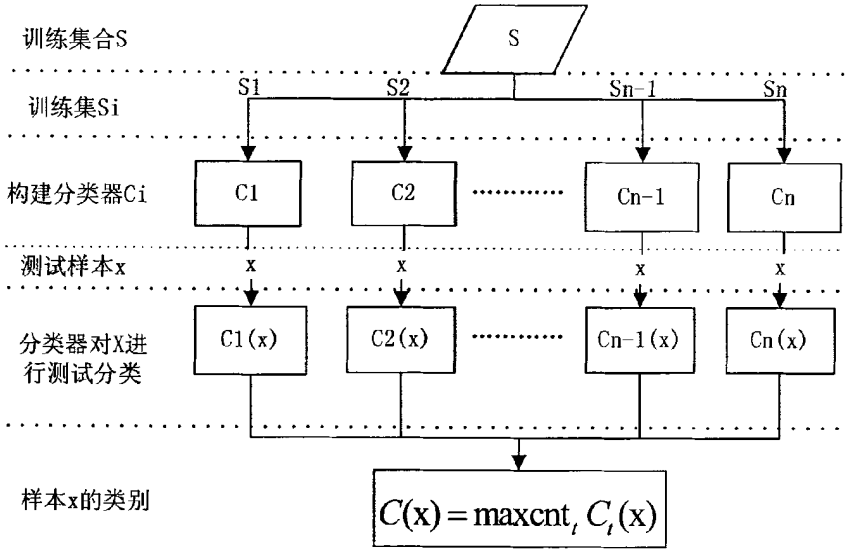

图 3-5 典型的 Bagging 组合过程

Fig.3-5 The process of Bagging

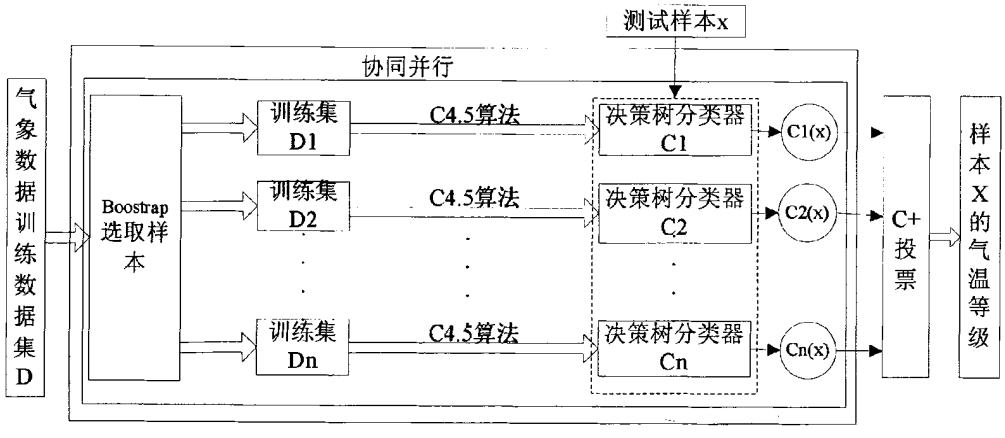

图 3-6 Bagging 气温预测模型

Fig.3-6 Bagging temperature prediction model

Bagging 气温预测算法的主要步骤描述如下：

步骤一： 初始化 Bagging 算法，设定 Baggging 算法的最大迭代次数 t，即弱分类器学习的个数，选定弱学习算法，本文就选用决策树 C4.5 作为基分类器算法，同时设定弱学习算法的训练参数，预处理气象数据集，将数据集分为训练集 D 和测试集 T 两部分；

步骤二：以指定的概率从训练集 D 中协同并行地选取训练样本子集 D_i，$i \in \{1,2,.....,t\}$，作为基分类器决策树的训练样本子集 D_i；

步骤三：将训练样本子集 D_i 输入到对象的弱分类器（决策树）算法中训练构建，得到对应的决策树分类器 Ci；

步骤四：检查当前集成学习 Bagging 算法是否达到算法的最大迭代次数 t，若已达到，则执行第五步，否则返回第二步；

步骤五：计算决策树序列 C1，C2，C3，...Cn 预测结果按投票方法组合的结果：$C(x) = f(C_1(x), C_2(x), ..., C_n(x))$，即为强学习机 C 的预测结果 C(x)。

3.4.3 用于气温预测的 AdaBoost 算法

AdaBoost 算法的基本思想是利用大量分类能力一般的弱分类器，通过一定的方法叠加起来构成一个分类能力比较强的组合分类器。典型的 AdaBoost 组合分类器的构造和分类预测过程如上图 3-7 所示。对每个训练样本都分配一个权重，表明它被当前分量分类器选入训练集的概率，每次迭代改变训练样本的分布；样本的分布取决于样本在上次的训练中预测函数的输出与期望数据的差异来对权重进行调整，被分类错误的样本权重得到提升，正确分类则减少权值；最终的结果是弱分类器的加权组合[56]。构建模型的关键点是对误判样本的权重调整。

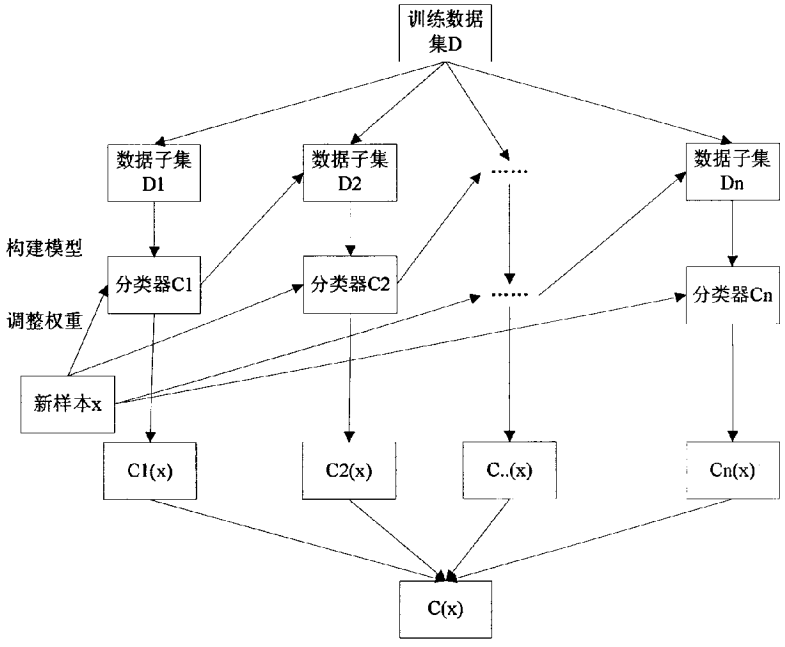

图 3-7 AdaBoost 组合分类器的构造过程和分类过程

Fig.3-7 Construction process and classification process of AdaBoost classifier

目前 AdaBoost 算法不多，主要有解决两类问题的 AdaBoost 算法、解决多类单标签问题的 AdaBoost.M1 算法和解决多类多标签问题的 AdaBoost.M2 算法。多类单标签的 AdaBoost.M1 算法描述如下[56]：

输入： 样本训练集 S，元学习算法（弱分类算法）C，训练迭代的次数 M

输出：组合分类器模型

训练步骤：

把样本训练集 S 中的每个样本的权重统一初始化为 1/N

For i=1 to M

获得训练集 S 中样本的权重 S_i

根据 S_i 权重情况选取新的训练样本训练一个 C 类型的基分类器 H_i

计算 H_i 的分类误差 $error(H_i)$

$error(H_i)=\sum_{j=1}^{N} w_j \times err(x_j)$ ，其中如果样本错误分到其他类中，$err(x_j)=1$，否则为 0

If $error(Hi)>0.5$ then

设置 M=i-1

取消循环

Endif

For S_i 中每个正确分类的样本 do

每一个样本的权重均乘以 $error(H_i)/(1-error(H_i))$

Endfor

将每个样本的权重进行规范化

$$w_i = \log\frac{1-error(H_i)}{error(H_i)}$$

EndFor

分类步骤：

输入一个待测样本 x

For i=1 to M do

$F = M_i(x)$

把训练过程中得到的 w_i 加到对应的类 f 的权重中

Endfor

返回具有最大权重的类（便是样本 x 的类别）

为表示统一性和便于实验结果的比较，弱分类器选为 C4.5 决策树算法。组合分类器算法步骤如下：

1. 从气象数据集中选择样本数据，从样本空间中随机选择 m 组训练数据，初始化测试数据的分布权重 $D_t(i)=1/m$，根据样本输入的维数和最终的结果类别构建决策树分类器。

2. 弱分类器分类预测。训练第 t 个弱分类器时，用训练数据训练 C4.5 决策树并且预测训练数据输出，得到预测序列 g(t)的预测误差和 e_t，误差和 e_t 的计算公式为：$e_t = \sum_t D_t(i)\, i=1,2,...,m(g(t)\neq y)$ 。其中，g(t)是预测分类的结果；y 为样本的实际类别。

