南开大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，进行研究工作所取得的成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：石有印

2007年11月26日

中文摘要

煤炭自燃是煤炭生产过程中的重大灾害之一，定量研究其发生机理可以为及时、准确地预测煤炭自燃和防治火灾提供科学依据。本文首先介绍了目前煤炭自燃预测预报的研究现状，进一步探讨了煤炭自燃的定性描述机理和过程。其次，针对大同煤矿的实际情况，应用支持向量机、粗糙集等现代数据处理技术一方面分析了用于判定煤层自燃情况的原有各项指标的重要性，得出了指标中的供风量、丢煤率和煤层厚度这三个属性是判定煤层自燃情况的主要因素；另一方面，通过对指标气体数据的计算、分析之后发现，CO 确实是煤炭自燃的最主要的指标气体，其次是甲烷，二氧化碳和乙烯。做好其监测分析并采取合理的数据熔合会为解决煤炭自燃的预测提供新的思路和方法。特别值得一提的是：在使用支持向量机方法之下，不论是依据 CO、CH_4 与 CO_2 三组数据，还是依据 CO、CH_4、CO_2 和 C_2H_4 四组数据，都得到了预测精度为 95%左右的好效果，没有任何区别。因此，提示我们不一定要更换老设备，这将节约巨额的成本开支。

关键词：煤炭自燃　指标气体　影响因素　粗糙集　支持向量机

Abstract

Spontaneous combustion of coal is one of the most serious disasters that can occur in coal production. Studying its mechanism by quantitative research can provide a timely and accurate scientific basis for forecasting and preventing the fire. This thesis introduces the current research on forecasting spontaneous combustion in coal, then, further, it discusses the qualitative descriptive mechanism and process of this phenomenon. Then, specifically targeting the current condition of the coal in Datong, using modern data processing techniques including support vector machines and rough sets, first, by analyzing the importance of the original norms, to determine the spontaneous combustion conditions in coal layers, it draws the conclusion that the air flow rate, the coal loss rate, and the thickness of the coal layer are three properties that can be used to tell the combustion state of coal layers, and second, by calculating and parsing the data from index gases, it determines that CO is indeed the most important index gas for spontaneous combustion of coal, followed by methane, CO_2, and acetylene. Therefore, properly monitoring and analyzing the data, and combining the data with other data will provide new ideas and approaches with which to solve the spontaneous combustion problem. It is very interested that prediction accuracies based on the data set for CO、$CH_4$and CO_2, or data set for CO、CH_4、$CO_2$and C_2H_4 do not change if we use Support Vector Machine approach, and both are the 95% accuracy. Thus, we may serve too much money if we continue to use the old equipments.

Key words: spontaneous combustion of coal index gases influence factors rough set support vector machine

目录

第一章 引言

第一节 选题的背景

我国目前已成为世界上最大的煤炭生产国，2005年全国原煤产量达到创纪录的21.9亿吨。同时我国也是世界上最大的煤炭消费国，年煤炭消费量约占全球的30%，国内主要耗煤产业的年煤炭消费量约占总消费量的70%。随着国民经济的发展，以煤为主要能源的生产和消费特征在今后相当长的时间内都不会改变，这就给煤炭企业提供了巨大的发展机会。

我国煤炭资源分布较广，但北方煤炭资源远远超过南方，很久以来有“南粮北调，北煤南运”之说。随着国际原油价格的飙升，作为原油主要替代品的煤炭也随之金贵。于是不分南北，到处都在大量地开采煤炭，导致矿难也屡有发生。引起煤矿重大灾害的因素和变化机制到底是什么？这成了摆在研究人员面前的挑战性问题。山西大同是著名的煤炭基地之一，也饱受矿难之苦。为此也引起了我们对引起重大灾害的因素和变化机制进行研究的兴趣。希望能够凭借在大同工作的地理优势，深入了解煤炭生产过程之后给出合理的数学模型。

在实际调查和查阅资料后发现，引起煤炭生产过程中的重大灾害很多，原因也十分复杂，题目很大。但是，我们发现，包括大同在内，全国煤炭自燃是煤炭生产过程中的重大灾害之一。我国每年因为煤炭自燃而造成的毁煤量达到数亿吨。在现有煤矿中，国有重点煤矿 54.9%的矿井有自然发火的危险，地方国有煤矿年产 3 万吨以上的矿井中 29.1%有自然发火的危险。由于自燃引起的火灾约占总火灾数的90%以上[1]，给国家和矿井带来了极大的危害及经济损失。同时，由矿井自燃火灾诱发的瓦斯、煤尘爆炸事故也时有发生，严重地威胁着人民的生命财产安全，阻碍了煤炭工业的可持续发展，煤炭自燃已成为煤炭行业广为关注并试图治理的主要灾害之一。因此，如果能及时并准确地预测预报煤炭自燃可以为防治火灾提供科学依据，达到防止或减少火灾损失的目的。这将是不可估量的贡献，这是一个具有诱惑力的课题。为此我们将兴趣锁定在煤炭自燃的预测课题上，这就是选题的背景。

第二节 煤炭自燃的常识以及可观测指标

煤是一种由多种官能团、多种化学键组成的复杂的有机大分子。煤的自燃是一个非常复杂的物理化学变化过程，是一种涉及物质、动量、能量和煤的化学结构在复杂多变的环境条件下相互作用的三维、多相、非定常、非线性、非平衡态的动力学过程。该动力学过程与外部环境及其它干预因素等相互耦合，是具有复杂性本质的科学研究对象。

山西大同煤田是双系煤田，上部侏罗系与下部石炭二叠系含煤地层。侏罗系各煤层有自然发火倾向，危险程度多属Ⅲ～Ⅳ级，自然发火期 6～12 个月。石炭二叠系各煤层属自燃～易自燃煤层。由于煤层的自然发火期短，井下开采时丢弃的残煤较多，导致井下火灾增加。虽然采用了均压灭火技术、加强通风管理等措施，但仍存在有安全隐患、生产费用高、管理困难等问题。其中有些自燃火区长期无法扑灭，大量煤炭资源被冻结，昂贵的生产设备毁于火区之中，严重干扰了矿井的正常生产秩序，还浪费了大量的人力、物力和财力，甚至造成人员的伤亡。

煤炭自然发火能够得到及时控制是防止煤炭自燃火灾的基本措施之一。目前所知，自然发火有以下特征：

1. 地点：煤矿井下绝大多数的自然发火都是发生在回采工作面的采空区内或相邻的老空区内。由于火源位置无法接近，往往无法确定火源的实际位置，正是由于采空区火源点的无法接近性和隐蔽性，且又是经常发生在难以进入的采空区或煤柱中，要想准确找到火源并非易事。给防灭火工作往往带来了很大的盲目性，一些有效的防灭火措施不能充分发挥作用。

2. 孕育期：煤层的自燃往往伴有一个孕育的过程，所以，煤炭自燃的早期预测预报技术对于矿井的安全生产是十分重要的。这就转化成需要解决的三个问题：

- 煤体自燃隐患点或火源位置；
- 煤体自燃隐患点或火源的温度；
- 自燃隐患发展到着火需要的时间。

在煤矿的实际生产中，往往是当自燃发展到一定程度时才有所发觉。国内外许多学者和煤炭生产、科研单位对此都十分重视，但由于这一问题的复杂性，至今没有得到很好的解决。主要原因有：

- 探测技术手段和途径不成熟，所采用的各种技术手段都无法准确确定高温火源点区域及其内部温度；
- 井下条件复杂，影响因素较多，给准确探测井下火源区域带来困难；
- 目前对这一问题的研究还不够深入。

对于生产矿井来说，煤炭自然发火的有效控制对矿井安全具有重要意义。而控制的成功与否，取决于对自燃火灾的超前发现程度，及时准确的预测预报是超前发现并采取有效措施的前提。所以，在实际的生产过程中，如果能够提前预测出采空区煤炭的自然发火趋势，例如温度、指标气体浓度的变化等情况。对于矿井的防灭火工作会有积极的意义。

关于煤炭自燃的起因及过程，是一个古老而又具有挑战性的问题。人们在17世纪就开始了探索研究，但迄今仍然未能得到圆满的解释。各国学者发表了各种学说以解释煤炭自燃的起因，其中主要有黄铁矿导因、细菌导因、酚基导因、煤氧复合导因等学说[2]。煤氧复合作用学说得到大多数学者的赞同，因为煤自燃的主要参与物一个是煤，一个是氧，煤对氧的吸附是经实验考察得到证实的。可以说，煤氧复合是煤自燃的最普遍的规律。因而仅此就可以得到许多可能与煤炭自燃相关的可观测指标。

吸氧量

煤氧复合作用这一学说是从1870年开始的，经过许多人的探索得如下一些结果：

1. 1870 年瑞克特(Rachtan,H.)经实验得出：一昼夜每克煤的吸氧量为 0.1～0.5ml,其中褐煤可达到0.12ml。
2. 1945 年姜内斯(Jones,E.R.)提出：常温下烟煤在空气中的吸氧量可达0.4ml/g。
3. 1951 年，前苏联学者维索沃夫斯基等提出：煤的自燃正是氧化过程自身加速的最后阶段，但并非任何一种煤的氧化都能导致自燃，只有在常温条件下，稳定、绝热、氧化过程能自身加速的才能导致自燃。这种氧化反应的特点是分子的基链反应：即每一个参加反应的团粒或者说在链上的原子团首先产生一个或多个新的活化团粒(活化链)，然后，又引起相邻团粒活化并参加反应。这个过程从开始要持续地进行一段时间，即通常所称的“煤的自燃潜伏期”。他们通过实验还发现，烟煤低温氧化后，着火点降低，活化度提高，易于点燃。低温氧化过程的持续发展使得反应过程的自身加速

作用增大，若最终生成的热量不能及时散发，就会进入自热阶段。

4. 20 世纪 60 年代，抚顺煤研所通过大量煤样分析测定：100g 煤样在 30℃的条件下经 96h，吸氧量小于 20ml 时属于不自燃的煤；超过 30ml 时属于易自燃的煤。这也说明，在常温时，煤的吸氧量愈大，愈易自燃。

温度

从上述的煤氧复合作用学说分析得到，温度也是监测自燃的一个重要可观测因素。因为在自燃之前必定有温度的渐变过程，自燃与爆炸不同，不是瞬时完成的。引起煤自燃的根本原因是煤岩体与某些物质的作用产生热量，这些热量使煤体温度升高而促进氧化反应最终导致自燃。

煤岩体

煤岩体并非一个均质体，它是含有多种成分的有机和无机物质，其化学结构、物理性质、煤岩体组分等各产地都有很大差别；不同的煤岩其氧化性和放热性亦不同。煤岩体中有机物和无机物含量，化学结构，物理性质，特别是煤岩体组分也可以作为候选的可观测指标。

放热效应

煤自燃的过程非常复杂，表现为有些条件下煤会自燃，而另一些条件下煤不会自燃。放热效应是煤体升温引起自燃的主要因素。能够引起煤岩放热的主要因素有：

- 破碎煤岩的外表面积和微孔隙对瓦斯和氧的吸附产生的热，
- 煤的有机质及某些无机矿物质与氧产生的反应热，
- 水对煤及含硫矿物质的水解热、润湿热以及汽化潜热，
- 物理吸附产生的热，
- 化学吸附和化学反应放出的热。

这些热量在一定条件下得以积聚，可使煤体温度上升，会促进煤自燃过程。

煤氧复合放出的热量

引起煤自燃的最初导因在不同条件下是不同的，但不管是哪种导因，都含有煤与氧复合产生热量的成份，并且，煤自燃过程的自动加速发展主要是煤氧复合作用的结果。根据专家们多年来实验研究得知：任何一种煤都存在煤氧复合放出热量的属性，只是不同的煤的煤氧复合能力和产生热量能力差别较大。

放热强度与散热强度

煤氧复合过程中能否引起自燃取决于煤体的放热与散热。当热量产生速率

大于散发速率时，煤温上升就会引起自燃；也可以说，任何一种煤都存在煤氧复合产生热量的属性，即都存在自然发火的属性，但能否发生自燃则取决于煤氧复合过程中的放热强度和环境散热强度。

第三节 国内外研究现状

煤炭自燃是一种复杂的物理、化学反应。煤炭自然发火机理的研究已有100多年的历史，先后提出了多种理论和假说。近些年，国内外学者从不同的角度、采用不同的方法对煤炭自燃的机理及预测预报进行了研究。我们简述如下：

1. 舒新前等[3]用热重分析方法分别研究了神府煤和汝箕沟无烟煤的氧化自燃动力学过程，认为煤炭低温氧化遵循阿仑尼乌斯定律。
2. 彭本信[4]对我国长焰煤、气煤、焦煤、贫煤及无烟煤等煤种的70个煤样进行TGA、DTA、DSC及热分析红外光谱试验，测定煤的自燃倾向性，研究煤的自燃机理，查清了变质程度浅的煤容易自燃的主要原因是其氧化放热量大于变质程度深的氧化放热量。
3. TeVrucht[5]采用FTIR检测脂类C—H吸收峰强度的变化求取煤氧反应的活化能和速度常数。Kelemen采用XPS考察了煤表面O—C原子化的变化；并据此求出在295-398K时表观活化能为11.45lkJ / mol。Martin[3]用SIMS研究了23℃、70℃和90℃时煤表面O_2浓度变化，计算结果说明70℃前后活化能差别很大，因此在氧化反应的第一周内，温度小于70℃；具有以吸附扩散过程为主的低活化能特征。刘剑等通过对煤活化能的理论研究，推算出了活化能的计算公式。
4. Shinn[6]提出的煤结构模型是由结构表征和反应性数据提出的，它的结构可以看作是由许多包含有杂原子的聚合环簇，由亚甲基键、醚键等连接在一起，它代表高挥发性烟煤的一般结构。Shinn给出了其模型在相对温度和条件下反应交链热解所得到的可溶性产物。
5. 徐精彩等在煤低温自然发火实验基础上，建立了松散煤体自然发火的动态数学模型，数值模拟煤的自燃过程，预测煤在不同条件下的自然发火期，以弥补实验之不足。较好地满足现场实际工作的需要。由于模型预测时需要实验室和现场的大量参数，因此工作量较大，且有很多参数不能准确确定，而这些参数对预测结果的影响较大，偶尔会出现参数设定不当，而使

预测结果与现场实际有较大误差的情况。

- 对于煤自然发火期的认定

煤体接触空气到开始发生自燃的时间称为煤自然发火期。在实际条件下，煤自然发火期并非定值，对特定地点而言，由于各种因素已定，煤自然发火期基本为定值。通常煤自然发火期可划分为实验自然发火期、煤层最短发火期和煤自燃实际发火期。具体讲：实验自然发火期即为在实验条件下，使松散煤体从供风开始到冒青烟所经历的时间，其意义在于确定各类煤体相对自然发火性的强弱。煤最短自然发火期为矿井现场中上述参数得以最佳组合地点的松散煤体从暴露于空气中到冒青烟所经历的时间。煤最短自然发火期有可能小于实验自然发火期。煤自燃实际发火期为特定地点的特定条件下相对应的自然发火期。井下防灭火工作最为关注的就是煤自燃实际发火期。

最早自然发火期预测主要凭经验和现场统计进行粗略估计，通常以月为单位，不能有效地指导现场工作。后来各国采用自然发火实验台模拟法预测煤自然发火期。但该方法还存在如下不足[7]：(1)实验周期较长，影响因素多，实验条件难以控制；(2)模拟煤量有限；(3)实验费用较高，不可能多次重复实验；(4)实验条件单一等。之后随着计算机的发展，许多学者开始采用数值模拟的方法预测煤的自然发火期。

