摘 要

癌症治疗面临的重大挑战是如何针对病原上各自独特的癌症类型制定具体的治疗方法，以达到最大疗效的同时降低药物的副作用。因此，癌症检测或癌症分类成为癌症治疗的中心环节。一直以来，癌症检测主要基于肿瘤的形态表观，但这种检测方式有很大的局限性，因为具有相似组织病理学表观的肿瘤可能表现出很不相同的临床发展过程，或者对同种治疗呈现出不同反应。近年来，DNA微阵列技术的发展产生了海量的基因表达谱数据，为寻找基因之间表达调控的复杂关系网络，研究功能基因组和癌症检测提供依据。目前，利用基因表达谱进行癌症检测成为癌症研究的重点之一。但是基因表达谱数据具有高维性，高噪声，高冗余，数据分布不均衡等特点，对基因数据分析方法提出了更高要求，对基于DNA微阵列基因表达谱的癌症检测带来了挑战。

本论文从基因表达谱数据的分析着手，以挖掘基因表达模式和癌症检测研究为主要目标，研究癌症检测中基因表达数据的预处理、特征基因的选取、癌症组基因表达模式的分析以及建立合适的基因诊断模型的问题。本文的主要工作归纳如下：

第一，针对基因表达数据的特点，提出一种基于CMST聚类方法的分步的特征基因选择方法，然后，在分步的特征基因选择方法中引入“Gap Statistic”理论，以确定特征基因数目，提出一种自适应的特征基因的选择方法，弥补目前的特征基因选择算法中缺乏较好的基因数目预置机制的不足。

第二，利用主分量分析方法（PCA）和独立分量分析方法（ICA）挖掘基因表达谱中隐含的基因表达模式，揭示癌症中基因的调控机制，通过抽样来选取特征基因子集以减少噪声对PCAP和ICAP的影响，并且根据基因子集中隐含模式的相似性来重构基因表达，提出一种基于隐含变量模型的癌症检测算法。

第三，利用癌症组基因表达存在的局部特征相关性的生物病理特点，提出DNA微阵列基因表达谱中癌症组关联空间的概念，抽取不同癌症组基于关联空间的基因特征模式，研究与癌症组相关联的基因表达模式在癌症组中的表达以及调控，并提出适合癌症组相关联的基因表达模式的癌症预测算法，有效缓解基因数据集中“维数灾难”的问题。

第四，由于不同的特征选择方法采用不同的搜索机制和评价策略，挑选出的特征基因偏向癌症特征的不同方面，因此不同方法选择的特征基因明显不同，导致分类器的识别结果不稳定。针对癌症组基因数据和基因组数据构建一组具有互补性分类器，提出一种组合分类算法提高癌症分类算法的泛化性能。

第五，从基因之间的协同表达来分析基因数据，研究具有可解释的基因表达模式。在显现模式的提取中增加虚拟样本以挖掘具有更高辨识能力的显现模式，并在候选分割点选择策略中通过高斯分布来模拟分割点的分布，提高分割点选择的可靠性，然后提出两种基于显现模式的癌症检测算法。

关键词： DNA微阵列；基因表达谱；癌症检测；特征基因；基因调控；基因表达模式

Abstract

The great challenge in cancer treatment is how to direct specific treatment to particular tumour in order to achieve the better therapy effect while the lower toxicity. So the cancer detection or cancer classification becomes one key point for cancer therapy. For a long time, the classification lies on the sample morphology, which is not efficient in many cases. Because tumours in different stages may present similar pathomorphism and tumours with similar pathomorphism may react differently to various therapies. Now cancer detection using gene expression data is an important aspect in cancer research. Recently, with the development of Microarray technology massive of gene expression data is produced, which is help for exploring complicated genetic regulating network, investigating functional genome and studying on cancer detection. However, there are characters in gene expression data, such as high dimensionality, huge noise, huge redundancy and nonequilibrium distribution, which imposes challenges for development of the associated data mining techniques and cancer detection.

In this dissertation, we emphasize on analysis of gene expression data. Our major goals are for gene expression mode mining and cancer detection. We explore the gene expression data pre-processing, the feature gene selection, analysis of gene expression model to cancer and building the cancer detection model. The main contributions of this dissertation are summarized as below:

Firstly, the characters of gene expression profile are analyzed and a CMST clustering based multi-step gene selection scheme is proposed, then "Gap Statistic" is introduced into this feature gene selection to determine the number of feature genes, so we develop a self-adaptive gene selection method, which makes a great improvement compared to the mechanism of setting the number of feature genes arbitrarily.

Secondly, PCA and ICA is applied to analyze the gene expression data and investigate the underlying regulating factor and gene regulating networking in cancer. Sampling is used to produce the gene subsets, and in the PCAP and ICAP of subsets the noninformative features are reduced, then the gene expression modes are reconstructed and a hidden gene expression model based cancer detection is presented.

Thirdly, the biological locality of gene expression to the cancer is explored, and a concept of relative space to a cancer is proposed, then the cancerogenic gene mode based on relative space is extracted, and the regulation with cancerogenic gene mode is discussed.

Then a cancer detection algorithm with relative gene expression mode is proposed, in which the problem of "curse of dimensionality" is relived.

Fourthly, when different feature selections are used, as the researching mechanism and evaluation strategy are different the distinct feature genes, which tend to different aspects of cancer, are selected. The classification results using these classifiers with distinct genes varied a lot. So a group of complemental gene classifiers are constructed, and a ensemble cancer classification algorithm is proposed.

Fifthly, the gene co-expression and explainable emerging pattern are explored. The virtual samples is added to to improve distinguishment of emerging pattern, and in the strategy of choosing cut point the distribution of cut point is assumed to be the Gaussian distribution for improving the reliability of emerging pattern and two emerging pattern based cancer detections are presented.

Keywords: DNA Microarray; Gene Expression Profile; Cancer Detection; Feature Gene; Gene Regulation; Gene Expression Mode

插图索引

附表索引

湖南大学

学位论文原创性声明

本人郑重声明：此处所呈交的论文《基于DNA微阵列基因表达谱数据的癌症检测研究》，是本人在导师的指导下独立进行研究所取得的成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律后果由本人承担。

作者签名：[signature]　　日期：2007年 11月 28日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权湖南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、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于

1、保密□，在____年解密后试用本授权书。

2、不保密☑。

(请在以上相应方框内打“✓”)

作者签名：[signature]　　日期：2007年 11月 28日

导师签名：[signature]　　日期：2007年 11月 29日

第1章 绪论

在多年的癌症（疾病）研究中，科学家和医学工作者们认识到，癌症并不只是某一种疾病，在它的背后，隐藏着形形色色，变化多端的种类，存在着几百种这样的癌症。它们为什么一直难以攻克呢？其主要的原因是由于每一种癌症都有自己的特点，一种药物并不能对各个不同组织的癌症都能产生疗效，有些能抑制住肿瘤细胞，但有些却毫无作用，甚至在病症上相同的癌症，也无法用一种药物达到治疗的目的。随着人类生命科学的发展，人们对于基因这一有关人类生长、发育、衰老、遗传的最重要和最本质的因素，有了新的认识，并逐渐开始将基因引入对疾病的诊断、治疗、药物研制、药物筛选等方面。因此，基因诊断、基因治疗、药物基因组图等应运而生。通过基因进行疾病诊治是对传统诊治方法提出的巨大挑战，成为人们关注的焦点。

20世纪90年代初开始实施的人类基因组计划（Human Genome Project，HGP）与20世纪40年代制定的曼哈顿原子弹计划（Manhattan Project）以及60年代制定的阿波罗登月计划（Apollo Project）并称为美国的三大国家计划。人类基因组计划是由美国科学家于1985年率先提出[1,2]，旨在阐明人类基因组30亿个碱基对（Base Pairs）的序列，发现所有人类基因，并搞清其在染色体（Chromosome）上的位置，破译人类全部遗传信息，让人类第一次在分子水平上全面地认识自我，该计划1990年正式启动。英、日、德、法随后相继加入该计划，值得关注的是1999年中科院基因组中心代表中国正式加入该计划，承担了1%人类基因组的测序任务。2001年2月，人类基因组草图宣布完成[3,4]。随着以测序为主的结构基因组计划（Structural Genomics Project）的完成，生命科学研究的重点也逐渐的转变为以对基因功能研究为主的功能基因组计划（Functional Genomics Project）。功能基因组计划的主要任务之一是寻找调控疾病的相关基因，研究与疾病相关基因功能，进行基因功能鉴定，研究通过基因表达实现疾病诊断和基因治疗。在“九五”“十五”期间，功能基因组计划研究已被列为国家高科技计划863和973重大专项。

一项类似于计算机芯片技术的新兴生物高技术－DNA微阵列（Microarray）技术，或称为生物芯片（Biochip）、DNA芯片（DNA Chip）、基因芯片（Gene Chip）[5,6]，随着人类基因组研究的进展应运而生。自从1991年Affymetrix公司的Fodor博士等人[7]提出基因芯片的概念后，已有多种不同功用的基因芯片问世，并在生命科学研究中开始发挥重要作用。近年来DNA微阵列技术[8-11]得到了迅猛发展，产生了大量基因序列和基因表达水平数据。如何利用DNA微阵列技术研究

基因的功能、基因的调控，以及在疾病中的基因变异和基因表达。因此，研究者提出了后基因组计划、蛋白组计划、疾病基因组计划以破译人类基因这部天书。

生物体发育、分化、生长相代谢的过程，始终是遗传信息从储有到表达、加工及传递的过程；实质上，主要是基因中mRNA（cDNA）信息的传递过程。而生物体的遗传、变异和进化问题则主要体现在遗传信息的复制、重组、变异和选择。DNA微阵列利用成千上万密集排列的基因探针，通过已知碱基顺序的DNA片段，并结合碱基互补的原则检测细胞基因mRNA（cDNA）表达水平。不同个体基因变异、不同组织、不同时间、不同生命状态等基因表达的分析是基因组计划、蛋白组计划和疾病基因组计划中最重要的一个环节。DNA微阵列基因表达数据在疾病诊断、基因治疗、药物筛选、给药个性化、新基因发现、DNA计算机研究等领域发挥重要的作用。

DNA微阵列具有高速度、高通量、集约化的特点，所以我们可以通过微阵列一次性对大量序列进行检测和基因分析，获取高维的基因表达数据[12-15]。通过基因表达数据研究人员能够在基因组层次上研究任何种类细胞在任何时间、任何给定条件下的基因表达模式，可以帮助我们深入研究和了解生物过程的本质。通过分析基因表达数据，我们可以了解疾病在基因级别的发病机理、疾病的诊断、基因级别的药物研制以及基因治疗。当前的肿瘤检测和分类技术高度依赖于病理学工作者对癌症组织的主观判断，而基于微阵列技术，即使一些组织没有显著变化，利用基因表达数据也可以对之做出早期诊断[16]。如何利用基因表达数据揭示基因在影响和调控癌症组织产生的变异？如何利用基因表达数据有效地识别癌症组织，并为人类最终战胜各种病魔提供有效武器？然而，基于微阵列数据的分析方法和基于微阵列数据的癌症检测的发展才刚刚起步，解决上述问题具有巨大的挑战[17, 18]。

1.1 DNA微阵列技术简介

DNA微阵列技术是融微电子学、生物学、物理学、化学、计算机科学于一体的高度交叉的新兴技术。DNA微阵列技术已被公认将会给21世纪的生命科学和医学研究带来一场革命，并因此成为学术界和工艺界研究的一个热点。美国总统克林顿在1998年1月的国情咨文演讲中指出：“在未来的12年内，基因芯片将为我们一生中的疾病预防指点迷津”。另外，美国商界权威刊物Fortune对其重大意义作了如下阐述：“微处理器在本世纪使我们的经济结构发生了根本改变，给人类带来了巨大的财富，改变了我们的生活方式。然而，生物芯片给人类带来的影响可能会更大，它可能从根本上改变我们的医学行为和生活质量，从而改变世界的面貌”[19]。由于生物芯片技术领域的飞速发展，美国科学促进协会于1998年底将

生物芯片评为1998年的十大科技突破之一[20]。

基因芯片就是利用点样技术、现代探针固相原位合成技术、照相平板印刷技术等微电子技术在有限的空间内，将成千上万种基因的DNA片段有组织的点在固相片基上作为可寻址识别的基因探针。在微阵列实验中，所有的RNA被反转录成带有放射性同位素或荧光标记的cDNA。然后，cDNA与由基因片段组成的、固相片基上的大型DNA文库杂交。最后，采用荧光或其他成像技术测定上千个基因在各种不同实验条件下的表达，以检测不同组织或不同细胞的基因表达情况，为疾病诊断和基因治疗提供大量的遗传变化信息[6]，如图1.1所示。按照芯片的制作原理，基因芯片可以分为很多类，但目前真正成熟的，且广泛应用的有使用原位合成和合成点样技术的微阵列（Microarray）。

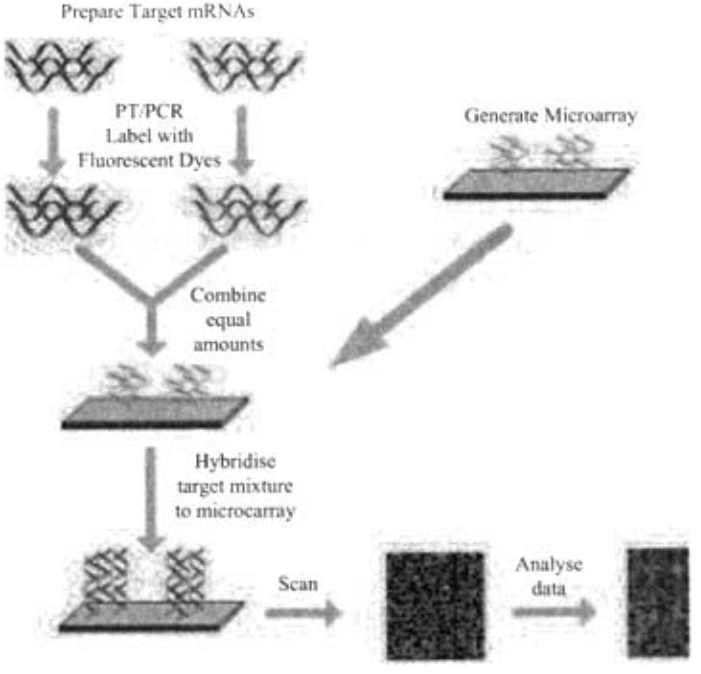

图1.1　DNA微阵列技术及应用

1.1.1 DNA微阵列的制备技术

DNA微阵列的实质是高度集成的寡核苷酸阵列，制造基因芯片首先要解决的技术是如何在芯片片基上定位合成高密度的核酸探针。目前，基因芯片的制备技术主要有以下几种：

1.1.1.1 原位合成法

原位合成（In Situ Synthesis）是指直接在芯片上用四种核苷酸合成所需探针的基因芯片制备技术。原位合成方法可以制作高密度基因芯片，但是，需要的技术设备复杂，成本较高并且合成的效率较低。主要包括：

1. 原位光刻

美国Affymetrix公司结合了半导体工业的光刻技术和DNA合成技术制造发展的一项高密度核酸阵列的基因芯片制备技术。它利用光保护基团修饰芯片片基表面碱基单体的活性羟基，通过设计特定的光刻掩膜和不断地更换曝光区域，直接在片基上合成所需高密度寡核苷酸阵列，探针数目在合成循环中呈指数增长。

2. 原位喷印合成

原位喷印合成原理与喷墨打印类似，不过芯片喷印头和墨盒有多个，墨盒中装的是四种碱基等液体而不是碳粉；采用的化学原理与传统的DNA固相合成一致，因此不需要特殊制备的化学试剂。

3. 分子印章多次压印

根据所需微阵列，设计有凹凸的微印章，然后根据预先设计在制备的各级印章上涂上对应的单核苷酸；按照设计的顺序将不同的微印章逐个依次压印在同一基片上，得到256×256阵列的高密度基因芯片。

1.1.1.2 合成点样法

合成点样法（Off-chip DNA Synthesis）是指将合成好的探针、cDNA或基因组DNA通过特定的高速点样机器人直接点在芯片片基上。制作基因芯片的密度低，需要的设备简单，成本较低，适用于多数实验室制作基因芯片。目前，除Affymetrix等研究和生产基因芯片的少数大公司使用原位合成法外，其他中小型公司和实验室研究中普遍采用合成点样法。

1. 微型机械点样法

该技术是由Shalon 和Brown于1995年发展起来的一类芯片制备技术，而后由美国Synteni公司开发出商品仪器。该方法通过毛细作用使用点样针将生化物质转移到固体基底表面（点样针与基底表面接触），每一轮结束后，清洗点样针进行下一轮操作，而且机器人控制系统可使其实现自动化生产。

2. 化学喷射法

将合成好的寡核苷酸探针定点喷射到芯片片基上来制作DNA芯片。该技术由Incyte Pharmaceuticals和Protogene公司等发展。该方法通过应用与压电接口相连的微型喷嘴将生化物质喷向基底，通过电流控制使样品体积得到精确控制。

1.1.2 DNA微阵列技术的主要特点

DNA微阵列技术将成千上万的核酸探针固定于芯片片基上与标记的样品分子进行杂交，通过检测每个探针分子的杂交信号强度获取样品分子中的基因表达水平。相对于传统的基因检测技术，DNA微阵列技术的具有以下特点：技术操作简单、自动化程度高、检测基因数量大、检测效率高、应用范围广、成本相对低。

1.2 DNA微阵列技术的应用

DNA微阵列技术将生命科学研究中许多不连续的分析过程，如样本制备、生化反应和定性、定量检测等，集中到指甲盖大小的芯片上，使基因分析过程全自动化，被称为“芯片实验室”（Lab-on-a Chip）。该技术成千上万倍提高基因分析效率的同时，大大减少了样品和试剂，并且实验结果更具全面性、直观性和可重复性。因此，DNA微阵列技术广泛地应用于分子生物学及医学研究的各个方面。

1.2.1 基因组测序

DNA微阵列的思想是在基因测序的早期提出的，由于传统的基因测序方法难以解决人类基因组计划如此繁重的工作，因此DNA微阵列早期主要用来研究基因组结构。DNA微阵列技术可在一次实验中利用探针与待测样本分子进行大量杂交反应，并分析杂交反应产生的杂交图谱而排列出待测样品的序列。Hacia在Nature Genetics上对用寡核苷酸微阵列进行基因重复测序和基因突变分析进行了较为详细的叙述[21]。

1.2.2 基因表达分析

基因表达（Gene Expression）是指储存遗传信息的基因经过一系列步骤表现出其生物功能的整个过程。典型的基因表达是基因经过转录、翻译，产生有生物活性的蛋白质的过程。DNA微阵列已被用来测定菌类、植物、动物和人类样品中的基因表达水平。

斯坦福大学的Schena于1995年首先使用DNA微阵列研究拟南芥（Arabidopsis Thaliana）基因表达，通过芯片杂交分析拟南芥根与叶两种组织中基因的差异表达[22]。

DeRisi等[23]应用酿酒酵母（Saccharomvces Cerevisiae）cDNA基因芯片研究孢子在有丝分裂状态下基因转录和表达水平的差异。斯坦福大学的Brown研究小组应用合成点样法制备酿酒酵母cDNA微阵列，获得酵母在不同细胞周期状态以及

在热休克冷休克处理后其2473个基因的表达谱，较直观地反应了不同条件和状态下基因转录调控水平，为寻找基因调控的机理提供了一条有效的途径[24]。

Tanaka等[25]利用基因芯片技术检测了1500只小鼠妊娠中期子宫及胚胎发育过程中基因的表达情况，从而了解到哺乳类动物胚胎发育过程中基因表达的动态变化。

Golub等[6, 26]分析了人类白血病的6817个基因表达谱，利用基因表达水平的差异将72个白血病样本分成AML和ALL两组，并取得了较好的准确性。Bull等[27]用包含前列腺癌、损害前身和正常组织的cDNA微阵列研究前列腺癌中基因表达。标记从前列腺电切术（TURP）或前列腺根治术得到的肿瘤样品的cDNA，分析其表达，揭示了许多上调转录和基因过表达

1.2.3 发现新基因

微阵列技术是一项发现新基因及分析各个基因在不同时空表达方面十分有用的技术，它具有样品用量极少，自动化程度高等优点，便于大量筛选新基因。Heler等[28]利用cDNA芯片比较了炎证性疾病类风湿关节炎和肠炎组织中基因表达的不同，并导致进一步发现了炎症相关基因IL-3，Gro-A等。Buates等[29]鉴定了由激活因子（S-28463）诱导表达的一系列基因。Schena等[30]用包含1056个cDNA的芯片与热休克作用和佛波酯处理的T细胞的cDNA杂交，发现了4个新基因。目前，人类基因数据库中有400000 个基因表达序列标签（Expressed Sequence Tag，EST）。成千上万的ESTs微阵列为人类基因表达研究提供强有力的分析工具，加速人类基因组的功能分析。

1.2.4 在疾病诊断中的应用

疾病的发生和发展实际是多种疾病相关基因表达失常或许多疾病抑制基因失活所致。利用基因微阵列技术，可以找到与该疾病相关的基因，实现对该疾病快速、简便、高效的诊断。人类恶性肿瘤的60%与人类P^{53}抑癌基因的突变有关，对癌症样本的基因突变和异常表达进行检测可以作为诊断的重要指标。

Kauraniemi等利用cDNA芯片技术和Northern杂交技术检测了BRCA1 乳腺癌患者MYB基因mRNA的表达情况，发现在BRCA1突变型中MYB基因表达较常见[31]。Okustsu等从76例急性细胞白血病（AML）患者获得23040个基因构成的肿瘤细胞基因表达谱，结果AML患者有63个基因过表达，372个基因的表达受到抑制，这些基因可能调控与AML发病分子机制有关的关键因子，也成为AML药物治疗的潜在靶基因。同时通过比较AML患者对化疗敏感者与对化疗不敏感患者的基因表达情况发现有28个基因的表达水平不同，基于此基础建立一个个体抗肿瘤药物的敏感系统，预示化疗最终走向个性化治疗的目标[32]。Kan等

把人食管癌细胞系和人食管组织点在cDNA微阵列上，把KYAZ和OE33（腺癌）从KYSE系（鳞状细胞癌）中区分出来，识别了在KYAZ和OE33中特征性表达的基因[33]。Lossos等在一项独立研究中发现，根据Ig基因超突变的有无，可将弥漫性大B细胞淋巴瘤（Difuse Large B-Cell Lymphoma，DLBCL）分为两个亚型[34]。Brown等运用DNA微阵列技术对脆X综合症（Fragile-X Syndrome，Fra X）的分子生物学机制及早期诊断进行了研究[35]。

DNA微阵列技术为临床疾病的诊断提供了一种全新的概念，它不仅使实验检测的高通量、高自动化、微量化得以实现，并且在临床上对使某些疑难疾病的准确诊断成为可能。

1.2.5 在药物研究中的应用

基因芯片技术在新药开发、药物靶标的发现，多靶位药物筛选、药物作用的分子机理研究、药物疗效及副作用等方面具有明显优势，还可以将药物的生物效应和基因变化密切相联系，从而为药物的研究和开发注入了新的生机和活力[36]。Kumar-Sinha等利用DNA芯片筛选发现脂酸合酶（FAS）基因及其相应的信号通路与乳腺癌的发生相关，提示该通路可能被用来作为治疗或药物筛选的新靶标[37]。Rogers等通过DNA芯片对氟康唑耐药及敏感的白色念珠菌株的基因比较发现：可能由于CDR1、ERG2、CRD2、GPX1、RTA3、IFD5等基因表达上调导致了耐药的产生，而这些基因显然也就成为今后新药筛选的候选靶标分子[38]。

1.3 DNA微阵列基因表达谱数据

基因表达谱数据是指基于DNA微阵列实验得到的反映mRNA丰度的数据。基因表达谱数据中蕴含着基因活动的信息，可以反映细胞当前的生理状态，例如细胞是处于正常还是恶化状态、药物对肿瘤细胞是否有效等；同时基因表达谱数据可以提供大量的遗传变化信息[6]。利用微阵列技术研究癌症的倡导者，美国国家癌症研究院主任理查得克罗森曾说“基因表达谱代表着一种新型的数据”。