3. 计算预测序列权重。根据预测序列 g(t)的预测误差和 e_t 计算序列的权重 a_t，权重计算公式为 $a_t = \frac{1}{2}\ln(\frac{1-e_t}{e_t})$ 。

4. 调整测试数据的权重。根据预测序列权重调整下一轮训练样本的权重，调整公

式为：

$$D_{t+1}(\mathrm{i})=\frac{D_t(\mathrm{i})}{Z_t}*\begin{cases}\exp[-\mathrm{a}_t] & if \quad y_i=h_t(\mathrm{x}_i)\\ \exp[\mathrm{a}_t] & if \quad y_i\neq h_t(\mathrm{x}_i)\end{cases} \qquad \mathrm{i}=1,2,...,\mathrm{m} \qquad (3.11)$$

式中，Z_t 是归一化因子，目的是在权重比例不变的情况下，使得分布权重值和为 1。

5. 确定强分类函数。训练 t 轮后得到 t 组弱分类函数，有 t 组弱分类函数组合得到强分类函数 H(X)。

$$\mathrm{H}(\mathrm{x})=\underset{y\in Y}{\operatorname{argmax}}\sum_{t:h_t(\mathrm{x})=y}\alpha_t \qquad (3.12)$$

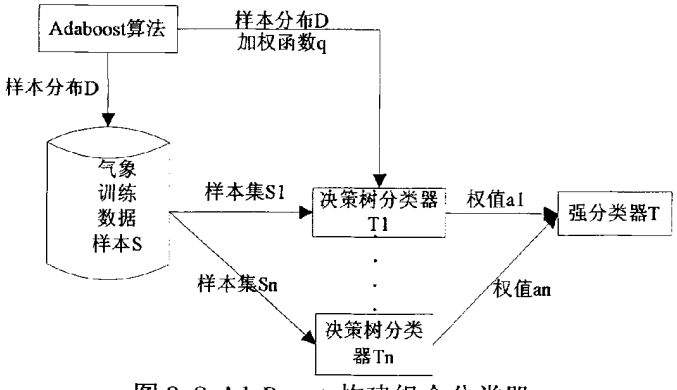

图 3-8 AdaBoost 构建组合分类器

Fig.3-8 Construct combination classifier with Adaboost

3.5 随机森林

随机森林模型[67]是一种组合分类器算法，随机森林分类算法（RF）是利用 Bagging 方法形成训练集，采用分类回归树法 CART 作为其元学习算法构建基分类器，最后用简单多数投票法（分类问题）或简单的平均方法（回归问题）确定最终的分类结果。

3.5.1 随机森林对噪声的容忍度

当样本数据量过大时，不管是数学中的统计分析还是模型的构建都可能要考虑到数据的噪声问题。在利用模型进行气温等级预测时，如果气象数据的输出标识别改变，RF 模型的测试准确性是否会受到影响，也就说随机森林模型对噪声的容忍度如何？Dietterich 曾指出，当随机改变训练数据集中的部分数据的输出标识时，(即给训练数据集中加入噪声数据)，AdaBoost 算法的准确率会显著下降，而 Bagging 算法和随机森林对噪声就有很好的容忍度。原因是因为 AdaBoost 算法是通过增加误分样本的权重来加大对误分样本的关注，提高训练数据的整体准确率，当训练数据带有噪声时，

AdaBoost 就会重点关注学习噪声数据，从而导致对噪声数据的过拟合现象，反而降低了模型的真实准确率。而 Bagging 方法和随机森林方法在特征分裂是含有很大的随机性，并不会对噪声数据进行加强学习，数据中的噪声数据能被学习的概率较小，因此这两种方法对噪声具有一定的容忍度。

气象数据不可避免不含噪声数据，如何有效的利用算法在数据挖掘的过程中不受噪声数据影响和干扰是智能学习领域中的焦点。而随机森林正好具备这种优势，因此其应用的领域将会更加广泛。

为了验证随机森林具有克服噪声这一特性，本文设计了如下实验：

（1）设定 ntree=10，mtry=3;随机的改变气象数据集中数据的目标变量“气温等级”使其成为噪声数据，改变数据样本量占训练样本集的比例分别为 0%，5%，10%，20%，30%，50%。测试集中的数据不改变。

（2）利用加入了噪声数据的 训练数据进行训练构建气温预测随机森林模型。

（3）用测试集的数据测试已构建的模型，检验其准确率。

（4）对比仿真结果进行分析。

仿真结果如下表 3-5 所示：

表 3-5 训练集数据的噪声变化与模型的误差率

Table 3-5 Noise in training set change and model's Error rate

训练集噪声数据比例（%）	0	5	10	20	30	50
训练误差率（%）	11.23	13.56	17.98	24.16	33.43%	53.45
测试误差率（%）	10.35	12.34	14.93	16.77	18.35	27.33

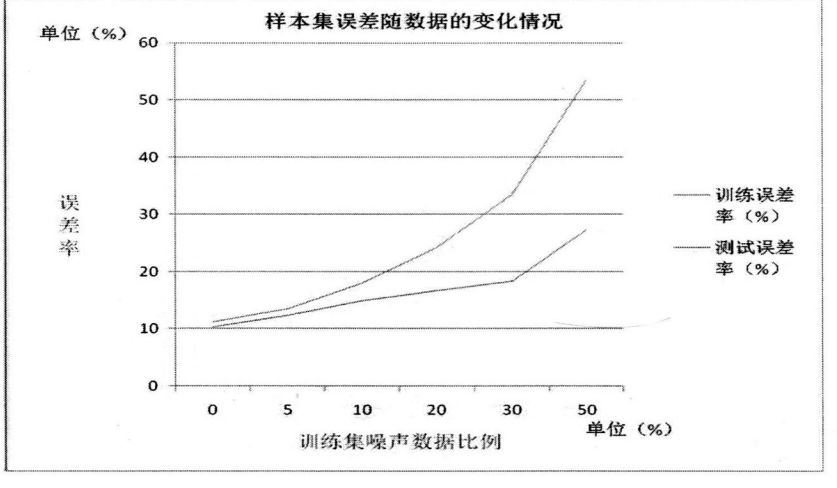

图 3-9 样本集误差随噪声数据的变化情况

Fig.3-9 Sample error with noisy data set changes

通过表和图的数据分析可知，当数据集中没有加入噪声时，训练集数据误差率为11.23%，测试集数据的误差率为10.35%。当数据集中加入了5%的噪声数据进行扰乱时，训练集数据误差率变为13.56%，测试集数据的误差率为12.34%，随着噪声的不断加大，训练集数据的误差率和测试集数据都有了提高，但是两者的变化率是不一样的，但噪声比例为50%时，训练集数据的误差率是53.45%，测试集数据的误差率为27.33%。由此可见，测试集受噪声的影响较小，随机森林算法对噪声有较好的容忍能力，所以利用RF建模时，可以不需要对样本数据进行过多的预处理，可以尽可能的保持样本数据的原始数据进行实验，以提高预测的准确率。

3.5.2 随机森林模型参数的选择

随机森林模型需要对模型中的一些参数进行调整，使得模型的误差率尽可能的小，以求达到最佳的性能。随机森林模型中含有的参数较多，主要的两个参数是ntree（森林中树的数目）和mtry（每个节点处候选特征的个数）。对这两个参数进行调整，就能使模型达到较好的性能。ntree是RF模型中树的数目，只要让森林的整体误差率趋于稳定状态就可以了，一般来讲，只有当ntree足够大时，才能确保模型的误差接近上界值。

参数mtry是指在构建随机森林模型中的决策树除了根节点、叶子节点外的其他节点处要随机选择属性特征的数目，mtry的默认值为sqrt(M)，M是属性的个数，在我们研究的气象数据中M=7。

为了确定模型中mtry的最优取值，对气象训练数据集设计了如下实验：

首先固定mtry的取值为缺省值sqrt(M)，由于气象数据集中的属性M个数为7，所以令mtry=2。调整ntree的值，选取ntree的不同取值为10、100、200、500、1000，建立随机森林模型，观测森林的整体误差及各类别的误差随RF中树的数目的变化。实验结果如下图3-10和图3-11所示，图的横坐标为随机森林中树的个数，纵轴为误差率。由图可知当ntree=300时，误差已经处于稳定状态了，所以实验选用的ntree的值为300。