- 煤炭自燃的预测预报的研究内容

近二十年来，世界各相关研究机构对煤炭自燃产生的机理及其化学物理间的相互作用进行了研究，以期开发出可靠的预测预报方法和火源探测技术，研制出性能可靠的检测仪器，最大限度的消除其危害。煤炭自燃的预测预报研究分两种情况：

  - ✧ 指处于低温氧化阶段，还未出现自然发火征兆之前，仅根据煤的氧化放热特性和实际开采条件，超前判断松散煤体自燃的危险程度、自然发火期及最易自燃的区域。
  - ✧ 指煤层开采后，松散煤体放热，引起升温，进入自热阶段，在煤体冒烟和出现明火之前，根据煤氧化放热时引起的标志性气体、温度等参数的变化情况较早发现自燃征兆，判断自燃的状态。

- 火灾预报的主要内容是：确定火源发生的空间位置，判断煤炭自热的程度。
- 对于火灾预报的要求是：及时、准确和可靠。这就需要正确的选择预报参

数和预报方法，使用高灵敏的传感器，且传感器安装的位置要合理，信息传输系统要可靠，应尽量防止漏报和误报。

- 本文的技术线路

本文的目的在于基于对伴随着自燃现象出现的这些特征指标的观测数据，应用支持向量机（SVM），粗糙集（RS）等现代数据处理技巧，基于可得到的不完备的大同煤矿煤炭自燃情况的可观测指标数据以及空气样分析数据，对指标体系进行约简，给出删除某些指标对于预报精度所付出的代价，给出最重要的观测指标的代表性，以及基于新老设备之下由于指标不同而引起的预测精度的差异。以期在煤炭自燃发生前，把握好自燃的发展趋势，对于火灾的早期预报和治理起到积极的作用。

第四节 目前所采用的煤炭自燃预测预报方法简介

目前所应用的预测预报技术方法主要有：自燃倾向性预测法，综合评判预测法，统计经验预测法，计算机模拟法，测温法，标志气体分析法以及神经网络的应用等等。

1. 自燃倾向性预测法[8]

该方法主要是根据煤自燃倾向性不同，划分煤层自然发火等级，以此区分煤层的自燃危险程度，从而采取相应防灭火措施。80 年代前，国外对煤自燃倾向性的测试方法主要以煤的氧化性为基础，大体可分为：化学试剂法和吸氧法。80 年代后，美国研究出利用绝热炉(adiabatic heating oven)测定煤炭最小自热温度，评估煤炭自燃倾向性。我国从 50 年代初期即开展对煤炭自燃倾向性鉴定方法的研究，先后研究了克氏法、着火点温度降低值法、双氧水法及静态容量吸氧法，但由于各种方法所存在的缺点和客观原因，我国一直沿用着火点温度降低值法。进入 90 年代，我国推广使用色谱动态吸氧法。以上煤自燃倾向性测试方法仅考察了煤与氧的作用速度和作用量，而没有考察其作用的效果，即热效应，也没有考察它们随煤温变化的动态趋势，因而无法真实全面地反映出煤的内在自燃性，与煤实际开采条件下自燃的实质相差甚远。因此，煤层自燃倾向性预测法仅能粗略判断出煤层的自然发火危险程度，不能确定实际条件下松散煤体的自燃危险程度。因此，近 20 年来世界各主要产煤国对煤层自燃的

预测，主要朝着准确预测实际条件下松散煤体自然发火期和自燃危险程度的方向努力。

2. 综合评判预测法

综合评判预测法是基于模糊数学建立起来的一种预测方法。为预测实际开采条件下煤层的自燃危险性，根据影响煤层自燃危险程度的内、外因素---煤的自燃倾向性、地质赋存条件、通风条件、开采技术因素和预防措施等，进行主观判断、分析评分，然后用模糊数学理论逐步聚类分析，根据标准模式计算聚类中心，对开采煤层自燃危险程度进行综合评判预测[9]。通常的也采用 BP 神经网络来预测煤层的自燃危险性。综合评判的特点是利用大量的统计资料，分析煤炭自燃主要因素的影响程度，粗略预测煤层自然发火危险程度，而对发火期以及可能发火的区域则无法进行预测。

3. 统计经验预测法

这种方法是建立在已发生自然发火事故统计资料基础上，分析预测巷道松散煤体开采条件下的自燃危险程度。该方法只能根据巷道实际情况和自然发火统计资料，粗略判断巷道可能的发火区域和高温点位置。

4. 数学模型模拟预测法

20 世纪 80 年代以后，世界各国针对采空区或地面煤堆的自燃条件，根据传热、传质学和流体力学等理论，分析煤体温度和各种影响因素之间的关系，建立了多种自然发火动态数学模型，模拟煤的自燃过程，预测采空区或煤堆的自然发火危险性[10]。“八·五”期间,西安矿业学院在煤低温自然发火实验模拟和综放面采空区大量现场跟踪观测的基础上，根据对采空区三带划分和综放面采空区自燃主要影响因素等理论分析结果，建立了综放面自然发火动态预测模型。该预测模型技术经过部分煤矿多个综放面的预测验证，其预测精度较高，能满足现场需要。

5. 标志气体分析法

标志气体分析法预报技术主要利用一氧化碳、乙烯、乙炔、烯烷比等做为指标气体预报煤自燃的发展过程。由检测到的气体浓度可以判断煤炭自燃的危险性。由于其分析技术和监测系统都已配套，故该技术在煤矿中得到了广泛的应用。但实际中，煤种不同则煤在氧化自燃过程中气体产物的组成和变化规律也有所不同；其次，有些指标气体是煤自燃发展过程中煤温升高产生的氧化气体，只能在煤已经自热或自燃时才能检测到，且气体产量较少，会随着风流流

动；另外，人工取样工作量较大，间隔时间长，不能进行连续实时检测。可见，该预报技术还存在许多问题有待进一步研究。

6. 测温法

测温法就是测定井下煤与周围介质的温度变化情况，因为巷道松散煤体及周围介质温度的升高直接反映着煤的氧化程度。测温法是发现煤炭自热和探寻高温点及火源的最直接、可靠的方法。但巷道松散煤体内部温度的测温技术尚未完全解决。目前，探测煤自然发火的测温仪主要有红外线测温仪和温度传感器两种。

上述方法对于开采煤层的自燃火灾发生情况的预测比较有效果，但是无法完成对采空区或工作面煤炭自燃的实时监控。测温法比较直观，而且受外界因素影响小，但是对预测系统的要求很高，特别是传感器的选择，这样就必然加重经济上的负担，且时常有误报的可能。计算机模拟法是一种理想化的模拟，与实际存在一定的差异。但是采用计算机对煤炭自燃情况的预测预报研究费用较低，可以综合多种预测手段，所以也一直颇受人们的关注。

7．应用神经网络预测预报煤炭自燃

如上所述，各种方法都有优缺点，由于煤炭自燃现象本质上是一种非线性系统，只有从非线性科学角度进行认识，才能更深刻地认识煤炭自燃的规律。

人工神经网络（Artificial Neural Network，ANN）具有强大的非线性映射能力，擅长处理模糊的、随机的、低精度的信息，是动态预测煤炭自然发火的有效方法。目前在国内见于报道的最早应用神经网络于煤层自燃预测方面的研究是1996年西安科技学院学报的黄梦涛等撰写的《神经网络在煤低温自燃实验炉建模中的应用》，该论文把神经网络应用到煤低温自燃实验炉的模式识别上。1997年蒋军成和王省身应用神经网络预测开采煤层的自燃危险性，对南屯煤矿某煤层进行了自燃危险性预测。1999年王德明和王俊应用无导师神经网络对煤炭自燃危险性进行了预测和分析，克服了有导师神经网络预测模型的不足。目前，西安科技学院的徐精彩教授应用人工神经网络做煤氧化性的预测。1999年，辽宁工程技术大学的宋志等人以采场自然发火为充要条件，依据时空一致的观点，分析自然发火因素空间，提出用人工神经网络来预测煤自然发火期，自燃地点和自燃时期的模型，把采场自燃预测模式识别问题看成是N个影响采场自燃因素所形成的N维空间到M 维空间的映射。但是在文献中只提供了预测模型，并未对具体采场自然发火做预测。2004年，邬剑明和王俊峰做了基于神经网络

的煤层自然发火的非线性预测的研究，当山东枣庄柴里煤矿 2313 工作面风流出现 CO 气体，通过对 2313 工作面所布置采样点取样分析检测到的乙烷、乙烯气体，应用人工神经网络进行预测。目前所做的自燃预测一般都是对待开采煤层所做的一次性推断，不能很好的满足工作面实时监控的要求。而传统的 BP 算法是一种典型的监督学习方法，其学习是由期望输出与实际输出之间的差别驱动的，需要积累足够的样本后统一进行训练，不能根据新样本渐进的调整网络参数。其次，经典的神经网络对于随时间而不断变化的时间序列多步预测问题具有很大的局限性。而 Sutton 于 1988 年提出的时差方法(TD Method)是由实际输出之间的差别驱动的，解决了监督学习中得不到期望输出这个难以回避的问题。在煤矿中，如果能用神经网络的方法对采空区的易燃区域进行实时预测，对于矿井的防灭火工作会有很大的帮助。目前，应用这种方法的有 2003 年中国矿大的曾勇和吴财芳对平顶山六矿的瓦斯突出进行的预测；安徽理工大学何启林老师对孔庄做的采空区自然发火的预测。但是，这些都是对一个生产矿井所做的整体性的预测[11]-[16]。如果能够开发出针对具体工作面的煤炭自燃预报系统，将对矿井的安全生产有很大的帮助。

本课题的研究内容

综上所述，本课题的研究计划按照如下技术线路展开：

1. 收集本文提到的文献中的观测数据和可观测指标。
2. 从煤自燃的机理出发，将现有使用过的可观测指标体系复原出来。
3. 利用粗糙集理论将收集到的指标进行约简。
4. 在指定某些指标不可实现的条件下的条件约简。
5. 将约简的指标集看作自变量，将其后固定时间内煤炭自燃或不自燃看作因变量，对于这样的高维非线性函数，使用*支持向量机*来模拟。或者将约简的指标集看作自变量，将距离煤炭自燃日的时间长度看作因变量，对于这样的高维非线性函数，使用神经网络模拟。
6. 检验现在已经发表的国内文献分别使用的方法中涉及到的可观测指标的不完备性。
7. 检验现在已经发表的文献分别使用的方法对于其它地区的不适应程度。
8. 针对约简后的可观测指标集，或者条件约简的指标集，重新评估各地区现在使用的关于煤自燃监测的设备的正确率。

9. 筛选出具有普适性的最小约简的指标集和各地可以接受的独特的条件最小约简的指标集，指出各地区监测设备应当重点注意的指标气体观测指标的精度。

本文的目的就是期望能应用现代数据处理技术，提前预测出相关地点的指标气体浓度、温度等，进而推断出可能发生煤炭自燃的区域以及自燃发展的趋势，对于井下煤炭自然发火的防治提供有益的帮助。

第二章 煤炭自燃的非线性特征

煤炭自燃(煤炭自然发火)，是指在没有外来热源的情况下，由于煤炭自身氧化积热，使煤的温度升高，而发生的燃烧现象。煤的氧化和自燃从本质上来说是基-链反应，即具有自燃倾向性的煤与空气接触，能吸附空气中的氧气从而在煤的表面生成不稳定的初级氧化物，使煤的化学活性增强，进而煤炭氧化速度加快，氧化放热量加大，当热量不能及时散发时，煤温就会逐渐升高，当达到煤的着火点，便开始自燃[17] 。

第一节 煤的自燃过程及影响因素

2.1.1 煤的自燃过程

煤的自燃过程一般可分为三个阶段(如图 2.1)：准备期(又称潜伏期)；自热期；燃烧期。各阶段的特征分别如下：

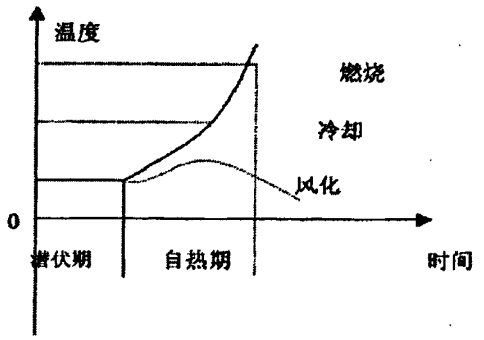

图 2.1 煤的自燃过程

准备期：具有自燃倾向性的煤与空气接触时，吸附空气中的氧(O_2)而生成不稳定的氧化物羟基(OH)与羧基(COOH)。此时氧化热量很少，观测不到煤体温度的变化，也看不到其周围环境温度上升。煤的氧化进程平稳而缓慢，但是煤的重量略有增加，化学活性增强，这个阶段通常称之为煤的自燃准备期，又称*潜伏期*。潜伏期长短取决于煤的变质程度和外部条件，因此具有高度的随机性。

自热期：经过潜伏期之后，被活化的煤的氧化速度增加，产生的热量较大，若来不及散发，煤温继续升高，超过临界温度(一般为 70～90℃)时，煤温上升

急剧加速，氧化进程加快，这时空气中的O_2减少，CO, CO_2含量显著增加，出现烃类气体(烷烃或烯烃)；煤中水分蒸发，空气湿度加大，并出现雾气，巷壁或支架上有水珠凝结，后期则可出现煤的干馏，可生成芳香族的碳氢化合物(CxHy)、一氧化碳(CO)等可燃性气体，在工作面可嗅到煤油或煤焦油气味，这就是煤的自热期。

燃烧期：进入自热期后，如果煤温继续上升达到着火温度(300-350℃)，就会导致煤的自燃，进入燃烧期，此时就可看到烟雾，明火的产生，并有大量的 CO（如上隅角检测通常可达到 1000～3000ppm)、CO_2等气体产生，目前被认为是自燃的标志性可观测气体。

风化期：如果在煤温上升到临界温度以前，采取措施，改变其供氧和散热条件，煤温会很快地降下来，这样便进入了风化状态，一般来说已经风化了的煤，不会再次发生自燃。如图 2.1 中虚线所示。

从煤的自燃发展可见，煤的自燃实质上就是自身氧化加速的过程，其氧化速度之快，以致产生的热量来不及向外界放散，从而导致了自燃。所以，在煤的氧化急剧加速之前，做出准确的预测预报，及时采取措施，就能避免煤的自燃。

如前所述，煤矿井下火灾，90%以上都是由于煤炭自燃引发的。煤炭自燃严重影响着矿井的正常生产，威胁井下人员的生命安全，烧毁煤炭资源。但由于井下环境恶劣复杂，往往难以直接观察到着火现象，特别是对于早期的自燃，不能及早的查出，往往只有等到有了较明显的特征时，才有所察觉，而这时煤炭已经燃烧到一定程度了。因此，提前预测自燃非常重要也很有挑战性。

在实际工作中，矿井某一个区域或采掘工作面出现下列情况之一时，便定义为自然发火：

1. 由于煤炭自燃出现明火，火灾烟雾，煤油味等现象；
2. 由于煤炭自燃使环境空气、煤炭、围岩及其他介质温度升高，并超过 70℃。
3. 由于煤炭自燃在采空区或回风中出现一氧化碳或者其他重碳指标气体，其浓度超过矿井实际统计的自燃发火的临界指标，并有上升趋势。