1.3.1 基因表达数据的获取

在制备样本时，使用两个样本，一个称为控制样本（Control Sample）或对照样本（Reference Sample），通常用绿色荧光素（Cy3）标记其cDNA，另一个为测量样本，用红色荧光素（Cy5）标记其cDNA 。这两个样本按照相同的实验方案分别制备不同荧光素标记的cDNA，并按1：1的比例混合，然后与cDNA微阵列杂交，用不同波长的激光扫描杂交后微阵列，分别获取荧光强度，并成像。来自两个样本的基因如果以相同水平表达则显示黄色，而如果表达水平有差异，则

图像显示红色或绿色。从DNA微阵列实验中得到的数值反映了基因的相对表达水平，即测量样本与对照样本之间荧光信号强度的比率或者对数化的比率。

基因芯片实验所产生的原始微阵列数据是图像，必须把经过图像处理后转化为基因表达谱数据，转化过程如图1.2所示。基因表达谱数据通常利用矩阵形式表示，称为基因表达矩阵，一般意义下的基因芯片数据分析都是基于该矩阵的。基因表达矩阵的行代表基因，列代表各样本及条件（如组织、实验条件、处理因素等），每个格子的数据表示特定的基因在特定的样本中的表达水平。

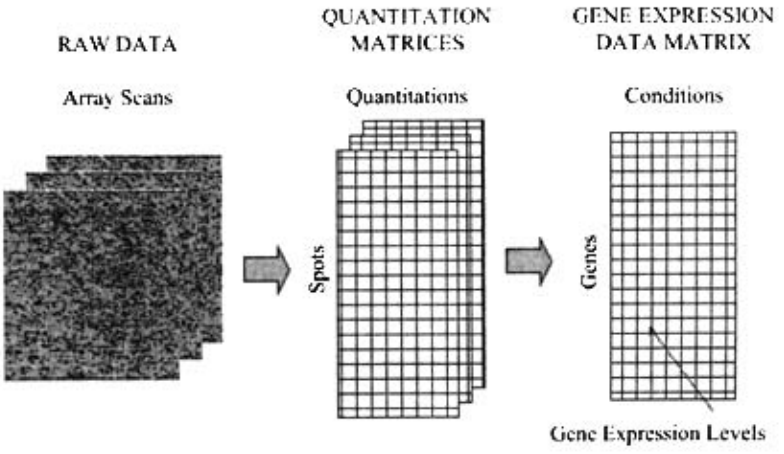

图1.2 基因表达谱数据

1.3.2 基因表达谱数据的特点

通过一系列基因微阵列杂交实验获取的基因表达数据具有以下特点：

1. 数据量巨大

伴随着DNA微阵列技术的发展与成熟，各种基因表达谱数据迅速增加。目前比较权威的基因芯片数据库有欧洲生物信息学研究所（European Bioinformatics Institute，EBI）的ArrayExpress[1]、美国斯坦福大学的SMD[2]（Stanford Microarray Database）以及美国国家生物技术信息中心（National Center for Biotechnology Information，NCBI）的GEO[3]（Gene Expression Omnibus）。以ArrayExpress数据库为例，截至到目前为止，该数据库共收录了涉及51个物种（Species）、46种实验

[1]http://www.ebi.ac.uk/arrayexpress/
[2]http://genome-www5.stanford.edu/
[3]http://www.ncbi.nlm.nih.gov/geo

类型（Experiment Type）以及64个实验因子（Experimental Factor）共16140次杂交的基因芯片数据。

2. 高维性

任何一个物种的基因组，都是由非编码蛋白质的核苷酸序列和编码蛋白质的核苷酸序列（基因）所组成。狭义基因只是有蛋白质产物的核苷酸序列，只占基因组很小的一部分。例如编码人类蛋白质的核苷酸序列大约占人类基因组的2%。从混杂有大量非编码核苷酸序列的基因组中找出基因非常困难。根据起始密码和终止密码所确定的非编码蛋白质的核苷酸序列是一种潜在基因。生物学家们把这两类基因都称为“开放阅读框”（Open Reading Frame，ORF）。通常一个基因组内的基因数目是指ORF的数目。秀丽线虫（C.ekegans）基因组的基因数为1.9万多个；人类基因组拥有的基因数目大约是在3万到4万个之间；水稻基因组的基因总数在4.6万到5.5万之间。如此高的维数给我们分析数据带来很大的困难，甚至会引起所谓的“维数灾难”（Curse of Dimensionality）。

3. 高噪声

在微阵列基因表达谱数据的获取过程中，由于实验设备和环境的影响，研究机构提交的大量基因数据在质量上无法保证充分的精确性，使得基因数据库中存在相当程度的噪声。比如样本制备过程中样本会有污染，杂交实验会因为实验条件的不一致产生误差，实验操作人员会产生偏差，实验仪器会带来偏差，这些都会引入噪声。因此，如何最大程度的去除基因芯片数据中的噪声，还原真实信号成为基因数据分析中必不可少的一步。

4. 高冗余

DNA数据库中的很多记录是属于同一基因家族（Gene Family），或在不同生物体上发现的同源基因。不同的研究机构可能向数据库发送了相同的序列数据，如果没有被检查出来，则这些记录或多或少地紧密相关。导致数据库中部分数据的冗余度太高。甚至某些基因有数千条EST与之对应。对基因数据分析时，一大堆无用的信息可能淹没了有用的信息。同时，由于功能相似的基因表达相近，有研究表明，在人类基因组中，细胞在任何发展阶段不同基因的相对冗余可以达到10000 以上。

5. 数据分布不均衡

通常的数据分析方法在数据分布均衡时可以取得好的结果，但遇到数据分布不均衡时会遇到麻烦，甚至引起错误。事实上，很多DNA微阵列数据中存在数据分布不均衡。例如，Gordon等人用基因芯片对肺恶性胸膜间皮瘤（Malignant Pleural Mesothelioma，MPM）和肺腺癌（Adenocarcinoma，ADCA）进行分类诊断，两种组织样本分别为31个和150个，后者数量是前者数量的5倍。这样的数

据对方法提出了更高的要求，否则很容易得出假阳性或假阴性的结果。

由以上分析可知，DNA微阵列数据的这些特点，对数据分析方法提出了更高要求，对基于DNA微阵列数据的癌症检测提出了挑战。

1.4 基因表达数据在癌症检测中的应用

科研工作者已经利用微阵列基因表达数据进行癌症检测和癌症分类研究。通过分析微阵列基因表达数据鉴定一些癌症的特定亚型，包括白血病、淋巴瘤、危险的恶性皮肤癌以及乳腺癌等。并且他们从中可以了解目前的治疗方法对哪些癌症是有效的，哪些是没有作用的。美国国家癌症研究所（National Cancer Institute，NCI）的露易斯斯图特曾预测由于DNA微阵列技术和基因表达谱数据的发展将带来癌症医学翻天覆地的变化－“癌症医学的教科书就需要重新编写了”[39]。

研究人员利用DNA微阵列技术可以高通量的检测癌症样本中基因的表达，分析样本中的基因表达数据以鉴定在癌症中哪些基因是表达上调或是表达下调；揭示出与癌症的病因和发展紧密相关的基因，发现激发癌症产生的致癌基因和抑制癌症的阻遏基因；并挖掘影响致癌基因、阻遏基因表达的干扰基因，阐明基因调控的相互关系。

麻省理工学院基因组研究中心的一个研究小组曾经宣称，用基因表达谱将癌症进行分类的设想是完全可以实现的[6]。他们从临床的标准病理检测中难以区分的急性髓性白血病（AML）和急性淋巴性白血病（ALL）的基因表达谱中挑选出了50个差异表达的基因，研究结果表明他们已经可以用这些基因表达对AML和ALL进行有效区分。

NCI的Staudt的研究小组与斯坦福大学医学院的Patrick Brown 和David Botstein小组合作，研究大面积扩散的B细胞淋巴瘤。在美国每年有一万五千多人的发病率，而且临床上变异大，只有40%的患者可以被治愈，60%的患者会死亡。研究人员制作了一种排列有18,000个基因的Lymphochip芯片，从40位患者的组织中分离出mRNA并与Lymphochip杂交。发现40位患者的基因表达有很大的差异，尽管他们都经过了相同的临床诊断。根据基因表达谱分析，可以将40位患者分成两组，一组患者中，发生免疫应答的脾脏和淋巴结的B细胞特征性表达了一组基因；而另一组患者中，这些基因却没有表达，但在受到抗原刺激后发生分裂的血液中的B细胞表达了另外一组基因。这两组患者的临床检测结果显示第一组有75%的患者多活了5年，而另一组恰恰相反。

MIT的研究小组在比较高转移黑色素瘤细胞和低转移黑色素瘤细胞的基因表达谱，找到了一组随着肿瘤的恶化表达明显上调的基因。这些基因中，许多都是

参与了癌细胞的转移，直接或间接的影响到细胞的移动和进攻性。其中一个研究小组深入的研究了一个称为RhoC的基因，这个基因曾经报道过与胰腺癌的转移有关。他们将这个基因转到人的黑色素瘤细胞（无扩散倾向）中，然后接种到小鼠上，他们发现这些细胞具有高度的转移性。

除了利用鉴定出哪些基因是对癌症的病因和发展紧密相关之外，研究人员利用微阵列技术和基因表达谱来研究那些激发肿瘤产生的癌基因和抑制肿瘤的癌基因是如何干扰其它基因表达的。微阵列技术可以阐明特定遗传缺陷及如何调控的相互关系。Staudt小组利用Lymphochip来研究BCL-6基因的异常活动的情况。他们发现BCL-6基因的活动会导致blimp-1基因表达的抑制，blimp-1基因的功能是促进B细胞的分化，变成一个可以分泌抗体的血细胞。另一个受到影响的基因是p27kip1，它会影响细胞的分裂周期。这两个基因对细胞产生双重作用，使细胞保持在在一个分化停滞，不断分裂的状态。在西雅图Fred Hutchinson癌症研究中心的Eisenman小组与MIT合作研究Myc基因的活性变化。他们发现，Myc基因的活性会引起27个基因的上调表达，这些基因中包括一些促进细胞分裂的基因，同时还引起另外9个基因的下调表达。

1.5 课题的研究意义

随着以测序为主的结构基因组计划（Structural Genomics Project）的完成（在该计划中中国承担了1%人类基因组的测序任务），生命科学研究的重点也逐渐的转变为以对基因功能研究为主的功能基因组计划（Functional Genomics Project）。功能基因组计划将系统研究人类3万多个基因在正常人体的生理功能；深入探讨这些基因变化导致人类疾病的分子机制；最终找到具有巨大医学价值的基因药靶。它正在掀起生物医学领域的一场新的革命，并全面推动生命科学领域的基础研究以及生物工程，农业和医药等领域应用研究的蓬勃发展。在“九五”“十五”期间，功能基因组计划研究已被列为国家高科技计划863和973重大专项。

功能基因组计划的主要任务之一是进行基因功能鉴定，研究与疾病相关基因功能，寻找调控疾病的相关基因，研究通过基因表达实现疾病诊断和基因治疗。通过微阵列基因表达谱数据的分析来检测癌症不仅能够预测患者样本的癌症类型，针对患者制定有效的治疗方案，帮助患者进行癌症治疗，还能够辅助研究人员开发与研制新的抗癌药物，分析癌症样本中的基因表达模式以确定所研制的新药的疗效。因此利用基因表达谱数据研究癌症检测具有非常重要的实际意义和应用价值。

癌症检测是从患者样本中检测患者是否患有癌症，以及识别样本的癌症类型

等，是疾病诊断和基因治疗非常关键的步骤。癌症识别方法包括无监督的聚类方法和有监督的分类方法，根据基因表达数据集如何建立有效的癌症识别模型，以预测患者样本的癌症类型，在基于微阵列基因表达数据的癌症检测中具有非常重要的地位。现有的癌症识别算法各有特点，有研究者证实没有哪种算法具有绝对的优势。建立在不同的特征子集上的分类模型得到的分类结果存在较大的差别，尤其在噪声较大和冗余严重的基因数据集上差别更为严重。因此如何建立适合高维、高噪和高冗余的基因表达数据的癌症识别方法具有非常重要的意义。

在微阵列基因表达谱中，样本数目一般为几十或上百例，而检测基因的数目往往高达几千甚至几万。在癌症检测中，容易导致“维数灾难”（Curse of Dimensionality）问题。特征选择法是一种有效的降维方法，但不同的特征选择法使用的搜索机制和评价策略不同，挑选出的特征基因明显不同。并且在高冗余的基因表达谱数据中选取数量有限的基因容易丢失有生物价值和分类意义的信息。如何选择合适的特征基因来识别患者样本的癌症类型，以及如何识别不同癌症组的有区别性的基因表达模式也是癌症检测中具有非常重要的意义。

基于以上的背景，本课题将着重研究癌症检测研究中的癌症类型识别、基因表达模式的挖掘和基因特征的抽取问题。

1.6 课题的研究内容

本课题主要研究高维基因表达数据中最具辨别能力的特征基因的选择方法；基因组和癌症组中隐含的潜在的基因表达模式；癌症基因组之间的相互调控关系，具有可解释性的显现模式；利用选取的特征基因和抽取的基因表达模式以及显现模式建立合适的癌症分类模型，预测患者样本的癌症模型；癌症预测中的多分类模型的组合；为基因的功能性研究、癌症的基因表达分析、疾病的临床诊断、抗癌药物的研制提供基础。

对于癌症检测中基因表达数据的预处理、特征基因的选取、癌症组基因表达模式的分析以及建立合适的基因诊断模型的问题。研究的具体内容如下：

1. 在癌症检测中自适应的基因选择方法研究

基因表达谱是一种高维、高噪、高冗余的数据。在癌症检测中，需要对这些数据进行除噪和降维。传统的特征选择算法在基因表达数据处理中存在不足，并且没有较好的特征基因数目的预置机制。因此，本课题研究适合基因表达数据的特征基因选择算法过滤误差数据、冗余信息和噪声信息，提出一种基于CMST聚类的分步的特征基因的选择机制，利用分步的方法消除不同方面的无价值信息，并在CMST中引入“Gap Statistic”理论以确定合适的特征基因数目，提出一种自适应的特征基因选择方法。

2. 基于隐含变量模型的癌症分类算法研究

在人类基因组中，基因表达之间存在相互影响，并且具有复杂的基因调控原理和机制。因此，本课题研究基因表达数据中隐含的基因表达模式，通过隐含的基因表达模式来分析不同基因之间的相互影响，以及在癌症中整个基因组的调控机制，利用抽样策略来选择基因子集，减少噪声和冗余信息对PCAP和ICAP的影响，在此基础上重构基因表达，提出基于隐含变量模型下的癌症分类算法。

3. 癌症组相关联的基因表达模式抽取与癌症识别研究

由于癌症组中样本分布不均衡，传统的分类算法在癌症识别中遇到遇到麻烦，并且难以克服高维的基因谱和低维的阵列谱（样本）带来的“维数灾难”问题。因此，本课题通过建立癌症组的关联空间来研究癌症组中的基因特征抽取和基因表达模式，讨论癌症组相关联的基因表达模式在癌症组中的表达以及调控，研究适合癌症组相关联的基因表达模式的癌症预测。

4. 基于组合分类算法的癌症检测算法研究

由于基因表达谱具有高维性、高噪声和高冗余的特点，使用不同搜索机制和评价策略的特征选择方法，挑选出的特征基因偏向于癌症病理特征的不同方面，存在明显区别。那么，建立在不同特征基因上的分类器的癌症识别结果差别很大，导致分类器缺少泛化性。因此，本课题主要研究癌症识别中的组合分类算法，综合不同分类器优点的同时，消除噪声和冗余信息的干扰，并在癌症组数据上和基因组数据上建立一组具有互补性的分类器，通过互补性的分类器组合以提高癌症检测效果。

5. 基于显现模式的癌症分类算法研究

以往的基因表达数据分析都是从简单的单个基因表达入手，而癌症的产生和发展过程是由部分基因共同调控和表达的结果，需要我们从基因之间的协同表达来分析基因数据。同时，在后续所服务的临床研究中，需要具有可解释的基因表达模式。因此，本课题主要研究适合于解释和分析的基因协同表达的显现模式，在显现模式的抽取过程中增加虚拟样本和利用高斯分布来模拟候选分割点的分布，以提高显现模式的癌症辨识能力，研究适合于显现模式的癌症检测算法。

1.7 论文的组织结构

本文的主要工作是研究基于DNA微阵列基因表达谱数据的癌症检测问题，由于微阵列表达数据高噪声、特征量大、样本数少和特征之间关系复杂等，对机器学习领域的现有分类算法的实现和可行性，提出了新的挑战。针对这些情况，本文主要对癌症检测研究中的癌症类型识别、基因表达模式的挖掘和基因特征的抽

取问题内容进行了研究。本论文的组织结构如下：

第1章 介绍了DNA微阵列技术、DNA微阵列基因表达谱数据的数据获取技术、DNA微阵列基因表达谱数据的特点和应用、以及相关的背景内容，为后续研究提供基础。

第2章 对基于微阵列基因表达数据的癌症检测研究中的癌症类型识别、基因表达模式的挖掘和基因特征的抽取问题的主要研究内容，方法及其特点作了综述，该领域的研究内容包括：无监督的聚类方法、有监督的分类方法及判别分析、特征选择和特征抽取等模式处理方法等。接着本文描述了我们在这些方面所完成的工作：

第3章 介绍本文提出的在癌症检测中一种自适应的基因选择方法，首先针对基因表达数据高维、高噪声、高冗余的特点，提出一种基于CMST聚类的分步的特征基因选择方法。在不同的特征分析步骤中分别去除基因数据中的高噪声和高冗余，达到特征基因选择降维的目的。然后，引入Tibshirani等提出的“Gap Statistic”理论，提出了一种自适应的特征基因的选择方法，解决目前的特征基因选择算法中缺乏较好的基因数目预置机制问题。

第4章 介绍本文提出的基于隐含变量模型的基因分类算法，利用主分量分析方法（Primary Component Analysis，PCA）和独立分量分析方法（Independent Component Analysis，PCA）分别挖掘基因表达谱中的隐含的基因调控因子，构建基于隐含变量的基因表达模式，以揭示基因表达之间存在的相互影响以及调控机制，并利用隐含变量对患者样本进行癌症预测。

第5章 介绍本文提出的基于癌症组关联空间的基因表达模式抽取与癌症识别算法。针对基因数据集中高维的基因谱和低维的阵列谱（样本）带来的“维数灾难”问题，利用癌症组基因表达存在的局部特征相关性的生物病理特点，抽取不同癌症组的特征模式和基因表达模式，讨论与癌症组相关联的基因表达模式在癌症组中的表达以及调控，并提出适合癌症组相关联的基因表达模式的癌症预测算法。

第6章 介绍本文提出的基于组合分类算法的癌症识别算法。在高维的基因表达谱中，不同的特征选择法被用来选取特征基因。由于基因数据高噪声和高冗余的特点，并且不同的方法采用不同的搜索机制和评价策略，挑选出的特征基因明显不同，导致分类器的癌症识别结果不稳定。针对癌症组基因数据和基因组数据提出一组具有互补性分类器，然后在互补分类器的基础上利用组合分类算法提高癌症分类效果。

第7章 介绍本文提出的基于显现模式的癌症分类算法。癌症的产生和发展过

程是由部分基因共同调控和表达的结果，需要我们从基因协同表达来分析基因数据。同时，临床研究和应用中需要具有可解释的基因表达模式。在显现模式的抽取过程中增加虚拟样本和利用高斯分布来模拟候选分割点的分布，挖掘具有更高辨识能力的显现模式，并提出显现模式的癌症检测算法。

最后，第8章对本文的研究工作进行了总结，指出了有待继续深入研究的问题，展望了未来的发展方向和进一步的研究工作。

第 2 章　基于DNA微阵列数据癌症检测的研究现状

通过分析DNA微阵列基因表达谱数据研究人类基因功能和基因表达调控机制，从而利用基因在患者样本中的表达识别样本的癌症类型，并使研究人员关注那些对癌症的发展、维持、扩散有重要影响的基因和那些可能的靶药物。这是生物信息学研究的重大挑战之一[40]。分析DNA微阵列基因表达谱数据可以从人类基因组来研究基因的表达和调控，改变了以往从单一基因来分析的局面。美国国家人类基因组研究院的保罗麦尔兹博士曾说："我们一直习惯于一次研究一个基因，可一旦这项工作（DNA微阵列技术）开始了，你不必再那样做了，一切都完全变了。"

围绕本文的工作，给出癌症检测研究中的癌症类型识别和基因模式抽取相关问题的文献综述：聚类（Clustering）分析是基因功能分析中使用最广泛的技术，通过分析在不同环境和条件下表达相似的基因，对基因功能进行鉴定。研究与疾病相关基因功能，以及其在疾病中的调控作用。同样，通过不同组织样本基因表达的聚类分析以检测和发现癌症模式。分类（Classfication）方法是一种有监督的机器学习方法，根据具有标识的基因表达数据集建立癌症识别模型，揭示基因功能和参与的生化途径基因之间调控关系。在DNA微阵列数据中包含了大量的基因，其数量远超过样本数量，这些基因中大部分与区分癌症类型无关，通常会降低癌症检测性能。所以在高维、高噪和高冗余的基因数据中进行有效的基因模式预处理，抽取合适基因表达模式也是基于DNA微阵列数据癌症检测的重要内容[36, 41-43]。

2.1 聚类方法

通过各种不同的数学模型，对具有相同统计行为的多个基因进行归类，归为一个类的基因在功能上可能相似或相联，并根据同类中已知基因的功能推测未知基因的功能。也可根据聚类结果对基因转录调控网络作进一步的分析。同样采用聚类分析对样本聚类可鉴定出以前未曾认识到的癌症。聚类方法是一种无监督的机器学习方法，主要优点是不需要事先标识好类别标签的训练集，在基因数据集中可以发现和预测未知的基因模式。聚类已经成为基因芯片数据模式分析的一个基本方法，当前，聚类算法很多，常见的包括K-means[44]、层次聚类（Hierarchical Clustering）[45]、自组织映射（Self-Organizing Map，SOM）等[46, 47]

2.1.1 K-means

K-means[44,48]是一种动态聚类方法[42]，从预先指定聚类数量开始，起始位置是每个聚类的中心。它由以下步骤组成:

1. 初始随机或通过先验知识选择k个基因作为聚类的中心；

2. 其他基因表达向量被分配到距离自己最近的聚类中心的聚类内，以此进行数据的分区，欧氏距离（Euclidean Distance）是最常用的向量对之间距离的测度；

3. 每个聚类计算本聚类内所有基因表达向量的平均值，以此作为该聚类中心新的值；

重复执行以上步骤2、3，直到目标函数收敛；

一般利用类内散度最小化作为目标函数，即$min(\sum_{i=1}^{k}\sum_{x\in C_i}|x-m_i|^2)$，其中，x是向量空间中某个基因表达向量，$m_i$是聚类$C_i$中各向量的平均值。

K-means算法运算速度快，内存开销小，比较适合于大样本量的情况，但是聚类结果受初始凝聚点的影响很大，不同的初始点选择会导致截然不同的结果；并且当按最近邻归类时如果遇到到两个凝聚点距离相等的情况，不同的选择也会造成不同的结果。因此动态聚类法具有很大的不确定性，并且初始类数的确定需要领域专家参与。

目前，Stanford大学的Botstein实验室和美国国家人类基因组研究所（National Human Genome Research Institute，NHGRI）的Trent实验室主要研究K-Means算法分析基因表达谱。Tavazoie等利用K-Means对酿酒酵母细胞周期（Cell Cycle in Saccharomyces Cerevisiae）中的基因表达数据进行聚类并搜索类中具有共同表达的DNA序列[44]，从12个类（Cluser）中发现了18个DNA序列，其中7个被实验证明在相应类中具有调控基因表达的功能。

2.1.2 层次聚类法

层次聚类法（Hierarchical Clustering）[45]和K-means算法是目前聚类分析中应用最多的两种方法，根据层次（系统）树的生成方法（自底向上或自顶向下），层次聚类法分为凝聚法（Agglomerative）和分裂法（Divisive）。凝聚法采用自底向上的策略：首先将每个对象作为一个簇，然后合并这些原子簇为越来越大的簇，直到所有的对象都在一个簇中，或者某个终结条件被满足。分裂法采用自顶向下的策略：首先将所有对象置于一个簇中，然后逐渐细分为越来越小的簇，直到每个对象自成一簇，或者达到了某个终结条件。