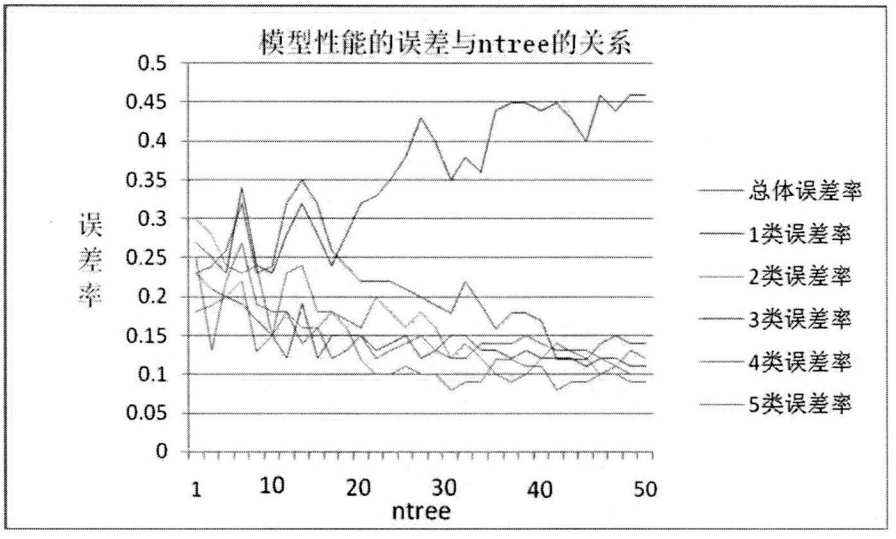

图 3-10 模型性能的误差率随 ntree 的变化

Fig.3-10 The model performance error changes with ntree

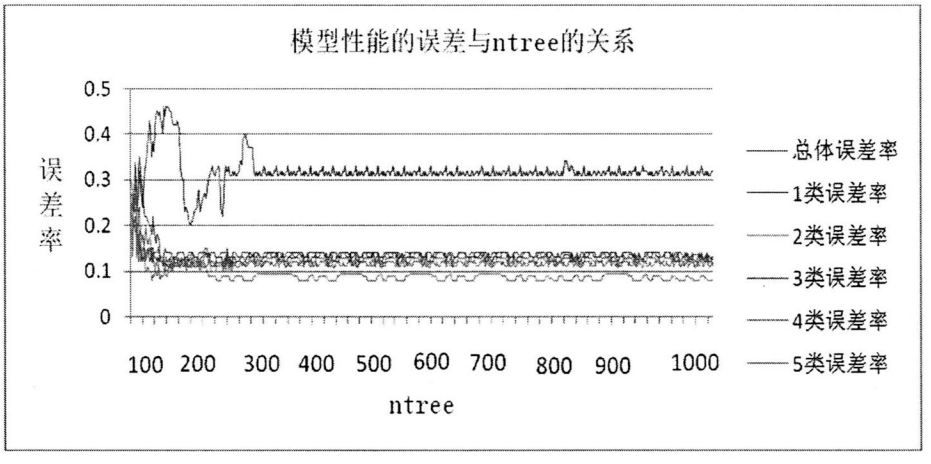

图 3-11 模型性能的误差率随 ntree 的变化

Fig.3-11 The model performance error changes with ntree

参数 mtry 的值是用于构建树时各个节点处候选属性的个数，mtry 是影响随机森林性能的敏感参数，mtry 取值的不同，所得的模型的准确率波动非常明显，一般 mtry 的默认值是 sqrt(M)。

为了是实验获得最佳的效果，我们先固定 ntree=300，调节 mtry 的取值，分别取 mtry=1，2，3，4，5，6，7。观测误差的变化情况，模型的误差情况如图 3-12 所示，由图可知，mtry=2 时，构建的随机森林的整体误差是最小的。因此选定 mtry=2。

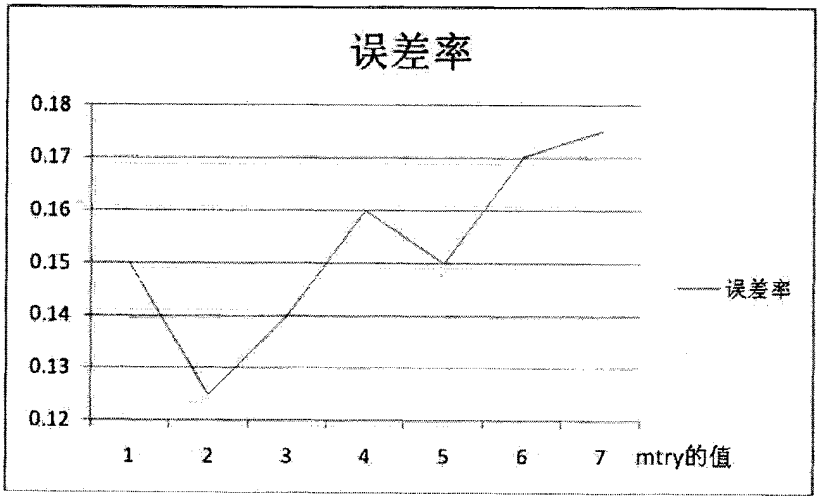

图 3-12　模型误差随 mtry 的变化

Fig.3-12 model error change with mtry

3.5.3 随机森林对不平衡分类问题的处理方法

在气象数据的气温预测中，由于地理位置的原因，该地的气温一般不低，普遍集中在温和天气，低温数据相对较少，这样会导致算法偏向高温样本类，而使得低温样本的预测准确率较低，这就表明，如果我们将气温低的样本极容易误判成高温类，这样就会给当地人民的生活带来不便，使得他们无法对自己或农作物做出合适的保暖措施，这种错误评估会带来一定的损失。于是需要对原始样本进行一定的处理方法，尽可能的平衡各类的分类准确性，

现实的分类问题绝大多数是不平衡分类问题，导致不平衡分类的原因可以分为两种：一是由于数据的不平衡分布导致不平衡分类，例如，至少有一类样本仅占总体数据的一小部分；另一个是由于数据本身的特点导致不平衡分类，例如，某类样本的分布相对较差，不利于正确分类。我们的实验数据兼有这两方面的特点，一是数据分布不平衡，气温低的气候类样本并不多，而气温高的样本几乎占居了总体的一半。如果在建模的过程中，对样本不进行任何处理，那么实验模型将会偏向热类的样本，二温度低的样本的预测准确率将会比较低，为了很好的处理不平衡分类问题，使得各类的准确率相近，RF 处理不平衡分类问题有两种方法：

1）基于样本等量采样技术的方法。即从原数据集中抽取等量的差异样本，已达到均衡类别样本分布。

2）权重法，即通过设定各类别的权重值来平衡各类的分类准确率。在 RF 算法中，

权重值有两个作用：一个是在单棵树构建过程中，将计算 Gini 指数融入到权重系数中；另一个是在对组合分类器的各个子分类树输出结果进行组合的过程中，将权重系数融入到最终票数中。加大被误分的类别的权重，同时降低误差较大的分类器的权重，可以提高该类别的预测准确率。

建模的过程中分别使用这 2 种方法来处理气象数据的不平衡问题，为了方便对比，另加一组不做任何处理的实验。实验设计为：

（1）对不平衡分类问题的样本不做任何处理。

（2）从数据集中等量的抽取各个类的样本，以低气温的个数为基准，随机对高气温样本进行采样，使得各个类别的训练样本数据量都相等。

（3）权重法，气象数据的总类别是 5，气象数据维数为 k=7，所以权重向量初始值为 1/7，用经过加权的频数代替 $Gini(t)=1-\sum_{j=1}^{k}[p(j|t)]^2$ 中的概率，其中 $p(j|t)$ 为类别 j 在 t 节点处的概率，计算 Gini 指数然后按权重倍数复制类别样本，减少类别间的样本数量的不平衡。

（4）最后，构建完分类树之后，比较各类别的加权票数，将票数最多的类别定位 RF 的分类结果。实验结果见表 3-6。

表 3-6 处理不平衡分类问题的方法比较

Table 3-6 Comparison between classifications for unbalanced problems

处理方法	预测误差率（%）					
	整体	1 类	2 类	3 类	4 类	5 类
不做任何处理	21.35%	40.34%	23.32%	12.56%	19.80%	8.14%
等量抽样法	30.17%	20. 32%	26.11%	31.98%	23.44%	37.97%
权重法	9.69%	10.28%	11.23%	11.55%	10.33%	6.04%

从表可以看出，不做任何处理而建构的模型，整体的预测误差了偏低，但是 1 类、2 类这些负向类的误差率却远远大于 3 类、4 类和 5 类的误差率；等量取样处理，虽然各个类别的误差大致接近，但是整体误差和各个类别的预测误差均较高；而用权重法进行数据抽样的方法构建的模型，不仅整体的预测误差率很低而且各个类别的预测误差率也较低和相差不大，保持了平衡性。所以通过实验结果表明，使用权重法不仅提高了总体准确率，同时也均衡了各类的准确率。

3.5.4 RF 气温预测模型

综合以上几小节的实验结果，基于随机森林的气温预测模型的建模过程如下：

（1）首先对原始的气象数据进行预处理，并将数据集按照 80%和 20%分割成训练集和测试集，得到实验数据集；

（2）从数据集的众多气象属性中筛选出与预测气温目标属性相关的属性特征作为模型的输入特征，具体的实践过程见第三章的实验预处理部分。特征选择后包括了 7 个属性特征：风速、相对湿度、气压、水汽压、露点温度、地面温度、时总蒸发量。