在开采过程中发生过煤炭自燃的煤层，则定为自然发火煤层。

2.1.2 煤炭自燃的影响因素

矿井自燃火灾形成的过程是许多因素综合作用的结果。影响煤层自然发火

的因素包括煤的自燃性能、开采技术、地质因素等[18] 。

(1) 自燃性能

煤的自燃性能主要受下列因素影响：

①煤化程度。煤化程度是影响煤炭自燃倾向性的决定性因素。就整体而言，煤的自燃倾向性随煤化程度增高而降低，即自燃倾向性从褐煤、长焰煤、烟煤、焦煤至无烟煤逐渐减少；局部而言，煤层的自燃倾向性与煤化程度之间表现出复杂的关系，即同一煤化程度的煤在不同的地区和不同的矿井，其自燃倾向性可能有较大的差异。

②煤岩成分。煤岩成分对煤的自燃倾向性表现出一定的影响。各种单一的煤岩成分具有不同的氧化活性，其氧化能力按镜煤>亮煤>暗煤>丝煤的顺序递减。镜煤受力后易形成碎屑，含氢、氧和挥发成分高，易氧化自燃；丝煤虽然本身氧化活性弱，自燃点高，但丝煤组分中的细胞空腔能增大煤的裂隙和反应面，为氧向深部扩散提供通路，促进烟煤氧化自燃。

③水分。煤的外在和内在水分以及空气中的水蒸汽对煤在低温氧化阶段起一定的影响，既有加速氧化的一面，也有阻滞氧化的因素。煤在自燃的过程中，只有其水分降低到一定值后，氧化速度才会加快，煤温才会急剧升高；煤湿水后再干燥与未湿水的干煤相比，化学活性增加；在低温时煤对水蒸汽的亲和性比对氧大，水蒸汽凝结成水时生热比氧化生热量大；因此稳定地保持采空区内空气具有较高的湿度，增加并保持煤本身的湿度，也可以抑制煤的低温氧化。

④煤的空隙率。煤的自燃倾向性随煤的空隙率大小的不同而有很大的差异。煤的空隙率越大，其吸附氧气的能力越大，所以，空隙率越大的煤炭越容易自燃。煤的脆性越大，就越容易破碎，产生大量的碎煤。使其接触氧的表面积大大的增加，且着火点温度显著降低，所以就越容易自燃。

⑤煤中硫和其他矿物质。煤中含有的硫和其他催化剂，则会加速煤的氧化过程。一般来说，硫在煤中有三种存在形式，其中对煤的自燃起主要作用的是硫化铁，硫化铁氧化时放出的热量比煤大的多。因此，硫化铁在煤中占的比重越大，越有利于煤的自燃。

(2) 开采技术

除了上述对于自燃性能的影响因素之外，在矿井的实际生产过程中，由于矿井的开拓方式、采区巷道布置、回采方法和回采工艺、通风系统和技术管理等开采技术和管理水平等因素也对自然发火起决定性作用。许多开采同一煤田、

相同煤层的不同矿井，甚至同一矿井不同的开采时期，在煤的性质、煤层赋存条件基本相同的矿井，由于上述因素的差异，造成自然发火的次数明显不同。开采技术对自然发火的影响主要表现在三个方面：

①矿井开拓方式和采区巷道布置既决定保护煤柱的数量、机器大小，又决定所留煤柱受压与破碎程度，既决定可燃物的分布和集中情况，又决定向这些可燃物供风的时间。

②回采方法和回采工艺决定的因素是回采率和工作面推进速度。其中，回采率如果丢煤过大，则容易引起自燃，所以在实际生产过程中，应该尽量减少丢煤。而工作面推进速度则对于煤的自燃有重要影响，一般来说，如果推进速度大于工作面散热带到采空区窒息带的最大距离和浮煤最短自然发火期之商时，就不容易引起自燃。所以，我们一般尽可能快的提高推进速度。

③漏风在煤炭氧化过程的热平衡关系中起两方面的作用：一是向煤提供氧化所必须的氧气，促进氧化发展；二是带走氧化生成的热量，降低煤温，抑制氧化过程发展。易燃风速的划分目前还没有一定的标准，对于采空区，有人认为当其单位面积上漏风量大于 1.2 $m^3/min.\ m^2$或小于 0.06 $m^3/min.m^2$，都不会自燃，最危险的漏风量是 0.4～0.8 $m^3/min.m^2$.

(3) 地质因素

地质因素也是影响自燃的重要成因，其中包括：

①煤层厚度。煤的厚度越大，自然发火的危险性越大。厚煤层容易自燃的主要原因有几点：

- 厚煤层回采率低，遗留大量浮煤和煤柱。
- 矿山压力较大，煤柱、煤壁被压裂而漏风。
- 煤层越厚，其回采时间越长，容易超过煤层的自然发火期。
- 分层开采时，上部煤层采空区的浮煤氧化产生的热量难于通过下部煤层散发，相对容易自燃。
- 煤层越厚，开采巷道越容易冒顶，形成冒顶空洞，容易引起自燃。
- 煤层越厚，遗煤分布越大，越容易自燃。

②煤层倾角。煤层倾角较大时，尤其对急倾斜煤层容易造成丢煤多，而且对采空区的封闭也很困难。在同等的矿山压力作用下，倾角越大的煤受压后破碎程度越大，所以煤的自燃倾向性越大。

③煤层埋藏深度。煤层埋藏越深，煤的原始温度越高，煤中含有的水分较

少，其危险程度越高。相反地， 而对于埋藏很浅的煤层，开采过程中容易与地表沟通，产生漏风通道，也容易引起自燃。也就说，太深或太浅都容易自燃。

④地质构造。在有断层和岩浆入侵等地质构造被破坏的区域，煤体较破碎，存在大量裂隙，而且在岩浆岩侵入时使煤层受到干馏，空隙率增加，强度降低，因此，自燃危险程度增大。

从上面的分析可以看出，煤炭的自燃影响因素众多，但根据经验及现场实际可知：煤的自燃倾向性，推进度，煤层倾角，漏风量，煤厚，空隙率等对煤自燃发展影响较大，是主要影响因素。我们也将通过数学的手段，证明它们是主要影响因素，且给出各自的贡献率。

2.1.3 煤的自燃机理

在研究的过程中，如果能弄清煤炭自然发火的机理，对于煤炭自燃的预测预报有着重要的意义。关于煤炭自燃的机理问题是当前各国都没有解决的大难题。各国的学者发表了各种学说以解释煤炭自燃的原因，本文在第一章中已做了阐述，概括起来讲主要有下面七个方面的学说：自然发火黄铁矿学说；细菌导因学说；磺化导因学说；酚基导因学说；活性煤质粘土岩导因学说；过氧化物催化导因学说；煤-氧复合物导因学说。

目前得到大多数学者认可的是煤-氧复合物导因学说，也是目前最为流行的煤炭自燃机理学说。但煤一氧复合反应发生的机理至今仍未真正了解。进入20世纪90年代，国内外学者又提出了一些更具体的学说[19]：

①李增华于1998年提出煤自燃的自由基作用学说[20]，该学说认为煤是一种有机大分子物质，在外力(如地应力、采煤机的切割等)作用下煤体破碎，产生大量裂隙，必然造成煤分子的断裂。分子链断裂本质就是链中共价键的断裂，从而产生大量自由基，为煤自然氧化创造了条件，引发煤的自燃。Unal 用ESR(顺磁共振)技术测量了煤的氧化过程中自由基浓度变化后指出：煤的低温氧化至少部分是通过自由基反应机理，支持了自由基链反应机理。

②IPXNM 等人于1990年提出了电化学作用学说，该学说认为，煤中含有铁的变价离子，组成氧化还原系 Fe^{2+}/Fe^{3+}，其在煤的氧化反应中起催化作用，在煤中引起电化学反应，产生具有化学活性的链根，从而极大地加快煤的自动氧化过程，引发自燃。

③Lopez 等人于1998年提出氢原作用学说，该学说认为，煤低温氧化过程

中，由于氢原子在煤中各大分子基团间的运动，增加了煤中各基团的氧化活性，从而促进煤的自燃。

④Wang HH.等人于 1999 年提出了基团作用理论，该理论利用孔模型模拟了煤中孔隙的树状结构，提出所有有效孔所构成的树状能够到达煤粒表面，使煤中各基间以及与氧气能充分作用，从而导致煤的自燃。

从上面的不同学说可知，在煤炭自燃的形成过程中，须具备的条件为：

- 具有低温氧化特性即自燃倾向性的煤成破碎状态堆积存在；
- 通风供氧维持煤的氧化过程不断发展；
- 在煤的氧化过程中生成的热量大量蓄积，难以及时放散。

只有同时具备了这三个条件，煤炭才可能在一定时间内引起自燃。

从煤的自燃发展可见，煤的自燃实质上就是自身氧化加速的过程，其氧化速度之快，导致产生的热量来不及向外放散，引起自燃。所以，在煤的氧化急剧加速之前，做出准确地预测预报，及时采取措施，就能避免煤的自燃。本文采用的就是这种“氧化机理”。

第二节 煤炭自燃的非线性特征

从本质上说，非线性系统动力轨道的拓扑和演化特性，是深刻认识系统灾害的科学基础。煤炭自燃是矿井灾害的一个典型的非线性现象。灾害现象的孕育、发生和发展，与构成灾害系统的各要素及其环境之间的非线性相互作用所产生的分叉、突变、孤立子、分形、自组织和失稳等非线性现象有着密切关系。矿井灾害系统内各要素之间与其环境系统的相互作用也具有强烈的非线性特征。只有从非线性科学的角度提高认识，才能对火灾过程规律的认识更完整、更深刻。

从前面内容的分析可以看出，煤炭自燃过程中受到了各种因素的影响，既有煤炭本身性质的影响，也有外部环境和开采技术的原因。煤体低温氧化产生的热量在系统中形成初始随机涨落，并通过煤体裂隙向周围煤体及环境释放，这是由煤炭本身的特性所决定的。其自燃倾向性等级高，则形成的热量就大，反之则相反。自燃带中的煤体氧化放热反应所产生的热量，除了加热周围的煤体之外，还加速其氧化放热反应速度。因为有部分热量反馈回自身，使自身氧化放热反应加速(自加速、自催化、自组织)，这是重要的非线性特征，我们在

本文的研究中将会展现它。如果开采技术合理，则向周围煤体传递的热量多，散发的快，向其自身反馈回的热量相应的减少，使自身氧化反应保持在原状态或者变化缓慢。由于氧化放热反应关系的非线性，以及氧化放热对温度的非线性，初始涨落被非线性的放大。这一反馈-放大过程继续下去，产生的热量若不能及时地散发，将使煤体内蓄积的热量越来越大，并超过煤炭的着火温度，使该区域煤炭产生燃烧反应。此时煤体内部处于阴燃状态。随着阴燃持续进行，煤体内部的热力风压将进一步增大，导致煤体内部漏风量进一步加大。热流体在运移过程中，使受其影响的煤体氧化放热反应加速，当得到充足的氧气时就转化为明火燃烧[21]。

目前，针对自燃的非线性特征的研究方法主要有：小波理论、遗传算法、分形理论。这些方法对于刻画自燃的特征有较大的贡献。但是针对本文的实际需要，我们将采用现代数据分析的技术将数值化的结果呈现出来。

本文的研究目的是：首先应用现代数据分析等数学知识对用于判定煤层自燃情况的各因素属性的大小进行研究，在此基础上，可进一步判断哪些煤层需要加强重点监测，从而为矿井的防灭火工作提供指导性意见。其次运用支持向量机、粗糙集等知识分析指标气体，进行约简，通过计算分析得到预测精度，为进一步做好指标气体的监测分析提供指导，为解决煤炭自燃的预测问题提供新的思路和方法。

第三章 大同煤炭自燃机理及指标气体的理论研究

第一节 煤炭氧化自燃机理的理论研究简介

大同是我国著名的煤炭生产基地和战略要地，煤炭开采过程中的安全问题始终是大家关注的大问题，针对大同矿区煤炭自燃现状，近年来有关专家从不同角度对煤炭氧化自燃的机理进行研究，通过这些基础研究为大同矿区煤炭自燃的防治工作提供指导。

1. 煤的分子结构研究

煤的分子结构研究主要是通过红外光谱实验分析，得到煤分子结构的官能团归属，总结前人的研究成果建立基于化学反应的煤化学结构基本单元模型，对构建的煤分子化学基本结构单元进行了优化，得到分子构型参数。

2. 通过热重实验分析煤氧化自燃的过程

基于煤氧复合作用学说，由于煤与氧发生化学反应，煤分子内部化学键重新分布，其内能也随之变化，放出热量，煤结构的变化反过来又会影响整个物理化学变化的进程[17]。有关研究人员通过对大同矿区煤样的热重实验分析，在程序升温条件下，分析煤氧化自燃的热重曲线，总结出煤的自燃机理：

1) 煤氧化自燃可分为三个过程

三个过程是指升温失水失重过程、升温增重氧化过程、升温燃烧失重过程。在失水失重过程中，从25～60℃左右为吸热反应，之后为放热反应，到大约400℃之后为吸热反应，如图3.1。在煤的自燃过程中，从25～60℃吸收热量，所吸收热量的来源是煤与氧发生物理化学吸附的吸附热。如果没有吸附使煤失水，煤就不能发生自燃。

2) 煤的着火活化能能够反应煤自燃的本质

分析热重曲线表明，煤体温度从25～110℃为失水阶段，从110-260℃左右氧化增重阶段，270℃以后为着火燃烧阶段。在失水结束后的增重阶段，是经过失水干燥后的煤大量吸附氧气，并发生复杂的化学反应，出现增重现象。煤中的官能团发生氧化反应放出热量。从增重转为失重的拐点温度为煤的着火温度，增重阶段对应的活化能为着火活化能。着火活化能能够反应煤自燃的本质。

前人对大同煤矿不同层位和工作面的多种煤样通过计算煤的着火活化能

得出的煤的自燃倾向性指数，本文收集的煤的自燃危险性指数就是用着火活化能来表示的，煤的自燃危险性指数越小越容易发生自燃，越大越不易发生自燃。

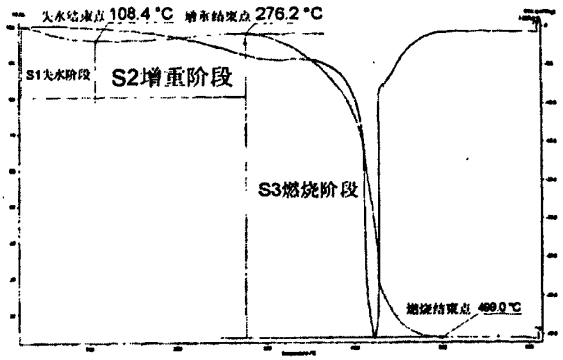

图 3.1 马矿 7#层实验煤样热重程序升温条件下煤的氧化自燃过程

3) 实验煤样的着火温度

从加热失重到失重结束转为增重阶段的温度为失水温度，从增重阶段转为失重阶段温度为着火温度。大同矿区部分实验煤样的着火与失水温度见表 3.1。

表 3.1 大同煤矿部分实验煤样的着火与失水温度

矿 名	采样地点	失水温度℃	着火温度℃
白洞矿	5#层	130.5	301.5
	11#层 307 盘区	129.0	298.2
晋华宫矿	11#层 2502 面	122.6	299.3
	12#层 8111 面	125.6	288.8
	10#层 8118 面	130.7	300.9
马脊梁矿	7#层	113.6	276.1
煤峪口矿	14#层 5809 面	130.1	296.8
	11、12#合并层	119.5	298.7
四台矿	12#层 8208 面	119.6	303.5
	11#层 8429 面	113.7	287.6
同家梁矿	11#层 410 面	122.0	301.5
	12#层 2808 面	121.5	310.1
永定庄矿	11#层 317^{-1} 面	119.1	289.5
	12#层 81715 面	119.1	290.1
云岗矿	12#层 8804 面	150.1	305.4
	8#层 51119 面	129.9	287.8
	7#层 87104 面	133.1	283.2