层次聚类算法比较简单，可以用谱系聚类图形象地显示聚类过程（图2.1），

并且不需要预置簇的数目，在任意层次上进行划分可产生不同的聚类结果。但是层次聚类经常面临合并或分裂点选择的困难。一旦一组对象被合并或者分裂，下一步的处理将在新生成的簇上进行。已做的处理不能被撤消，聚类之间与不能交换对象。如果在某一步没有很好地选择合并或分裂的决定，可能会导致低质量的聚类结果。并且合并或分裂的决定需要检查和估算大量的对象或簇，不具有很好的可伸缩性。

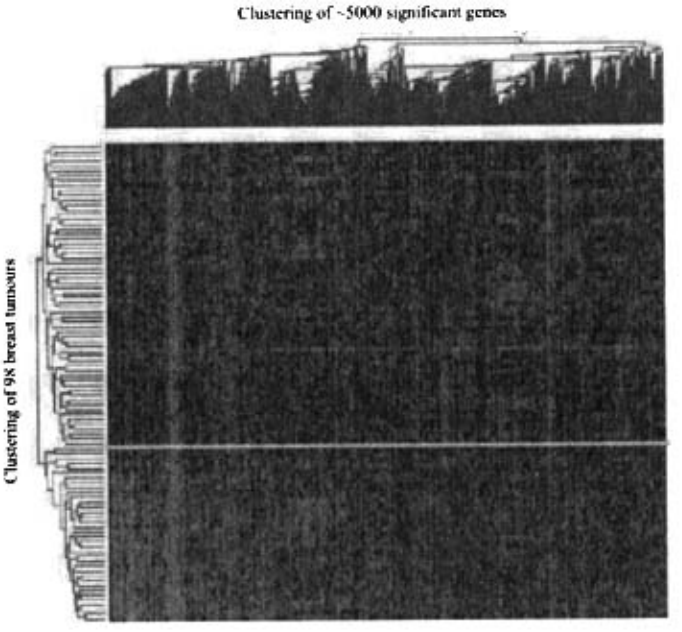

图2.1 双向层次聚类图

在利用凝聚法研究DNA微阵列的层次聚类方面，Eisen等利用皮尔逊相关系数（Pearson Correlation Coefficient）构造了在酿酒酵母数据集和饥饿的人类基本细胞在血清刺激下的基因表达数据集（Gene Expression of Primary Human Fibroblasts Stimulated with Serum Following Serum Starvation）中的基因的层次结构聚类树，从两个数据集中都得到在相同组中的基因存在着功能相似性的特性[45]。Alizadeh等[49]研究了96 例正常和恶性浸润型大B 细胞淋巴瘤样品，层次聚类分析显示可能存在两种新的肿瘤亚型，分别代表B 细胞分化的不同阶段。研究结果表明这种新分型与肿瘤病人生存率有很好的相关性。Alon等[50]利用层次聚类方法分析6500条基因在40例大肠癌和20个正常大肠组织中的表达谱，同时做基因和样品的聚类，这种方法称为“双向聚类”，发现一些基因群体是和某些肿瘤类别相关联的。在利用分裂法研究DNA微阵列的层次聚类方面，Herrero等提出了一

种基于神经网络的自组织树算法（Self-Organising Tree Algorithm, SOTA）[51]。

2.1.3 SOM

自组织映射（Self-Organizing Map，SOM）是Kohonen教授在1981年提出的一种竞争式神经网络[52]。它模拟大脑神经系统自组织特征映射的功能，通过调整权系数，使神经网络收敛于一种表示形态，在这一表示形态中的一个神经元只对某种输入模式特别匹配或特别敏感，从而实现输入对象的自组织聚类（图2.2）。SOM神经网络是一种基于模型的聚类方法[53]。Tamayo等[46]利用SOM研究了肿瘤细胞株HL-60、U937、Jurkat和NB4，将6000多个基因分成在表达水平有较大的差异且具有生物表达意义的基因簇。Wang等[53]提出了一种基于SOM的双层（Two Level）基因表达分析法，有效减少了弥漫大B细胞淋巴瘤（Diffuse large B-cell Lymphoma，DLBCL）基因数据的多维性。

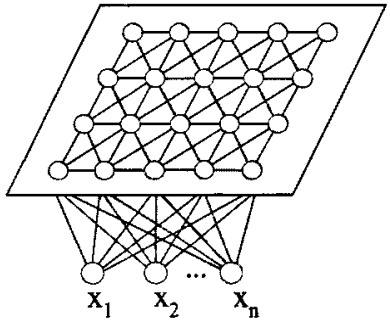

图2.2 自组织映射（SOM）

2.2 分类方法

对于DNA微阵列数据的分类，是通过已知类别的样本训练集来训练预测模型，建立能区分给定类别中成员和非成员的分类器，用该分类器来预测未知组织样本的类别。为建立分类器，需要用已知类别的数据来训练，在训练数据集中每个示例不仅包含属性（特征），即样本在所有基因中表达值，还有各样本所属的类别。与聚类的根本不同在于分类是有监督学习（Supervised Learning）方法[54, 55]。主要的分类方法包含G-S法、KNN、SVM和决策树等[36, 56, 57]。

2.2.1 G-S法

Golub和Slonim等[6, 58]首次在《Science》上发表利用微阵列基因表达谱数据分析和识别急性白血病的两种亚型AML与ALL的研究成果，通过分析DNA微阵列

数据成功鉴别出27例ALL及11例AML样本，在此称之为G-S法。首先构造了一种“基因表达的理想模式”（Idealized Expression Pattern），认为理想的基因表达模式是在一种癌症类别中表达水平高，而在另外的癌症类别中表达水平低，并通过“信噪比”$P(g,c)$度量基因对样本分类贡献大小，利用邻域分析方法选择特征基因。

$$P(g,c)=\frac{\mu_1(g)-\mu_2(g)}{\sigma_1(g)+\sigma_2(g)} \tag{2.1}$$

其中，$\mu_1(g)$和$\mu_2(g)$分别为基因g在$class_1$和$class_2$上表达的平均值，σ_1、σ_2为对应的方差。然后设定投票决策分界线（Decision Boundary），用V(g)表示，并联合分类贡献度权重通过投票方法构建分类模型（图2.3）。

$$\begin{aligned} V\{ALL,AML\}&=\sum_i V(g_i)Weight(g_i)\\ Weight(g_i)&=P(g_i,c)\\ V(g_i)&=g_i-(\mu_{AML}(g_i)+\mu_{ALL}(g_i))/2 \end{aligned} \tag{2.2}$$

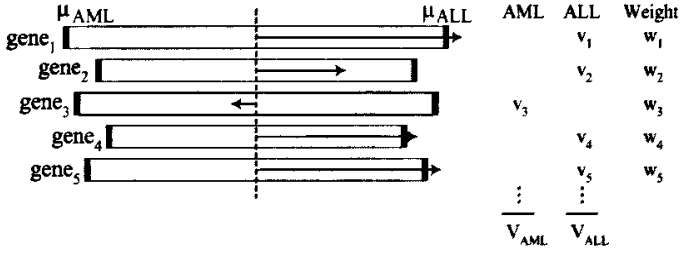

图2.3 G-S法

2.2.2 K近邻分类

K近邻分类算法（K-Nearest Neighborhood，KNN）是懒散的分类算法。假设每个样本代表空间中的一个点，给定一个未知样本，KNN搜索模式空间，找出最接近未知样本的k个训练样本。这k个样本是未知样本的k个“近邻”，“临近性”一般用欧几里德距离定义。未知样本被分配到k个最近邻中最公共的类[59, 60]。

Kuramochi等[55]利用KNN对酵母数据集[45]中2462个基因进行功能分类，功能基因组取自MIPS[61, 62]中50个最大基因家族，不同基因家族包含的基因变化很大，发现分类正确率随功能组不同变化很大。由于KNN的分类性能依赖于相似性度量，在处理异类基因数据（Heterogeneous data）时分类性能不理想，Yao等[60]结合不同

的相似性度量采用回归分析方法（Regression Method）产生优化的相似性度量，从而获取更相似的“近邻”。

2.2.3 SVM

SVM建立在计算学习理论的结构风险最小化原则之上。其主要思想是针对两类分类问题，在高维空间中寻找一个超平面作为两类的分割，以保证最小的分类错误率[13, 14, 63]。SVM能够有效处理高维的基因数据集。

Brown等[63]利用SVM为从MIPS获取的的六个功能基因组建立一组二分法分类器（Binary Classifier），对Eisen[45]选取的有功能标识的2467 个酵母基因进行功能分类。Furey等[13]等利用SVM对31例卵巢癌（Ovarian Cancer）组织数据集进行了研究，分离出了卵巢癌组织和正常卵巢组织。Semolini等[64]引入直推式学习（Transductive Inference），根据已知样本对特定的未知样本建立一套进行识别的方法和准则，提出了一种基于支持向量机的渐进直推式分类学习算法（Transductive Inference with Support Vector Machines，TSVM），有效扩展了SVM的泛化性能。通过专家建立基因训练集非常困难，并且十分耗时，易导致概念漂移（Concept Drift），Liu等[65]在癌症分类中提出了一种基于主动学习（Active Learning）的SVM算法，有效地减少了训练样本数量，并且提高了识别率。

2.2.4 决策树

决策树（Decision Tree）是一个类似于流程图的树结构，其中每个内部结点表示在一个属性（基因）上的测试，每个分枝代表一个测试输出，每个树叶节点代表类（癌症类型）。决策树一般都是自上而下来生成的。从根到叶子节点都有一条路径，这条路径就是一条“规则”。

决策树可以生成可以理解的基因表达规则，可以清晰的显示哪些基因比较重要，哪些基因表达激发或抑制癌症的生成，计算量不太。但是当类别太多时，错误会增加的比较快，并且在分类的过程中，只是根据一个属性来分类，由于基因表达数据具有高噪声的特点，噪声易淹没有用的基因表达规则。

Gerald等[66] 利用决策树研究结核菌素皮下测试（Tuberculin Skin Test）的情况，挖掘易导致测试呈阳性的属性。Feinglass等[67]利用决策树研究了间歇性跛行症（Intermittent Claudication）治疗情况，建立了诊治间歇性跛行症的方案模型。Gaudart等[68]基于CART算法（Classification and Regression Trees）提出了一种倾斜决策树算法（Oblique Decision Tree model，ODT）研究疟疾（Malaria）识别问题。组合分类算法可以有效的扩展样本容量，并综合不同分类器识别结果弥补不足，文献[69, 70]提出基于组合策略（Bagging 或Boosting）的组合决策树分类算

法，并在七种肿瘤细胞株进行识别，取得了比单一分类器更好的识别率。

2.3 基因模式预处理

在基因表达数据中挖掘基因表达模式，识别癌症类型所面临的如下问题：首先，基因表达数据中包含大量基因数目，然而，在多数情况下，这些基因中只有一小部分包含与分类相关的信息。识别这部分基因将有助于揭示生物学过程。另外，与分类无关的基因将会干扰相关基因的信息。其次，数据的高维数不可避免的会使模型推导或估计复杂化，这往往会引起机器学习的“过学习”（Overfiting）问题，使学习机器泛化能力降低。并且基因数量远高于样本数量，导致维数灾难问题（Curse of Dimensionality）。因此，对数据预处理并减少基因模式数量对于癌症检测和分类显得尤其重要。

2.3.1 特征选取

排序法（Sorting）是应用最广泛的特征基因选择方法，首先通过某种度量来衡量基因对样本识别的贡献度，然后依据度量值进行排序。常用的度量有信噪比（Signal to Noise Ratio，SNR），相关系数法（Correlation Coefficient）、信息增益（Information Gain）和互信息（Mutual Information）等。Golub等[6]提出一种以信噪比为基础的“邻域分析法”（Neighborhood Analysis），首先构造一种能够区分不同癌症类型的基因的“理想表达模式”，然后根据基因与“理想表达模式”的皮尔逊相关系数排序基因。Guyon等[14]提出了递归特征减少法（Recursive Feature Elimination，RFE），借助支持向量机递归去除分类函数中关联权重绝对值最小的基因，得到基因集合的排序来选择鉴别基因。Cho等[71]讨论了基于不同度量函数的排序方法选择特征基因的问题，发现不同方法提取的特征基因存在很大的不同，对癌症类型的识别也有很大的影响。

基于模型的特征选择方法也是一种常见的特征基因选择方法。Keller等[72]根据基因在同一样本组中具有的分布相似性，提出一种基于似然性的基因选择策略（Likelihood Selection）。通过基因的相关对数似然值（Relative Log Likelihood Score）来选择特征基因以提高NB（Native Bayes）算法的预测精度。Lee等[73]针对特征基因选择不稳定的情况提出分层贝叶斯基因选择模型（Hierarchical Bayesian Model），结合马尔可夫链蒙特卡罗方法和Gibbs抽样来估计每个基因在模型中出现的概率，根据概率的大小来选择鉴别基因。基因表达数据通常含有噪音，而且不同数据间的差异很大。因此，很难构造一个适用于所有基因表达数据的数据模型。

不同的特征选择法利用不同的搜索机制和评价策略，挑选出的特征基因也明

显不同，导致样本识别结果差别较大。因此，研究人员提出了组合特征基因选择方法。Jaeger等[74]提出基于聚类的特征基因选择方法，Goh等[75]提出组合皮尔森相关系数（Pearson Correlation Coefficient，PCC）和信噪比（SNR）的特征基因选择方法，利用PCC产生基因功能组，在基因组中采用SNR提取选取特征基因。李博士等[76, 77]提出一种基于支持向量机（SVM）的两步特征基因选取方法，利用“分类信息指数”（Information Index to Classification，IIC）作为一种新的类别可分性判据以滤除分类无关基因，并采用两两冗余分析及基于支持向量机分类模型评价特征基因的分类性能剔除冗余基因。

特征选择方法的主要优点是计算复杂度较低、速度快。但是挑选出的特征基因子集缺乏稳定性，并且选取数量有限的特征基因容易丢失有生物价值和分类意义的信息。

2.3.2 特征变换法

特征变换法通过变换基因特征以抽取隐含的基因表达模式。Conde等[78]提出了一种基于聚类的特征基因抽取方法，利用SOTA层次聚类方法产生一系列组之间基因表达非冗余的基因簇，用簇的平均值作为基因特征以建立癌症识别模型。

Raychaudhuri等[79]利用主分量分析方法（Principal Components Analysis，PCA）揭示了在酵母细胞周期（Yeast Cell Cycle）数据[80]的隐含变量（Underlying Factors），发现2个隐含变量可以压缩数据集中7个测试变量90%的信息量，3个隐含变量可以压缩95%以上的信息量。Khan等[33]结合PCA处理基因数据，结合人工神经网络研究儿童小圆形蓝色细胞恶性肿瘤（Small Round Blue-Cell Tumors，SRBCTs）4种癌症亚型的识别问题。Liebermeister等[81]利用ICA抽取酵母细胞周期（Yeast Cell Cycle）中的25个基因表达模式和淋巴癌中12基因表达模式。Hori等[82, 83]利用ICA盲分离酵母细胞周期基因表达数据，比PCA取得更好的分离效果。特征变换法的主要优点是可以揭示在基因数据集中潜在的有价值的隐含变量。

2.4 本章小结

围绕本文的研究内容，本章综述了癌症检测研究中的癌症类型识别和基因模式抽取等工作的研究现状及相关问题。主要介绍了常用的聚类方法（包括K-means、层次聚类、自组织映射等）、分类方法（包括G-S法、KNN、SVM和决策树等）和特征预处理方法（包括排序法、基于模型的特征选择方法、特征变换法等）以及在基因数据处理和癌症检测中的研究现状。

第3章 在癌症检测中一种自适应的基因选择方法

DNA微阵列（Microarray）技术的发展使我们能利用基因表达谱进行癌症预测。然而，在维数巨大的基因表达数据中含有大量的噪声和冗余数据，而具有强分辨能力的特征基因相对很少。如何选择特征基因，以及选择的特征基因数目困扰利用基因表达数据对癌症检测的研究。本章在癌症检测中提出了一种基于最大相似树聚类算法（CMST）的分步基因选择方法。通过这个方法，我们得到一组基因簇，在每个基因簇中的基因之间具有相似性，而在不同基因簇中的基因之间冗余较小。然后从基因簇中选择最具分辨能力的特征基因，并将这些特征基因作为输入数据训练感知器模型，产生一个用于癌症检测的分类器。最后在CMST中，引入间隙统计量（Gap Statistic）选择最佳相似度阀值，提出一种最优自适应的CMST（OS-CMST）算法，并通过OS-CMST选择特征基因进行癌症检测。实验结果显示，在癌症检测中利用CMST进行基因模式预处理不仅能明显地降低数据特征维数，还能有效提高分类的准确性；OS-CMST利用Gap Statistic预测基因表达模式数目，并选取了具有强分辨力的基因。

3.1 概述

在利用DNA微阵列基因表达谱数据进行癌症检测中，由于基因表达数据含有大量噪声和冗余信息，有用的特征基因容易淹没在噪声和冗余信息之中。研究人员经常面临噪声和冗余信息对癌症识别决策的干扰。因此，特征基因模式的选择和合适数目的基因模式的决策问题在癌症检测中具有非常重要地位。

特征选择是在建模的时候选择最适当特征的过程[52]。在癌症检测中，许多特征选择方法被用来获取最适当的基因模式：Veer等根据基因和疾病类别关联程度排序基因，选择了乳腺癌（Breast Cancer）中70个特征基因[54]；Shipp等利用信噪比选择了弥漫性大B细胞淋巴癌（Diffuse Large B-cell Lymphoma，DLBCL）中30个特征基因[84]；Cho等系统地分析和评价了不同特征选择方法的性能[71]。特征选择方法可以减少噪声信息，然而特征选择方法不能有效地消除冗余信息，从整个数据集中选择的少量特征基因之间存在较大的冗余[75, 85]。

Mateos和Conde等利用聚类的特征处理方法消除基因特征之间的冗余[78, 86, 87]，首先利用一个基于自组织映射（SOM）的分层算法，即自组织树算法（SOTA）将基因分为若干基因簇，然后把基因簇中基因的平均表达水平作为感知器的输入参数，并对感知器加以训练后对样本进行分类[78, 87]。然而，由于基因簇的平均表达水平没有去除噪声数据的影响，同样也只是利用经验值决定特征基因数目。

本章提出一种分步的特征基因选择算法，第一步，利用提出的最大相似树算法（CMST）对基因进行聚类，产生功能相似的基因簇，并降低基因簇之间的冗余度；第二步，利用特征选择方法IG或SNR从基因簇中选取得最具辨别能力的基因，以消除簇中的噪声和不相关数据。最后，利用这些特征基因训练感知器模型并进行样本分类。这种分步方法可以有效的排除冗余和噪声信息，提高被选特征基因的质量。

在如何决定特征基因数目的问题中，我们在CMST中利用相似度量的间隙统计量（Gap Statistic）[88]选取最佳相似度临界值，提出最优自适应CMST（OS-CMST）聚类算法，从而通过非参数输入分步选择最优的特征基因。实验表明基于CMST的分步特征基因选择算法有效地消除了噪声和冗余信息的影响，提高癌症检测的准确度；基于OS-CMST的分步特征基因选择有效地预测基因表达模式数目，并选取了具有强分辨力的特征基因。

3.2 相关方法

3.2.1 IG

信息增益（Information Gain，IG）[71]是指信息熵的有效减少量，是一种有效衡量属性变量与输出类别关系的方法，可以用来度量基因变量在检测癌症类别中的贡献程度，根据IG排序基因以选择特征基因。设数据集C划分为正例样本（Positives，P）和负例样本（Negatives，N），则信息熵为：

$$Entropy(C) = -\frac{P(p)}{P(p)+P(n)}log\frac{P(p)}{P(p)+P(n)} - \frac{P(n)}{P(p)+P(n)}log\frac{P(p)}{P(p)+P(n)} \quad (3.1)$$

根据基因g_i对于样本的诱导（Induced）性和抑制（Regressed）性将C分为两个子集C_1和C_2，期望信息熵为：

$$E(g_i, C) = \frac{P(C_1)}{P(C_1)+P(C_2)}Entropy(C_1) + \frac{P(C_2)}{P(C_1)+P(C_2)}Entropy(C_2) \quad (3.2)$$

基因g_i对样本进行分类的信息增益为：

$$IG(g_i) = Entropy(C) - E(g_i, C) \quad (3.3)$$

其中，$P(\cdot)$是样本子集中的样本数。

3.2.2 SNR

信噪比（Signal to Noise Ratio，SNR）[71]是一个描述变量区分不同类别能力的统计量，在基因选择中，可以用来度量基因含有样本分类信息多少：

$$SNR(g_i) = \frac{\mu(g_i, P) - \mu(g_i, N)}{\sigma(g_i, P) + \sigma(g_i, N)} \tag{3.4}$$

其中$\mu(g_i, P)$和$\mu(g_i, N)$分别代表子集P和N中基因g_i表达水平的平均值，$\sigma(g_i, P)$和$\sigma(g_i, N)$则表示子集P和N中基因g_i表达水平的标准方差。

3.2.3 SOTA

自组织树算法（Self-Organizing Tree Algorithm，SOTA）[51]（图3.1）结合了自组织图（Self-Organizing Map，SOM）和层次聚类（Hierarchical Clustering），是一种自顶向下的分裂聚类法（Divisive Clustering）。自组织树的拓扑结构为动态的二叉树。和自组织图聚类相似，每个基因向量都依次迭代地作为组织图节点的输入，并且定位在离它最近的节点上。这个节点和他的临近节点向输入的基因向量移动。将所有基因向量输入到组织图中直到组织图收敛。收敛之后，将节点中距离最大的两个基因向量分解，形成两个新的节点，这就是动态二叉树的生长。整个过程再重新开始，反复这个过程直到每个节点中基因向量的距离达到设定的阈值。从而构造一个分层的基因树状图，基因属于哪个基因簇由它所处的树的层次决定。在参考文献[78]中，在对基因表达数据样本进行分类的过程中，基因簇的基因平均表达水平被用来训练感知器模型。

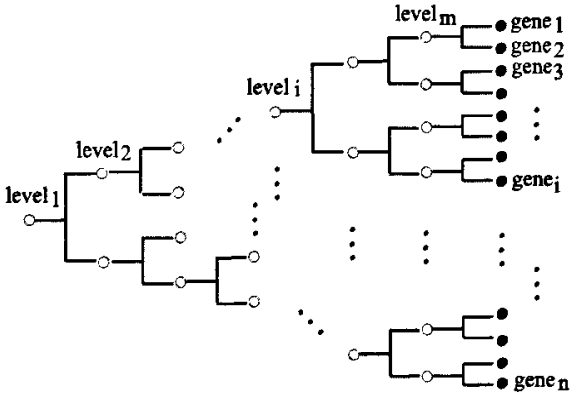

图3.1 自组织树算法（SOTA）

3.3 癌症检测中基于CMST的特征基因选择方法

这一节，在癌症检测中我们提出一种分步的特征基因选择方法。第一步，利用提出的CMST的聚类算法来对基因进行聚类分析，将所有基因划分成功能相似的基因簇；第二步，在每个基因簇中利用IG/SNR排序基因，选择具有分辨力的特征基因。在不同步骤混合不同基因特征预处理方法，利用CMST来消除基因簇之间的冗余，同时利用IG/SNR选取基因簇中的特征基因，消除噪声信息。结合了不同预处理方法在特征基因选择中的优点，在不同步骤中有效地消除冗余和噪音给研究带来的影响，从而提高癌症检测的准确性。

3.3.1 基于CMST的基因聚类方法

本节根据基因功能的相似性，结合集合论和图论中等价关系和等价类的理论基础，利用基因表达谱数据在基因聚类分析中提出一种最大相似树的基因聚类算法（Clustering Based on Most Similarity Tree，CMST）。

定义 3.1:（相似度量）设g_i和g_j分别为第i个基因和第j个基因的表达水平，则两基因g_i与g_j之间的相似度量（Similarity Measurement，SM）为：

$$SM(g_i,g_j)=\frac{\sum_{k=1}^{m}|g_{ik}-\overline{g}_i||g_{jk}-\overline{g}_j|}{\sqrt{\sum_{k=1}^{m}(g_{ik}-\overline{g}_i)^2\cdot\sum_{k=1}^{m}(g_{jk}-\overline{g}_j)^2}} \tag{3.5}$$