（3）建模过程按 3.5.2 节中所述方法选择 ntree 和 mtry 的参数值，并按权重法进行样本抽样，设定各类别的权重。

它的框架如下图 3-13 所示。

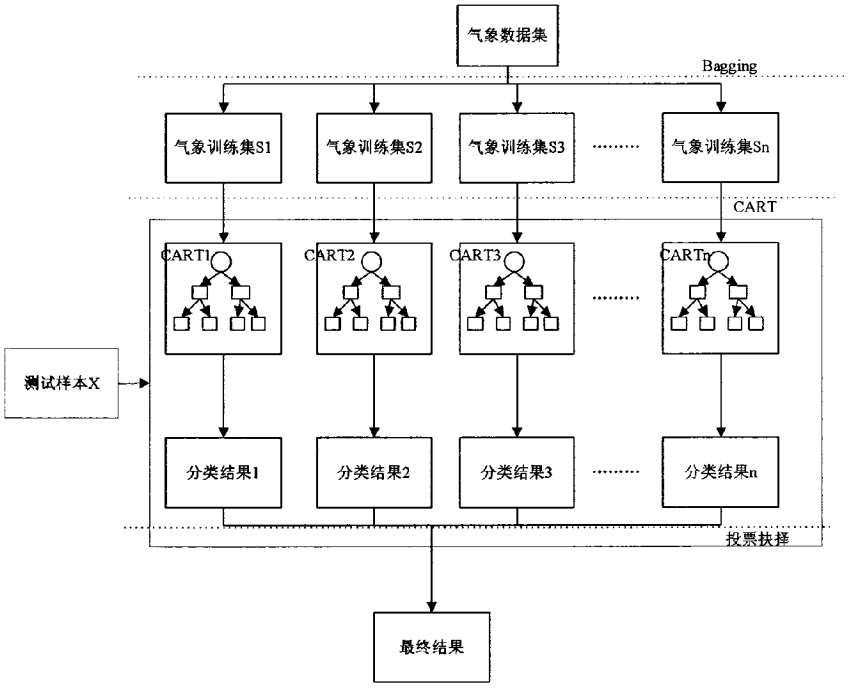

图 3-13 随机森林基本框架图

Fig.3-13 Random Forests basic framework map

随机森林模型的根节点是所有的气象训练样本数据集 D，每个 CART 决策树内部节点包括的数据是气象样本数据集 D 的子集。

每棵决策树都是 CART 算法构建的，构建的过程中，每棵树的中间节点进行属性分裂，从气象训练数据的所有属性中搜索最佳的属性分裂方式来完成；然后依次对后续节点进行类似分裂，到叶子节点时结束。

详细而言，本文随机森林算法过程可以详细的描述如下：

1. 用 ntree 来表示决策树的个数，M 表示每个样本的属性个数。

2. 采用 Bagging Sampling 形成各个与样本数目与原始样本 D 相等的气象训练集 S_i：从原始气象数据集 D 中有放回的选取与原始数据集中的样本数量个数相同的气象样本数据，这些新选的样本数据构成一个训练集 S_i。

3. 构建决策树时内部节点分裂时随机选取特征进行分裂：由 3.5.2 节中的可知，内部节点进行分裂时，随机从 7 个气象数据集中的特征中选取 2 个特征，然后在所选取的 2 个特征中选择最优秀的分裂方式对内部节点进行分裂。

4. 对森林中的每棵树的形成过程都不进行修剪，完整成长。

3.6 本章小结

本章主要是对气象数据进行了预处理，同时针对真实的气象数据对组合分类器的三种模型进行了构建，Bagging 模型、AdaBoost 模型和随机森林模型，对于以 C4.5 决策树作为基分类器的 Bagging 组合分类器的构建过程中加入了协同并行的思想，而 AdaBoost 组合分类器由于训练样本的选择是根据前一轮的结果来进行调整权重的，所以只能串行运算，通过重视误判的样本的再次分类来提高其分类准确率。随机森林中的两个参数直接影响模型的准确率，为此经过统计分析得出了两个参数的最佳取值。并且对随机森林的强容忍能力和对不平衡分类问题的处理方法进行仿真实验，进一步在真实上数据集上验证了随机森林的这些特点。

第四章 组合分类器对局部区域气温的预测

在前一个章里，利用 C4.5 决策树作为弱分类学习算法构建了多分类器模型，研究了针对气象数据怎样用 Bagging、AdaBoost 和随机森林三种算法分别建立气温预测模型。结合上一章的算法思想，我们在本章对三个实验：Bagging 组合分类器、AdaBoost 组合分类器以及随机森林，分别将这三种组合分类器应用于某地局部区域的气象数据中，对当地的气温进行预测，然后对得出的结果进行比较和分析。

4.1 实验测试环境

本文所做具体的实验环境为：

（1）硬件环境：Intel(R) Core(TM) i5 CPU M 480 @2.67GHz ， 4G 内存，320G 硬盘， 操作系统是 WIN 7 版。

（2）软件环境：实验程序编码平台是 Microsoft Visual Studio 2010， 实验平台是 matlab 7.0，采用的开发语言是 C++等。

Matlab 平台集成了机器学习领域中的很多的算法，同时也提供了统计学中的检验方法和相对应的可视化分析工具，这样对实验的结果分析有很大的帮助，提供了很大便利。

4.2 实验数据分析

这一节中我们给出本次实验用到的数据集，本文的样本数据集是广州市某地区局部区域气象站的实际地面观测数据，气象数据是从 2011 年 1 月份到 12 月份全年的气象数据，气象站收集的是当地全年每天 24 小时每小时的即时数据共 8744 条，任何一条数据的数据项包括风向、风速、时总雨量、最高气温、湿球温度、水汽压、本站气压、最高本地气压、最低本地气压、草面温度、5cm 草面温度、10cm 草面温度等共 54 多项。

由第二章节中的气象数据预处理阶段提取出了气温主要受风速、相对湿度、气压、水汽压、露点温度、地面温度、时总蒸发量 7 个因素的影响。因此，从某地区气象数据集中提取对气温有影响的 8 个属性，去掉一些异常或者属性缺失的气象数据，生成原始数据集。将风速、相对湿度、气压、水汽压、露点温度、地面温度、时总蒸发量

作为分类属性，气温等级作为类属性。部分原始数据如下：

表 4-1 部分原始数据

Table 4-1 Portion of the original data

气温等级类别	风速	相对湿度	水汽压	露点温度	地面温度	时总蒸发量	本站气压
1	38	86	64	6	39	2	10122
3	16	81	147	127	158	0	10100
5	13	91	268	222	240	1	10045
2	11	69	61	1	36	1	10173
5	21	82	242	205	229	1	10068
2	46	67	60	-3	43	0	10150
4	27	58	128	107	258	0	10123
3	31	40	73	24	299	5	10159
5	20	91	268	222	239	1	10044
4	28	45	99	69	257	2	10142
2	27	52	46	-38	21	2	10178
4	0	94	208	181	190	1	10083
4	15	93	206	179	192	1	10097
3	43	54	98	67	211	4	10140
5	20	83	245	207	233	2	10046
5	13	92	271	224	242	3	10050
1	35	84	62	3	38	1	10126

通过采访当地气象部门的工作人员了解到本地区近 50 年来为发生过地震、洪涝等地质灾害，数据变化平稳，经过数据预处理后。按气温等级进行分类预测，将气温作为模型的目标因子，风速、相对湿度、气压、水汽压、露点温度、地面温度、时总蒸发量作为属性因子。

经过处理后，气温分为 5 个等级，数据集中含有等级为寒的数据有 50 条，等级为凉的数据有 1613 条，等级为温的数据有 1023 条，等级为暖的数据有 1268 条，等级为热的数据有 4786 条。为了实验的随机性和实验结果的完整性，将样本训练集和测试集按 4:1 的比例进行。

4.3 实验结果与分析

将组合分类器应用于局部气象数据中预测气温，应用引入组合算法 Bagging、组合算法 AdaBoost 和随机森林三种组合分类器，分别分析三种组合分类器对局部区域气温预测性能。这些都是我们实验所期待的，为了验证这些问题，我们依据上一章的模型，分别设计和实现了三个实验来对比预测性能和组合规模。实验都使用了同一个数据集，即上一节所介绍的局部区域的气象数据集。

1. Bagging 组合器模型

训练样本数分别为 n=1000，测试样本为 250；训练样本数为 n=4000，测试样本数为 1000；训练样本数为 n=6000，测试样本为 1500。由于组合分类器的最终结果是采用投票法进行决策，所以基分类器的个数为奇数比较合适，为此在基分类器个数 k=3，5，7，9，11，13，15，21，25 的条件下，分别进行实验，不同情况下的准确率如下图 4-1 所示。通过比较可知，训练集越多，构建的模型的准确率越高，基分类器的个数在 k=9 的时候基本趋于稳定，增加基分类器个数对最终结果影响不大，最好的效果是 k=15，n=6000，准确率为 91.73%。