第二节 煤氧化自燃指标气体的研究

大同矿区煤矿井下火灾的早期发现很早以来一直主要依靠观测烟、味、温度及人工采样检测 CO 浓度等来判断，但近年来气体分析法被广泛应用。气体分析法的实质是研究、检测煤氧化升温过程中所释放的气体成分与矿井空气的区别，并将足以能够用来判断自然发火征兆的气体作为自然发火标志气体，根据标志气体变化的规律推断煤自热温度的一种预测煤自燃过程的方法。针对大同煤矿煤炭自燃的现状，有关技术人员通过分析煤样氧化自燃生成气体红外 3D 光谱图对指标性气体进行研究，得到指标性气体显现峰和强峰的温度。通过标准浓度气体的 FTIR 实验，建立产生指标性气体的浓度与温度的关系。

- **标志性气体浓度与温度关系**

应用傅利叶变换红外光谱仪实验测定大同煤矿煤样在氧化自燃过程中的不同温度下生成的气体，生成气体红外光谱图峰值的大小与气体浓度成正比。部分实验煤样在不同温度下生成气体红外光谱图峰值的大小见表 3.2。

表 3.2 大同矿区部分实验煤样气体 3D 光谱图峰值表

矿名	采样地点	出现 CO 温度(℃)		出现 C_2H_4、CH_4 温度(℃)	
		显现峰	强峰	显现峰	强峰
晋华宫矿	11#层 2502 面	54	124	99	124
	12#层 8111 面	45	138	109	138
	10#层 8118 面	47	120	113	120
同家梁矿	11#层 410 面	38	133	83	133
	12#层 2808 面	32	105	70	105
四台矿	12#层 8208 面	49	119	104	123
	11#层 8429 面	61	129	106	129
忻州窑矿	11#层材料斜井	70	140	114	140
云岗矿	9#层 408 盘区	53	129	87	129
	11#层 2908 掘面	44	128	93	128
	7#层 87104 面	58	153	98	153
	8#层 51119 面	49	128	81	128

可以看出，煤种不同出现显现峰和强峰的温度是不同的。较低的温度下出现显现峰和强峰的煤种说明其容易生成 CH_4、CO、C_2H_4 等指标性气体，也就是说这种煤容易发生氧化自燃。

自燃很显然与温度的变化直接关联，但由于实际井下监测得不到煤层（特

别是采空区）的温度，因此使用与温度有关的标志性气体来监测自燃。下面列出某矿实验煤样氧化自燃生成标志性气体的浓度与温度关系表，见表3.3。

表3.3 四台矿11#层煤样氧化自燃生成标志性气体的浓度与温度关系表

温度(℃)	CH_4浓度 (ppm)	CO浓度(ppm)	C_2H_4浓度(ppm)
101.1916	——	95.815	——
106.9886	——	98.156	——
112.7848	——	110.487	——
118.5818	20.811	123.945	3.749
124.3789	26.095	191.329	5.521
130.176	35.290	146.073	4.339
135.9731	60.068	185.116	32.193
141.7702	26.858	257.016	3.970
147.5673	77.026	283.201	3.617
153.3643	51.419	330.966	1.469
159.1614	70.617	459.954	5.053
164.9585	38.130	557.779	1.440
170.7556	24.455	762.962	1.822
176.5517	53.955	944.896	2.150
182.3488	74.685	1248.761	10.890
188.1459	43.584	1574.899	4.732
193.943	58.163	1929.301	3.815
199.74	65.629	2385.884:	1.959
205.5371	30.727	2937.680	3.233
211.3342	64.991	3585.721	6.449
217.1313	25.315	4687.299	3.604
222.9284	15.164	6497.773	3.640
228.7254	2935.436	14470.829	68. 015
234.5225	6800.887	38539.617	285. 398
240.3187	6021.176	37909.853	240.909
246.1157	2868.176	43475.730	72.0166
251.9128	2906.745	42955.836	42.179
257.7099	3017.730	39434.737	30.038
263.507	3055.726	42653.103	28.016
269.3041	2872.985	57154.531	103.014
275.1011	2875.229	67152.504	73.583
280.8982	2918.721	58022.961	66.382
286.6953	2957.483	51073.853	58.736

利用上述数据,我们可作图来看CO浓度与温度的关系:

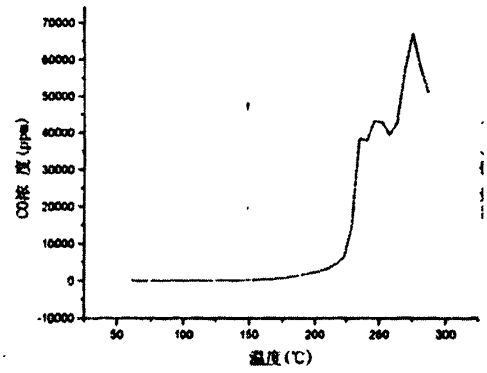

图 3.2 60～300℃ CO浓度与温度关系曲线

由图 3.2 可大致看出CO气体浓度随温度的变化规律：在60～150℃范围内CO浓度缓慢增大；150℃后CO浓度增大趋势明显，200～240℃CO浓度呈指数形式迅速增大；240℃后CO浓度仍有变大趋势，但波动较大规律不明显。各煤种不同，CO气体浓度变化的拐点温度不同，但总的变化趋势是一致的。

同理对于CH_4、C_2H_4也可作图观察，如图 3.3、3.4。

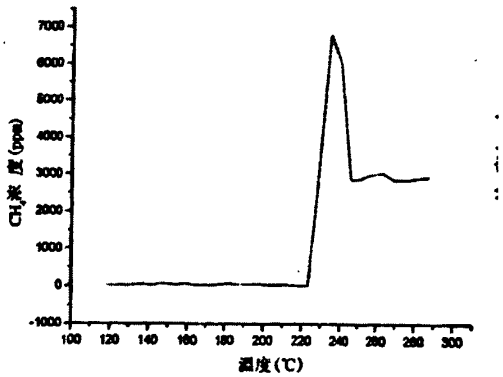

图 3.3 120～300℃ CH_4浓度与温度关系曲线

由图 3.3 可大致看出CH_4气体浓度随温度的变化规律：在120～220℃范围内CH_4浓度维持在很低的浓度范围内上下波动； 220℃后CH_4浓度明显增大，出现浓度峰值；250℃后CH_4浓度又维持在一个较高的浓度范围内上下波动。各煤种不同，CH_4气体浓度变化的拐点温度不同，但总的变化趋势是一致的。

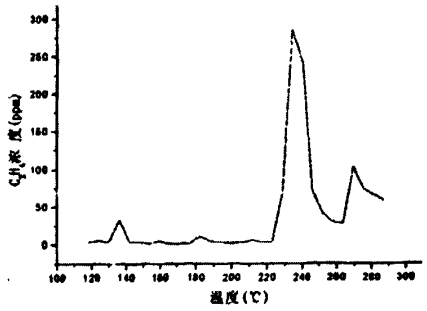

图 3.4 120～300℃ C_2H_4浓度与温度关系曲线

由图 3.4 可大致看出 C_2H_4气体浓度随温度的变化规律：在 120～220℃范围内 C_2H_4浓度维持在一个很低的浓度范围内上下波动； 220℃左右 C_2H_4浓度明显增大，出现浓度峰值；250℃后 C_2H_4浓度波动很大，不具有明显的规律性。各煤种不同，C_2H_4气体浓度变化的拐点温度不同，但总的变化趋势是一致的。

我们由有关实验数据及图形曲线看到了煤样在氧化升温过程中产生的部分气体随温度的变化规律。但实际上，如前面章节所述，直接测温法的应用有很大的局限性，可是指标气体可以通过多种途径获得。因此针对收集到的不完备的大同地区部分矿井的空气样数据，找出温度与指标性气体的变化关系是一个非常具有应用价值的工作。但表面上看来，如果我们将CH_4浓度、CO浓度、C_2H_4浓度看成变量x_1, x_2, x_3，将温度视为变量的函数，即令$V=F(x_1, x_2, x_3)$，那么我们就可以应用人工神经网络数据处理技巧，找出函数 F 的逼近。如果网络训练达到要求，就可以根据观测的指标气体的数据在煤炭自燃发生前，把握好自燃的发展趋势，对于矿井火灾的防治起到积极的作用。但是，仔细分析确不那么简单，因为CH_4浓度与CO浓度、CO浓度与C_2H_4浓度、CH_4浓度与C_2H_4浓度的相关系数分别为：0.793948、 0.56877、 0.908019。因此，一方面将三种变量装入 3－维欧几里德空间的假设不成立，也即函数关系$V=F(x_1, x_2, x_3)$不一定成立。另一方面，由上述相关系数表明，三种气体对于温度都提供了一定的信息量，如果抛弃其它指标，仅仅选取其中一项指标，比如只选取CO，那么信息丢失较为严重，预测准确率必然很低。此外，由于煤种不同，煤在氧化自燃过程中气体产物的组成和变化规律也有所不同[22]；又由于矿井作业现场条件复杂，影响因素多，特别是目前许多矿井仍采用人工取样，工作量大，不能面面俱到，且间隔时间长，不能进行连续实时检测，导致指标气体的作用往往

不能正常发挥，影响了煤炭自燃的准确的预测预报。所以必须另外寻求新的数据熔合方法，使得所有信息都充分发挥作用。

- 其它与自燃相关的因素

基于神经网络的煤炭自燃预测问题已有了许多的研究[12]~[16]，大致可以分为两类:

1) 利用已知数据，建立各参数之间的非线性关系;

2) 利用前一时期数据来预测后一阶段的状况。

从前人的研究结果可以看出：在煤矿的生产过程中，一般情况下，尽管有适宜煤炭自燃的条件存在，但只要推进度足够快，大于安全推进度，就可以避免火灾的发生。图 3.5 是某矿一个工作面在不同推进速度下，煤体温度的变化趋势。

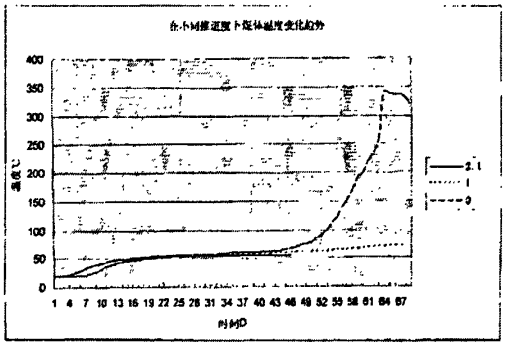

图 3.5 煤体温度变化趋势与推进度

从图上可以明显看出，当推进度为 2.1m/d 时，第 30 天之后的变化几乎为常数（实线所示）；当推进速度为 1 m/d 的时候，煤体温度升高到某一个数值以后就不再继续升温（点虚线所示），而当推进速度小于安全推进速度或者停采，煤温则不断升高，并可达到着火温度（虚线所示）。

研究中拟解决的是煤层自燃的危险性等级，以及采空区煤炭自燃趋势的预测预报问题，故确定输入、输出层的关键是探讨影响采空区煤炭自燃的因素。对于这两个问题而言，前者的输出变量只有一个，即煤层自燃的危险性等级，后者的输出变量不确定，对于不同矿井，可能需要预测的指标气体浓度个数不同，所以不能一概而论。

为了预测采空区煤炭自然发火的变化趋势，必须采用已有的历史资料来预测今后一定时间内的变化。这些历史资料包括指标气体的浓度变化、或者预埋测温探头的温度变化、或者温度最高点的坐标变化，以及在应用时间序列进行预测时与这些变化有关的参数，包括煤的吸氧速率、煤的着火温度降低值、工

作面的推进速度以及漏风情况。采用哪一种变化来表征自燃的发展过程应具体分析，考虑到矿山的实际，在不增加矿山设备的情况下，尽可能依据现有数据做好预测，在方法上下功夫。还有其它一些重要因素我们将在第五章中针对计算结果再给予解释。

第四章 煤炭自燃预测预报的数学知识

从前三章的介绍我们可以看出，建立自燃的预测模型非常复杂但十分有意义，是一个非常具有挑战性的问题。我们所面临的因素十分庞大，各地和各种矿山监测设备的功能不同使得报告的观测数据也不尽相同。前面提到的数据有三种：

- 温度与指标气体的试验数据
- 矿井自燃倾向性与矿山结构数据
- 单个矿井在封闭后的指标气体的实时观测数据（包括新老仪器提供的）

为了从这些数据中挖掘信息，我们将使用支持向量机（SVM）来分析矿山地质结构与自燃倾向的分类；利用粗糙集(RS)理论来约简指标集，最后达到评价新老监测设备的功效。在本章中我们分别对涉及到的理论知识做出简述。

第一节 ANN 简介

人工神经网络(Artificial Neural Network，ANN) 可以概括地定义为：由大量简单的高度互联的处理元素（神经元）所组成的复杂网络计算系统[23]。我们对其发展历史简要回顾如下：

- 1943 年美国心理学家 Warren McCulloch 和数学家 Walter Pitts 提出第一个神经网络模型后，神经网络科学理论研究进入了新时代。他们在分析、总结神经元基本特性的基础上，首先提出了神经元的数学模型，简称为 MP 模型。
- 1957 年美国康奈尔大学心理学家 Rosenblatt 提出了“感知器(Preceptor)”神经网络模型，第一次把神经网络的理论研究付诸于工程实践，掀起了研究人工神经网络的高潮。
- 1962 年，Bernard Widrow 和 Marcian Hoff 提出了自适应线性元件网络，简称为 Adaline。它实质上是一个两层前馈感知机型网络。它成功地应用于自适应信号处理和雷达天线控制等连续可调过程。
- 20 世纪 70 年代后，神经网络处于相对低潮时期，主要在对于人脑的高容错性、高度并行性的特征和构造相应的学习算法的困难性，以及随后冯·诺依曼体系的计算机串行处理的迅速发展和美好前景认识不足，降低了人们

对神经网络的热情。研究神经网络第二次高潮到来的标志和揭开神经网络计算机研制序幕的是美国加州工学院的 John Hopfield，他提出了模仿人脑的神经网络模型，即著名的 Hopfield 模型。

- 1984～1985 年间，Terrence Sejnowski , Hinton 和 Ackley 用统计物理学的概念和方法研究神经网络，提出了 Boltzmann（玻兹曼）机。
- 1986 年 McClelland 和 Rumelhart 提出了多层网络的误差反传算法 BP (Back Propagation)。
- 90 年代以后，神经网络的应用取得了令人瞩目的进展，特别在人工智能、自动控制、计算机科学、信息处理、模式识别等方面都有重大的应用实例。

在预测领域，1987 年 Lapedes 和 Farber 首先应用神经网络进行预测，开创了人工神经网络预测的先河。用人工神经网络预测，其基本思想是:首先收集数据去训练网络，然后用人工神经网络的算法去建立数学模型，进行预测。自此之后，神经网络方法在时间序列预测中的应用越来越受到重视。目前，已有多种不同形式的神经网络被用于工业、经济等的预测中。