式中$\overline{g}_i=\frac{1}{m}\sum_{k=1}^{m}g_{ik}$，$\overline{g}_j=\frac{1}{m}\sum_{k=1}^{m}g_{jk}$，$g_{(\cdot k)}$为$g_{(\cdot)}$的第$k$个元素。

定义 3.2:（相似度量距阵）根据相似度量，我们可以建立基因组的相似度量距阵SMM（Similarity Measurement Matrix），

$$SMM(i,j)=SM(g_i,g_j) \tag{3.6}$$

SMM(i,j)是SMM的第i行第j列元素。

下面我们构造基因之间功能相似性的一种等价关系，并通过等价关系生成其代表元素的等价类对基因进行聚类。

定义 3.3:（等价关系）假设R是基因簇C上的等价关系，表示为：

$$R=\{<g_i,g_j>|g_i,g_j\in C\} \tag{3.7}$$

那么R具有下列三个充要条件。

定义 3.4: （自反性）对于C中的任意基因g_i，满足：

$$< g_i, g_i >\in R, if g_i \in C \tag{3.8}$$

定义 3.5: （对称性）对于C中的任意两个基因g_i和g_j，满足：

$$< g_j, g_i >\in R, if < g_i, g_j >\in R \wedge g_i, g_j \in C \tag{3.9}$$

定义 3.6: （传递性）对于C中的三个基因g_i，g_j和g_k，满足：

$$\begin{aligned} &< g_i, g_k >\in R, if < g_i, g_j >\in R \\ &\wedge < g_j, g_k >\in R \wedge g_i, g_j, g_k \in C \end{aligned} \tag{3.10}$$

令$G = \{g_1, g_2, ..., g_n\}$是一组基因，依照等价关系R，设$g_i$是基因簇中具代表性的基因，从基因集合G中可以获取等价类$[g_i]|R$。由等价关系的性质可知，等价类$[g_i]|R$可以将基因组进行划分产生基因簇。

下面我们通过建立最大相似树的方法来构造基因组G中满足上述三个条件的等价关系。最大相似树是根据相似度量距阵SMM 建立起来的带权图$T<v,e>$，其中，v是树中的结点，e是连接树中相邻结点的边。首先，初始化$T<v,e>$，$v = \phi$，$e = \phi$。然后从SMM中选择相似度最大的两个基因g_i和g_j加入v，并将$< g_i, g_j >$加入e，e的权重为$SMM(i,j)$，并且保证$T<v,e>$中不出现环。最后，删除SMM中元素$SMM(i,j)$和$SMM(j,i)$。重复上述步骤，直到$v = G$。

在$T<v,e>$中，我们定义基因之间功能的相似性如下：

定义 3.7: （相似性）给定相似性阀值λ，定义G中基因之间功能的相似性$R(\lambda, G)$，

$$R(\lambda, G) = < g_i, g_j > |PW(g_i, g_j) > \lambda \wedge g_i, g_j \in G \tag{3.11}$$

其中，

$$\begin{aligned} PW(g_i, g_j) = \ & min\{SMM(m,n)\}|g_m, g_n \in path\, of\, (g_i, g_j) \\ & \wedge g_m, g_n\, are\, neighbors \end{aligned} \tag{3.12}$$

不难证明，基因之间功能的相似性$R(\lambda, G)$满足自反性，对称性和传递性，因此$R(\lambda, G)$是基因簇中的一种等价关系。然后我们根据$[g_i]|R$对基因组进行划分，生成功能相似的基因簇，在每个基因簇中，任意两基因g_i和g_j满足$SM(g_i, g_j) > \lambda$。一个简单的聚类方法是删除最大相似树中权值小于λ的边，从而基因被划分成

具有相似性$R(\lambda, G)$的基因簇。

3.3.2 基于CMST的特征基因选择

本节结合CMST聚类方法和传统的特征选择法，在癌症检测中提出了一种分步选择最具辨别能力特征基因的方法。

首先，利用CMST聚类方法把基因分为功能基因簇，在簇内部的基因之间功能非常相似，而基因簇之间的基因相似性较小，减少基因簇之间的冗余；然后利用IG/SNR排序每个基因簇内部的基因，选择每个簇中最具辨别能力的特征基因。这些基因具有能最大程度分离不同样本类别的特征，最后这些挑选出来的基因被用来训练分类模型。由于该方法各步骤凝聚了不同数据预处理方法的长处，因此，利用它来进行特征基因选择具有最少数据冗余、不受噪声和不相关数据干扰的优点。

基于CMST特征基因选择算法详细步骤见算法3.1。

算法 3.1 基于CMST的特征基因选择算法

Require: $G = \{g_1, g_2, \ldots, g_n\}$，相似性阈值$\lambda$
Outputting: 特征基因FG
1: 根据基因之间相似度量方法，构造相似度量矩阵SMM，其中$SMM(i,j) = SM(g_i, g_j)$;
2: 令$FG = \phi$，初始化$T<v, e>$，$v = \phi$，$e = \varphi$;
3: **repeat**
4: 从SMM中选择相似度最大的两个基因g_i和g_j加入v，并将$< g_i, g_j >$加入e，e的权重为$SMM(i,j)$，保证$T<v, e>$中不出现环;
5: 删除SMM中元素$SMM(i,j)$和$SMM(j,i)$;
6: **until** $v = G$
7: 删除最大相似树$T<v, e>$中权值小于λ的边，生成具有相似性$R(\lambda, G)$的基因簇，C_i，设$1 \le i \le k$;
8: **for all** C_i **do**
9: 计算$IG(g_j)$或$SNR(g_j)$，$g_j \in C_i$;
10: 选择$g_{j'}$加入FG，其中$IG(g_{j'}) = Max(IG(g_j))$或$SNR(g_{j'}) = Max(SNR(g_j))$;
11: **end for**

3.4 癌症检测中基于OS-CMST的特征基因选择方法

在癌症检测中，特征基因的数量对样本的分类具有非常重要的意义。在本节，我们提出一种最优自适应CMST（Optimal Self-adaptive CMST，OS-CMST）基因聚类方法，在CMST中利用间隙统计量（Gap Statistic）[88]选择最优相似性阈值，生成最优的基因功能簇，然后，从基因功能簇中选择最具分辨能力的特征基因。

3.4.1 基于OS-CMST的基因聚类方法

假设根据CMST算法将基因组G 划分为k个基因簇$C_1, C_2, ..., C_k$，C_r代表第r个基因簇，且$n_r = |C_r|$。

定义 3.8: （簇内相似度）定义基因簇C_r的簇内相似度为C_r中不同基因对之间相似度量的平方之和，即

$$WS_r = \sum_{g_i, g_j \in C_r \wedge g_i \neq g_j} SM^2(g_i, g_j) \tag{3.13}$$

定义 3.9: （平均簇内相似度）定义基因组的平均簇内相似度为$AS(\lambda)$，

$$AS(\lambda) = \sum_{i=1}^{k} \frac{1}{n_r} WS_i \tag{3.14}$$

利用蒙托卡罗（Monte Carlo）法产生B组虚拟基因集$G_b^* = \{g_{b1}^*, g_{b2}^*, ..., g_{bn}^*\}$ $(1 \leq b \leq B)$，其中g_{bi}^*为在$[min(g_i), max(g_i)]$上均匀分布。引入Tibshirani 等提出的间隙统计量（Gap Statistic）[88]的方法，利用基因表达数据的统计规律提出基因的间隙统计量$Gap(\lambda)$，求解基因簇的阀值λ，寻找最优基因功能分簇。

定义 3.10: 对于相似度量λ，定义$Gap(\lambda)$

$$Gap(\lambda) = E^*(log(AS(\lambda))) - log(AS(\lambda)) \tag{3.15}$$

其中

$$\begin{aligned} E^*(log(AS(\lambda))) &= \frac{1}{B} \sum_b log(AS(\lambda)_b^*) \\ AS(\lambda)_b^* &= \sum_{i=1}^{k} \frac{1}{n_r} WS_i | g_i, g_j \in G_b^* \end{aligned} \tag{3.16}$$

然后利用公式3.17和公式3.18获取B个副本的标准方差和模拟偏差，并通过循环计算得到满足3.19式的最大的相似度量值，即为最优基因簇模式阀值λ。

$$sd(\lambda) = \sqrt{\frac{1}{B} \sum_b \left(log(AS(\lambda)_b^*) - \frac{1}{B} \sum_b log(AS(\lambda)_b^*) \right)^2} \tag{3.17}$$

$$s(\lambda) = sd(\lambda) \sqrt{1 + \frac{1}{B}} \tag{3.18}$$

$$Gap(\lambda_i) \geq Gap(\lambda_{i+1}) - s(\lambda_{i+1}) \tag{3.19}$$

3.4.2 癌症检测中基于OS-CMST的特征基因选择

本节结合OS-CMST聚类方法和传统的特征选择法，在癌症检测中提出了一种分步的具有自适应能力的选择最具辨别能力基因的方法。

首先，提出基因表达的间隙统计量$Gap(\lambda)$，在CMST在基础上提出一种自适应的OS-CMST基因聚类方法，将基因划分成若干最优的基因功能分簇。最大程度地去除基因簇之间的冗余；然后利用IG/SNR排序基因簇内的基因，选择簇中最具辨别能力的特征基因。这种特征基因选择方法在不需要输入参数的情况下，具有自适应能力挖掘基因组中最佳数量的特征基因。同时在特征基因子集中有效地消除了冗余数据和噪声数据对患者样本检测地影响。最后这些挑选出来的基因被用来训练分类模型。

基于OS-CMST的基因选择算法的详细描述见算法3.2。

3.5 实验结果与分析

本节所进行的实验分析分为两部分。第一部分为基因的聚类分析，分别利用CMST/OS-CMST算法在芽殖酵母数据集（Budding Yeast Dataset）[80]和酵母功能基因组数据集（Yeast Functional Genome）进行基因聚类[85]，并利用评价函数（Minskowski Sore）对划分的基因簇进行分析。第二部分为患者样本的的癌症检测，利用提出的基于CMST/OS-CMST的分步方法选择特征基因，然后利用这些特征基因建立一个感知器模型对Alizadeh等整理的弥漫性大B细胞淋巴瘤（Diffuse large B-cell lymphoma，称之为Alizadeh's dataset）进行癌症识别，并和其它方法的识别结果进行比较。实验硬件环境主要包括Intel P4 1.8G，256M内存，软件环境包括Windows 2000 Server，Matlab 6.5。

3.5.1 数据集

3.5.1.1 Budding Yeast Dataset

酵母菌通过孢子形成（Sporulation）繁殖方式将二倍体营养细胞（Diploid Cells）分裂成单倍体营养细胞（Haploid Cells）。Chu等发现在孢子形成过程中大约有500多个诱导基因（Induced Genes）[80]，它们的mRNA水平在孢子形成过程中产生了显著的变化。可以分为七种不同的基因模式（Temporal Patterns，TP），分别是Metabolic、Early I、Early II、Early-middle、Middle、Middle-late和Late。如表3.1所示，第一列是基因模式的名称，第二列是基因模式包含的基因数目。

算法 3.2 基于OS-CMST的特征基因选择算法

Require: $G = \{g_1, g_2, \ldots, g_n\}$
Outputting: 特征基因FG
1: 根据基因之间相似度量方法，构造相似度量矩阵SMM，其中$SMM(i,j) = SM(g_i, g_j)$;
2: 令$FG = \varphi$，初始化$T<v,e>$，$v = \varphi$，$e = \phi$;
3: **repeat**
4: 从SMM中选择相似度最大的两个基因g_i和g_j加入v，并将$< g_i, g_j >$加入e，e的权重为$SMM(i,j)$，保证$T<v,e>$中不出现环;
5: 删除SMM中元素$SMM(i,j)$和$SMM(j,i)$;
6: **until** $v = G$
7: 生成最大相似树$T < v,e >$，升序排列边权值;
8: 产生B组虚拟基因数据集$G_b^* = \{g_{b1}^*, g_{b2}^*, \ldots, g_{bm}^*\}$
9: 计算标准方差$sd(\lambda_1)$和计算模拟偏差$s(\lambda_1)$
10: **for** $\lambda_i = \lambda_2$ to λ_n **do**
11: 利用CMST算法，生成B组虚拟基因数据集具有平均簇内相似度$AS_b^*(\lambda_i)$的基因簇;
12: 计算标准方差$sd(\lambda_i)$和模拟偏差$s(\lambda_i)$
13: **if** $Gap(\lambda_{i-1}) \geq Gap(\lambda_i) - s(\lambda_i)$ **then**
14: return;
15: **end if**
16: **end for**
17: 删除最大相似树$T<v,e>$中权值小于λ_{i-1}的边，生成具有平均簇内相似度$AS(\lambda_{i-1})$的基因簇C_i，设$1 \leq i \leq k$;
18: **for all** C_i **do**
19: 计算$IG(g_j)$或$SNR(g_j)$，$g_j \in C_i$;
20: 选择$g_{j'}$加入FG，其中$IG(g_{j'}) = Max(IG(g_{j'}))$或$SNR(g_{j'}) = Max(SNR(g_{j'}))$;
21: **end for**

表3.1 Budding Yeast Dataset

TP	Induced Genes
Metabolic	52
Early I	62
Early II	47
Early-middle	95
Middle	158
Middle-late	61
Late	5

3.5.1.2 Yeast Functional Genome

在酵母功能基因组数据集中存在6221个基因[85]。在树状的功能组中，根据基因功能家族（Function Family）可以将分为不同的功能基因数据集。利用基因

功能家族，我们产生了6个功能基因数据集（C2,C3,C4,C5,C6,C7）。表3.2给出了功能基因数据集的产生方法以及基因簇中的基因数目。比如，数据集C3表示包含3个基因功能簇，分别为ATP synthesis、mitosis和vacuolar protein targeting，并且每个簇中包含19个基因。

表3.2 Yeast Functional Genome

Function Families	Genes	Cluster Sets
ATP synthesis	19	C3
mitosis	19	
vacuolar protein targeting	19	
silencing	20	C5
fatty acid metabolism	20	
meiosis	21	
phospholipid metabolism	21	
TCA cycle	22	
protein processing	27	C7
DNA repair	29	C6
protein folding	30	C4
nuclear protein targeting	31	
signaling	31	
major facilitator superfamily	32	C6
mRNA splicing	34	
chromatin structure	42	C2
DNA replication	42	

3.5.1.3 Alizadeh's Dataset[49]

弥漫性大B细胞淋巴癌（Diffuse Large B-Cell Lymphoma，DLBCL）是成人最常见的非霍奇金淋巴癌（Non-Hodgkin's Lymphoma）。分为生发中心来源（Germinal Centre B-like DLBCL，称为GC B-like DLBCL）和活化B细胞来源（Activated B-like DLBCL）两种淋巴癌症亚型。Alizadeh等根据患者淋巴细胞（Lymphocytes）的发展分为九种细胞株（Cell Line），如表3.3所示。

3.5.2 评价标准

明考斯基分值（Minkowski Score，MS）是一种聚类结果的衡量方法。在此，

表3.3 Alizadeh's Dataset

Cell Line No.	Cell Line	Samples
No.1	DLBCL	46
No.2	Germinal Centre B	2
No.3	Normal Lymph Node/Tonsil	2
No.4	Activated Blood B	10
No.5	Resting/Activated T	6
No.6	Transforme D Cell lines	6
No.7	Follicular lymphoma	9
No.8	Resting Blood B	4
No.9	Chronic Lymphocytic Leukemia	11
Total		96

聚类结果用$n \times n$矩阵C表示，其中

$$C_{ij} = \begin{cases} 1 & \text{if } g_i \text{ and } g_j \text{ in the same cluster} \\ 0 & \text{otherwise} \end{cases}$$

假设T是参照基因簇的矩阵表达形式，同矩阵C。则T和C的明考斯基分值用式3.21表示：

$$MS(T,C) = \frac{||T-C||}{||T||} \tag{3.21}$$

其中，

$$||T|| = \sqrt{\sum_i \sum_j |T_{ij}|}. \tag{3.22}$$

明考斯基值是矩阵T和C之间标准化的距离。从式3.21可知，如果MS越小，那么聚类产生的基因簇越接近参照基因簇，反之则与参照基因簇的差异越大。如果$MS=0$，那么生成的基因簇最优。

3.5.3 基于CMST的基因聚类

利用K-Means、SOM和CMST分别对Budding Yeast Dataset和Yeast Functional Genome两个数据集进行聚类。并对聚类产生的基因簇进行比较，然后利用Minkowski Sore评价各聚类方法。

在Budding Yeast Dataset基因数据集上K-Means和SOM的聚类数目设置为7，CMST上的阀值λ设置为0.56。表3.4给出了Budding Yeast Dataset数据集上K-Means、SOM和CMST聚类后产生的基因簇与参照基因簇的重叠基因数目。表3.5给出了在Budding Yeast Dataset数据集上K-Means、SOM和CMST聚类后生成

基因簇所获得的MS值。在Yeast Functional Genome的C2、C3、C4、C5、C6和C7数据集上K-Means和SOM的聚类数目分别设置为2、3、4、5、6、7，CMST上的阀值λ则分别设为0.25、0.42、0.26、0.38、0.48、0.52。表3.6则分别给出了Yeast Functional Genome中C2、C3、C4、C5、C6和C7数据集上K-Means、SOM和CMST聚类后生成基因簇所获得的MS值。从表3.4、表3.5和表3.6可以看出在两类数据的7个数据集（TCs、C2、C3、C4、C5、C6和C7）上，CMST都取得了优于K-Means和SOM的基因簇。

表3.4 在Budding Yeast Dataset数据集上的基因聚类结果比较

TC	Induced Genes	K-means	SOM	CMST
Metabolic	52	40	36	43
Early I	62	51	48	54
Early II	47	37	35	39
Early-middle	95	78	72	81
Middle	158	125	111	130
Middle-late	61	42	38	46
Late	5	2	1	4

表3.5 在Budding Yeast Dataset上的聚类结果（MS）

Evaluation	K-means	SOM	CMST
MS	1.165	1.335	1.022

表3.6 在Yeast Functional Genome上的聚类结果（MS）

Cluster Set	Average Number	K-means	SOM	CMST
C2	42	0.902	0.995	0.883
C3	19	0.890	0.931	0.772
C4	30.25	1.180	1.194	1.030
C5	20.8	1.207	1.241	1.014
C6	31.17	1.288	1.355	1.285
C7	30.57	1.326	1.301	1.255

3.5.4 基于CMST的癌症检测

应用基因选择算法3.1选择Alizadeh数据集上的特征基因，然后利用单层感知器模型[78]对Alizadeh数据集上的患者样本分类。单层感知器模型中输入层节点数

目为特征基因数目，单层感知器模型中输出层有9个节点，每个节点对应不同的细胞株。模型中的权重更新如式3.23：

$$\Delta w_{ij} = \eta * (D - Y) * x_i \tag{3.23}$$

式中Δw_{ij}为结点i到结点j之间权值的变化，η是学习率，设为0.5，D是期望输出，Y是实际输出，x_i是结点i的输入值。

这个实验包括两个部分。第一部分中，在算法3.1分步基因选择方法中第二步采用IG进行基因簇内基因选取，然后利用单层感知器模型识别患者样本。并与经SOTA预处理方法（见文献[78]）和IG预处理方法后的感知器模型识别结果进行比较。根据特征基因数目，我们分别进行了三次实验。其中，SOTA和IG中被选择的特征基因数量分别定为15、40、75，CMST则选择基因簇数量与SOTA/IG中被选择的特征基因数量接近。第二部分中，在算法3.1分步基因选择方法中第二步采用SNR进行基因簇内基因选取。并利用经分步预处理、SOTA预处理和IG预处理Alizadeh数据集后训练单层感知器模型。在SOTA/SNR中被选择基因数量定为12、35、72，同样CMST选择的基因簇数量与SOTA/SNR中被选择的特征基因数量接近。在两部分实验中，采用“留一交叉检验法”（Leave One Out Cross Validation，LOOCV）来识别癌症样本。即在Alizadeh数据集中选取一个样本作为测试数据，其余的样本作为训练数据集训练单层感知器模型，然后利用感知器模型来识别测试数据。如此重复选择测试数据，并且保证每次挑选的样本与前面的测试数据不同，直到Alizadeh中所有的样本都有一次机会被感知器模型作为测试数据识别。表3.7和表3.8给出了分类结果。

表3.7 经不同基因预处理（SOTA，IG，CMST）后的癌症识别结果

Cell Line	SOTA (15)	IG (15)	CMST (14)	SOTA (40)	IG (40)	CMST (38)	SOTA (75)	IG (75)	CMST (76)
No.1	41	39	43	42	40	45	42	40	43
No.2	1	1	2	2	1	2	2	1	2
No.3	1	1	1	1	1	2	1	1	1
No.4	8	6	8	9	8	10	8	6	9
No.5	5	5	6	4	5	6	5	4	6
No.6	5	5	5	5	4	5	4	5	5
No.7	8	7	8	9	8	8	8	6	8
No.8	3	3	3	3	3	3	3	2	4
No.9	9	8	10	10	8	9	9	7	9
Total	81	75	86	85	78	90	82	72	87

表3.8 经不同基因预处理（SOTA，SNR，CMST）后的癌症识别结果

Cell Line	SOTA (12)	SNR (12)	CMST (11)	SOTA (35)	SNR (35)	CMST (36)	SOTA (72)	SNR (72)	CMST (70)
No.1	40	37	42	41	39	45	41	38	44
No.2	1	1	2	2	1	2	1	2	2
No.3	2	1	1	1	2	2	1	1	1
No.4	8	6	9	9	7	9	7	8	8
No.5	4	6	5	5	5	6	5	4	5
No.6	4	4	5	5	4	5	5	4	5
No.7	8	7	8	9	7	8	8	4	7
No.8	3	3	4	3	4	4	2	2	4
No.9	9	7	9	8	8	10	10	8	10
Total	79	72	85	83	77	91	80	71	86

从表3.7和表3.8可以看出，利用基于CMST的特征选择方法进行基因选择后再分类的准确度要比基于SOTA，IG/SNR的分类准确度要高。主要原因在于：首先，在Alizadeh数据集具有4026个基因，其中包含了很多冗余信息，如果在整个数据集上运用IG/SNR方法进行基因选择，那么在选取的特征基因中存在有冗余信息，并且还有一些具有强辨别能力的基因被漏选和淹没；其次，数据集中同样包含了很多噪声和不相关数据，如果用SOTA来对数据进行预处理，利用簇的平均值作为特征基因，那么噪声基因降低了特征基因的辨别能力。从而影响IG、SNR和SOTA预处理后感知器对Alizadeh数据集中样本分类的准确性。而我们的方法分两步进行特征基因选取，结合了各种不同处理方法的优点，它能在降低冗余信息的同时消除噪音，所以它能提高癌症检测的准确性。

3.5.5 基于OS-CMST的基因聚类

本节利用OS-CMST分别对Budding Yeast Dataset和Yeast Functional Genome两个数据集进行聚类，验证OS-CMST是否可以自适应挖掘最优的基因簇模式。

OS-CMST在Budding Yeast Dataset基因数据集上的聚类结果如图3.2所示。图中圆圈旁的数字是该点对应的基因簇数目。从OS-CMST判断原则可知，OS-CMST在Budding Yeast Dataset基因数据集上挖掘的最优基因簇数目为6，如果我们忽略数据集中最小的基因模式Late（表3.4），则正好是6组基因表达模式。与Chu等人工挑选的基因簇数目一致[80]。同时发现潜在的最优基因簇数目为7。

OS-CMST在Yeast Function Genome数据集上的聚类结果如图3.3所示，图中子图C2、C3、C4、C5、C6、C7 分别是OS-CMST在C2、C3、C4、C5、C6和C7基

因数据上的聚类结果。图中圆圈旁的数字同样是该点对应的基因簇数目。从图3.3可知，OS-CMST在C2、C3、C4、C5、C6和C7基因数据上挖掘的最优基因簇数目分别2、4、2、5、6、7。除了在C3和C4数据集上OS-CMST生成的最优基因簇与参照基因簇存在偏差外，OS-CMST在其余五个数据集上生成了与参照基因簇一致数量的的基因簇。