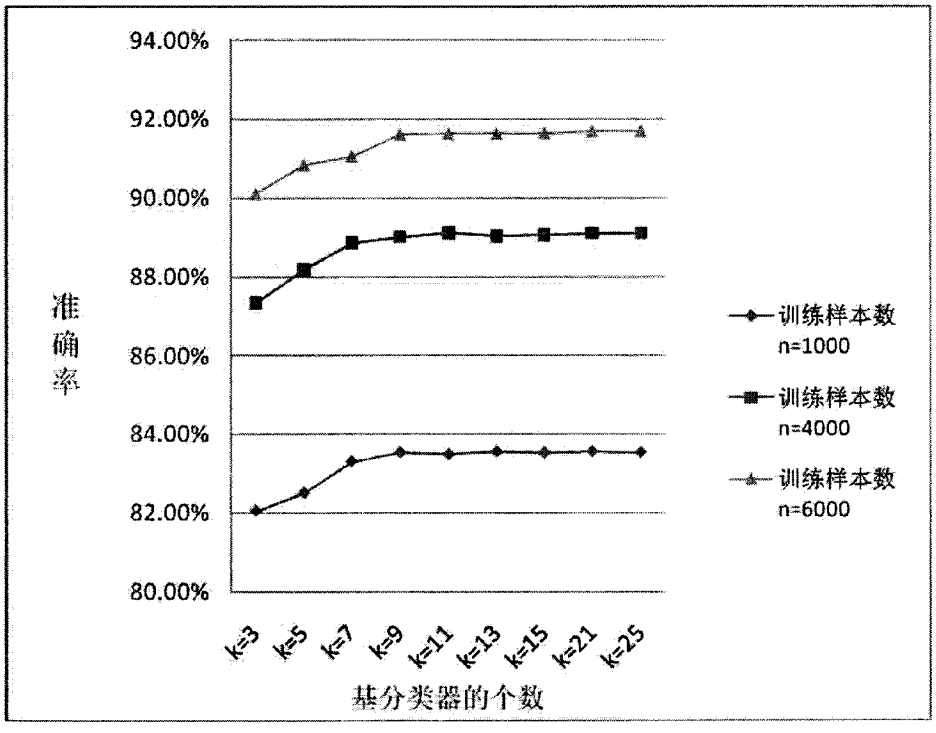

图 4-1 K 的不同取值情况下的准确率

Fig.4-1 Accuracy rate changes with different values of K

在样本集为 n=1000 的情况时，进行了多次实验，从中随机选取了次数为奇数的实验结果记录下来，结果如图 4-2，当基分类器个数 k=3 时，实验的准确率较不稳定，时高时低，准确率波动较大，基分类个数 K=9 时，趋近平稳。

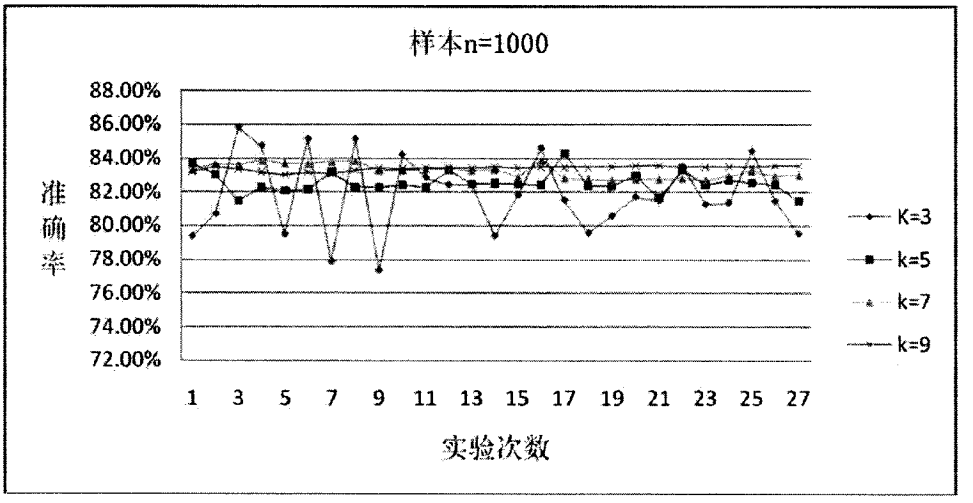

图 4-2 样本数 n=1000

Fig.4-2 Samples n=1000

样本数 n=4000 的情况下，进行多次实验，奇数次抽取的实验结果见图 4-3，较样本数 n=1000 来说，准确率较稳定，没有那大的波动性，同时准确率也提高了不少，基分类器个数 k=9 时，波动最小，几乎趋于直线。

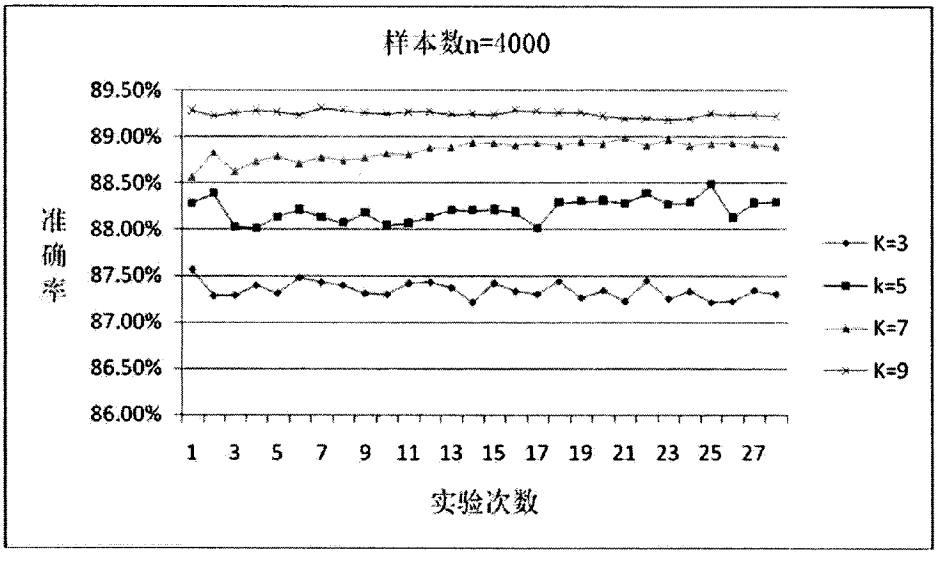

图 4-3 样本数 n=4000

Fig.4-3 Samples n=4000

样本数 n=6000 的情况下，进行多次实验，部分的实验结果见图 4-4，较样本数 n=1000 和 n=4000 来说，准确率也提高了，基分类器个数 k=9 时，波动最小，几乎趋于直线。样本数也多，基分类器个数越多，则整个组合分类器的性能越稳定，准确率波

动较小，准确率较高。

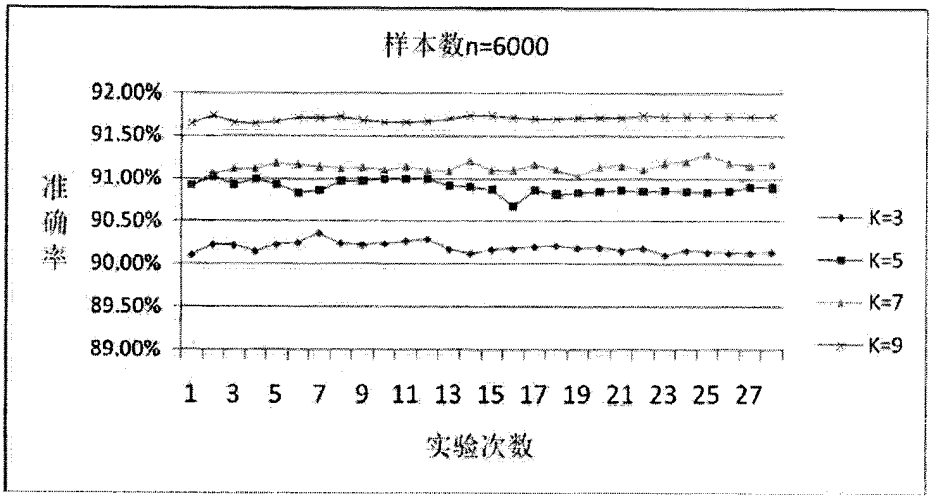

图 4-4 样本数 n=6000

Fig.4-4 Samples n=6000

2. AdaBoost 组合器模型

按三种情况来训练组合器模型，训练样本数分别为 n=1000，测试样本为 250；训练样本数为 n=4000，测试样本数为 1000；训练样本数为 n=6000，测试样本为 1500。在基分类器个数 k=3，5，7，9，11，13，15，21 的条件下，进行实验，不同情况下的准确率如图 4-5 所示。模型的准确率跟基分类器的个数和样本数的多少有关，训练集为 6000，基分类器个数 k=11 时，其模型的准确率达 98.45%。