网络的信息处理由神经元之间的相互作用来实现并以大规模并行分布方式进行，信息的存储体现在网络中神经元互联分布形式上，网络的学习和识别取决于神经元间联接权系数的动态演化过程。神经网络用数学语言可以这样表示:

设有 N 个神经元互联，每个神经元的状态 $x_i(i=1,2,...,N)$ 取 0 或 1，分别代表抑制或兴奋。每一个神经元的状态按下述规则受其他神经元的制约

$$x_i = f(\sum_{j=1}^{N} w_{ij}x_j - \theta_i) \quad i=1,2,3,...,N \tag{4.1.1}$$

式中：w_{ij}——神经元 i 和神经元 j 之间联接权值；

θ_i——神经元 i 的阈值；

$f(\bullet)$——取阶跃函数 $step(\bullet)$，有 $step(a)=\begin{cases}1 & a\geq 0\\ 0 & a<0\end{cases}$ (4.1.2)

式(4.1.1)也可理解为神经元 i 的输入/输出关系。x_j 为第 j 个神经元向第 i 个神经元的输入；x_i 为第 i 个神经元的输出；$f(\sum_{j=1}^{N} w_{ij}x_j - \theta_i)$ 为第 i 个神经元

的净输入。若将阈值也看作一个权值，则式(4.1.1)可改写成

$$x_i = f(\sum_{j=0}^{N} w_{ij} x_j) \quad i = 1,2,3,...,N \tag{4.1.3}$$

此时有 $w_{i0}x_0 = -\theta_i, x_0 = 1$，这就是最初的 MP 模型。更一般的 MP 模型的数学表达式为：

$$\sigma_i = \sum_{j=1}^{N} w_{ij} x_j + S_i - \theta_i \tag{4.1.4}$$

$$u_i = f(\sigma_i) \tag{4.1.5}$$

$$y_i = g(u_i) = h(\sigma_i) \tag{4.1.6}$$

式中 w_{ij}, x_j, θ_i 的含义与式(4.1.1)一样；

σ_i——第 i 个神经元的净输入；

S_i——第 i 个神经元外部输入(应用中，它可控制神经元保持在某一状态)；

u_i——第 i 个神经元的活化状态；

y_i——第 i 个神经元的输出；

$f(\bullet)$——神经元的激活规则（激活函数）；

$g(\bullet)$——神经元的输出规则（传递函数）。

在某些模型中，假设神经元没有内部状态，可以令 $f = 1$ (恒等映射)，此时 $y_i = g(u_i) = g(\sigma_i)$。不同的系统对活化值作了不同的假设，它可以是连续的，也可以是离散的。若是连续的，它可以取任意实数，也可以取某个最大值与最小值之间的数；若是离散的，则可取二值、三值或一个小的有限数集。

为了反映神经元状态连续变化的性质，用一阶非线性微分方程来模拟它随时间变化的规律

$$\tau \frac{du(t)}{dt} = -u(t) + \sum_{j=1}^{N} w_{ij} x_j(t) - \theta \tag{4.1.7}$$

式中 τ——时间常数。

$g(\bullet)$ 一般是一个非线性函数，它反映了神经元的非线性特性。常用的神经元非线性特性有：

1) 阈值型(Threshold)。无内部状态且函数$g(\bullet)$是一个阶跃函数，即

$$g(u_i)=step(u_i)=\begin{cases}1 & (u_i>0)\\ 0 & (u_i\le 0)\end{cases} \tag{4.1.8}$$

这是最早提出的二值离散神经元模型，如图 4.1(a)所示。

2) 子阈累积型(Sub-threshold Summation)。无内部状态且响应，如图 4.1(b)所示。当产生的活化值超过阈值 T 时，它将产生一个反应。在线性范围内系统的反应是线性的。子阈的作用可以抑制噪声，即对小的随机输入不发生反应。

3) 线性饱和型或分段线性型。在线性范围内是线性变化，输出值达到最大值 M 后不再增大，如图 4.1(c)所示。

4) S 型。无内部状态且连续取值。其输入/输出特性为一个有最大输出值（饱和特性）的非线性 Sigmoid 曲线，如

$$g(u_i)=1/(1+\exp(u_i)) \tag{4.1.9}$$

或

$$g(u_i)=(1+(\tan(u_i/u_0))/2 \tag{4.1.10}$$

这类曲线也反映了神经元的饱和特性，在有限范围内有抑制噪声的作用，如图 4.1(d)所示。

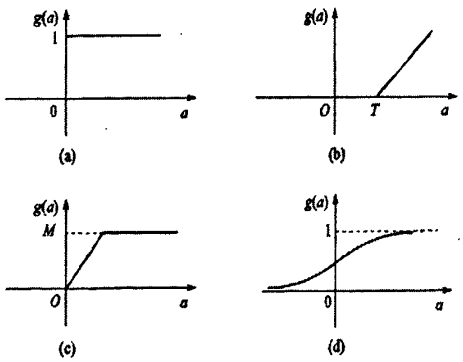

图 4.1 神经元的输入/输出特性

一般地，称一个神经网络是线性或非线性是由网络神经元中所具有的传递函数的线性或非线性来决定的。

第二节 支持向量机

支持向量机（Support Vector Machine，SVM）是美国 Vapnik 教授于 20 世纪 90 年代提出的，现在已成为了很受欢迎的机器学习方法。它与多层感知器类似，被视为高维空间中的函数的另一种表达形式。它的结构酷似三层感知器，是构造分类规则的通用方法。SVM 方法的贡献在于：它使得人们可以在非常高维的空间中构造好的分类规则，也为分类算法提供了统一的理论框架。作为副产品，SVM 从理论上解释了多层感知器的隐蔽层数目和隐节点数目的作用，也是一种将输入样本集合变换到高维空间使其分离性状况得到改善。因此，将神经网络的学习算法纳入了核技巧范畴。

所谓核技巧，它是在这样的背景下产生的。对于非线性分类，一般是先找一个非线性映射 ϕ 将输入数据映射到高维特征空间，使之分离性状况得到很大改观，此时在该特征空间中进行分类，然后再返会原空间，就得到了原输入空间的非线性分类。这看起来很合理了，但相应问题是如何找这样的映射以及如何减少在特征空间中内积 $(\phi(x),\phi(y))$ 的运算量？于是，在泛函分析的知识指导下，想到了找一个核函数 $K(x,y)$ 使其满足 $K(x,y)=(\phi(x),\phi(y))$，这就是核技巧。我们对特征空间为 Hilbert 空间的情形，对核函数给出更多一点的叙述。

设 $K(x,y)$ 是定义在输入空间 R^n 上的二元函数，设 H 为对应的特征空间且具有规范正交基 $\phi_1(x),\phi_2(x),...,\phi_n(x),...$ 使得

$$K(x,y)=\sum_{k=1}^{\infty}a_k{}^2(\phi_k(x),\phi_k(y)) \tag{4.2.1}$$

那么取 $\phi(x)=\sum_{k=1}^{\infty}a_k\phi_k(x)$ 即为所求的非线性映射，且 $K(x_i,x_j)=(\phi(x_i),\phi(x_j))$。于是，计算核函数 $K(x,y)$ 的值就是在原来输入空间中进行的，而不是在高维特征空间中。这样魔术般地避开了计算高维内积 $(\phi(x),\phi(y))$ 所需付出的计算代价。

在实际计算中，我们只要选定一个 $K(x,y)$，并不再去重构 $\phi(x)=\sum_{k=1}^{\infty}a_k\phi_k(x)$。所以寻找核函数 $K(x,y)$（对称且非负）就是主要任务了。满足以上条件的核函

数很多，例如：

- 可以取为 d-阶多项式：$K(x,y)=(1+x\cdot y)^d$，其中 y 为固定元素。
- 可以取为径向函数：$K(x,y)=\exp\left(-\|x-y\|^2/\sigma^2\right)$，其中 y 为固定元素。
- 可以取为神经网络惯用的核函数：$K(x,y)=\tanh\left(c_1(x\cdot y)+c_2\right)$，其中 y 为固定元素。

一般地，核函数的存在性只依赖于如何寻找一个平方收敛的非负序列$\{a_k\}$。这样的序列在l^2空间的正锥中的序列都满足。但哪一个最佳还是有待于进一步讨论的问题。由于分类问题对于核函数不太敏感，且构造一个核函数也不是一个简单的事，因此，实际操作中往往就在上述三类中挑出一个来使用就可以了。

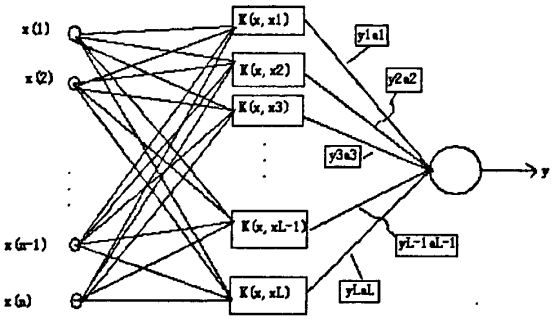

图 4.2 支持向量机结构示意图

其中输入层是为了存贮输入数据，并不作任何加工运算；中间层是通过对样本集的学习，选择$K(x,x_i)$, $i=1,2,3,...,L$；最后一层就是构造分类函数

$$y=\mathrm{sgn}(\sum_{i=1}^{L}y_i a_i K(x,x_i)+b) \tag{4.2.2}$$

整个过程等价于在特征空间中构造一个最优超平面。为了实现支持向量机的算法，我们从头来给出它的数学推理。

设样本集为$\{(x_i,y_i)\,|\,x_i\in R^n;y_i\in\{-1,+1\},i=1,...,I\}$，我们的目的是寻找一个最优超平面$H$使得标签为$+1$ 和-1 的两类点不仅分开且分得间隔最大。由于n维欧几里德空间中，超平面H的数学表达式应该是一个线性方程

$<w,x>+b=0$， 其中，w为权系数向量，x为n维变量，$<w,x>$为内积，b为常数。由空间中点到直线距离公式（推广），可得 $d(x_i,H)=\frac{|<w,x_i>+b|}{\|w\|}$。欲使得$d(x_i,H)$最大，等价于$\frac{1}{2}\|w\|^2$最小。于是，得到一个在约束条件下的极值问题：

$$\begin{cases}\min\frac{1}{2}\|w\|^2 \\ y_i(<w,x_i>+b)\geq 1,\quad i=1,2,...,I\end{cases} \tag{4.2.3}$$

引入 Lagrange 乘子$\alpha=(\alpha_1,\alpha_2,...,\alpha_I)$，可以解得关于该参变量的方程

$$Q(\alpha)=\sum_{i=1}^{I}\alpha_i-\tfrac{1}{2}\sum_{i,j=1}^{I}\alpha_i\alpha_j y_i y_j<x_i,x_j> \tag{4.2.4}$$

称之为 Lagrange 对偶函数。在约束条件

$$\sum_{i,j=1}^{I}\alpha_i y_i=0,\ \alpha_i\geq 0, i=1,2,...,I \tag{4.2.5}$$

之下，使得$Q(\alpha)$达到最大值的α的许多分量为 0，不为 0 的α_i所对应的样本x_i就称为支持向量。

当在输入空间不能实现线性分离，假设我们找到了非线性映射ϕ将样本集$\{(x_i,y_i)|x_i\in R^n;y_i\in\{-1,+1\},i=1,...,I\}$映射到高维特征空间$H$中，此时我们则考虑在$H$中对$\{(\phi(x_i),y_i)|x_i\in R^n;y_i\in\{-1,+1\},i=1,...,I\}$进行线性分类，即在$H$中构造超平面，其权系数$w$满足类似的极值问题。由于允许部分点可以例外，那么可以引入松弛项，即改写为：

$$\begin{cases}\min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{L}\xi_i \\ y_i(<w,x_i>+b)\geq 1-\xi_i,\quad \xi_i\geq 0,\ i=1,2,...,I\end{cases} \tag{4.2.6}$$

最终转化为一个二次型在约束条件下的二次规划问题：

$$\begin{cases}\min_{\alpha}\dfrac{1}{2}\alpha' D\alpha + c'\alpha \\ y'\alpha = 0,\ 0 \le \alpha = (\alpha_1,...,\alpha_l)^T \le A = (C,...,C)^T\end{cases} \tag{4.2.7}$$

其中，$y=(y_1,...,y_l)^T$，$c=(-1,...,-1)^T$，$D=\left(K(x_i,x_j)y_iy_j\right)_{1\le i,j\le l}$ 为矩阵。$K(x,s)$ 是核函数。

分类问题其实还包括了一个极端但很有用的情形。设 $\left\{x_i \mid x_i \in R^n, i=1,...,l\right\}$ 为空间 R^n 的有限个观测点，那么如何找一个以 a 为心，R 为半径的包含这些点的最小球体？这实际上是对于只有一类点的分类问题。此问题是对于求一个化合物成分的最小包络曲面的最佳方法。与前面完全相同的手法，设 ϕ 是由某个核函数 $K(x,s)$ 导出的从输入空间到特征空间中的嵌入映射，最后可以得到二次规划问题

$$\begin{cases}\min_{\alpha}\dfrac{1}{2}\alpha' D\alpha + c'\alpha \\ y'\alpha = 0,\ 0 \le \alpha = (\alpha_1,...,\alpha_l)^T \le A = (C,...,C)^T\end{cases} \tag{4.2.8}$$

其中，$y=(y_1,...,y_l)^T$，$c=(-1,...,-1)^T$，$D=\left(K(x_i,x_j)y_iy_j\right)_{1\le i,j\le l}$ 为矩阵。$K(x,s)$ 是核函数。此时

$$f(x)=K(x,x)-2\sum_{i=1}^{L}\alpha_iK(x,x_i)+\sum_{j=1}^{L}\sum_{i=1}^{L}\alpha_i\alpha_jK(x_i,x_j) \tag{4.2.9}$$

这时几乎所有的点满足 $f(x)\le R^2$。参数 C 起着控制落在球外点的数目，变化区间为：$1/L<C<1$.