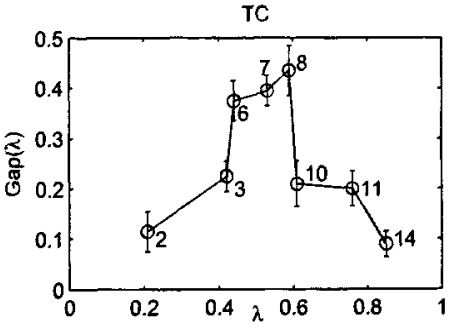

图3.2 OS-CMST在Budding yeast dataset上的实验结果（Gap Statistic）

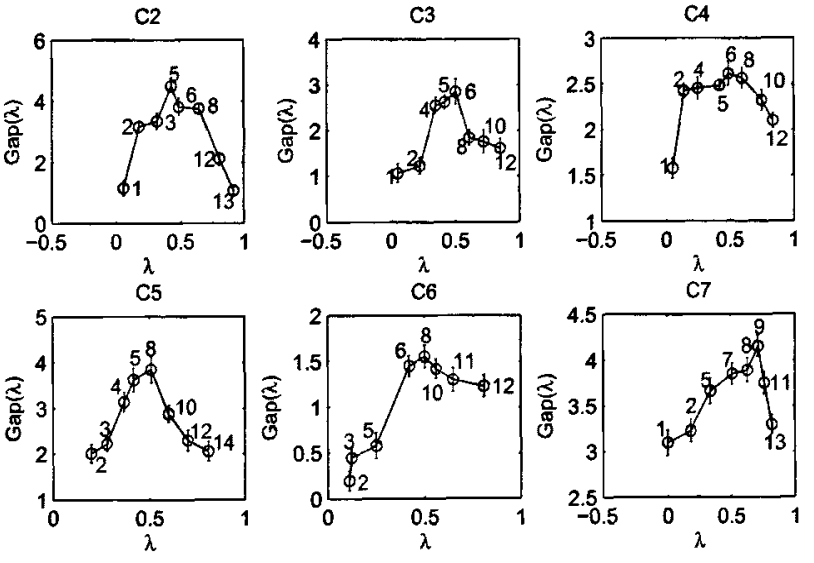

图3.3 OS-CMST在Yeast functional genomet上的实验结果（Gap Statistic）

3.5.6 基于OS-CMST的癌症检测

从表3.7和表3.8可以看出，癌症样本的分类准确度并不随特征基因数量的增

加而提高。当特征基因数量在30-70之中时，分类准确度达到峰值。到目前为止，特征基因数目是根据研究者分析数据之前根据经验值预先设置。在本节，我们利用OS-CMST选取Alizadeh数据集中最具辨别能力的特征基因子集，特征基因之间的相似性和冗余度都达到最小。从图3.4可以看出，最优的特征基因数量是46。文献[78]通过穷举搜索法找到在SOTA基因预处理中的44个最优基因模式。在此实验中，基于OS-CMST的特征基因选择法取得与穷举法一致的最具辨别性的特征基因，同时克服了穷举法需要大量时间消耗的不足。

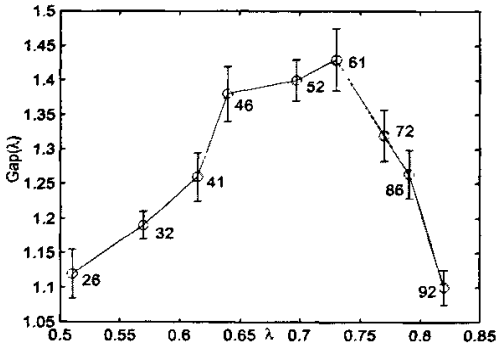

图3.4 OS-CMST在Alizadeh上的实验结果（Gap Statistic）

然后，利用IG/SNR方法在每个基因簇中选择最具辨别能力的基因作为单层感知器的输入特征。在此单层感知器模型的输入层有46个结点，输出层有9个结点。并利用LOOCV来检验测试结果。另外，经SOTA预处理后的44个特征基因模式同样用来被用来训练感知器模型，也利用LOOCV来检验测试结果。单独的IG/SNR作用在Alizadeh数据集上选取46个特征基因，然后训练感知器模型，并采用LOOCV的检验方法。表3.9 列出了上述实验的实验结果。利用OS-CMST和IG的分步预处理方法在96个样本中正确识别了94个样本，具有最高分类准确度。高于从整体数据集上单独利用IG或SNR筛选特征基因再识别癌症样本的准确度，也高于利用SOTA预处理后再识别癌症样本的准确度。利用SNR预处理后的癌症识别的准确率仅有82/96，但当结合OS-CMST后，大大提高了识别的准确率，达到了92/96。

表3.9 分别利用SNR、IG、SOTA和OS-CMST分步方法预处理后的分类结果

Total	SNR	IG	SOTA	OS-CMST (SNR)	OS-CMST (IG)
96	82	86	91	92	94

3.6 本章小结

在维数巨大的基因表达数据中有辨别能力的特征基因容易淹没在大量的噪声和冗余数据信息中。本章在癌症检测中提出了一种基于最大相似树聚类算法（CMST）的分步基因选择方法。首先通过CMST产生一组功能相似的基因簇，并且在不同基因簇中的基因之间冗余较小。然后从基因簇中选择最具分辨能力的特征基因，并将这些特征基因作为输入数据训练感知器模型，产生一个用于癌症检测的分类器。最后在CMST中，引入间隙统计量（Gap Statistic）选择最佳相似度阀值，提出了一种最优自适应的CMST（OS-CMST）算法，并通过OS-CMST选择特征基因进行癌症检测。由于在分步中结合了不同数据预处理方法的优点，实验结果显示，该特征基因选择方法可以有效的消除数据冗余和噪音，提高分类的准确性；基于OS-CMST的特征基因选择方法可以有效预测特征基因数目，并在无输入参数的情况选取了具有强分辨力的基因。

第4章 基于隐含变量模型的癌症分类算法

基因表达（Gene Expression）是指储存遗传信息的基因经过一系列步骤表现出其生物功能的整个过程。典型的基因表达是基因经过转录、翻译，产生有生物活性的蛋白质的过程。对这个过程的调节称为基因表达调控（Regulation of Gene Expression）。基因调控是现代分子生物学研究的中心课题之一。掌握了基因调控机制，就等于掌握了一把揭示生物学奥秘的钥匙。

在人类基因组中，基因表达之间存在相互影响，并且具有复杂的基因调控原理和机制。主分量分析方法（Primary Component Analysis，PCA）和独立分量分析方法（Independent Component Analysis，ICA）是基于统计的数据分析方法。本章利用PCA和ICA分别挖掘癌症表达数据中隐含的基因表达模式（EPCAP和EICAP），通过隐含的基因表达模式来分析不同基因之间的相互影响，以及整个基因组的调控机制，并研究隐含变量模型下的基因检测算法（CDHV），对患者样本进行癌症预测。实验结果表明，EPCAP和EICAP有效地描述了Budding Yeast Dataset中的基因表达模式，CDHV算法在Leukemia Dataset上取得很好的预测性能。

4.1 概述

基因组学和分子生物学研究的飞速发展使人们已经不再满足于孤立地研究单个基因，而是希望能同时研究多个基因及它们之间的相互关系。近年来基因芯片技术的发展产生了大量的大规模基因表达谱数据，让我们研究和分析复杂的基因调控原理和机制成为可能。目前，在癌症研究中基因调控分析以及基因之间的相互作用和相互影响已经成为这一领域的研究热点。

聚类技术广泛应用于基因表达谱数据分析，通过聚类将海量的基因表达数据划分成数量相对较少的基因簇，揭示具有生物意义的功能组。比如K-Means[44]、SOM[46,47,89?]、分层聚类[45,51]、双向聚类等[50]。特征变换法通过变换基因特征以抽取隐含的基因表达模式。Conde等[78]提出了一种基于聚类的特征基因抽取方法，利用SOTA层次聚类方法产生一系列组之间基因表达非冗余的基因簇，用簇的平均值作为基因特征以建立癌症识别模型。Raychaudhuri等[79]利用主分量分析方法（Principal Components Analysis，PCA）分析酵母细胞周期（Yeast Cell Cycle）数据[80]，发现2个主要分量可以压缩数据集中7个测试变量90%的信息量，3个主要分量可以压缩95%以上的信息量。Khan等[33]利用PCA压缩基因数据，结合人工神经网络研究儿童小圆形蓝色细胞恶性肿瘤（Small Round

Blue-Cell Tumors，SRBCTs）4种癌症亚型的识别问题。Wall[90]等利用奇异值分解（Singular Value Decomposition，SVD）分析基因表达数据，得到了与PCA相同的分析结果。Liebermeister等[81]利用独立分量分析（Independent Components Analysis，ICA）[91-94]抽取酵母细胞周期（Yeast Cell Cycle）中的25个基因表达模式和淋巴癌中12个基因表达模式。Hori等[82,83]利用ICA盲分离酵母细胞周期基因表达数据，取得比PCA更好的分离效果。

特征变换法可以有效压缩基因表达数据，但是特征变换法在压缩基因数据的同时也压缩了噪声和冗余数据，在癌症识别中同样导致有价值的基因特征信息被淹没。本章通过无放回取样选取基因子集，减少基因之间的冗余和噪声的影响，并利用PCA和ICA分别挖掘癌症表达数据中隐含的基因表达模式（PCAP和ICAP），构建基于隐含表达模式的基因表达模型，分析不同基因之间的相互影响，通过在不同子集之上的PCAP和ICAP重构基因表达模式EPCAP和EICAP，研究整个基因组的调控机制，并提出隐含变量模型下的癌症检测算法（CDHV），对患者样本进行癌症预测。实验结果表明，EPCAP和EICAP有效地描述了Budding Yeast Dataset中的基因表达模式，CDHV算法在Leukemia Dataset上排除了噪声和冗余的影响，取得很好的预测性能。

4.2 相关知识

4.2.1 PCA

主分量分析（Principal Component Analysis，PCA）是由Pearson[95]最早在1901年提出，并由Hotelling[96]于1933年加以发展的一种多元数据分析方法，它利用正交分解原理将一组相关变量转化成为另一组互不相关的综合变量（即主分量）。

设由p个随机变量组成一个p维随机向量$Y = (y_1, y_2, \ldots, y_p)^T$，记第i，j个分量$y_i$，$y_j$间的协方差为

$$\sigma_{ij} = Cov(y_i, y_j) \tag{4.1}$$

那么协方差矩阵$\sum$为:

$$\sum = (\sigma_{ij})_{p\times p} \tag{4.2}$$

则

$$U^{-1}\sum U = diag(\lambda_1, \lambda_2, \ldots, \lambda_p) \tag{4.3}$$

其中$\lambda_1, \lambda_2, \ldots, \lambda_p$是$\sum$的特征值，$U = (U_1, U_2, \ldots, U_p)$，$U_1, U_2, \ldots, U_p$是$\lambda_1, \lambda_2, \ldots, \lambda_p$

相应的标准正交的特征向量，则Y的第i个主分量是：

$$y_i' = U_i^T Y \quad (i = 1, 2, .., p) \tag{4.4}$$

4.2.2 ICA

独立分量分析（Indepedent Component Analysis，ICA）方法是近些年发展起来的一种高效盲信号分离方法（Blind Source Separation，BSS）[97,98]。它最早是用来解决“鸡尾酒会”问题的，现在在语音识别、人脸识别和医学信号处理等方面有着广泛的应用。ICA和PCA一样，同属多变量数据分析的线性方法。PCA根据信号数据的二阶统计特性去除源信号之间的相关特性，并未考虑信号数据的高阶统计特性，所以变换后的数据间仍有可能存在高阶冗余信息，实际上信号的高阶统计特性同样包含许多重要的特征信息。ICA是PCA的一种延伸，经ICA处理得到的各个分量不仅去除了相关性，并且分量之间相互统计独立的。

ICA模型可以用式4.5描述：

$$X = AS \tag{4.5}$$

式中$X = (x_1, x_2, \ldots, x_n)^T$，$S = (s_1, s_2, \ldots, s_m)^T$，其中$x_i (i = 1, 2, \ldots, n)$表示n个观测信号，每个观测信号是由m个信源信号$s_j (j = 1, 2, \ldots, m)$混合而成，$A_{n \times m}$为混合矩阵。独立分量分析研究的目的就是找出混合矩阵A或者分解矩阵W，使其满足下式：

$$\begin{aligned} I &= WX = WAS \\ A &= W^{-1} \end{aligned} \tag{4.6}$$

从矩阵分析角度讲，ICA的目的是寻找分解混阵W来实现多维观测信号的独立分量提取，一旦求得W，混合阵也就可以求出。分解矩阵W的求解过程是一个优化过程，需要建立一个描述分离结果独立程度的优化判据，称为“目标函数”，从信息论角度考虑，使输出分量独立统计则要求输出分量的互信息为零；然后设计优化算法，寻求W的最优解。这是ICA算法的两个核心问题。文中利用Hyvarinen给出的快速定点算法（FastICA）[91]来求分离矩阵W。

FastICA算法是基于负熵最大化推导出来的。负熵的定义如下：

$$\begin{aligned} J(x) &= H(x_{gauss}) - H(x) \\ H(x) &= -\int P(x) log(P(x)) dx \end{aligned} \tag{4.7}$$

其中x_{gauss}是与x具有相同协方差的高斯随机变量，$P(x_{gauss})$、$P(x)$是x_{gauss}和x的

概率密度函数。

对n维随机向量$S=(s_1,s_2,\ldots,s_n)^T$，联合概率密度函数为$P(S)$，互信息定义如下:

$$\begin{aligned} I(S) &= \int P(S)log\frac{P(S)}{\prod_{i=1}^{n}P(s_i)} \\ &= J(S)-\sum_{i=1}^{n}J(s_i) \end{aligned} \tag{4.8}$$

其中，$J(S)$是S的负熵，$P(s_i)$为分量s_i的边缘概率密度。

在ICA分析中，$S=WX$，由于负熵对于所有可逆线性变换保持不变，故$J(S)$为常数。FastICA的最终目的是互信息$I(S)$最小，如果互信息$I(S)$越小，则分离的源信号独立性就越强。由此问题进一步简化为：寻求一个适当的w_i，$s_i=(w_i)^TX$，最大化各自的边缘负熵，即使得负熵$J(s_i)$最大。

由上述知，FastICA思想中，W的求解问题转化为互信息的最小化问题，进一步简化为边缘负熵的最大化问题。在此给出基于高阶累计量的负熵估计：

$$J(x)\approx\frac{1}{2}k_3(x)^2+\frac{1}{48}k_4(x)^2 \tag{4.9}$$

式中$k_3(x)$和$k_4(x)$分别为x的3阶和4阶累积量（Kurtosis）

$$\begin{aligned} k_3(x) &= E(x^3) \\ k_4(x) &= E(x^4)-3(E(x^2))^2 \end{aligned} \tag{4.10}$$

4.3 癌症分析中的隐含变量模型

设基因表达矩阵X，X中的行是基因在不同实验样本中的表达水平，列表示不同实验环境下的实验样本，元素x_{ij}是第i个基因在第j个实验环境下的基因表达水平。本节通过对X的PCA和ICA分析，分别给出了基于PCA的癌症基因表达模型和基于ICA的癌症基因表达模型这两种癌症中的隐含变量表达模型。

4.3.1 基于PCA的癌症基因表达模型（PCAE）

在基因表达矩阵X中存在m个基因和n个癌症样本，设$X=(S_1,S_2,\ldots,Sn)$。我们将样本$S_i(1\leq i\leq n)$看作是一个m维的随机向量，$S_i=(s_{i1},s_{i2},\ldots,s_{im})^T$。根

据PCA分析，利用式4.1-4.4求得S_i的m个主分量，$s'_{i1}, s'_{i2}, \ldots, s'_{im}$。由式4.4可知，

$$\begin{aligned} S_i = & (U_1, U_2, \ldots, U_m)(s'_{i1}, s'_{i2}, \ldots, s'_{im})^T \\ = & U(s'_{i1}, s'_{i2}, \ldots, s'_{im})^T \end{aligned} \tag{4.11}$$

在此，设第i个主分量的贡献率α_i，

$$\alpha_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} \tag{4.12}$$

将贡献率将主分量降序排列，前q个主分量的累积贡献率为α，

$$\alpha = \sum_{j=1}^{q} \alpha_j \tag{4.13}$$

在主分量个数的确定时，我们设定$\alpha \geq 0.85$为累积贡献率阀值，并尽量减少主分量的个数。

假设选定了满足条件的k个主分量，即利用PCA挖掘出的K个隐含变量，这K个隐含变量通过各自的影响系数来调控基因表达，我们通过下面的公式来描述这种基于PCA的癌症基因表达模型（PCAE）：

$$\begin{aligned} s''_{i1} &= s'_{i1}u_{11} + s'_{i2}u_{21} + \cdots + s'_{ik}u_{k1} \\ s''_{i2} &= s'_{i1}u_{12} + s'_{i2}u_{22} + \cdots + s'_{ik}u_{k2} \\ &\cdots\cdots \\ s''_{ik} &= s'_{i1}u_{1k} + s'_{i2}u_{2k} + \cdots + s'_{ik}u_{kk} \end{aligned} \tag{4.14}$$

那么，

$$s''_{ij} = \sum_{t=1}^{k} s'_{it}u_{tj} \tag{4.15}$$

式中$s'_{i1}, s'_{i2}, \ldots, s'_{ik}$为样本S中的隐含变量，控制基因在样本中的表达。$U'_t = (u_{t1}, u_{t2}, \ldots, u_{tk})$是隐含变量$s'_{it}$在样本S中的影响系数，调控隐含变量$s'_{it}$在样本中的表达。$s''_{ij}$则是隐含变量在样本中的表达，是$s'_{it}$在其影响系数$u_{tj}$的综合调控下的表达。

4.3.2 基于ICA的癌症基因表达模型（ICAE）

ICA是从盲源信号分离技术发展而来的数据分析方法，根据信号数据的高阶统计特性分离出统计独立的源信号。假设在样本数据的基因表达中存在k个独立的基因源信号，源信号的向量形式为$S=(s_1,s_2,\ldots,s_k)^T$，源信号之间满足：

$$P(S)=\prod_{i=1}^{k}P(s_i) \tag{4.16}$$

在基因表达矩阵中有m个观测基因信号$G=(g_1,g_2,\ldots,g_m)^T$，这些信号是由源信号混合而成，设混合矩阵为$A_{m\times k}$，则有：

$$G=AS \tag{4.17}$$

我们利用ICA分析癌症组织样本中基于独立分量的基因表达的混合模型，如图4.1所示，其中G为癌症样本组成的基因表达矩阵，列为样本中观测基因的表达，行为观测基因在样本中的表达；S为癌症样本的独立分量（ICs），列对应为癌症样本，行对应为独立分量。

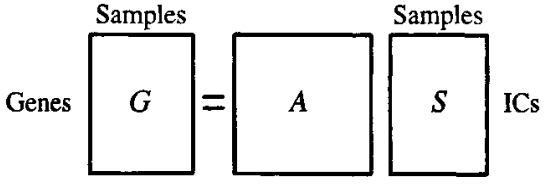

图4.1 癌症组织中基因表达的混合模型

基因观测信号是基因源信号的混合，信号之间不具有独立性，

$$P(G)\neq\prod_{i=1}^{m}P(g_i) \tag{4.18}$$

利用4.2.2节的FastICA算法，求得分解矩阵$W_{k\times m}$，

$$S'=WG \tag{4.19}$$

观测基因变量在分解矩阵$W_{k\times m}$作用下产生源变量S'，$S'=(s'_1,s'_2,\ldots,s'_k)$，通过观测基因变量的高阶统计特性使得$s'_i$之间尽可能独立。癌症组织中基因表达

的解混模型如图4.2所示。

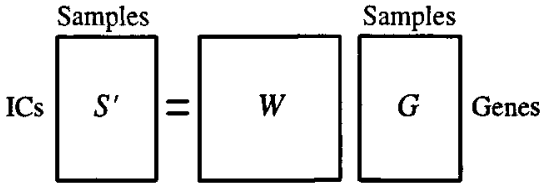

图4.2 癌症组织中基因表达的解混模型

为了进一步描述基因表达的混合模型，我们将混合模型表述成如下的向量形式。

$$G_i = \sum_j s_{ij} A_j \tag{4.20}$$

其中G_i是第i例癌症样本，A_j是A的第j列向量，s_{ij}是第i例癌症样本的第j个独立分量。假设A_j是基因表达谱中第j个隐含模式，关系式4.20将基因谱表示为隐含模式的一个线性组合，称之为基于ICA隐含变量的基因表达模型（ICAE）（图4.3）。

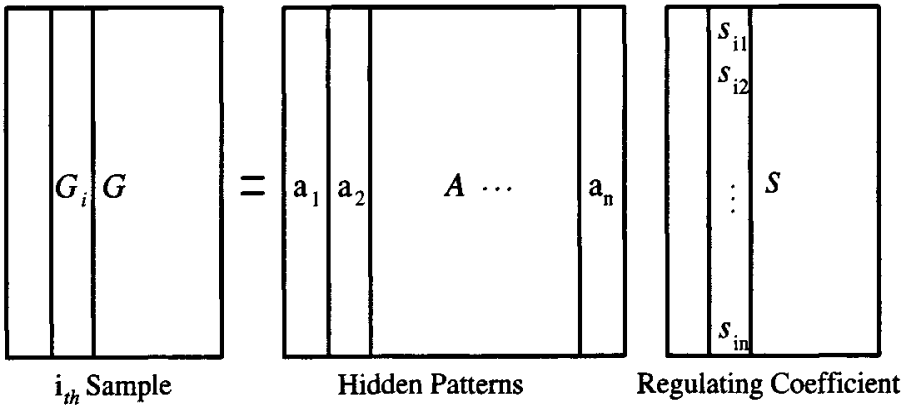

图4.3 基于ICA隐含变量的基因表达模型

在癌症基因表达中假设存在k种遗传控制因子，第i个隐含模式可看作是仅由第i 个控制因子贡献的模式。混合矩阵A 的行和列分别对应癌症样本中基因和控制因子，a_{jk}表示第j个基因中第k个控制因子的数量。另一方面，矩阵S 的行和列分别对应控制因子和样本，元素s_{ik}表示第k个癌症样本中的第i个控制因子的调控

系数。基于ICA隐含变量的基因表达模型可以理解为基因表达谱是基因隐含变量与相对应调控系数贡献的线性组合。

我们利用另外的形式来表示表达模型4.20，

$$g_{ij} = \sum_{k} a_{ik} a_{kj} \tag{4.21}$$

那么，在第i个样本的第j个基因的表达即元素g_{ij}为遗传控制因子a_{ik}乘以其对应的贡献程度s_{kj}的总和。

4.4 基于隐含变量模型的癌症检测（CDHV）

4.4.1 分类器

4.4.1.1 支持向量机

支持向量机（Support Vector Machine，SVM）是由Vapnik提出的一种基于结构风险最小化原则的模式分类学习算法，通过结构风险最小化理论建构最优分割超平面，每一类数据与超平面距离最近的向量与超平面之间的距离最大。

设$X = \{x_i, i = 1,2,\ldots,n\}$是样本集，$c = (c_1, c_2, \ldots, c_n)$为类别向量，$c_i = \{1, -1\}$，最优分类函数为：

$$L(x_i) = \sum_{j=1}^{n} c_j \alpha_j K(x_i, x_j) + b \tag{4.22}$$

其中，$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$是Langrange乘子，核函数$K(x_i, x_j)$采用高斯核函数，

$$K(x_i, x_j) = exp(-\|x_i - x_j\|^2/(2\sigma^2)) \tag{4.23}$$

4.4.1.2 KNN

K近邻分类算法（K-Nearest Neighborhood，KNN）是一种懒散的分类算法。假设每个样本代表空间中的一个点，给定一个未知样本，KNN搜索模式空间，找出最接近未知样本的k个训练样本。这k个样本是未知样本的k 个“近邻”。未知样本被分配到k个最近邻中最公共的类[55, 60, 61]。“近邻性”采用余弦距离度量，设v_i和v_j两个样本的基因表达向量，余弦距离为

$$\cos(v_i, v_j) = \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|} \tag{4.24}$$