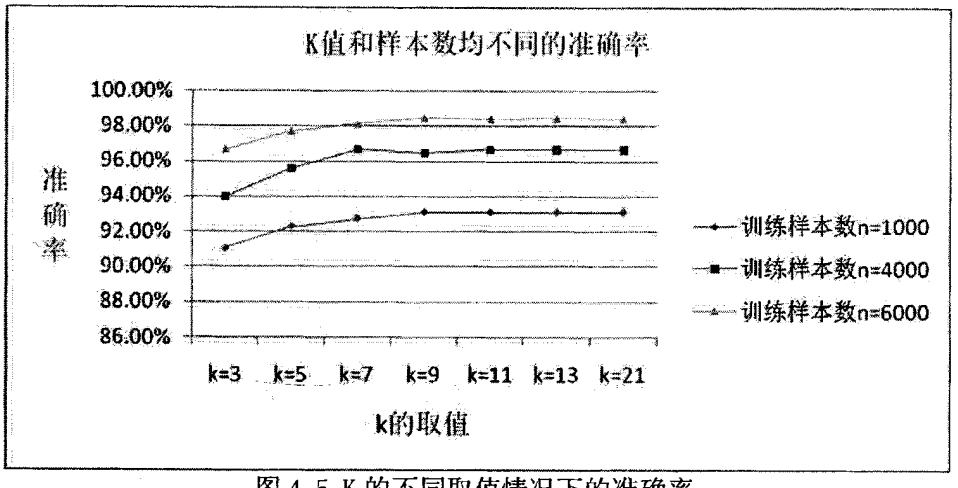

图 4-5 K 的不同取值情况下的准确率

Fig.4-5 The accuracy under different values of K

样本数 n=1000 时，在不同的基分类器个数的条件下，模型的准确率情况如图 4-6

所示。基分类器个数为 3 时，模型的准确率波动较大，不稳定，随着基分类器个数的提升，模型的准确率稳定性趋于平稳状态，k=11 时，基本上达到稳定。

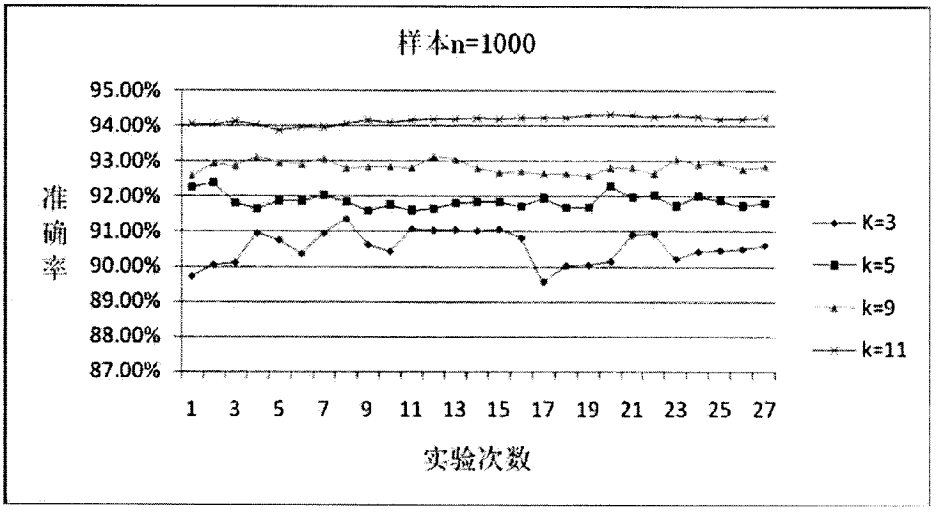

图 4-6 样本数 n=1000

Fig.4-6 Samples n=1000

样本数 n=4000 时，基分类器个数不同的情况下模型的整体准确率如图 4-7 所示。由于样本数较多，所以模型在基分类器不同的情况下，它的性能波动比 n=1000 时较小，基分类器越多，分类的准确性越高、越平稳。

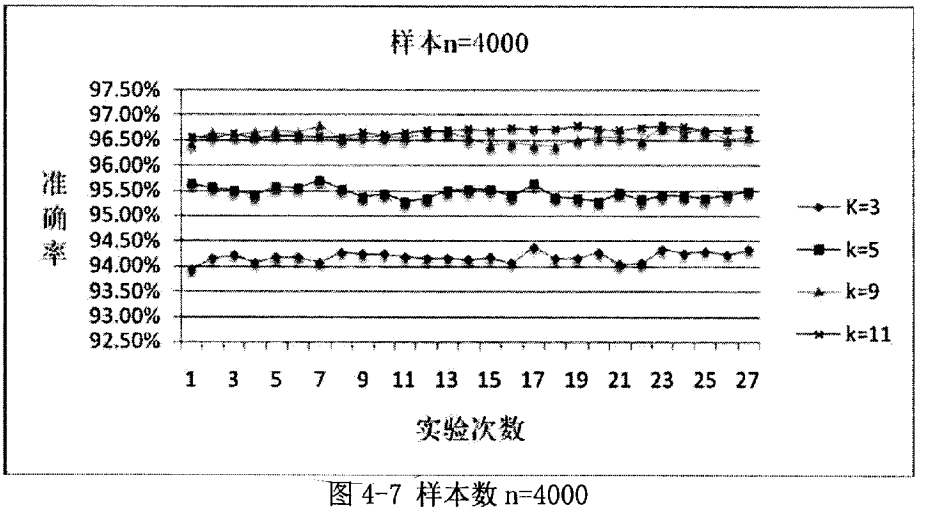

图 4-7 样本数 n=4000

Fig.4-7 Samples n=4000

样本数 n=6000 时，基分类器个数不同的情况下模型的整体准确率如图 4-8 所示。模型在基分类器不同的情况下，它的性能波动比 n=1000 和 n=4000 两种情况都较小，基分类器越多，分类的准确性越高。

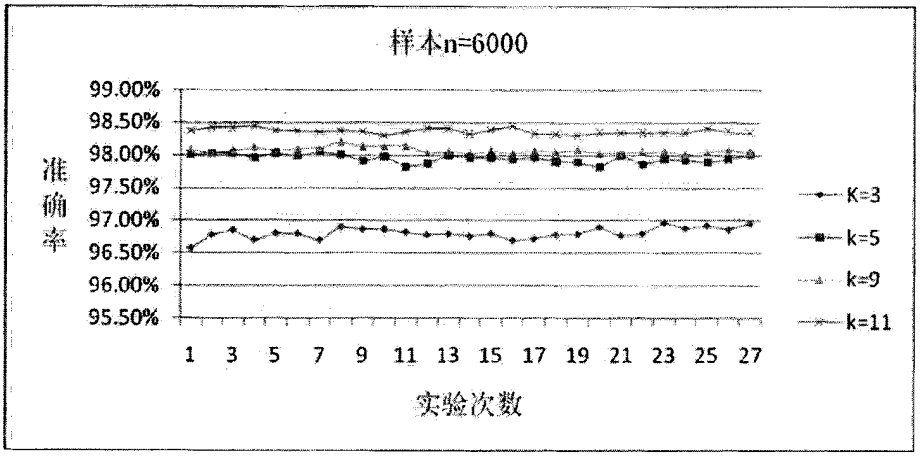

图 4-8 样本数 n=6000

Fig.4-8 Samples n=6000

3. 随机森林组合器模型

取 mtry=2 和 ntree=300 构建随机森林组合模型，按三种情况来训练组合器模型，训练样本数分别为 n=1000，测试样本为 250；训练样本数为 n=4000，测试样本数为 1000；训练样本数为 n=6000，测试样本为 1500。其准确率如图 4-9 所示：