当然，真正的分类问题是多类分类问题，至少是分成两类的问题，简称为二分类。如果分成多类，就简称为多分类。多分类问题可以用至少两种手法将其分解为若干个二分类 SVM：

- $1:(n-1)$，即对于待分的 n 个类，任取其一，而将其它 $n-1$ 个类捆成一个类看待。这种手法实现 n 分类需要运行 n 次二分类 SVM。
- C_n^2，即对于待分的 n 个类，每次任取其中两类，而将其它 $n-2$ 个类忽略。

这种手法实现n分类需要运行$n(n-1)/2$次二分类SVM。

- SVM决策树法。即与二叉树结合构造多类的分类SVM。
- 多目标函数法。即直接在目标函数上进行改进，变成同时兼顾的多目标优化问题。
- 有向图方法。
- 误差纠错码方法。
- 基于聚类思想（一分类）开发的多分类算法。即，将多分类输入样本在高维特征空间中求出每个类的最小包络超球即中心点位置，对于每个待测试的样本，按照到各个超球的中距离最小贴标签。这样处理方法在面对大规模样本时，会使得计算量大为简化。

由于分类问题的自然推理过程都会归结到二次规划求解，计算复杂度相对较高。为此想简化为线性规划，且不会有较大的误差。于是提出了基于线性规划的SVM分类。此方法经过数学严格推理，是合理的。因此产生了基于线性规划一分类、二分类、多分类（涉及泛函的知识较多）。此处我们仅给出基于线性规划的SVM分类的最终形式：

$$\begin{cases} \min\left(-\rho+C\sum_{i=1}^{L}\xi_i\right) \\ s.t \\ \sum_{i=1}^{L}\alpha_i K(x_i,x_j)\geq\rho-\xi_j,\ j=1,...,L;\ \sum_{i=1}^{L}\alpha_i=1;\alpha_i,\xi_i\geq 0 \end{cases} \tag{4.2.10}$$

解出α与ρ则得决策函数$f(x)=\sum_{i=1}^{L}\alpha_i K(x_i,x_j)$以及阈值。参数$C$控制着满足条件$f(x)\geq\rho$的样本数量。特别核函数取为径向函数时，参数$\sigma^2$越小，精度越高。

另外，要提醒注意的是，在求解大规模分类问题得SVM算法实现时，需要以下辅助手段：

停机准则：由于分类问题等价于求对偶问题在约束条件下的极值

$$\begin{cases} \max\sum_{i=1}^{L}\alpha_i-\sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \\ s.t.\ \sum_{j=1}^{L}\alpha_i y_i=0,\ 0\leq\alpha_i\leq C,\ i=1,2,...,L \end{cases}$$

而 KKT 条件：
$$\begin{cases}\alpha_i[y_i(<w,\phi(x_i)>+b)-1+\xi_i]=0\\(C-\alpha_i)\xi_i=0,\ i=1,2,...,L\end{cases}$$

是收敛的充分必要条件。 因此通过监控 KKT 条件来得到停机条件

$$\begin{cases}\sum_{j=1}^{L}\alpha_i y_i=0,\ 0\le\alpha_i\le C,\ i=1,2,...,L\\ y_i(\sum_{j=1}^{L}\alpha_i y_i K(x_i,x_j)+b)\begin{cases}\ge 1, & \alpha_i=0,\forall i\\ =1, & 0<\alpha_i<C,\forall i\\ \le 1, & \alpha_i=C,\forall i\end{cases}\end{cases}$$

这个条件中的不等式不必严格成立，只要在一定误差条件下成立就可以用了。

选块算法＋分解法：

1. 给定参数 $M>0$，$\varepsilon>0$，$k=0$。 选取初始工作集 $W_0\subset T$，记其对应的样本点的下标集为 J_0。令 $W_k\subset T$ 第 k 次更新的工作集，其对应的样本点的下标集为 J_k。

2. 基于工作集 $W_k\subset T$， 由优化问题

$$\begin{cases}\max\sum_{i=1}^{L}\alpha_i-\sum_{i=1}^{L}\sum_{j=1}^{L}\alpha_i\alpha_j y_i y_j K(x_i,x_j)\\ s.t.\ \sum_{j=1}^{L}\alpha_i y_i=0,\ 0\le\alpha_i\le C,\ i\in J_k\end{cases}$$

求出最优解 $\{\hat{a}_j,\ j\in J_k\}$，构造 $\alpha^k=(\alpha_1^k,...,\alpha_L^k)$ 按照如下方式：

$$\alpha_j^k=\begin{cases}\hat{\alpha}_j^k, & j\in J_k\\ 0, & j\notin J_k\end{cases}$$

3. 若 α^k 已在精度 ε 内满足停机准则，则以此权系数构造决策函数即可。否则继续下一步。

4. 在 $T\setminus W_k$ 中找出 M 个最严重破坏条件

$$y_i(\sum_{j=1}^{L}\alpha_i y_i K(x_i,x_j)+b)\begin{cases}\ge 1, & \alpha_i=0,\forall i\\ =1, & 0<\alpha_i<C,\forall i\\ \le 1, & \alpha_i=C,\forall i\end{cases}$$

加入 W_k 得出新的工作集 W_{k+1}，相应的下标集记为 J_{k+1}。

5. 重复2）—3），直到样本集耗完为止。

序列最小优化算法（SMO）：

1. 给定精度$\varepsilon>0$，令$\alpha^0=0$，$k=0$；

2. 由更新公式

$$\alpha_2^{k+1}=\alpha_2^k+\frac{y_2(E_1-E_2)}{\|\phi(x_1)-\phi(x_2)\|}$$

$$\alpha_1^{k+1}=\alpha_1^k+y_1y_2(\alpha_2^k-\alpha_2^{k+1})$$

如果第k步时达到停机要求，取近似解$\hat{\alpha}=\alpha^k$，否则继续迭代，直到满足停机为止，取为近似解。详细内容请参考现代数据处理技术[24]。

第三节 粗糙集（Rough Set，RS）

如果我们将研究对象看成是现象，那么我们可以将这些现象分类，并且看看粗糙集到底是处理些什么现象。如下所示：

现象
- 确定现象
- 不确定现象
  - 随机系统，0—1律，多种可能性满足分布规律。
  - 模糊系统，律属度$\hat{1}\in(0,1)$，不是非此即彼。
  - 粗糙集系统，研究那些因为信息不充分而导致的不确定性

相对于前两种系统的处理对象，粗糙集的确侧重面不同，它是基于不完全的信息或知识去处理不分明的现象，基于观测或者测量到的部分信息对数据进行分类。直观地讲，它是基于一系列既不知道多了还是少了，也不知道有用还是没用的不确定、不完整乃至于部分信息相互矛盾的数据或者描述，对数据进行分析甚至推测未知信息。下面我们对粗糙集的基本思想、特征、以及数学符号进行简述。

4.3.1 粗糙集理论的基本思想及特点：

粗糙集理论中的一个重要要素就是“知识”，指的是那些属性特征和决策特征。该理论的哲学基础就是：将所有研究的“对象”的集合称为论域，凭借掌握的知识对论域进行分类，而对论域中的对象的认知程度取决于所拥有的知识的多少。知识越多，则分类能力越强；反之，区分能力越差。粗糙集理论使用

了下近似与上近似的精确概念来描述每个不精确概念。下近似和上近似的差也是一个集合，称为边界集。它包含了所有那些不能确切判定是否属于给定的类的对象。影响分类的属性很多，其中某些属性起决定性作用。粗糙集理论的核心就是在所能得到的属性中找出起决定性作用的属性集，也称知识的约简，以减少结构化数据的维数，从而消除重复、冗余的属性和属性值，达到简化数据集合的目的，实现了对知识的压缩和再提炼。论域中的对象可用多种属性来描述。当两个不同的对象出相同的属性来描述时，这两个对象在系统中被归于同一类，它们的关系称之为不可分辨关系或等价关系。不可分辨关系是粗糙集理论的基石。

粗糙集理论的特点是利用不精确、不确定、部分真实的信息来得到易于处理、鲁棒性强、成本低廉的决策方案，属于软算法。因此更适合于解决某些现实系统，比如，中医诊断，统计报表的综合处理等。粗糙集理论还有一个重要特点就是它只依赖于数据本身，不需要样本之外的先验知识或者附加信息。因此可以避免主观性。粗糙集理论能够处理的数据类型包括确定性的、非确定性的、不精确的、不完整的、多变量的、数值的、非数值的。粗糙集使用上、下近似来刻画不确定性，使得边界有了清晰的数学意义并且降低了算法设计的随意性。

4.3.2 粗糙集的基本概念

粗糙集要涉及论域U（这与模糊系统相似），还要涉及属性集合$R=C\cup D$（这被认为是知识，或者知识库）。当然，也要有属性值域V，以及从$U\times R$到V的信息函数f。因此，一个信息系统S可以表示为一个四元组$S=\{U,R,V,f\}$。在不混淆的情况下，简记为$S=(U,R)$，也称为知识库。

等价关系（通常用来代替分类）是不可或缺的概念，根据等价关系可以划论域中样本为等价类。而每个等价类被称为同一个对象。但是，等价关系又是建立在不可分辨概念之上的，为了便于描述这里的等价关系，我们首先介绍不可分辨性。

设$B\subseteq R$为一个非空子集，如果$x_i,x_j\in U$，均有$f(x_i,r)=f(x_j,r)$，$\forall r\in B$成立，那么，我们称x_i和x_j关于属性子集B不可分辨。B不可分辨关系，简记为$Ind(B)$，是一种等价关系（易验证它满足等价关系的数学公理），于是$Ind(B)$可以将论域U中的元素分成若干等价类，每一个等价类称为知识库的知识颗粒。

全体等价类组成的集合记为$U/Ind(B)$，称之为基本集合。若集合 X 可以表示成某些基本集的并时，则称 X 是 B 精确集，否则称为 B 粗糙集。

粗糙集中的“粗糙” 主要体现在边界域的存在，而边界又是由下、上近似来刻画的。对于任意$X \subset U$，X关于现有知识R的下、上近似分别定义为：

$$R\_(X)=\{x \in U,[x]_R \subseteq X\} \tag{4.3.1}$$

$$R^-(X)=\{x \in U,[x]_R \cap X \neq \phi\} \tag{4.3.2}$$

X的确定域$Pos(X)=R_(X)$，是指论域U中那些在现有知识R之下能够确定地归入集合X的元素的集合。反之，$Neg(X)=U-R_(X)$被称为否定域。边界域是某种意义上论域的不确定域，即在现有知识R之下U中那些既不能肯定在X中，又不能肯定归入$\bar{X}=U \setminus X$中的元素的集合，记为$Bnd_R(X)$。

样本子集 X 的不确定性程度可用粗糙度$a_R(X)$来刻画，粗糙度的定义为：

$$a_R(X)=\frac{Card(R\_(X))}{Card(R^-(X))} \tag{4.3.3}$$

式中$Card$表示集合的基数（集合中元素的个数）。显然，$0 \le a_R(X) \le 1$，如果$a_R(X)=1$，则称集合 X 关于 R 是确定的；如果$a_R(X)<1$，则称集合 X 关于 R 是粗糙的，$a_R(X)$可认为是在等价关系 R 下逼近集合 X 的精度。

为了使得上述概念具体化，下面我们举一个例子说明如何理解和计算以上相应的概念和对应量。

例 1. 针对以下医学信息表我们来理解前面所提到的概念。

表 4.1 某医疗信息表

属性 / 对象	条件属性 C			决策属性 D
	头疼 r_1	肌肉疼 r_2	体温 r_3	流感
x_1	是	是	正常	否
x_2	是	是	高	是
x_3	是	是	很高	是
x_4	否	是	正常	否
x_5	否	否	高	否
x_6	否	是	很高	是

依据此表，如果取属性子集$R=\{头疼,肌肉疼\}=\{r_1,r_2\}$，$X=\{x_1,x_2,x_3\}$。那么我们下面给出X的上近似集、下近似集、确定域、边界域、粗糙度。

解：①计算论域U的所有R基本集：$U/Ind(R)=\{\{x_1,x_2,x_3\},\{x_4,x_6\},\{x_5\}\}$

令 $R_1=\{x_1,x_2,x_3\}$ $R_2=\{x_4,x_6\}$ $R_3=\{x_5\}$

②确定样本子集X与基本集的关系

$$X\cap R_1=\{x_1,x_2\}\neq\phi;\ X\cap R_2=\phi;\ X\cap R_3=\{x_5\}\neq\phi$$

③计算$R^-(X)$、$R_-(X)$、$Pos(X)$和$Bnd(X)$：

$$R^-(X)=R_1\cup R_3=\{x_1,x_2x_3,x_5\};\quad R_-(X)=R_3=\{x_5\}$$

$$Pos(X)=R_-(X)=\{x_5\};\quad Bnd(X)=R^-(X)-R_-(X)=\{x_1,x_2,x_3\}$$

④计算近似精确度：

$$a_R(Z)=\frac{Card(R_-(X))}{Card(R^-(X))}=1/4=0.25$$

与粗糙度类似，在给出了两个知识集（特征属性）的相对肯定域的概念$Pos_P(Q)$之后，我们也可以用一个量来刻画两个知识集的依赖度。设$K=(U,R)$为一个知识库，$P,Q\subseteq R$为两个知识集。令 $k=r_P(Q)=Card(Pos_P(Q))/Card(U)$，称为知识$Q$依赖于知识$P$的依赖度。特别，当$k=1$时称为完全依赖；$0<k<1$时，部分依赖；$k=0$时，$Q$完全独立于知识$P$。

4.3.3 知识约简

知识约简是粗糙集理论的核心内容之一，它是研究知识库中哪些知识是必要的，以及在保持分类能力不变的前提下，删除冗余的知识。在粗糙集理论应用中，约简与核是两个最重要的基本概念。

（1）一般约简

设P,Q是属性集，Q中的每一个属性都是不可省略的。如果$Q\subseteq P$且$Ind(Q)=Ind(P)$，则称Q是P的一个约简（Reduce），记为$Red(P)$。另外，若以$Core(P)$记P中所有不可省略的属性集合称为P的核（Core），那么所有约简

$\mathrm{Red}(P)$的交正好等于P的核，即$Core(P)=\cap \mathrm{Red}(P)$。该式的意义在于，不仅体现了核与所有约简的关系直接由约简得到，而且也表明了核是知识库中最重要的部分，是进行知识约简的过程中不能删除的知识。

（2）相对约简

一般地，考虑一个分类相对于另一个分类的关系，这就导出了相对约简与相对核的概念。在粗糙集中，相对约简的概念是条件属性相对决策属性的约简。我们需要给出如下的概念：

设P和Q为论域U上的两个等价关系，定义Q关于P的相对肯定域，记为$Pos_P(Q)$，为论域U中的所有那些对象构成的集合，它们可以在分类U/P的知识指导下，被正确地划入到U/Q的等价类之中。即

$$Pos_P(Q)=\cup P_-(X) \qquad (X\in U/Q) \tag{4.3.4}$$

其中，$P_-(X)$是集合X的下近似。

设P和Q为论域U上的两个等价关系，$r\in P$。 如果

$$Pos_P(Q)=Pos_{(P-\{r\})}(Q) \tag{4.3.5}$$

那么称r关于Q可省略，否则称为Q不可省略。特别，当$P-\{r\}$为P中的独立子集（即它的每个元素都再不可省略），且$Pos_P(Q)=Pos_{(P-\{r\})}(Q)$。 那么称$P-\{r\}$为$P$的关于$Q$的相对约简，记为$Ind_Q(P)$。$P$的所有关于$Q$的相对约简之交称为$P$的关于$Q$的核，记为$Core_Q(P)$。此时有$Core_Q(P)=\cap Ind_Q(P)$。

比较相对约简与一般约简的定义，我们能够发现，前者是在不改变决策属性Q的前提下对特征属性集P的约简，而后者是不改变对于论域中对象的分辨能力的前提下对于特征属性集的约简。

（3）决策表的约简

决策表约简的重要内容之一是简化决策表中的条件属性使得约简前后的决策表具有相同的功能。同样的决策可以通过基于更少量的条件，便于我们借助一些简单的手段就能获得同样要求的结果，这是事半功倍的好事。

① 分辨矩阵和分辨函数

分辩矩阵是粗糙集理论中又一个重要概念，它将决策表中关于属性区分的

信息浓缩进一个矩阵当中，可用于决策表的属性约简。

一信息系统$S=(U,R,V,f)$中，$U=\{x_1,x_2,...,x_n\}$为论域，$R=C\cup D$是属性集合，$C=\{a_i,i=1,2,...,m\}$与$D=\{d_j,j=1,2,...,l\}$分别为条件属性集和决策属性集，$a_k(x_j)$是样本x_j在属性a_k上的取值。该系统的分辨矩阵定义为一个$n\times n$阶矩阵$M(S)=[m_{ij}]_{n\times n}$，其中第$i$行$j$行处元素

$$m_{ij}=\begin{cases}a_k\in C,\quad a_k(x_i)\neq a_k(x_j)\wedge D(x_i)\neq D(x_i)\\ \phi,\quad D(x_i)=D(x_j)\end{cases}\qquad i,j=1,2,...,n \tag{4.3.6}$$

也即，分辨矩阵中元素m_{ij}是能够区别对象x_i和x_j的所有属性的集合。但若x_i和x_j属于同一个决策类时，则分辨矩阵中元素m_{ij}的取值为空集ϕ。由定义可见，$M(S)=[m_{ij}]_{n\times n}$是一个对称矩阵，主对角线上的元素是空集。因此只要考虑上半或者下半三角部分足以。

每一个分辨矩阵M(S)，可以诱导出一个分辨函数$f_{M(s)}$，如下：

$$f_{M(s)}(a_1,a_2,...,a_m)=\wedge\{\vee m_{ij},1\le j<i\le n,\ m_{ij}\neq\phi\} \tag{4.3.7}$$

它实际上是一个具有m元变量$a_1,a_2,...,a_m,(a_i\in C,i=1,2,...,m)$的布尔函数，它是$(\vee m_{ij})$的合取，而$(\vee m_{ij})$是矩阵项$m_{ij}$中的各元素的析取。

根据分辨函数与约简的对应关系，可得到计算信息系统S约简Red(S)的方法为：

1) 计算信息系统S的分辨矩阵$M(S)$;

2)计算分辨矩阵$M(S)$对应的分辨函数$f_{M(s)}$;

3)计算分辨函数$f_{M(s)}$的最小析取范式,其中每个析取分量对应一个约简.