其中，$v_i \cdot v_j$是向量v_i和v_j的内积，$\|v\|$是向量v的模。

4.4.2 性能评估

每个数据集划分为正例样本（Positives）和负例样本（Negatives），在分类问题中普遍使用的性能评估指标有准确度（Accu）和精确度（Prec），定义如下：

$$Accu = \frac{N_{true\,positives}}{N_{true\,positives} + N_{false\,positives}}$$

$$Prec = \frac{N_{true\,positives}}{N_{true\,positives} + N_{false\,negtives}}$$

其中，$N_{true\,positives}$为识别成正例样本中分类正确的样本数，$N_{false\,positives}$为识别成正例样本中分类错误的样本数，$N_{false\,negtives}$为识别成负例样本中分类错误的样本数。准确度是所有判断的样本中与标识结果吻合的样本所占的比率；精确度是标识结果应有的样本中与分类系统吻合的样本所占的比率。

4.4.3 基于隐含变量模型的癌症检测（CDHV）

主分量分析和独立分量分析可以挖掘基因表达谱中的隐含变量，以揭示基于隐含变量的基因表达模型。主分量分析和独立分量分析通过将原变量进行转换，使少数几个新变量是原变量的线性组合，排除众多信息共存中相互重叠的信息，将数据降维，基于隐含变量的基因表达模型可以有效减少基因表达谱中的冗余；同时，新变量尽可能多地表征原变量的数据结构特征而不丢失信息。但是，基因表达数据是一种高噪声数据，过多地保留原始基因信息不利于消除噪声，从而影响癌症检测效果。本节通过无放回抽样方法生成基因子集，减少噪声，并提出一种基于隐含变量模型的癌症检测选方法（Cancer Detection with Hidden Variable，CDHV）。步骤如下：

STEP 1.（**初步处理**）对每个基因在每个样本上的表达值中心化，缺失数据补零，然后再归一化。然后利用信噪比（式4.25）进行基因的基本筛选，去除干扰基因。

$$SNR(g) = \frac{\mu_1(g) - \mu_2(g)}{\sigma_1(g) + \sigma_2(g)} \tag{4.25}$$

其中，$\mu_1(g)$和$\mu_2(g)$分别为基因g在$class_1$和$class_2$上表达的平均值，σ_1、σ_2 为对应的方差。

STEP 2.（**聚类**）利用K-means对STEP 1. 初步筛选的基因进行聚类，划分成功

能基因簇$GC_1, GC_2, \ldots, GC_k$。

STEP 3.（**生成基因子集**）假设生成的基因子集数为m，簇GC_i中基因数目为$|GC_i|$，通过无放回抽样方法产生基因子集$GS_j(1 \leq j \leq m)$，我们进行m轮无放回抽样，每轮随机抽样从GC_i中抽取$|GC_i|/m$个基因，获得GCS_i，设$GS_j = GCS_1 \cup GCS_2 \cup \cdots \cup GCS_k$。

STEP 4.（**建立基因表达模型**）根据基因子集$GS_j(1 \leq j \leq m)$构造基因表达矩阵XS_j，利用PCA或ICA分析基因表达矩阵XS_j，生成隐含的基因表达模型$PCAE_j$或$ICAE_j$。

STEP 5.（**重构基因表达**）根据隐含模式的相似性来重构基因表达，在$PCAE_j$中，排序后λ_l对应的U_l获取的$s'_j(l)$，

$$s'_j(l) = U_l^T(XS_j) \tag{4.26}$$

则重构基因表达，

$$s''(l) = \frac{1}{m}\sum_{j=1}^{m} s'_j(l) \tag{4.27}$$

在$ICAE_j$中，隐含模式a_l对应的调控系数为$s'_j(l)$，然后利用不同$ICAE_j$中隐含模式的皮尔逊相关系数选择最相似关联隐含模式并根据式4.27来重构基因表达。

STEP 6.（**癌症识别**）利用重构后的基因表达来建立分类器（SVM或KNN），通过癌症识别策略训练分类器，并识别未知癌症样本。

STEP 7.（**性能分析**）分别计算准确度Accu和精确度Prec。

4.5 实验及分析

4.5.1 数据集

4.5.1.1 Budding Yeast Dataset

酵母菌通过孢子形成（Sporulation）繁殖方式将二倍体营养细胞（Diploid Cells）分裂成单倍体营养细胞（Haploid Cells）[80]。Chu等发现在孢子形成过程中大约有500多个诱导基因（Induced Genes），它们的mRNA水平在孢子形成过程中产生了显著的变化。可以分为七种不同的基因模式（Temporal Patterns，TP），分别是Metabolic、Early I、Early II、Early-middle、Middle、Middle-late和Late。

4.5.1.2 Leukemia Dataset

急性白血病基因表达谱数据集（Leukemia Dataset）[6]。共有72例急性白血病样本，每例样本均含7129个基因的表达数据。其中47例样本被诊断为急性淋巴性白血病（Acute Lymphoblastic Leukemia，ALL），25例被诊断为急性骨髓性白血病（Acute Myeloid Leukemia，AML）。

4.5.2 实验结果与分析

本节在Budding Yeast Dataset和Leukemia Dataset分别进行实验。虽然Budding Yeast Dataset不是癌症基因表达数据集，但是基因表达原理和癌症基因表达相似，同样可以利用PCAE和ICAE进行分析。实验环境同3.5。

Budding Yeast Dataset包含酵母基因组中6118个基因表达数据，这些数据是在酵母发芽过程中0.0、0.5、2.0、5.0、7.0、9.0和11.5小时7个时间点获得。Chu等将该数据集中的酵母基因分为7个类[80]，手工选择了每个类中具有代表意义的基因，获得了7个类别的平均表达谱（图4.4）。在Budding Yeast Dataset上，利用4.4.3节的STEP 1～STEP 5来分析基因表达模式，设利用PCAE在GC_i中生成的对应基因模式为$U_i(l)$，则EPCAP为平均基因表达模式U_l，令$U_l = \frac{1}{m}\sum_i U_i(l)$，其中$1 \le l \le 7$。同样，利用ICAE在$GC_i$中生成的对应基因模式为$a_i(l)$，则EICAP为平均基因表达模式$a_l$，令$a_l = \frac{1}{m}\sum_i a_i(l)$，其中$1 \le l \le 7$。然后分别由PCAE和ICAE生成Budding Yeast Dataset的基因模式PCAP和ICAP，参见文献[82]。图4.5－图4.8分别描述了这些基因模式。并进行比较。

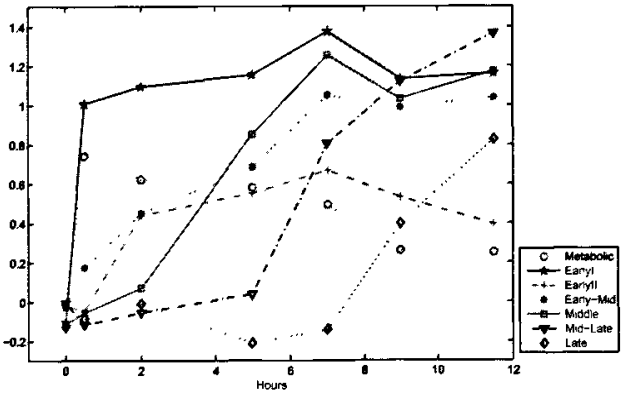

图4.4 在Yeast中七类基因的平均表达谱

从图4.4和图4.5可以看出，在基因表达模式ICAP中，虽然ICAP1，ICAP3，ICAP4可以很好地反映基因表达谱Middle，Early I，Early II。

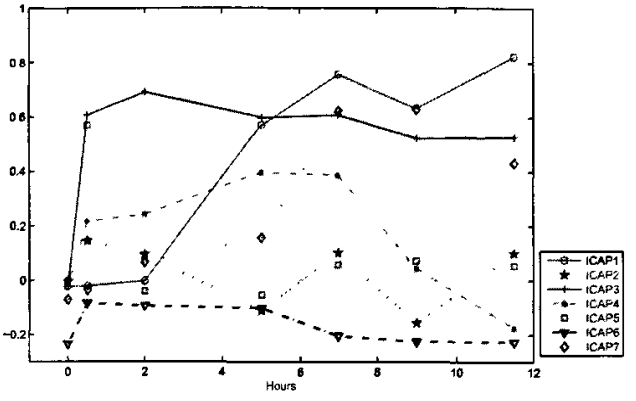

图4.5　在Yeast中ICAE模型的基因表达模式ICAP

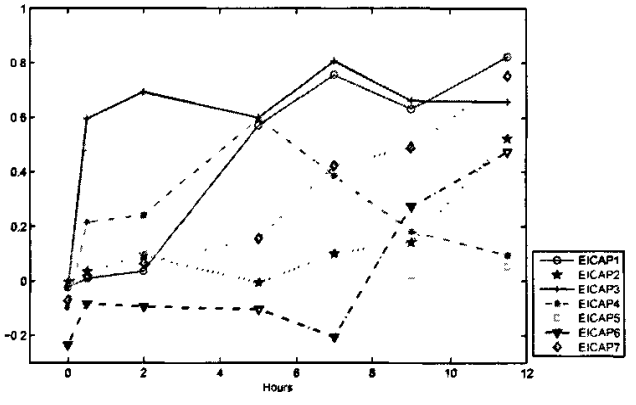

图4.6　在Yeast中的基因表达模式EICAP

但在其余4个模式中则相差比较大。从图4.4和图4.6中可以算出，在EICAP中，不但发现了Middle，Early I，Early II三个基因模式，也可以较好地反映Early-Mid模式，其余3个基因模式也在一定程度上得以体现。

从图4.7和图4.8可以看出，模式PCAP1和EPCAP1 分别表示了酵母菌芽殖过程中的基因表达的平均发展水平，PCAP2和PCAP3以及EPCA2和EPCA3 都表现了基因表达的基因调控的两个不同方面，但EPCA2和EPCA3在Mid-Late、Late上更具有辨识性。

在Leukemia Dataset上，利用基于隐含变量模型的癌症检测算法（CDHV）识别癌症样本。在CDHV中，分类器分别采用SVM和KNN，SVM采用径向基

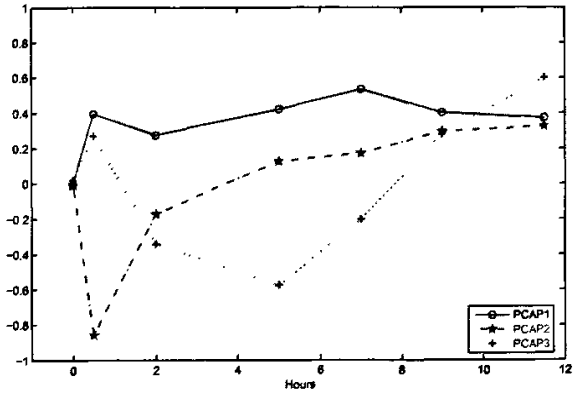

图4.7 在Yeast中PCAE模型的基因表达模式PCAP

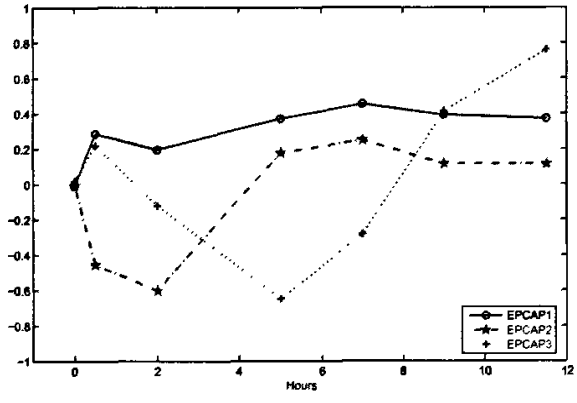

图4.8 在Yeast中的基因表达模式EPCAP

函数（RBF）作为核函数，KNN相似性度量函数采用Pearson相关系数，K = 5。用CDHV（PCAE）和CDHV（ICAE）分别表示基于PCAE和ICAE基因表达的CDHV癌症检测算法。在CDHV中，利用SNR初步处理后保留4000个基因，K-means的基因聚类数设为20，基因子集数目设为10，隐含变量的数目设为15。检测方法采用“留一交叉检验法”（Leave-One-Out Cross Validation，LOOCV），每次从数据集中挑选一个不同的样本作为测试样本，其余样本作为训练数据集训练CDHV模型。重复该过程，直到每一个样本作为测试样本时为止。统计所有被正确识别的样本，分别计算ALL和AML的分类性能评价指标Accu、Prec。上述分类实验重复10次，计算平均性能。同样，利用PCA和ICA预处理基因表达

数据，生成PCAE模型和ICAE模型，然后识别癌症样本。初步处理和检测方法与CDHV相同。利用SNR初步处理后保留1000个基因，隐含变量的数目设为15。表4.1和表4.2给出了实验结果。

表4.1 LOOCV测试实验结果（SVM）

	Accu%		Prec%	
	ALL	AML	ALL	AML
PCAE	78.6	72.5	70.3	73.1
ICAE	90.7	86.4	80.4	76.8
CDHV(PCAE)	86.1	85.2	81.3	82.2
CDHV(ICAE)	96.5	94.7	95.8	92.3

表4.2 LOOCV测试实验结果（KNN）

	Accu%		Prec%	
	ALL	AML	ALL	AML
PCAE	72.2	68.1	71.3	70.7
ICAE	80.3	81.4	82.2	79.8
CDHV(PCAE)	85.6	82.4	92.1	90.1
CDHV(ICAE)	94.2	93.9	92.8	90.5

从表4.1 和表4.2可以看出，基于PCAE模型的癌症检测准确度和精确度偏低，KNN分类器在AML上的精确度只有68.1%，主要原因是PCAE模型没有很好地消除噪声信息，噪声信息严重影响了癌症识别效果。ICAE模型从高阶统计量上消除噪声信息，提高了癌症检测准确度和精确度。并在SVM分类器检测癌症的准确度超过CDHV（PCAE）检测的准确度。CDHV 通过抽样方法减少噪声对模型的干扰，CDHV(PCAE)、CDHV(ICAE)在SVM和KNN分类器都取得了很好的准确度和精确度。

4.6 本章小结

本章利用主分量分析方法和独立分量分析方法分别构建基因数据中的隐含变量表达模型，揭示不同基因之间的相互影响和整个基因组的调控机制，并利用抽样方法消除在隐含变量表达模型中的噪声和冗余，提出了一种基于隐含变量模型下的基因检测算法（CDHV）。实验结果表明，EPCAP和EICAP有效地描述了Budding Yeast Dataset中的基因表达模式，在对患者样本进行癌症预测中，CDHV算法在Leukemia Dataset上取得很好的预测性能。

第5章 基于关联空间的基因特征抽取与癌症识别研究

利用微阵列基因表达谱可以识别患者样本的癌症类型，有利于针对患者制定有效的治疗措施。研究高维、高相关和高噪基因表达谱数据中致癌因子存在的局部相关性，利用基因特征的变换法构建癌症组的关联空间，提取使得癌症组具有最小组能量的最小扩展空间，提出一种基于最小扩展空间的癌症分类算法（Cancer Classification with Least Spread Space，CCLSS）。理论证明在最小扩展空间上可以有效识别癌症模式。在急性白血病和结肠癌数据集上进行实验，探讨最小扩展空间的秩与癌症识别率的关系，然后分别选取最佳辨别AML和ALL的最小扩展空间，结肠癌组织与正常结肠组织的最小扩展空间，利用CCLSS通过LOOCV实验方法进行癌症识别，结果表明CCLSS上比传统算法具有更好的分类精确度。

5.1 概述

近年来，基因微阵列（Microarray）技术的发展使得研究人员可以在同一实验中获得成千上万个基因的表达水平（Expression Level），产生了大规模基因表达谱数据，对癌症诊断及治疗的研究具有非常重要的意义。

由于非特异性杂交等原因，基因表达谱数据中含有较大的实验误差。同时，样本数目一般为几十或上百例，而检测基因的数目往往高达几千甚至几万。另一方面，由于功能相似的基因的表达谱高度相关，在大规模的基因表达数据中存在大量在分类学意义上的冗余基因。因此，利用微阵列基因表达谱数据进行癌症识别是典型的高维、高相关（冗余）、高噪问题，导致维数灾难（Curse of Dimensionality）[55, 99]。如图5.1所示，在I维空间中，单位长度内分布的样本密集；在II维空间中，单位面积内分布的样本较适中；然而在III维空间中，单位容积内分布的样本很稀疏。因此，在样本维数非常高的基因数据集中，有限的对象被分散到高维空间，造成对象（样本）的分布非常稀疏，基因特征的扰动对识别样本类型没有任何影响[100]。

降维方法经常被利用来处理高维、高噪问题[101]。对于癌症检测与分类，一类主要的降维方法是特征选择法。文献[6]以信噪比（Signal to Noise Ratio，SNR）为基础的邻居分析法（Neighborhood Analysis）选取了50个急性髓细胞白血病（Acute Myeloid Leukemia，AML）与急性淋巴细胞白血病（Acute Lym-

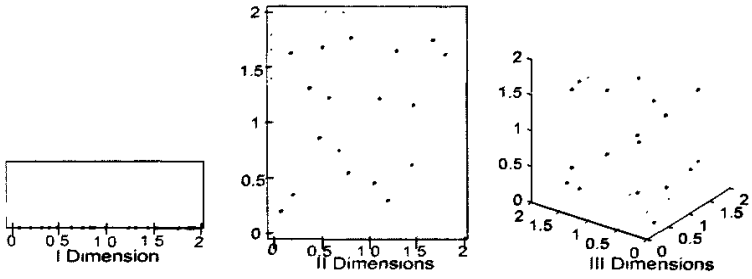

图5.1 样本在I维，II维和III维空间下的分布情况比较

phoblastic Leukemia，ALL）最具代表意义的基因。文献[54]利用排秩法（Rank）提取70个基因作为乳腺癌（Breast Cancer）的分类指标。文献[71]研究了相关系数法（Correlation Coefficient）、信息增益（Information Gain）和互信息（Mutual Information）在不同癌症数据集上的基因选择问题。另一类降维方法是特征变换法。Conde等利用聚类方法对基因进行聚类[75, 85]，然后用簇（Cluster）的平均值作为基因特征训练检测癌症的感知器模型[78]。Khan等结合主分量分析法和人工神经网络，研究儿童小圆形蓝色细胞恶性肿瘤（Small Round Blue-Cell Tumors，SRBCTs）的4种亚型的癌症识别[15]。但是目前的癌症分类模型都是利用统一的基因特征建立分类器，没有充分考虑在生物意义上致癌因子存在局部相关性，即不同癌症组具有不同的致癌因子，某一组癌症与一些致癌因子的表达非常相关，而另一组癌症则与另外一些致癌因子的表达非常相关。如图5.2所示，数据集合中有两个癌症模式P、Q，癌症P中样本在x-y平面图中相互邻近，癌症Q中样本在x-z平面图中相互邻近，但在x-y平面图中并不能发现癌症Q，在x-z平面图中也不能发现癌症P。同时，在所有维空间中也不能发现这两个模式。

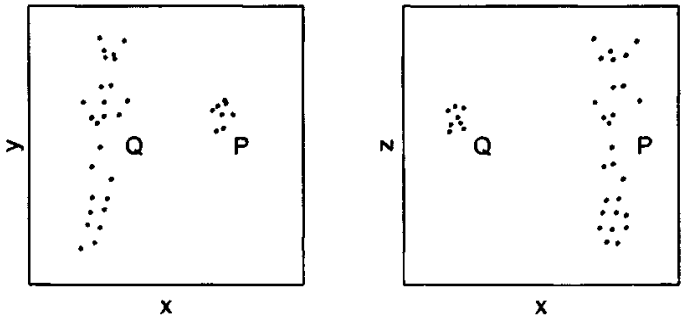

图5.2 癌症模式P和Q中致癌因子的局部相关性

本章研究高维、高相关和高噪基因表达谱数据中致癌因子存在的局部相关

性，利用特征变换法揭示每一组癌症表达数据中隐含变量的空间结构，并提取使得癌症组具有最小组能量的最小扩展空间，理论证明在最小扩展空间上可以有效识别癌症模式。每组癌症的最小扩展空间揭示出隐含的致癌因子和基因在癌症组中的表达，从癌症病理上有效地排除噪声和冗余的干扰。然后提出一种基于最小扩展空间的癌症分类算法，不同于传统的建立在统一基因特征子集之上的分类模型，基于最小扩展空间的分类模型中每组癌症都具有各自相关的基因特征，不同癌症组的分类器建立在不同的基因特征子集之上。最后在急性白血病和结肠癌数据集上进行实验，探讨最小扩展空间的秩与癌症识别率的关系，并分别选取最佳识别AML和ALL，结肠癌组织和正常结肠组织的最小扩展空间，利用CCLSS采用LOOCV实验方法进行癌症识别。

5.2 预备知识

5.2.1 相关定义

设$\widetilde{X}$是一组对象集合，$\widetilde{X}=\{\bar{x}_1,\bar{x}_2,\cdots,\bar{x}_n\}$，$\bar{x}_i=(x_{i1},x_{i2},\cdots,x_{im})^T$，$X$是由集合$\widetilde{X}$中对象组成的矩阵，见式5.1。

$$X=(\bar{x}_1,\bar{x}_2,\cdots,\bar{x}_n)=\begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1m} & x_{2m} & \cdots & x_{nm} \end{pmatrix} \tag{5.1}$$

定义 5.1: （质心）对于集合$\widetilde{X}$，$M(\widetilde{X})$表示$\widetilde{X}$的质心，$M(\widetilde{X})=\frac{1}{n}\sum_{i=1}^{n}\bar{x}_i=\frac{1}{n}(\sum_{i=1}^{n}x_{i1},\sum_{i=1}^{n}x_{i2},\cdots,\sum_{i=1}^{n}x_{im})^T$

定义 5.2: （迹）如果X是方阵，即$m=n$时，$TR(X)$表示X的迹，$TR(X)=\sum_{i=1}x_{ii}$

5.3 基于关联空间的特征抽取和癌症识别

传统的利用微阵列基因表达谱的分类方法通过降维法来排除“维数灾难”（图5.1）的影响，但是都没有考虑在病理上致癌因子存在的局部相关性（图5.2）。本节通过基因特征变换法构造癌症组的关联空间，以揭示基因特征与癌症之间的相关性以及基因在癌症样本中的表达，并提取具有最小癌症组能量

的最小扩展空间以压缩微阵列基因表达谱数据，提出一种基于最小扩展空间的癌症分类算法。

5.3.1 基于关联空间/最小扩展空间的特征抽取

假设基因微阵列表达谱中n个样本属于k种不同的癌症，第i类癌症样本集合是$\widetilde{C}_i$，$\widetilde{C}_i$的样本数目为n_i，$\sum_i n_i = n$。C_i是$\widetilde{C}_i$中样本的基因表达矩阵，则C_i为m行n_i列的矩阵。基于关联空间/最小扩展空间的特征抽取方法是首先通过本节提出的方法来获取$\widetilde{C}_i$的关联空间ε或最小扩展空间$\hat{\varepsilon}$，然后将$\widetilde{C}_i$中的样本在$\varepsilon/\hat{\varepsilon}$的方向上映射以抽取样本中有意义的基因特征。基于最小扩展空间的特征抽取可以有效地压缩癌症样本的维数。我们通过以下方法来构造$\widetilde{C}_i$的关联空间ε或最小扩展空间$\hat{\varepsilon}$。对C_i进行转置变换获得C_i^T，设$C_i^T = (\bar{g}_1, \bar{g}_2, \cdots, \bar{g}_m)$，其中$\bar{g}_j$是第$j$个基因在$\widetilde{C}_i$样本中的基因表达向量。$C_i^T$的协方差矩阵为$Cov(C_i^T)$，$Cov(C_i^T)$是半正定的$m$维方阵，可以进行如下矩阵分解：

$$Cov(C_i^T) = \sum \lambda_r \bar{p}_r \bar{p}_r^T = P \wedge P^T \tag{5.2}$$

其中，λ_r是$Cov(C_i^T)$特征值，$\wedge$是非负的对角矩阵，对角线上元素由$\lambda_r (1 \leq r \leq m)$组成，$\bar{p}_r$是$\lambda_r$对应的特征向量，$P = (\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_m)$。

定义 5.3：（关联空间）设癌症$\widetilde{C}_i$和表达矩阵C_i，$\lambda_1, \lambda_2, \cdots, \lambda_m$是$Cov(C_i^T)$的特征值，$\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_m$是$\lambda_1, \lambda_2, \cdots, \lambda_m$对应的特征向量。对于$d \leq m$，则由$\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_d$组成了$\widetilde{C}_i$上秩为$d$的关联空间$\varepsilon$，$\varepsilon = \{\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_d\}$，$\bar{p}_i$为$\varepsilon$的第$i$维方向，$\lambda_i$称为方向$\bar{p}_i$的方向扩展系数，$P = (\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_d)$为$\widetilde{C}_i$的关联空间矩阵。