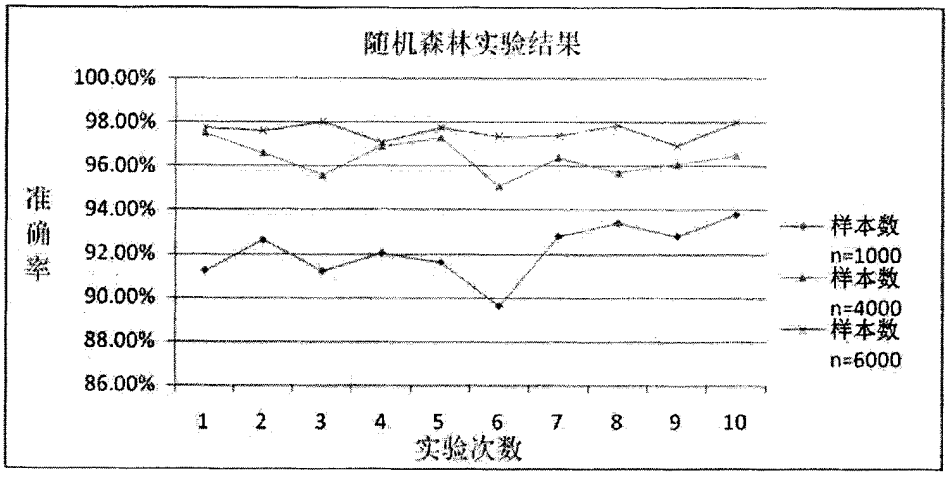

图 4-9 RF 模型的准确率

Fig.4-9 Accuracy of Random forest model

4. 各种模型的比较

比较不适用组合分类器仅使用单一的 C4.5 决策树和使用三种组合方式进行实验，各模型的准确率等性能情况如表 4-2 所示。

表 4-2 三种气温预测模型的性能比较

Table 4-2 Performance Comparison of three temperature prediction models

模型	准确率(%)						稳健性(%)	耗时(s)
	整体	一类	二类	三类	四类	五类		
C4.5	83.23	70.25	86.34	73.45	75.63	90.67	10.41	1.43
Bagging	91.93	83.33	90.27	80.25	82.82	96.88	6.57	20.41
AdaBoost	97.47	87.22	96.43	92.45	92.65	98.56	6.72	46.29
RF	97.57	87.56	97.39	91.79	93.64	99.14	6.59	14.94

注：稳健性指模型在训练集和测试集上分类精度的差异

从上面的表格中可知，利用组合的方式构建组合分类器对单一决策树模型的分类预测进度有了很大的提高，C4.5 决策树模型对气温预测的准确率是 87.23%，Bagging 算法的组合分类器模型对气温预测的准确率是 91.93%，比单一模型提升了 4.7%；AdaBoost 模型的准确率是 97.47%，比单一模型提升了 10.24%；RF 模型是 97.57%，，比单一模型提升了 10.34%。AdaBoost 和 Bagging 相比，准确率有了很大的提高，但是 AdaBoost 的计算量比较大，并且不能实行并行运算导致其运行时间较长。AdaBoost 与 RF 相比较，发现二者的准确率基本上相媲美，由于 RF 构建节点时并不是选所有的属性作为候选，所以 RF 的计算量比 AdaBoost 少了不少，因此其运算耗时较少。

从稳定性上来看，三种组合方法对气温预测有很好的稳定性，由于 C4.5 决策树是“不稳定的”分类算法，极容易出现过拟合现象，即在训练集上有很好的分类效果，但是对于测试集进行分类判别时准确率会明显下降。在经过 Bagging 和 Boosting 算法的组合后，C4.5 决策树的准确率得到了很大的提升。

气象数据集一般是复杂多样的，而且经常会出现缺省值等数据不完整性，基于这样的特点，所以对气象数据继续预测的时候，RF 模型是较为合适的。

4.4 本章小结

本章在上一章节针对气象数据建立的组合分类器模型的基础上，以广州某区的实际气象数据为数据集进行了实验，根据当地气温的实际情况将当地的气温划分成五个等级，并对原始气象数据集用第二章所介绍的方法进行相应的预处理得到实验数据集，然后用上一章建立的三个组合分类器对当地的气温进行了预测，对实验结果进行了分析。

总结与展望

本章主要对论文所做工作进行总结，在已有的工作和取得的效果上，对未来的工作进行了展望。

本文概述了数据挖掘的基本概念、数据挖掘常用的方法在气象数据中的挖掘研究的现状和气象数据的特点，并且重点介绍了组合分类器的构建过程，即对气象数据进行数据挖掘。研究了组合分类器的三种构建模型（Bagging、AdaBoost、随机森林）的原理和实现过程。然后针对广州市某区的实际气象数据的特点，分别构建了三种组合分类器模型，分别各个模型中的参数进行了选取，实验结果表明三种模型的预测效果令人满意，组合分类器模型在预测准确性和稳健性上都比单一模型表现好，通过分析比较不同模型之间的预测效果，针对气象数据的特点，对模型进行了推荐。

本文只是将针对局部气象数据建立了组合分类器模型，未来的方向就是将模型使用的范围进行推广和验证，同时扩大模型的功能，不单单是气温预测，可以涉及降雨、灾害天气等气象方面的研究。

本文选择了决策树作为元学习方法，运用 Bagging、AdaBoost 组合思想和随机森林三种方式建立气温预测组合模型，证明了模型的有效性。因此，将其他数据挖掘算法 KNN、SVM、NN 等分类方法进行组合对气象因素进行预测是今后研究的一个方向。

本文对森林中的各个决策树的结果进行简单投票确定随机森林模型的最终决策结果，那么如何选择更加有效的方式确定模型的最终结果也可以是未来的研究方向。

参考文献

[1] 蒋盛益，李霞，郑琪 编著.数据挖掘原理与实践[M] .北京：电子工业出版社，2011.

[2] 杨宸铸.基于 HADOOP 的数据挖掘研究[D] .重庆:重庆大学硕士学位论文，2010.

[3] Jiawei Han，Micheline Kamber，et cl.Data Mining Concepts and Technologyes (Third Edition)[M] .China Machine Press， Agu.2012.

[4] Kevin P， Murphy. Machine Learning[M] .The MIT Press， Agu.2012.

[5] 张倩，沈利，蔡焕杰等.基于灰色理论和回归分析的需水量组合预测研究[J] .西北农林科技大学学报(自然科学版)，2010，38 (8)： 223-227.

[6] 安俊琳，王跃思，朱彬.主成分和回归分析方法在大气臭氧预报的应用[J] .环境科学学报，2010，30 (6)：1286-1294.

[7] 张福丽，景天忠，王志英，等.气象因素与杨干象虫口密度的多元线性回归和判别分析[J] .安徽农业科学，2011， 39 (15)：9000-9001.

[8] 刘洪兰，张强，赵小强，等.张掖湿地公园水域结冰厚度预报的 BP 神经网络与统计回归方法对比[J] .干旱气象，2013，31 (2)：425-431.

[9] Sajib Barua， PReda Alhajj.A Parallel multi-scale region outelier mining algorithm for meteorological data[J] .ACM，2007，23 (1)：1-4.

[10] 陈德花，陈创买，周学鸣，等.福建汛期降水主分量逐步回归预测模型研究 [J] .气象，2013，39 (9)：1190-1196.

[11] 杨淑群，冯汉中，芮景析.支持向量机(SVM)方法在降水分类预测中的应用[J].西南农业大学学报（自然科学版)，2006，28 (2)：252-258.

[12] Vallippa Lakshmanan， Timothy W.A Map Reduce Technique to Mosaic Continental-Scale Weather Radar Data in Realtime[J] .IEEE Journal of Select Topics in Applied Earth obsevations and Remote Sensing，2013：1230-1238.

[13] C.Piani，et al. Statistical bias correction for daily precipition in regional climate models over Europe[J] .Theoretical and Applied Climatology，2010，99 (1)：187-192.

[14] 陈少斌，苏彦.气象信息数据挖掘技术的应用[J] .河南科技，2013，7：200-206.

[15] Siva Venkadesh, Gerrit Hoogenboom, et al. A genetic algorithm to refine input data selection for air temperature prediction using artificial neural networks[J] .Elsevier Science Publishers B.V, 2013: 2253-2260.

[16] 徐亮亮，付德胜.基于模糊支持向量机的夏季雨型的预报方法研究[J] .四川大学学报（自然科学版），2013，50 (6)：1230-1234.

[17] Linli Jiang , Jiansheng Wu. Hybrid PSO and GA for neural network evolutionary in monthly rainfall forecasting[J] .Springer-Verlag, 2013: 79-88.

[18] 黎玉芳，李志鸿.桂林地区气温与降水量的时间序列预测模型[J] .广西科学，2013，20 (2)：107-110.

[19] 王定成，汪春秀，等.基于SVM的灾害天气预测方法的研究[J] .武汉理工大学学报，2010，32 (24)：121-124.

[20] 滕少华，樊继慧，陈潇，等.基于KNN的多组合器协同挖掘局部气象数据[J] .广东工业大学学报，2014，31 (1)：25-31.

[21] 姜文瑞，王玉英等.决策树方法在气温预测中的应用[J] .计算机应用与软件，2012，29 (8)：141-145.

[22] 黄静华，刘小魏，袁玫，等.K-Means 聚类在气象数据分析中的应用[J] .计算机工程与应用，2009，45(专刊)：98-99.

[23] 何伟，孔梦荣，赵海青. 基于贝叶斯分类器的气象预测研究 [J] .计算机工程与设计，2007，28(15)：3780-3782.

[24] 马廷淮，穆强，等.气象数据挖掘研究[J] .武汉理工大学学报，2010，32 (6)：110-105.