下面举例说明如何利用分辨矩阵及分辨函数求约简及核。

例 2. 设有信息系统$S=(U,R),U=\{x_1,x_2,...,x_6\},R=\{a,b,c,d\}$,其数据表格如下表所示。

表 4.2 信息系统数据表

U/R	a	b	c	d	U/R	a	b	c	d
x_1	0	0	0	0	x_4	1	2	1	2
x_2	0	2	1	1	x_5	1	0	0	1
x_3	0	1	0	0	x_6	1	2	1	2

解：根据式(4.3.6)，分辨矩阵 M（S）为：

表 4.3 分辨矩阵

U	x_1	x_2	x_3	x_4	x_5	x_6
x_1						
x_2	b, c, d					
x_3	b	b, c, d				
x_4	a, b, c, d	a, d	a, b, c, d			
x_5	a, d	a, b, c	a, b, d	b, c, d		
x_6	a, b, c, d	a, b	a, b, c, d	—	b, c, d	

再根据式(4.3.7)，分辨函数为

$$
\begin{aligned}
f_{M(s)}(a,b,c,d) = {} & \{b \vee c \vee d) \wedge (b) \wedge (a \vee b \vee c \vee d) \wedge (a \vee d) \\
& \wedge (a \vee b \vee c \vee d) \wedge (b \vee c \vee d) \wedge (a \vee d) \\
& \wedge (a \vee b \vee c) \wedge (a \vee d) \wedge (a \vee b \vee c \vee d) \\
& \wedge (a \vee b \vee d) \wedge (a \vee b \vee c \vee d) \wedge (b \vee c \vee d) \\
& \wedge (b \vee c \vee d) \\
& = b \wedge (a \vee d) = ab \vee bd
\end{aligned}
$$

因此，该信息系统有两个约简$\{a,b\}$和$\{b,d\}$，核是$\{b\}$。由此得到两个约简的数据表格，如表 4.4 所示。

表 4.4 两个约简数据表

U	a	b
x_1	0	0
x_2	0	2
x_3	0	1
x_4	1	2
x_5	1	0
x_6	1	2

U	b	d
x_1	0	0
x_2	2	1
x_3	1	0
x_4	2	2
x_5	0	1
x_6	2	2

② 决策表

决策表是一类特殊而重要的知识表达系统，它是指当满足某些条件时，决策应该怎样进行，多数决策问题都可以用决策表形式来表达，决策表根据知识表达系统定义如下：

设$S=(U,R)$为一知识表达系统，若 R 可划分为条件属性集 C 和决策属性集 D，即$C\cup D=R, C\cap D=\phi$。则称为 CD 决策表，改记为$T=(U,R,C,D)$。$Ind(C)$的等价类称为条件类，$Ind(D)$的等价类称为决策类。

决策表可分为一致决策表和非一致决策表。当 D 完全依赖于 C（$C\Rightarrow D$）时，称为一致的；当部分依赖（$C\Rightarrow kD(0<k<1)$）时，称决策表是不一致的。

特别指出，决策表是否能够约简，取决于它是否为一致决策表。这是因为不同原因可以产生相同结果，但同一个原因则不允许导致多种结果。对于一个决策表，一般首先将其分解为一个一致决策表与一个不一致的决策表，然后再对一致决策表进行约简。约简的方法还是使用分辨矩阵的方法。此时，属性特征集C相对于决策属性集D的核$Core_D(C)$恰为分辨矩阵中所有m_{ij}为单元素集决定。也即

$$Core_D(C)=\{a_k \mid \exists m_{ij}\ s.t.\ m_{ij}=\{a_k\},\ 1\le i,j\le n\} \tag{4.3.8}$$

特别，如果$C'\subset C$是满足条件：$C'\cap m_{ij}\neq\phi$，$\forall m_{ij}\neq\phi$，且C'是最小的，那么称C'为C相对于决策属性集D的约简。

约简之后的决策表具有更少的条件属性，但却没有损失知识含量。同样对

于决策属性也可以约简。但是约简后的决策表还是不能直接看出条件与决策属性之间的关系，因此还需要挖掘出决策生成规则。为此我们引入另外一个量来刻画条件属性子集$X_i \subset C$与决策属性子集$Y_j \subset D$之间关联强度。令

$$Des(X_i)=\{(x,v_x)\mid \exists x\in U, v_x\in V \ \ st.\ f(x,a)=v_x, \forall a\in X_i\} \tag{4.3.9}$$

$$Des(Y_j)=\{(x,v_x)\mid \exists x\in U, v_x\in V \ \ st.\ f(x,d)=v_x, \forall d\in Y_j\} \tag{4.3.10}$$

分别称为属性子集$X_i \subset C$与决策属性子集$Y_j \subset D$的描述集。此时，$Des(X_i)$与$Des(Y_j)$同在空间$U\times V$中，于是可以作如下两件事：

1. 当$Des(X_i)\cap Des(Y_j)\neq\phi$时，定义$Des(X_i)$到$Des(Y_j)$的映射$T_{ij}$。

2. 当$Des(X_i)\cap Des(Y_j)\neq\phi$时，定义确定因子度

$$\mu(X_i,Y_j)=Card\left(Des(X_i)\cap Des(Y_j)\right)/Card(Des(X_i)) \tag{4.3.11}$$

特别，当$\mu(X_i,Y_j)=1$时，T_{ij}是确定的规则；当$\mu(X_i,Y_j)<1$时，T_{ij}是不确定的规则。$\mu(X_i,Y_j)$的大小反映的是满足属性子集$X_i \subset C$的对象中又能够满足决策属性子集$Y_j \subset D$的对象所占的比例。

4.3.4 基于分辨矩阵的启发式最小约简算法

以上介绍的约简的理论方法虽然简单，但只有通过计算机程序实现才有应用意义。对于决策表比较复杂，条件属性较多的情况下，由于对存储空间的要求过大，单纯使用分辨矩阵很难实现。下面建立一种基于分辨矩阵的启发式最小约简算法，可以较好地解决这个问题。

基于约简是必须能够区分所有对象的最小属性集合可以推出：一个约简与分辨矩阵的每一项的交都不空（假如与m_{ij}不相交，那么对象i与j关于该约简是不可分辨的）。于是得出如下“准约简”算法：

输入：决策表$(U,C\cup D)$，其中$C=\{a_i,\ i=1,2,...,m\}$

输出：准约简

1）选取初始约简R_0。

2）对于所有i,j检查分辨矩阵的每一项m_{ij}和候选约简集合R_0的交

$m_{ij} \cap R_0$，判断：

a. 如果$m_{ij} \cap R_0$为空集，随机从m_{ij}中选择一个属性，加到候选约简集合R_0中；

b. 若不空，检查下一项。

3）重复这一过程，直到分辨矩阵中的每一项都检查过了。

程序完成之后，我们可以得到$R = C \cup D$中的一个条件属性子集R'。但是，一般R'仅仅是一个“准约简”。因为上面的算法没有考虑它的最小性。为了使之变成真的约简，我们需要如下的启发知识。

一个简单而有效的方法是根据$|m_{ij}|$来对条件属性进行排序。我们知道，如果m_{ij}中只有一个属性，该属性一定是约简的成员。从分辨矩阵的定义可以看出，分辨矩阵中某项的长度越短，该项就对分类所起的作用越大。而且该项出现的越频繁，该项越重要。因此，我们对分辨矩阵排序时，除了按长度外，在长度相同的情况下，出现频率高的属性更重要。归结为两个重要的启发式思想：

- 属性在分辨矩阵中出现的次数越多，属性的重要性越大；
- 属性出现在分辨矩阵中的项越短，属性的重要性越大。

于是提出了一种新的基于分辨矩阵的计算属性重要性的方法。对于一个分辨矩阵M=$(m_{ij})_{n\times n}$，相应的属性a的重要性计算公式为：

$$f(a) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\lambda_{ij}}{|m_{ij}|},\quad \lambda_{ij} = \begin{cases} 0, & a \notin m_{ij} \\ 1, & a \in m_{ij} \end{cases} \tag{4.3.12}$$

式中$|m_{ij}|$为m_{ij}包含的属性的个数。由此得到了基于分辨矩阵的启发式约简算法如下：

输入：决策表$(U, C \cup D)$，其中$C = \{a_i,\ i = 1,2,...,m\}$

输出：约简(Reduction)。

1）计算分辨矩阵M，并找出所有不包含核属性的属性组合S；

2）将所有不包含核属性的属性集合表示为析取范式的形式，即：

$$P = \wedge\{\vee a_{ik}, i = 1,2,...,s, k = 1,2,...,m\}$$

3）将P转化为析取范式的形式，并按(4.3.12)式计算属性的重要性；

4）取初始约简为R'（由准约简程序提供），对于重要性最小的属性a，

令 $R(1) = R' \setminus \{a\}$，

5）判断 $Ind(R(1)) = Ind(R')$ 是否成立：

- 若成立，用 $R(1)$ 代替 R'，重复步骤 4）～5）；
- 否则，以 R' 作为约简，结束程序。

6）以 $R(i+1) = R(i) \backslash \{a\}$ 更新，重复 4）～5）直到程序中止。

第五章 现代数据处理技术在煤炭自燃分析中的应用

通过前面对煤炭自燃条件的分析，我们知道煤炭自燃必须满足的几个条件：即具有自燃倾向性的浮煤呈破碎状态大量堆积；持续供氧条件；蓄热条件；维持煤氧化过程不断发展的时间等。在矿井实际作业中，呈破碎状态的浮煤所表现出来的特性就是煤的自燃倾向性、煤层厚度及坚固性系数；持续的供氧条件就是工作面供风量及漏风量；蓄热条件和维持煤氧化过程不断发展的时间就是回采丢煤率以及工作面推进度等等。另外，煤矿开采过程中，煤层的倾角也对煤炭自燃有影响，一般来说，急倾斜煤层在同等的条件下，比较容易引起自燃。结合经验与现场实际，我们利用收集到煤的自燃倾向性、煤层厚度、倾角、工作面供风量及漏风量、回采丢煤率、工作面推进度等几个主要指标数据，针对自燃情况，我们应用支持向量机、粗糙集做具体分析。

第一节 支持向量机的应用

5.1.1 支持向量机两分类算法

输入：训练样本矩阵 $X_{n\times l}$，训练样本决策向量 $Y_{n\times 1}$，测试样本矩阵 $TX_{m\times l}$，

核函数 ker 和参数的上限 C，参数 $actfunc$（其中 n 为训练样本个数，m 为测试样本个数，l 为属性个数，核函数 ker 默认值为线性函数 $linear$，C 默认值为 ∞，$actfunc$ 默认值为0，$actfunc=0$ 为硬间隔，

$actfunc=1$ 为软间隔）

全局变量：p_1 和 p_2

输出：支持向量的个数 nsv，参数 α，偏差 b_0 和测试样本决策向量 TY

初始值：支持向量检验度 ε：若 $C=\infty,\varepsilon=10^{-5}$，否则 $\varepsilon=C\times10^{-6}$

矩阵 H：n 维零矩阵

1. 核函数为*linear*时，$H_{ij}=Y_i\times Y_j\times\left(\sum_{k=1}^{l}X_{ik}\times X_{jk}\right)$

   核函数为*poly*时，$H_{ij}=Y_i\times Y_j\times\left(\sum_{k=1}^{l}X_{ik}\times X_{jk}+1\right)^{p_1}$

   核函数为*rbf*时，$H_{ij}=Y_i\times Y_j\times\exp\left(-\frac{\sum_{k=1}^{l}\left(X_{ik}-X_{jk}\right)^2}{2\times p_1^{\,2}}\right)$

   核函数为*erbf*时，$H_{ij}=Y_i\times Y_j\times\exp\left(-\frac{\sqrt{\sum_{k=1}^{l}\left(X_{ik}-X_{jk}\right)^2}}{2\times p_1^{\,2}}\right)$

2. 给矩阵H对角各项添加小的扰动项：$H_{ii}=H_{ii}+10^{-10\times n},i=1,2,...,n$

3. 最优化方法解出参数（直接调用*qp*函数）

4. 计算$w^2=\alpha^T\times H\times\alpha$

5. 对于给定的很小的ε，判断α中大于ε的为支持向量，并确定支持向量的个数*nsv*，即：$sv1=\left\langle\alpha_i\middle|\alpha_i>\varepsilon,i=1,2,...,n\right\rangle$。

6. 计算偏差b：

   当核函数为*linear*或*sigmoid*，则支持向量为：

   $$sv2=\left\langle\alpha_i\middle|\varepsilon<\alpha_i<C-\varepsilon,i=1,2,...,n\right\rangle。$$

   如果*sv*2非空，即$nsv\neq 0$时，$b=\frac{1}{nsv}\left(\sum_{i,\alpha_i\in sv2}Y_i-\left(\sum_{j,\alpha_j\in sv1}H_{ij}\times\alpha_j\right)\times Y_i\right)$，

   否则b无法计算。

   当核函数为其他时，$b=0$。

7. 对测试样本进行分类预测：

（1）核函数为*linear*时，$H_{ij}=Y_j\times\left(\sum_{k=1}^{l}TX_{ik}\times X_{jk}\right)$

核函数为 *poly* 时，$H_{ij}=Y_j\times\left(\sum_{k=1}^{l}TX_{ik}\times X_{jk}+1\right)^{p_1}$

核函数为 *rbf* 时，$H_{ij}=Y_j\times\exp\left(-\frac{\sum_{k=1}^{l}\left(TX_{ik}-X_{jk}\right)^2}{2\times p_1^2}\right)$

核函数为 *erbf* 时，$H_{ij}=Y_j\times\exp\left(-\frac{\sqrt{\sum_{k=1}^{l}\left(TX_{ik}-X_{jk}\right)^2}}{2\times p_1^2}\right)$

$i=1,2,...,m \quad j=1,2,...,n$

（2）$YY_i=\sum_{j=1}^{n}H_{ij}\times\alpha_j+b_i$

$actfunc=0$ 为硬间隔时 $TY_i=\begin{cases}1 & YY_i>0\\ -1 & YY_i<0\end{cases} \quad i=1,2,...,m$