定义 5.4：（最小扩展空间）对于癌症$\widetilde{C}_i$和表达矩阵C_i，假设$\hat{\varepsilon}$是$\widetilde{C}_i$的d维关联空间，$\lambda_1, \lambda_2, \cdots, \lambda_d$是$\hat{\varepsilon}$的方向扩展系数，称$\hat{\varepsilon}$为$d$维最小扩展空间，if $\lambda_1, \lambda_2, \cdots, \lambda_d$ 满足：

$$max(\lambda_1, \lambda_2, \cdots, \lambda_d) \leq min(\lambda_d, \lambda_{d+1}, \cdots, \lambda_m)。$$

针对癌症病理上致癌因子存在的局部相关性，利用下面定义的映射方法来抽取癌症样本中有价值的特征信息。

定义 5.5：（映射）设癌症$\widetilde{C}_i = \{\bar{s}_1, \bar{s}_2, \cdots, \bar{s}_{n_i}\}$，$\bar{s}_l = (s_{l1}, s_{l2}, \cdots, s_{lm})^T$，则$P(\bar{s}_l, \varepsilon)$表示$\bar{s}_l$在关联空间$\varepsilon$的映射，$P(\bar{s}_l, \varepsilon) = (\bar{s}_l \cdot \bar{p}_1, \bar{s}_l \cdot \bar{p}_2, \cdots, \bar{s}_l \cdot \bar{p}_d)^T$，其中，$\bar{p}_j$为$\varepsilon$的第$j$维方向，$\bar{p}_j = (p_{j1}, p_{j2}, \cdots, p_{jm})^T$，$\bar{s}_l \cdot \bar{p}_j = \sum_{k=1}^{m} s_{lk} p_{jk}$。

定理 5.1：对于癌症$\widetilde{C}_i$和d维关联空间ε，$\widetilde{C}_i$中癌症样本在$\bar{p}_j (\bar{p}_j \in \varepsilon)$上映射的方差等于方向$\bar{p}_j$的扩展系数，在不同方向上的映射之间互不相关。

证明: 设癌症$\widetilde{C}_i$的基因表达矩阵C_i，$C_i = (\bar{s}_1, \bar{s}_2, \cdots, \bar{s}_{ni})$。癌症样本$s_l(1 \leq l \leq n_i)$在$\varepsilon$上的映射为$P(\bar{s}_l, \varepsilon)$，设$P(\bar{s}_l, \varepsilon) = (p_1, p_2, \cdots, p_d)^T$，其中$p_j = \bar{s}_l \cdot \bar{p}_j = \bar{p}_j^T \bar{s}_l$，$1 \leq j \leq d$。不妨假设$P(\bar{s}_l, \varepsilon)(1 \leq l \leq n_i)$为$d$维正态随机列向量，显然，$Var(p_j) = \bar{p}_j^T Cov(C_i^T)\bar{p}_j$，由于$Cov(C_i^T)\bar{p}_j = \lambda_j \bar{p}_j$，所以$Var(p_j) = \bar{p}_j^T \lambda_j \bar{p}_j = \lambda_j$。由向量之间的相关系数可知，$Cov(p_j, p_k) = Cov(\bar{p}_j^T \bar{s}_l, \bar{p}_k^T \bar{s}_l) = \bar{p}_j^T Cov(C_i^T)\bar{p}_k = \lambda_k \bar{p}_j^T \bar{p}_k$，由于$\bar{p}_j$与$\bar{p}_k$正交，所以$Cov(p_j, p_k) = 0$，那么$P(\bar{s}_i, \varepsilon)(1 \leq i \leq n_i)$在$\bar{p}_j$与$\bar{p}_k$方向上不相关。 □

从定理5.1可知，当癌症样本在ε上映射后产生的基因特征之间互不相关，可以有效地消除微阵列表达谱数据的冗余和噪声。如果$d < m$，则可以压缩样本维数，是一种有效的特征抽取方法。同时方向扩展系数λ_j描述了$\widetilde{C}_i$中样本在ε的$\bar{p}_j$方向上的相异度，即λ_j越小，$\widetilde{C}_i$中样本在方向$\bar{p}_j$上映射后的相异程度越小，相似程度越大。由定义5.4可知，在d相同的情况下，$\widetilde{C}_i$中样本在$\hat{\varepsilon}$上比在ε上具有更大的相似性。

相对于传统的特征选取方法，基于关联空间的特征抽取方法除了可以有效地压缩样本维数，并具有如下优点：

①传统的特征选取方法作用于整体的数据空间，而关联空间针对癌症组选择局部相关的特征空间，适应于高维的基因数据。

②癌症样本在关联空间上映射后基因特征之间不相关，可以消除基因特征之间的冗余和噪声信息。

③样本在选取的最小扩展空间的方向上具有最大的相似度，可以发现癌症模式。

④关联空间是基于线性组合的思想将基因特征进行压缩，不会导致大量特征信息的丢失。

⑤可以发现隐含在基因表达谱中的导致癌症产生的控制因子。

5.3.2 控制因子的基因调控

假设在癌症$\widetilde{C}_i$中隐含d个控制因子，这些控制因子调控所有基因在$\widetilde{C}_i$中的表达。由定义5.5可知，对于$\bar{s}_l \in \widetilde{C}_i$，$\bar{s}_l$在$\varepsilon$的映射为$P(\bar{s}_l, \varepsilon) = (\bar{s}_l \cdot \bar{p}_1, \bar{s}_l \cdot \bar{p}_2, \cdots, \bar{s}_l \cdot \bar{p}_d)^T$，可用如下矩阵形式表示：

$$P(\bar{s}_l, \varepsilon) = P^T \bar{s}_l \tag{5.3}$$

那么

$$\bar{s}_l = P \cdot P(\bar{s}_l, \varepsilon) \tag{5.4}$$

其中，P为$\widetilde{C}_i$的关联空间矩阵，P^T为P的转置矩阵。

在此，我们可以这样理解癌症$\widetilde{C}_i$的基因表达：一方面，关联空间矩阵P的行和列分别对应基因和控制因子，其元素p_{jk}表示该癌症中第j个基因中第k个控制因子的控制量。另一方面，$P(\bar{s}_l,\varepsilon)$的行和列分别对应控制因子和样本。$P(\bar{s}_l,\varepsilon)$的第k个元素表示第k个控制因子在该样本中的调控度。因此，在样本$\bar{s}_l$中第j个基因的表达水平即C_i的元素c_{jl}为所有遗传控制因子p_{jk}和其对应的在该样本中调控度（$P(\bar{s}_l,\varepsilon)$的第k个元素）乘积之和。

5.3.3 基于最小扩展空间的癌症分类算法（CCLSS）

定义 5.6: （关联空间距离）设癌症样本$\bar{s}_j$，$\bar{s}_{j'}$和关联空间ε，其中$\bar{s}_l = (s_{l1}, s_{l2}, \cdots, s_{lm})^T (l = j, j')$，则$D(\bar{s}_j, \bar{s}_{j'}, \varepsilon)$表示样本和在$d$维关联空间$\varepsilon$上的距离，$D(\bar{s}_j, \bar{s}_{j'}, \varepsilon) = \|P(\bar{s}_j,\varepsilon) - P(\bar{s}_{j'},\varepsilon)\| = \sqrt{\sum_{k=1}^{d} (\bar{s}_j \cdot \bar{p}_k - \bar{s}_{j'} \cdot \bar{p}_k)^2}$，其中$\bar{p}_k$是$\varepsilon$的第$k$维方向。

将癌症样本分为训练集（Training Set）和测试集（Test Set），训练集中样本的癌症类型都已知，而测试集中样本的癌症类型未知。基于最小扩展空间的癌症分类算法的基本思想是首先通过提出的特征抽取方法抽取每一类型癌症的关联空间$\widetilde{C}_i$，并提取对应的d维最小扩展空间$\hat{\varepsilon}_i$；然后将测试样本$\bar{t}$与训练样本$\bar{s}$（$\bar{s} \in$ Training Set）的关联空间距离$D(\bar{s}, \bar{t}, \hat{\varepsilon}_j)$作为判定$\bar{t}$为癌症$\widetilde{C}_j$的权重，其中$\bar{t} \in \widetilde{C}_j$，并构建$\bar{t}$的判定权重向量$w = (w_1, w_2, \cdots, w_k)$；最后识别$\bar{t}$为$w$中分量最小值对应的癌症类型。

算法CCLSS的详细描述见算法5.1。

5.4 算法分析与实验

5.4.1 算法分析

定义 5.7: （组能量）对于癌症$\widetilde{C}_i$，$M(\widetilde{C}_i)$为$\widetilde{C}_i$的质心，则$E(\widetilde{C}_i,\varepsilon)$表示$\widetilde{C}_i$在关联空间$\varepsilon$的组能量，$E(\widetilde{C}_i,\varepsilon) = \frac{1}{n_i} \sum_{\bar{s}_l \in \widetilde{C}_i, l=1\ldots n_i} \{D(\bar{s}_l, M(\widetilde{C}_i), \varepsilon)\}^2$，其中$n_i$是$\widetilde{C}_i$的样本数目。

从组能量的定义可知，癌症$\widetilde{C}_i$在ε上的组能量$E(\widetilde{C}_i,\varepsilon)$反映的是$\widetilde{C}_i$中的样本和$\widetilde{C}_i$的质心在$\varepsilon$上映射后的距离，是$\widetilde{C}_i$中的癌症样本在$\varepsilon$上相似性的度量，组能量越小，组内样本的相似程度越大，反之组内样本的相异程度越大。我们利用组能量$E(\widetilde{C}_i,\varepsilon_i)$来评价癌症$\widetilde{C}_i$的关联空间$\varepsilon_i$，有如下定理：

算法 5.1 基于最小扩展空间的癌症分类算法（CCLSS）

Require: 训练集（Training Set），测试集（Test Set），最小扩展空间的秩（d），其中训练集中有k种不同类型的癌症，第i类癌症样本集合是$\widetilde{C}_i$。

Outputting: $\bar{t}$的癌症类别，$\bar{t} \in$Test Set

```
for i = 1 to k do
    获取癌症C̃表达谱的协方差矩阵Cov(C̃_i^T);
    获取Cov(C̃_i^T)的特征值λ_1, λ_2, ..., λ_m，特征向量p̄_1, p̄_2, ..., p̄_m;
    选取d个最小的λ'_1, λ'_2, ..., λ'_d对应的p̄'_1, p̄'_2, ..., p̄'_m构成C̃_i的最小扩展空间ε̂_i;
    for each t̄ in Test Set do
        赋值t̄属于C̃_i的权重w_i = 0(i = 1...k);
        for each s̄ in Training Set do
            获取C̃_i的最小扩展空间ε̂_i，if s̄ ∈ C̃_i;
            计算s̄和t̄在的关联空间ε_i上的映射距离D(s̄, t̄, ε̂_i)(i = 1...k);
            计算t̄属于C̃_i的权重w_i = w_i + D(s̄, t̄, ε̂_i)(i = 1...k);
        end for
            识别t̄ ∈ C̃_i'，if w_i' = min(w_i)(i = 1...k);
    end for
end for
```

定理 5.2: 对于癌症样本集合$\widetilde{C}_i$和表达矩阵C_i，假设ε_i是$\widetilde{C}_i$的关联空间，ε_i的秩为d，$\lambda_r(1 \le r \le d)$是ε_i的方向扩展系数，则$E(\widetilde{C}_i, \varepsilon_i) = \sum_{r=1}^{d} \lambda_r$。

证明: 对于癌症样本集合$\widetilde{C}_i$，$\widetilde{C}_i = \{\bar{s}_1, \bar{s}_2, \cdots, \bar{s}_{n_i}\}$，$E(\widetilde{C}_i, \varepsilon_i) = \frac{1}{n_i} \sum_{\bar{s}_l \in \widetilde{C}_i, l=1...n_i} \{D(\bar{s}_l, M(\widetilde{C}_i), \varepsilon_i)\}^2$，其中$M(\widetilde{C}_i)$为$\widetilde{C}_i$的质心，由于$D(\bar{s}_l, M(\widetilde{C}_i), \varepsilon_i) = \left\| p(\bar{s}_l, \varepsilon_i) - p(M(\widetilde{C}_i), \varepsilon_i)) \right\| = \sum_{r=1}^{d} (\bar{s}_l \cdot \bar{p}_r - M(\widetilde{C}_i) \cdot \bar{p}_r)^2$，那么$E(\widetilde{C}_i, \varepsilon_i) = \frac{1}{n_i} \sum_{l=1}^{n_i} \sum_{r=1}^{d} (\bar{s}_l \cdot \bar{p}_r - M(\widetilde{C}_i) \cdot \bar{p}_r)^2 = \frac{1}{n_i} \sum_{r=1}^{d} \sum_{l=1}^{n_i} (\bar{s}_l \cdot \bar{p}_r - M(\widetilde{C}_i) \cdot \bar{p}_r)^2$，又由定理5.1知，$\sum_{l=1}^{n_i} (\bar{s}_l \cdot \bar{p}_r - M(\widetilde{C}_i) \cdot \bar{p}_r)^2 = n_i \lambda_r$，因此$E(\widetilde{C}_i, \varepsilon_i) = \sum_{i=1}^{d} \lambda_r$。 □

推论 5.1: 设癌症样本集合$\widetilde{C}_i$和表达矩阵C_i，假设ε_i是$\widetilde{C}_i$的关联空间，ε_i的秩为d，如果$d = m$，则$E(\widetilde{C}_i, \varepsilon_i) = TR(Cov(C_i^T))$。

证明: 对于癌症表达矩阵C_i，由公式5.2知，$Cov(C_i^T) = P \wedge P^T$。那么，$\wedge = P^T Cov(C_i^T) P$。由于$i \neq j$时，$\bar{p}_i$与$\bar{p}_j$正交，则$\sum_{i=1}^{d} \lambda_r = TR(Cov(C_i))$。因为$d = m$，由定理5.2得，$E(\widetilde{C}_i, \varepsilon_i) = TR(Cov(C_i^T))$。 □

定理 5.3: 对于癌症样本集合$\widetilde{C}_i(1 \le i \le k)$，假设$\varepsilon_i$是$\widetilde{C}_i$的$d$维关联空间，$\hat{\varepsilon}_i$是$\widetilde{C}_i$的$d$维最小扩展空间，则$\sum_{i=1}^{k} E(\widetilde{C}_i, \hat{\varepsilon}_i) \le \sum_{i=1}^{k} E(\widetilde{C}_i, \varepsilon_i)$。

证明: 由定理5.2可知，$E(\tilde{C}_i,\varepsilon_i)=\sum_{r=1}^{d}\lambda_r$，$E(\tilde{C}_i,\hat{\varepsilon}_i)=\sum_{r=1}^{d}\hat{\lambda}_r$，$\lambda_r$和$\hat{\lambda}_r$分别为$\varepsilon_i$，$\hat{\varepsilon}_i$的方向扩展系数。由定义4有$\sum_{r=1}^{d}\hat{\lambda}_r\leq\sum_{r=1}^{d}\lambda_r$，从而$E(\tilde{C}_i,\hat{\varepsilon}_i)\leq E(\tilde{C}_i,\varepsilon_i)$。因此，$\sum_{i=1}^{k}E(\tilde{C}_i,\hat{\varepsilon}_i)\leq\sum_{i=1}^{k}E(\tilde{C}_i,\varepsilon_i)$。 □

由推论5.1可知，癌症的关联空间的秩等于癌症样本的维数时，即$d=m$，癌症$\tilde{C}_i$具有最大的组能量，样本的相似度最小。由定理5.3，癌症$\tilde{C}_i$在最小扩展空间$\hat{\varepsilon}_i$上的组能量小于在秩相同的关联空间ε_i上的组能量，即在最小扩展空间$\hat{\varepsilon}_i$上癌症$\tilde{C}_i$中的样本比在秩相同的关联空间ε_i上具有更大的相似性，并且癌症$\tilde{C}_i(1\leq i\leq k)$在最小扩展空间$\hat{\varepsilon}_i(1\leq i\leq k)$上的组能量之和最小，所以在$\hat{\varepsilon}_i$上可以识别癌症模式$\tilde{C}_i(1\leq i\leq k)$。

5.4.2 数据集

本文的分析对象为急性白血病基因表达谱数据集（Leukemia Dataset）[6]和结肠癌基因表达谱数据集（Colon Dataset）[71]。实验环境同3.5。

急性白血病数据集共有72例急性白血病样本，每例样本均含7129个基因的表达数据。其中47例样本被诊断为急性淋巴性白血病（Acute Lymphoblastic Leukemia，ALL），25例被诊断为急性骨髓性白血病（Acute Myeloid Leukemia，AML）。结肠癌数据集共有62例样本，其中包括40例结肠癌组织（Tumor Colon Tissue，TCT）和22例正常结肠组织（Normal Colon Tissue，NCT），每例样本均含2000个基因的表达数据。

5.4.3 噪声基因的过滤

利用“分类信息指数”（Information Index to Classification，IIC）[76, 102]指标来度量基因包含样本分类信息量，见式5.5。

$$d(\bar{g})=\frac{1}{2}\frac{|\mu_{\bar{g}+}-\mu_{\bar{g}-}|}{\sigma_{\bar{g}+}+\sigma_{\bar{g}-}}+\frac{1}{2}\ln\left[\frac{\sigma_{\bar{g}+}^2+\sigma_{\bar{g}-}^2}{2\sigma_{\bar{g}+}\sigma_{\bar{g}-}}\right] \tag{5.5}$$

其中，$\mu_{\bar{g}+}$，$\mu_{\bar{g}-}$分别为基因$\bar{g}$在ALL和AML两个癌症类别所有样本中表达水平的均值，$\sigma_{\bar{g}+}$和$\sigma_{\bar{g}-}$分别为表达水平的标准差。

IIC由两部分构成：第1项实际上是“信噪比”指标；第2项则体现了表达水平分布的方差对样本分类的贡献。利用IIC选择了急性白血病数据集中分类信息指数大于0.8的196个基因，结肠癌数据集中分类信息指数大于0.65的93个基因作为进一步分析的基础。

5.4.4 实验结果与分析

分别将Leukemia Dataset和Colon Dataset随机划分训练集和测试集，训练集和测试集分别占70%和30%，即Leukemia Dataset中训练样本有51例，其中ALL33例，AML18例，测试样本有21例，其中ALL14例，AML7例，Colon Dataset中训练样本有43例，其中TCT28例，NCT15例，测试样本有19例，其中TCT12例，NCT7例。然后利用算法5.1识别测试样本的癌症类型，统计类型识别正确的的样本数目，并计算分类正确率。在Leukemia Dataset上，将算法5.1中d分别取值为$1,2,\ldots,196$；在Colon Dataset上，分别取值为$1,2,\ldots,93$，对于d的每个值将以上分类过程执行10次，计算平均正确率。Leukemia Dataset和Colon Dataset的平均正确率随d的变化情况分别如图5.3、图5.4所示。从图5.3中可以看出，在急性白血病数据集中当$d \leq 20$时，平均正确率随增加而增加；当$d = 20$时，平均正确率达到最大值97.3%；当$d < 120$时，CCLSS取得了较好的分类效果；而当$d > 120$时，CCLSS的平均正确率比较低，尤其当$d > 182$时，平均正确率急骤下降。图5.3说明在急性白血病数据集中，20维的最小扩展空间可以反映急性白血病整个基因组的空间结构，揭示出的隐含致癌因子非常准确地识别ALL癌症和AML癌症，有效地消除了噪声和冗余信息的干扰。然而当最小扩展空间的秩较高时，由于受到噪声的影响，平均正确率比较低。同样从图4可以看出在结肠癌数据集中，当$d = 18$时，平均正确率达到最大值，18维的最小扩展空间同样可以反映结肠癌整个基因组的空间结构，隐含致癌因子很好地识别了结肠癌组织和正常组织。

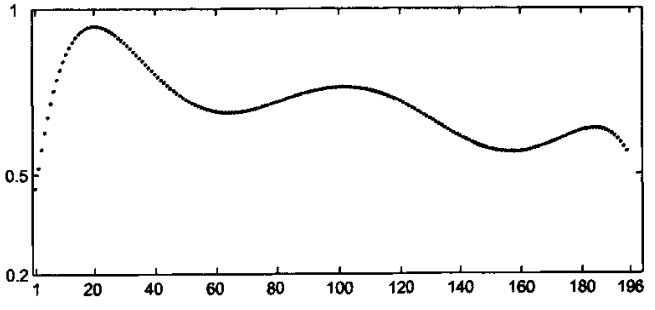

图5.3 Leukemia Dataset中平均正确率随d的变化情况

下面进一步讨论Leukemia和Colon中致癌因子的局部相关性，分别提取Leukemia Dataset中ALL（47例）和AML（25例）的最小扩展空间$\hat{\varepsilon}_{ALL}$和$\hat{\varepsilon}_{AML}$（$d = 20$）以及Colon Dataset中结肠癌组织（40例）和正常组织（22例）的最小扩展空间$\hat{\varepsilon}_{TCT}$和$\hat{\varepsilon}_{NCT}$（$d = 18$）。图5.5和图5.6分

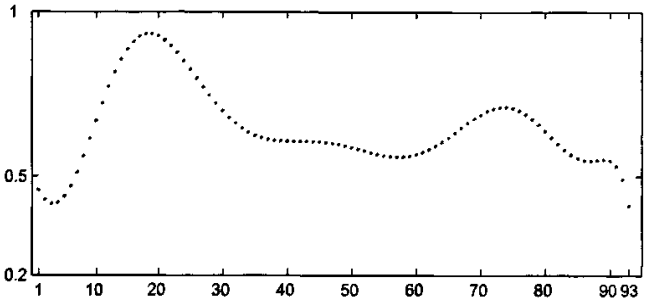

图5.4 Colon Dataset中平均正确率随d的变化情况

别给出了Leukemia Dataset中ALL和AML在$\hat{\varepsilon}_{ALL}$、$\hat{\varepsilon}_{AML}$下的分布情况。我们将$\hat{\varepsilon}_{ALL}$和$\hat{\varepsilon}_{AML}$中的方向按其扩展系数进行递增排序，设$\hat{\varepsilon}_{ALL}$和$\hat{\varepsilon}_{AML}$对应的方向为$dim1, dim2, \ldots, dim20$。图5.5(a)中选取的方向是$\hat{\varepsilon}_{ALL}$的$dim1 - dim3$，(b)选取的是$dim5 - dim7$，(c)选取的是$dim8 - dim10$，(d)选取的是$dim11 - dim13$，(e)选取的是$dim15 - dim17$，(f)选取的方向是$dim18 - dim20$。图5.6中$\hat{\varepsilon}_{AML}$方向的选择与图5.5相同。在$\hat{\varepsilon}_{ALL}$下，可以发现癌症模式ALL，图5.5(a)-(f)中所有的ALL样本分布都非常集中，虽然图(f)中有一个AML样本混淆在癌症ALL中，图(d)中癌症ALL和癌症AML的界线比较模糊，但是图(a)、(b)、(c)都可以有效的识别ALL。同样在$\hat{\varepsilon}_{AML}$下也可以发现癌症模式AML，图5.6(a)-(f)中所有的AML样本分布都非常集中，虽然在图(a)、(b)、(d)中癌症ALL和癌症AML的界线比较模糊，但在图(c)、(e)、(f)中可以有效地识别AML。

图5.7和图5.8则分别给出了Colon Dataset中TCT和NCT在$\hat{\varepsilon}_{TCT}$、$\hat{\varepsilon}_{NCT}$下的分布情况。同样分别将$\hat{\varepsilon}_{TCT}$和$\hat{\varepsilon}_{NCT}$中的方向按其扩展系数进行递增排序，对应的方向为$dim1, dim2, \ldots, dim18$。图5.7(a)中选取的方向是$\hat{\varepsilon}_{TCT}$的$dim1 - dim3$，(b)选取的是$dim4 - dim6$，(c)选取的是$dim7 - dim9$，(d)选取的是$dim10 - dim12$，(e)选取的是$dim13 - dim15$，(f)选取的方向是$dim16 - dim18$。图5.8中$\hat{\varepsilon}_{NCT}$方向的选取与图5.7相同。图5.7中在$\hat{\varepsilon}_{TCT}$下，(a)-(f)中所有的TCT样本分布都非常紧密，NCT则分布非常分散，可以识别TCT模式。同样，图5.8中在$\hat{\varepsilon}_{NCT}$下，则可以识别NCT模式。