[25] Qin Wang, Wei Fan. Cumulonimbus forecasting based on rough set and artificial immune algorithm[C] .Natural Computation (ICNC), 2010 Sixth International Conference on, 2010, 6: 2856-2860.

[26] Tsegaye Tadesse, Donald Awilhite, et al. Drought Monitoring Using Data Mining Techniques: A Case Study for Nebraska[J] .USA Natural Hazards, 2004(33): 137-159.

[27] Asanobu Kitamoto. Spatio-temporal Data Mining for Typhoon Image Collection[J] .Journal of Intelligent Information Systems, 2002, 19 (1):

25-41.

[28] Cheng Tao, Wang Jiaqiu. Application of a Dynamic Recurrent Neural Net work in Spatio-temporal Forecasting[C] // Information Fusion and Geographic Information Systems Proceedings of the Third International Workshop. New York: Springer, 2007: 173-186.

[29] 陈旭辉. 沙尘暴资料的数据挖掘算法分析及系统实现[D] . 兰州: 兰州大学硕士学位论文, 2008.

[30] 蒋芸, 陈娜, 周泽寻, 等. 基于 Bagging 的概率神经网络集成分类算法[J] . 计算机科学, 2013, 40 (5): 242—246.

[31] F. Roli, J.K. 7th Int' l Workshop on Multiple classifier Systems (MCS 2007) [C] . 2007.

[32] F. Roli, J.K. 8th Int' l Workshop on Multiple classifier Systems (MCS 2009) [C] . 2009.

[33] Lior Rokach. Ensemble-based classifiers[J] . Kluwer Academic Publis -hers, 2010, 33 (2): 1-39.

[34] 陈鑫. 基于决策树技术的遥感影像分类研究[D] . 南京: 南京林业大学硕士学位论文, 2006.

[35] 李章吕. 贝叶斯决策理论研究[D] . 天津: 南开大学博士学位论文, 2012.

[36] 付忠良. 分类器线性组合的有效性和最佳组合问题的研究[J] . 计算机研究与发展, 2009, 46 (7): 1206-1216.

[37] Simon S. Haykin. Neural networks and learning Machines[M] . Prentice Hall, 2008.

[38] Guohua Liang, Chenqi Zhang. Empirical Study of Bagging predictors on medical data[J] . Australian Computer Society, Inc. 2011: 31-40.

[39] Guosheng Lin, Chunhua Shen. David Suter. Fast training of effective multi-class boosting using coordinate descent optimization[J] . Springer-verlag, 2012: 782-795.

[40] Paulo Fernandes, Lucelene Lopes. Duncan D. A. Ruiz. The impact of random samples in ensemble classifiers[J] . ACM, 2010: 1002-1009.

[41] Chih-Fong Tsai, Yuah-Chiao Lin, David C. Yen. Predicting stock returns

by classifier ensembles[J]. Applied Soft Computing, 2011 (11): 2452-2459.

[42] Alexey Tymbal, Mykola Pechenizkiy, Padraig Cunningham. Dynamic integration of classifiers for handling concept drift[J]. Information Fusion, 2008, 9: 56-68.

[43] 郑春颖，王晓丹，郑全弟. 基于模糊积分的支持向量机动态集成方法[J]. 系统工程与电子技术，2011，33 (6)：1429-1432.

[44] Baoguo Yang, Yang Zhang, Xue Li. Classifying text streams by keywords using classifier ensemble[J]. Data and knowledge Engineering, 2011, 7 (1): 775-793.

[45] Yong Wang, Xiaolei Ma, et cl. A fuzzy-based customer clustering approach with hierarchical structure for logistics network optimization[J]. Expert Systems with Applications :An International Journal, 2014, 41(2): 521-534.

[46] 埃塔尔季奇 著，王晓海，吴志刚译. 数据挖掘：概念、模型、方法和算法（第 2 版）[M]. 北京：清华大学出版社，2013.

[47] 周涛，陆惠玲. 数据挖掘中聚类算法研究进展[J]. 计算机工程与应用，2012，48(12)：100-112.

[48] 李伶俐. 数据挖掘中分类算法综述[J]. 重庆师范大学学报（自然科学版），2011，28 (4)：44-48.

[49] 王令剑，滕少华.聚类和时间序列分析在入侵检测中的应用[J]. 计算机应用，2010，30 (3)：699-701,714.

[50] 王爱平，王占凤，等. 数据挖掘中常用关联规则挖掘算法[J]. 计算机技术与发展，2010，20 (4)：105-109.

[51] 朱吉龙. 孤立点检测在移动通信数据分析上的研究与应用[D]. 广州：广东工业大学硕士学位论文，2013.

[52] 沈家芬，张凌，等. 广州市空气污染和气象要素的主成分与典型相关分析[J]. 生态环境，2006，15 (5)：1018-1023.

[53] 唐旋. 数据挖掘技术在气象资料分析中应用研究[D]. 内蒙古：内蒙古科技大学硕士学位论文，2011.

[54] 游晓黔，黄小红，秦靖. 基于级联结构 AdaBoost 的入侵检测算法[J]. 计算机工程，2011，37 (3)：134-136.

[55] 蒲元芳，张巍，滕少华，等.基于决策树的协同网络入侵检测[J]．江西师范大学学报（自然科学版），2010，34（3）：302-308.

[56] 康恒政.多分类器集成技术研究[D]．成都：西南交通大学硕士学位论文，2011.

[57] 王小强.基于随机森林的亚健康状态预测与特征选择方法研究[J]．计算机应用与软件，2014，31（1）：296-298，307.

[58] 李俊磊，滕少华，张巍．基于决策树的多组合协同分类器在局部区域中的气温预测[J]．广东工业大学学报，2014，已录用.

攻读学位期间从事的科研项目及发表的论文

从事的科研项目

[1] 基于SVM和决策树的协同入侵检测，滕少华，教育部重点实验室开放基金，2万，110411，2011年.

[2] 2011年广州市交警支队警务车辆维护管理系统.

[3] 2012年中国电子第七研究所软件质量缺陷管理平台.

[4] 2012年中国电子第七研究所软件质量评估管理平台.

发表的论文

[1] Shaohua Teng，Junlei Li，Wei Zhang，Rigui Li. The Calculation of Similarity and its Application in Data Mining.ICPCA/SWS2013.

[2] 李俊磊，滕少华，张巍. 基于决策树的多组合协同分类器在局部区域中的气温预测. 广东工业大学学报，2014. [本文第四章部分内容]

学位论文独创性声明

本人郑重声明：所呈交的学位论文是我个人在导师的指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明，并表示了谢意。本人依法享有和承担由此论文所产生的权利和责任。

论文作者签名：李俊磊 日期：2014.6.3

学位论文版权使用授权声明

本学位论文作者完全了解学校有关保存、使用学位论文的规定，同意授权广东工业大学保留并向国家有关部门或机构送交该论文的印刷本和电子版本，允许该论文被查阅和借阅。同意授权广东工业大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印、扫描或数字化等其他复制手段保存和汇编本学位论文。保密论文在解密后遵守此规定。

论文作者签名：李俊磊 日期：2014.6.3

指导教师签名：陈力禾 日期：2014.6.3

致 谢

毕业论文撰写完成之际，标志着硕士研究生三年的学习即将画上圆满的句号，三年的学习生活对个人来说虽是短暂的，但三年的学习生活将是我一生受益的宝贵财富。在论文完成之际，借此机会，我要用简单而又真诚的话语对在学习和生活中给予我帮助和支持的人表示衷心的感谢。

首先，我要感谢我的导师滕少华教授，能在滕教授的指导下学习和研究数据挖掘领域的知识是我的荣幸，滕教授为人谦和，治学严谨，责任心强，感谢他在我的研究生学习生涯中给予我悉心指导，使得我学业能够顺利完成。他渊博的知识和严谨的治学态度使我受用终身。在此我要真诚地说：谢谢您，能够成为你的学生是我的荣幸，同时祝你家庭幸福，身体健康！

其次，我要感谢我最爱的父母，感谢他们二十多年来含辛茹苦把我养育成人，他们是我人生前进的动力和精神支柱，感谢他们对我无微不至的支持和关心。同时也感谢每一位亲朋好友，你们在我低落和遇到挫折的时候，能够开导我，给我提出宝贵的意见，让我走出一个个困境，再次表示深深地感谢！

再次感谢计算机学院协同软件开发实验室（511 实验室）的傅秀芬教授、张巍副教授、刘冬宁副教授等教授对我悉心指导，以及 511 实验室的伙伴们，三年里我们并肩作战共同经历了酸甜苦辣，一直以来对我付出我都铭记在心，谢谢你们的相伴，愿我们友谊天长地久。

最后向百忙中评审本文的各位专家和教授们表示衷心的感谢，向所有给予我帮助的人，致以诚挚的问候和深深的谢意。

再一次感谢以上所有帮助我的人，祝你们幸福安康！