$actfunc=1$ 为软间隔时 $TY_i=\begin{cases}1 & YY_i>1\\ -1 & YY_i<-1\end{cases} \quad i=1,2,...,m$

5.1.2 计算求解

根据所收集数据（样本数为 21），将容易自燃和可能自燃用 1 和-1 来表示，利用支持向量机进行两分类计算。分别将第 i 个样本作为测试样本，其余 20 个作为训练样本，$i=1,2,...,21$。记 $correct=\frac{n_{correct}}{n}$ 为测试的准确度，其中 $n_{correct}$ 为测试准确的次数，n 为测试总次数，即 21 次。对参数 $p_1=0,1,...,5$ 时，比较分别调用三种核函数 *poly*，*rbf* 和 *erbf* 的预测效果。结果如下表 5.1 所示：

表 5.1 调用三种核函数 *poly*，*rbf* 和 *erbf* 的预测结果

	poly		rbf		erbf	
p1	correct	nsv	correct	nsv	correct	Nsv
0.001	0.5714	20	0.8095		0.8095	
1	0.8095		0.1905	20	0.8095	
2	0.5238		0.1905	20	0.8095	
3	0.619	0	0.2857	20	0.8095	
4	0.4762	0	0.2857	20	0.8095	
5	0.4762	0	0.2857	20	0.8095	

表 5.1 中： poly 表示核函数取为：$H_{ij}=Y_j\times\left(\sum_{k=1}^{l}TX_{ik}\times X_{jk}+1\right)^{P_1}$，

rbf 表示核函数取为：$H_{ij}=Y_j\times\exp\left(-\frac{\sum_{k=1}^{l}\left(TX_{ik}-X_{jk}\right)^2}{2\times p_1^{\,2}}\right)$

erbf 表示核函数取为：$H_{ij}=Y_j\times\exp\left(-\frac{\sqrt{\sum_{k=1}^{l}\left(TX_{ik}-X_{jk}\right)^2}}{2\times p_1^{\,2}}\right)$

时对应的支持向量机。nsv 表示作 Jackknife 检验时例外点的数量。P1 是核函数中参数在不同取值下效果对比。由第 6、7 列结果表明，用核函数 erbf 时，效果稳定且精度高。

将指标的属性值按其所在区间调整为正整数，并将决策值变为正整数（数据处理程序见程序 datalselect.m）。将所得新的属性值矩阵和决策值向量代入粗糙集算法中，求出各指标的重要性如下表 5.2 所示：（数据处理程序见程序 roughset.m）

表 5.2 各项指标的重要性

指标	自燃危险指数	煤层厚度（m）	倾角(度)	供风量 (m^3/min)	漏风量%	丢煤率%	推进度 m/d
指标重要性	13.883	16.283	11.867	17.567	15.9	17.4	15.1

上述结果表明，指标中的供风量、丢煤率和煤层厚度这三个属性的影响较大（红色），是决定煤层自燃情况的主要因素。为了进一步对矿井实际作业提供

指导，我们通过粗糙集约简程序对上述指标进行约简，约简后得到的指标集为：

{煤层厚度，供风量，漏风量，丢煤率，推进度}

这不仅符合煤矿的实际情况，也反映了预测开采煤层的自燃情况必须依据个矿的自身条件来做，避免人为因素的干扰，依据上述五项指标全面抓好煤炭开采的组织和管理，一定会为矿井火灾的防治起到积极的作用。

第二节 应用粗糙集分析各指标气体的重要性

对于一个矿井工作面，其煤层厚度、倾角、开拓方式、供风及通风方式一般都是确定的，影响采空区煤炭自燃的因素主要是煤炭本身的自燃特性。基于第二章介绍的煤的自燃氧化机理和第三章的内容，我们针对收集到的大同煤矿井下空气样指标气体数据，应用粗糙集分析各指标气体的重要性，进而选择出最重要的指标气体重点监测，为煤炭自燃的预测预报提供指导。

首先将所给数据按照其属性（即指标气体）个数的不同，分别合并后分为两类：X_5和X_7。由于气体含量是小数，不能直接应用于粗糙集中，因此我们先进行数据处理。其次，对于每个属性，将属性值按照相应区间整数化，并使所得属性值均为正数。我们依据“燃烧”与“未燃烧”两种结果决定每个样本的属性值，设当CO含量达到1%及以上时，气体发生自燃现象，定义决策值为1，而CO含量未达到1%时，气体不发生燃烧，定义决策值为2。（数据处理程序见程序data2select.m）

将X_5和X_7代入以上程序中得到处理后的数据XX_5和XX_7，并同时确定决策向量Y_5和Y_7。由于样本X_5中的数据只有一种决策值，故无法应用粗糙集求解。我们将样本X_7及其决策值向量Y_7代入到分析属性重要性的粗糙计算法中（数据处理程序见程序roughset.m），得到结果如下表5.3所示：

表5.3 各指标气体属性重要性

指标 气体	O_2	N_2	CO	CO_2	CH_4 (甲烷)	C_2H_4 （乙烯）	C_2H_6 （乙烷）
属性 重要性	46.393	43.143	53.41	50.793	51.01	50.476	44.776

上述结果表明，在多种指标气体中，CO的属性的重要性最大，这与以往的经验值一致。另外，对由于引入新设备而采样分析得到的乙烯和乙烷两种气体，

它们的属性的重要性均没有CO大，但乙烯的属性的重要性相对偏高，接近CO的属性的重要性，因此可作为考虑因素。而乙烷的影响很小，可不作考虑。精确的取舍是根据粗糙集约简程序得出的，约简后得到了四项基本指标，如下表5.4。

表5.4 约简指标集

指标气体	CO	CO_2	CH_4	C_2H_4

- **新老设备的作用评估**

针对上述表中的结果分析，结合数据样本我们发现， 前面三项是新老设备都能够提供的，而第四项(C_2H_4)只有新设备能够提供。为此，我们希望得到新老设备的价值的评判。令人意外，但也在期待之中的是，对于收集到的数据样本，采用支持向量机，选取三种不同的核函数poly、rbf 及erbf，以及三种不同的参数 $p_1=1$、 $p_1=2$ 和 $p_1=3$，我们得到结果如下表5.5：

表5.5 选用三种核函数 *poly*， *rbf* 和 *erbf* 的结果比较

核函数	参数	CO 、CO_2 、CH_4		CO 、CO_2 、CH_4 、C_2H_4	
		correct	nsv	correct	nsv
poly	$p_1=1$	0.9589		0.9589	
	$p_1=2$	0.9589		0.9589	
	$p_1=3$	0.9589		0.9589	
rbf	$p_1=1$	0.9454		0.9454	
	$p_1=2$	0.9452		0.9452	
	$p_1=3$	0.9315		0.9315	
erbf	$p_1=1$	0.9452		0.9452	
	$p_1=2$	0.9315		0.9315	
	$p_1=3$	0.9315		0.9315	

此表说明：基于三种指标气体CO、CO_2、CH_4与基于四种指标气体CO 、

CO_2、CH_4、C_2H_4得到的精度没有区别。因此，如果单纯考虑数据有代表性的话，新老设备的作用基本没有差别，某种意义上讲，这也提示我们不一定要更换老设备。但在实际中，目前大部分矿井特别是一些老矿仍采用人工检测手段，即采用人工取样，通过色谱仪分析，给出指标气体浓度，以此判断自燃程度。该方法虽投入设备少，但人工取样工作量大，难以避免人为因素，又不能面面俱到且间隔时间长。而少数新矿井安装了束管监测系统，通过铺设聚乙烯管将气体抽到地面，经气样分选器送往色谱仪分析，这样减少了工作量且能更好地避免人为因素，但有时线路长，管理较困难，为改善这些现象，采用计算机技术与气体分析技术相结合效果较好。所以，各矿井可综合考虑经济等各方面因素来决定设备的更新。

基于前面的分析结果，若不考虑其它因素，在新老设备下如果能够得到大量的气体数据样本，再应用支持向量机等现代数据处理技术，将会达到令人满意的预测精度（约95%左右），这将能够很好的发挥指标气体的作用，对于煤炭自燃的预测和火灾的防治有积极的意义。由此，我们也看到支持向量机等现代数据处理技术在实际应用中所发挥的巨大作用。

第六章 结论与展望

第一节 结论

煤炭自燃是影响煤矿安全生产的主要灾害之一，探索煤炭自燃的早期预测预报技术，对于矿井火灾的防治具有积极的意义。本文从煤炭自燃的时序特点出发，应用现代数据处理技术，探讨矿井煤炭自燃的预测问题。通过研究，我们可以得出如下结论：

1. 通过对影响煤炭自燃的过程及影响因素的分析，进一步明确了煤炭自燃的非线性特征。我们知道影响煤炭自燃的因素很多，只有从非线性科学角度出发，才能更深刻的认识煤炭自燃的规律，从而做好自燃的预测预报工作。
2. 根据分析知道了煤层自燃情况的影响因素较多，但通过应用现代数据分析等数学知识对有关数据进行计算分析后发现：指标中的供风量、丢煤率和煤层厚度这三个属性的影响较大，是判定煤层自燃情况的主要因素。在此基础上，可以判断某煤层是否需要加强重点监测，对于矿井的防灭火工作提供了指导性意见。
3. 指标气体分析法是应用于煤层自然发火的预测预报的方法之一，但煤种不同，煤在氧化自燃过程中气体产物的组成和变化规律也有所不同。本文通过对大同煤矿井下空气样数据应用现代数据处理技术进行计算分析，得出CO是指标气体中最重要的一项，其次是CH_4、CO_2和C_2H_4，CO可作为煤自然发火的指标气体的首选，但如果综合考虑其次气体，再应用支持向量机方法，可使得精度达到95%左右，这就会使指标气体的作用得到充分发挥。
4. 粗糙集、支持向量机等现代数据处理技术是实用性很强的学科，其中核技巧提供了一种将非线性问题转化为线性问题的有效方法。从对样本数据的计算分析结果可以看出，这些方法的应用为煤炭自燃的预测提供了科学依据和新的思路及方法。也避免了人为因素干扰，对于减少煤炭自燃时的取样工作量有很大的好处。

第二节 展望

由于笔者的精力和知识水平有限，本文还存在一些不足之处，需要在今后的研究和实践中进一步完善。具体讲有下面三点：

1. 煤炭自燃的本质有待于进一步深入研究。
2. 本次研究中由于相关数据收集的不易，所以学习样本显得不足。如果能够把各种不同矿井的自燃样本收集起来形成数据库，对于矿井煤炭自燃的预测预报将起到很大的作用。
3. 对于现代数据处理技术研究不精，用于计算分析的数据项目较简单，如果能将影响指标气体浓度发展变化的所有因素综合进行考虑，将涉及更深的数据熔合理论，也会使数学模型的实用性大大增强。这是今后努力的方向。

致谢

本文是在导师阮吉寿教授的悉心指导下完成的。从最初的选题、现场调研、信息的收集、文稿的修改、直至论文的定稿，阮老师都时刻关注并给出了很有建设性的启发和建议，同时也付出了大量的心血。在指导我完成论文的同时，阮老师给我提出了工作和个人发展方面很多有益的指点。在三年的读研过程中，阮老师渊博的学识，高尚的人品，严谨的治学态度，平易近人的作风都给我留下了深刻的印象，使我受益匪浅，并将在今后的人生中给我带来积极的影响。借此谨向尊敬的阮老师表示我最衷心的感谢。

同时，我要感谢学习期间各课程的老师，老师们精彩的讲解、严谨的学风、渊博的知识使我终生受益。在学习期间，还得到了数学学院领导、院办及有关老师的关心和帮助。特别是为本文提供数据的大同煤矿技术科与通风区，以及郑晨光同学在程序方面给予的帮助，在此向他们表示深深的谢意！

感谢所有支持、关心我的家人和朋友，是他们的支持和鼓励使我得以潜心研究、完成论文的写作。

感谢各位老师专家在百忙之中评阅本人的论文，并提出宝贵的建议！今后我会更加努力以回报大家的关爱。

参考文献

[1] 汪巍. 煤炭自燃模拟试验监测系统的研究. 安全技术及工程，2005.
[2] 邓军，徐精彩，陈晓坤.煤自燃机理及预测理论研究进展[J]. 辽宁工程技术大学学报，2003，22(4)：455-459
[3] 舒新前. 煤炭自燃的热分析研究[J]. 中国煤田地质，1994，25(2)：25-29
[4] 彭本信. 应用热分析技术研究煤的氧化自燃过程[J]. 煤炭工程师，1992,(2)：1-10
[5] Tevrucht M L E, Griffiths P R. Energy & Fuel [J]. Fuel, 1989, (3):522
[6] Shinn J H. Study of coal molecular structure [J]. Fuel, 1984, (63):83
[7] 邓军等，煤自然发火预测理论及技术[M]. 西安：陕西科学技术出版社，2001，30(4):366-370
[8] 邓军等，“国内外煤炭自然发火预测预报技术综述”，《西安矿业学院学报》，1999-4
[9] 许波云，范明训. 运用模糊聚类分析法综合预测煤层自燃危险性[J]. 煤炭学报，1990-4
[10] 邓军，徐精彩，文虎等. 综放采煤法中沿空巷道煤层自然发火预测模型研究[J]. 煤炭学报，2001,26(1):62-66
[11] 谢振华，金龙哲，刘向来. 煤炭自燃预测预报技术及其应用[J]. 湘潭师范学院学报，2004，26(1)：59-62
[12] 何启林，王德明. 采空区遗煤自然发火预测的模糊分形神经网络研究[J]. 湖南科技大学学报，2004，19(2)：18-20
[13] 肖红飞，刘黎明，王海桥. 改进 BP 算法在开采煤层自燃危险性预测中的应用[J]. 煤炭学报，2001，26(6)：650-653
[14] 杨忠超，徐精彩. 神经网络在采场浮煤自燃危险性预测中的应用[J]. 西安科技学院学报，2001，21(3)：197-204
[15] 王德明，王俊. 基于无导师神经网络的煤炭自燃危险性聚类分析[J]. 煤炭学报，1999，24(2)：147-150
[16] 王洪德，狄英全,肖佳新. 用 BP 算法实现待开采煤层自然发火危险性预测[J]. 辽宁工程技术大学学报，2002，21(3)：281-283
[17] 王省身，张国枢. 矿井火灾防治[M]. 中国矿业大学出版社，1990
[18] 张国枢等. 通风安全学[M]. 中国矿业大学出版社，2000
[19] 邓军，徐精彩，陈晓坤.煤自燃机理及预测理论研究进展[J]. 辽宁工程技术大学学报，2003，22(4)：455-459
[20] 李增华. 煤炭自燃的自由基反应机理[J]. 中国矿业大学学报，1996，3(25)
[21] 谢之康等. 矿井灾害现象的非线形动力学[J]. 煤炭学报，1999
[22] 许延辉等. 煤自燃火灾指标气体预测预报的几个关键问题探讨[J]. 矿业安全与环保,2005，32(1):16-24
[23] 高隽. 人工神经网络原理及仿真实例[M]. 北京:机械工业出版社,2003.
[24] 吴今培，孙德山. 现代数据分析[M]. 机械工业出版社,2006.

个人简历

石有印，男，1971年11月23日出生。

1989年8月～1993年7月，就读于兰州大学数学系，获得理学学士学位。

1993年8月～至今，山西大同大学工学院，工作。

2005年至今，南开大学数学科学学院，学习。