最后采用“留一交叉检验法”（Leave One Out Cross Validation, LOOCV）[103]进行样本的癌症类型识别，即在样本集（Leukemia/Colon Dataset）上每次保留一个不同的样本作为测试样本，其余样本作为训练数据集。重复该过程，直到每一个样本都有一次作为测试样本时为止。统计所有被正确识别的样本数作为LOOCV的识别正确数。表5.1和表5.2给出了Leukemia Dataset和Colon Dataset中的基因特征数目及其CCLSS的分类性能，并与Golub提

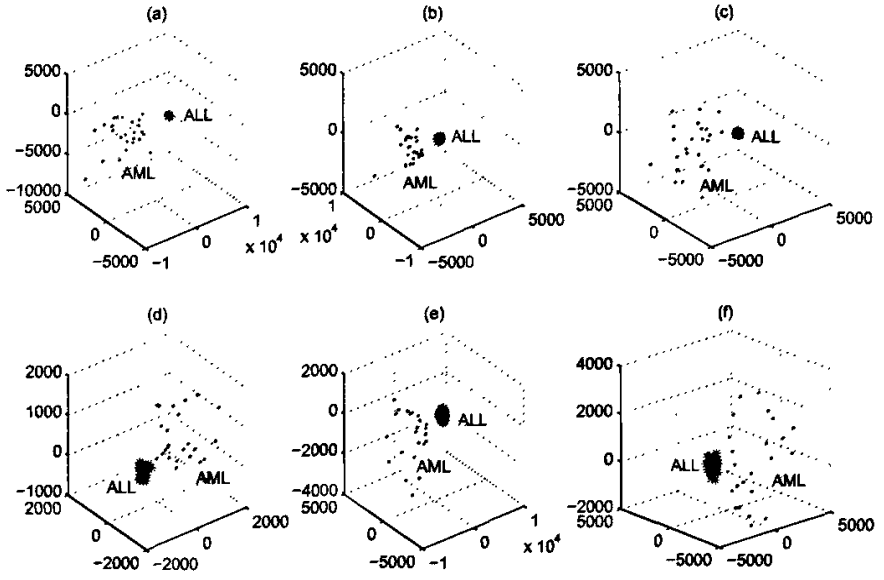

图5.5 ALL和AML在$\hat{e}_{ALL}$下的分布

出的加权投票法（Weighted Voting）[6]、Conde提出的基于聚类（Clustering）的感知器模型[78]、SVM和KNN所得结果进行了比较。感知器模型设置见第3.5.4节，$\eta = 0.5$，SVM采用径向基函数（RBF）作为核函数，KNN相似性度量函数采用Pearson相关系数，$K = 5$。在Leukemia数据集上，CCLSS中最小扩展空间的秩为20，在其余方法中，特征基因的数目设为50。在Colon数据集上，CCLSS中最小扩展空间的秩为18，在其余方法中，特征基因的数目设为40。从表5.1和表5.2可以看出，本文方法提取的最小扩展空间方法更加有效地压缩了基因表达数据维数，CCLSS提高了样本的识别正确率。

表5.1 Leukemia Dataset中LOOCV的实验结果比较

Method	Features	Accuracy
CCLSS	20	71
Weighted Voting	50	65
SVM	50	69
KNN	50	62
Clustering	50	63

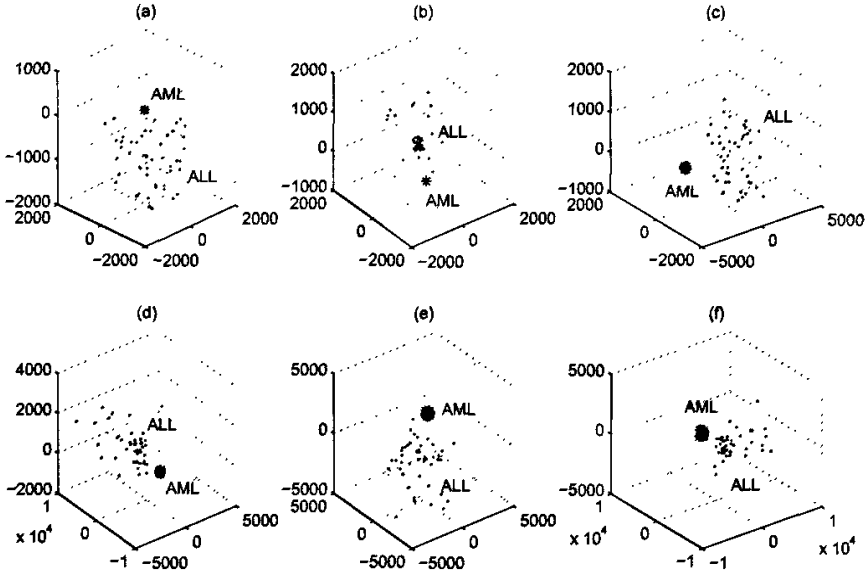

图5.6 ALL和AML在$\hat{\varepsilon}_{AML}$下的分布

表5.2 Colon Dataset中LOOCV的实验结果比较

Method	Features	Accuracy
CCLSS	18	61
Weighted Voting	40	58
SVM	40	57
KNN	40	55
Clustering	40	53

5.5 本章小结

针对癌症检测中的“维数灾难”问题，讨论致癌因子存在的局部相关性，利用基因特征的变换法构建癌症组的关联空间，提取使得癌症组具有最小组能量的最小扩展空间，提出了一种基于最小扩展空间的癌症分类算法（CCLSS）。经理论和实验分析在最小扩展空间上可以有效识别癌症模式。最后，在急性白血病和结肠癌数据集上进行实验，分别选取最佳辨别AML和ALL的最小扩展空间，结肠癌组织与正常结肠组织的最小扩展空间，利用CCLSS通过LOOCV实验方法进行癌症识别，结果表明CCLSS上比传统算法具有更好的分类精确度。

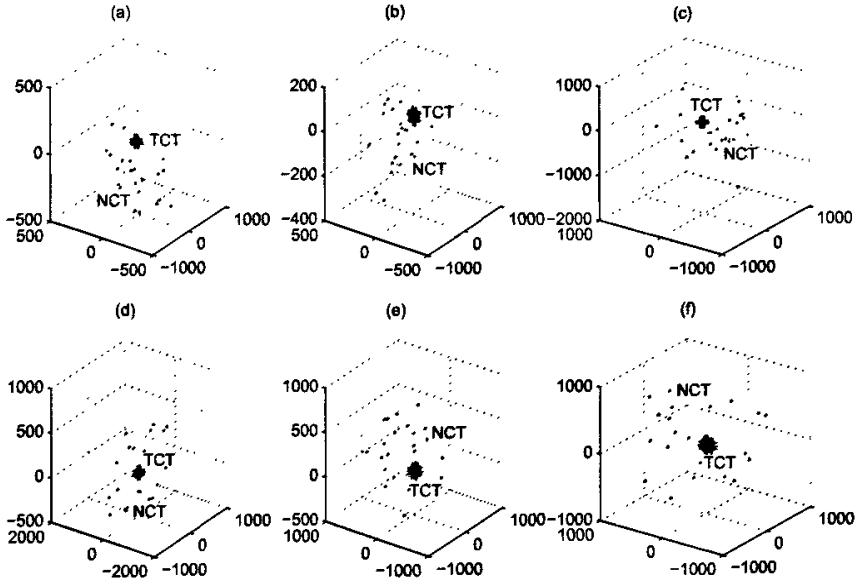

图5.7 TCT和NCT在$\hat{\varepsilon}_{TCT}$下的分布

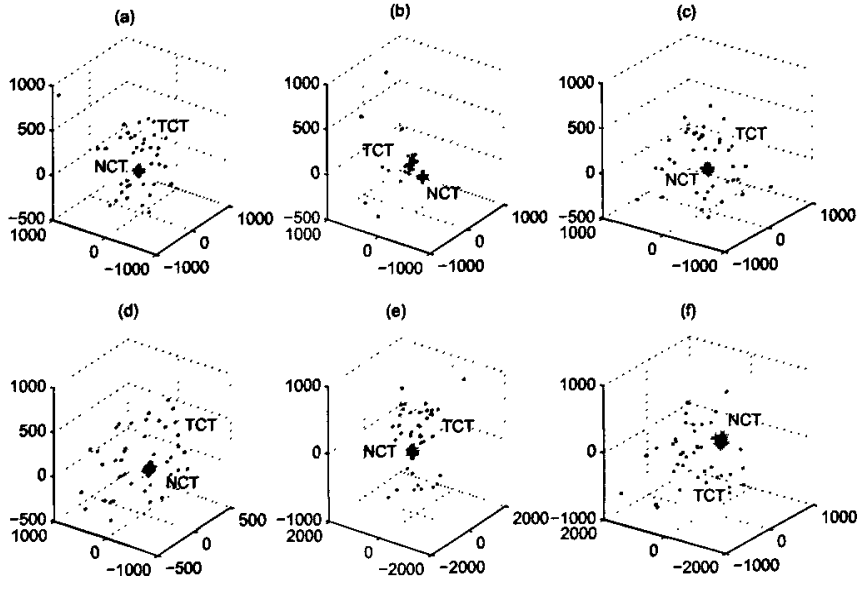

图5.8 TCT和NCT在$\hat{\varepsilon}_{NCT}$下的分布

第 6 章　基于组合GCM和CCM的癌症分类算法

在利用基因表达谱进行癌症识别中，建立在不同的特征基因子集之上的分类器的癌症分类精确度也不同，缺乏较好的泛化性。根据基因表达谱数据特点提出全局分量模型（GCM）和癌症组分量模型（CCM）两种癌症识别模型，并结合GCM模型和CCM模型的互补性，利用基于权值的投票组合策略提出一种基于组合GCM和CCM的癌症分类算法（EAGC）。在Leukemia、Breast、Prostate、DLBCL、Colon、Ovarian 等六个数据集上进行独立测试实验和交叉测试实验。结果表明EAGC有效综合了GCM和CCM识别模型的解决方案，弥补了单个分类器的不足，具有很好的泛化性，在所有数据集上都取得很好的分类性能。

6.1 概述

在癌症的诊断和治疗过程中，对癌症的精确分类是提高诊断准确率和癌症治愈率至关重要的一个环节[55, 99]。近年来，基因微阵列（Microarray）技术的发展使得研究人员可以从基因的表达水平（Expression Level）层次上分析癌症，揭示癌症的病理和基因调控，对癌症诊断及治疗的研究具有非常重要的意义。

利用微阵列基因表达谱数据进行癌症分类是典型的高维、高噪和高冗余问题[76, 101]。特征选择法是一种主要的基因表达谱数据预处理方法，如信噪比（Signal to Noise Ratio, SNR），排秩法（Rank），信息增益（Information Gain）等[6, 54, 76, 84]。文献[71]分析和比较不同特征选择法在癌症分类中的特征基因选取情况，发现在相同数据集中不同方法挑选出的特征基因明显不同，导致经过不同特征选择方法预处理之后的癌症识别效果亦不相同。主要原因是不同的特征选择法基于不同的搜索机制和评价策略，挑选出来的特征基因偏向于致癌病理的一个方面或多个方面中的一部分，而不是全面地反映癌症病理因素。对于一种癌症识别分类器，如果选取合适的特征子集则会获得较好的分类结果，反之则分类结果不理想。这样就导致分类结果不稳定，缺乏泛化性。

一种有效的解决方法是进行分类器组合[104]。文献[71]采用多数投票法（Majority Voting）组合4种不同的分类器进行癌症识别。文献[70]提出一种基于装袋（Bagging）的组合决策树的癌症分类算法。目前在组合分类算法中都是首先采用不同的特征选择方法选择不同的基因子集，如图6.1(a)所示，然后利用这些基因子集来训练分类器以进行分类器组合，如图6.1(b)所示。这些方法的共同不足是不同子集之间存在较多的重叠特征，导致分类器训练时输入较多的冗余信

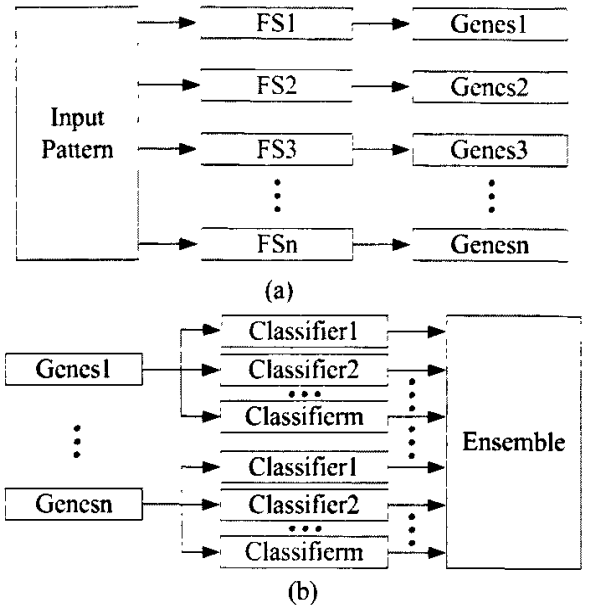

图6.1 基因特征选择和分类器组合，FS表示特征选择方法

息；同时没有充分考虑特征基因子集选取时的互补性以及分类器之间的差异性，具有互补性的分类器组合可以弥补单个分类器的缺点的同时也保持它的优点，有利于优化分类器的组合结果。

神经网络是一种有效的模式识别模型[15, 105]，但是由定量数据建立的单一神经网络模型往往缺乏泛化能力。结合组合分类算法的优点，本文提出了一种基于组合神经网络的癌症分类算法。首先利用主分量分析法（Principal Component Analysis, PCA）选取基因特征空间中大于分量累积贡献率阀值的r个主要分量，通过这些主分量训练识别癌症的神经网络模型以构造全局分量模型（GCM）；然后针对每一组癌症类型抽取癌症组分量，利用癌症组分量训练识别癌症的神经网络模型以构造癌症组分量模型（CCM）；最后利用基于权值的投票组合策略提出一种基于组合GCM和CCM的癌症分类算法（EAGC）。并在Leukemia、Breast、Prostate、DLBCL、Colon、Ovarian等六个数据集上分别进行独立测试实验和交叉测试实验。由于在基因特征的抽取和癌症识别模型的构造上，GCM和CCM都具有很强的互补性。EAGC综合了GCM和CCM识别模型的解决方案，有效扩展了算法的解决方案，以弥补单个分类器的不足，提高了整个系统的泛化能力。

6.2 相关知识

6.2.1 BP网络

BP网络是一种应用非常广泛的神经网络模型，在模式识别、智能控制和信号处理等领域都有大量的应用。BP 网络实质就是多层感知器（Multi-Layer Perceptron，MLP）。BP 网络是由输入层、输出层和若干隐层互相连接构成。BP 网络结构为：前后相邻层的任意两节点均连接，同层和非相邻层的节点均无任何藕合，从输入层开始逐层连接，到输出层连接结束。

6.3 基于组合GCM和CCM的癌症识别（EAGC）

根据基因表达谱数据的特点，本节将构建两种神经网络的癌症识别模型，并依据抽取的输入变量称之为全局分量模型（Global Component Model，GCM）和癌症组分量模型（Cancer Component Model，CCM）。

6.3.1 全局分量模型（GCM）

对于基因表达矩阵$X_{m\times n}$，不妨设$X^T=(\bar{g}_1,\bar{g}_2,...,\bar{g}_m)$，利用主分量分析法（PCA）抽取基因表达谱中的$r(r\leq m)$个隐含变量，用如下公式表示：

$$\begin{aligned}&\bar{h}_i^T=a_1\bar{g}_1+a_2\bar{g}_2+\cdots+a_m\bar{g}_m=\bar{a}_i^TX^T\\&Var(\bar{h}_i)=\bar{a}_i^T\sum\bar{a}_i=\lambda_i\end{aligned}\tag{6.1}$$

其中$\bar{a}_i^T=(\bar{a}_1,\bar{a}_2,\ldots,\bar{a}_m)$，$\sum=(v_{ij})_{m\times m}=(Cov(\bar{g}_i,\bar{g}_j))$，$Var$表示方差，$Cov$表示协方差，$\bar{h}_i$表示第$i$个主分量，即$PC_i$，$\lambda_i$是$\sum$的第$i$个特征值，$\bar{a}_i$是$\lambda_i$对应的特征向量，表示观察基因变量在$\bar{h}_i$上的载荷。

定义 6.1：（分量贡献系数）在基因表达数据中定义λ_i为隐含分量PC_i的分量贡献系数。

定义 6.2：（累积贡献率）不妨设$\lambda_1\geq\lambda_2\geq...\geq\lambda_m$，给定$r(1\leq r\leq m)$，分量$\bar{h}_1,\bar{h}_2,...,\bar{h}_r$的累积贡献率为$CR=\frac{\sum_{i=1}^{r}Var(\bar{h}_i)}{\sum_{i=1}^{m}Var(\bar{h}_i)}$。

定义 6.3：（全局分量空间）对于$r\leq m$，设$min(\lambda_1,\lambda_2,\cdots,\lambda_r)\geq max(\lambda_{r+1},\lambda_{r+2},\cdots,\lambda_m)$，则由$\bar{a}_1,\bar{a}_2,\cdots,\bar{a}_d$组成了基因表达数据的$r$维全局分量空间$\varepsilon_g$，$\varepsilon_g=\{\bar{a}_1,\bar{a}_2,\cdots,\bar{a}_d\}$。

假设基因表达谱中分为k个癌症类别$C_1, C_2, ..., C_k$，我们抽取$CR \geq$阀值的r个主分量，并利用BP网络构建识别癌症类型的全局分量模型（GCM），如图6.2所示。GCM包含输入层（I层）、隐层（H层）和输出层（O层）三层，I层有r个神经元节点，O层有k个神经元节点，设H层有q个神经元节点。

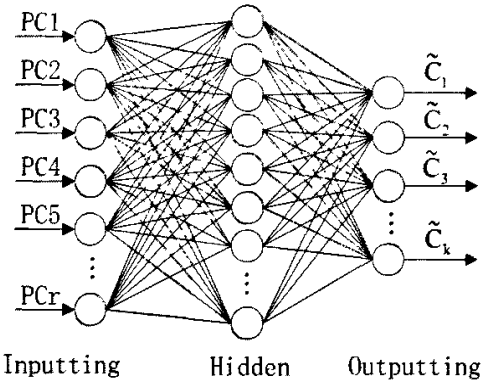

图6.2 癌症识别中的全局分量模型

全局分量模型用下面的数学公式6.2描述：

$$net_j = \begin{cases} \sum_i w_{ij} I_i - \theta_j & if\ j \in H \\ \sum_i w_{ij} H_i - \theta_j & if\ j \in O \end{cases} \qquad (6.2)$$
$$out_j = f(net_j)$$

其中out_j是神经元节点j的输出，w_{ij}是节点i到j的权值，I_i是I层节点i的输入，H_i是H层节点i的输出，θ_j是节点j的激活阀值，节点的特性函数是S型函数$f(x) = \frac{1}{1+e^{-x}}$。权值和激活阀值的调节如式6.3和式6.4所示：

$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij} = w_{ij}(t) - \eta \delta_j out_i$$
$$\delta_j = \begin{cases} -(\hat{O}_j - O_j) f'(net_k) & if\ j \in O \\ f'(net_k) \sum_k \delta_k w_{jk} & if\ j \in H \end{cases} \qquad (6.3)$$

$$\theta_j(t+1) = \theta_j(t) + \eta \delta_j \qquad (6.4)$$

其中w_{ij}，out_i，θ_j和net_k与式6.2相同，η是增益因子，$0 < \eta \leq 1$，O_j是O层节点j的输出，$\hat{O}_j$是对应的期望输出。对于癌症样本$\bar{s}$，如果$\bar{s} \in C_{i'}$，那么

$$\hat{O}_j = \begin{cases} 1 & if\ j = i' \\ 0 & else \end{cases} \tag{6.5}$$

6.3.2 局部分量模型（CCM）

假设基因微阵列表达谱中第i类癌症样本集合是$\widetilde{C}_i(1 \leq i \leq k)$，$\widetilde{C}_i$的样本数目为$n_i$，$C_i$是$\widetilde{C}_i$的$m \times n_i$基因表达矩阵。对于每一个$\widetilde{C}_i(1 \leq i \leq k)$，我们首先获取$\widetilde{C}_i$的最小扩展空间$\hat{\varepsilon}$，并将癌症样本在$\hat{\varepsilon}$上映射抽取有价值的基因特征，称之为癌症组分量（Cancer Component, CC），然后利用BP网络构建识别癌症类型的癌症组分量模型（CCM）。获取$\widetilde{C}_i$的最小扩展空间$\hat{\varepsilon}$以及样本的CC分量如下所示。

设$C_i^T = (\bar{g}_1, \bar{g}_2, \cdots, \bar{g}_m)$，其中$\bar{g}_j$是第$j$个基因在$\widetilde{C}_i$样本中的基因表达向量。$C_i^T$的协方差矩阵为$Cov(C_i^T)$，$Cov(C_i^T)$是半正定的$m$维方阵，可以进行如下矩阵分解：

$$Cov(C_i^T) = \sum_r \lambda_r \bar{p}_r \bar{p}_r^T = P \wedge P^T \tag{6.6}$$

其中，λ_r是$Cov(C_i^T)$特征值，$\wedge$是非负的对角矩阵，对角线上元素由$\lambda_r(1 \leq r \leq m)$组成，$\bar{p}_r$是$\lambda_r$对应的特征向量，$P = (\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_m)$。

定义 6.4: （关联空间）设癌症$\widetilde{C}_i$和表达矩阵C_i，$\lambda_1, \lambda_2, \cdots, \lambda_m$是$Cov(C_i^T)$的特征值，$\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_m$是$\lambda_1, \lambda_2, \cdots, \lambda_m$对应的特征向量。对于$d \leq m$，则由$\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_m$组成了$\widetilde{C}_i$上秩为$d$的关联空间$\varepsilon$，$\varepsilon = \{\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_d\}$，$\bar{p}_i$为$\varepsilon$的第$i$维方向，$\lambda_i$称为方向$\bar{p}_i$的方向扩展系数，$P = (\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_d)$为$\widetilde{C}_i$的关联空间矩阵。

定义 6.5: （最小扩展空间）对于癌症$\widetilde{C}_i$和表达矩阵C_i，假设$\hat{\varepsilon}$是$\widetilde{C}_i$的d维关联空间，$\lambda_1, \lambda_2, \cdots, \lambda_d$是$\hat{\varepsilon}$的方向扩展系数，当$\lambda_1, \lambda_2, \cdots, \lambda_d$满足$max(\lambda_1, \lambda_2, \cdots, \lambda_d) \leq min(\lambda_d, \lambda_{d+1}, \cdots, \lambda_m)$时，则称$\hat{\varepsilon}$为$d$维最小扩展空间。

定义 6.6: （癌症组分量）对于癌症$\widetilde{C}_i$和d维最小扩展空间$\hat{\varepsilon}$，设$\widetilde{C}_i = \{\bar{s}_1, \bar{s}_2, \cdots, \bar{s}_{n_i}\}$，$\bar{s}_l = (s_{l1}, s_{l2}, \cdots, s_{lm})^T$，那么$\bar{s}_l$在$\bar{p}_j$上的癌症组分量为$CC_j = \bar{s}_l \cdot \bar{p}_j$，其中$\bar{p}_j$为$\varepsilon$的第$j$维方向，$\bar{p}_j = (p_{j1}, p_{j2}, \cdots, p_{jm})^T$，$\bar{s}_l \cdot \bar{p}_j = \sum_{k=1}^{m} s_{lk} p_{jk}$。

我们通过构造$\widetilde{C}_i$的最小扩展空间$\hat{\varepsilon}$，抽取样本的d个癌症组分量，然后利用BP网络构建识别癌症的癌症组$\widetilde{C}_i$分量模型（CCM），如图6.3所示。CCM包含输入层（I层）、隐层（H层）和输出层（O层）三层，I层有d个神经元节点，O层

有2个神经元节点，设H层有q'个神经元节点。CCM模型中神经元节点的输入，输出，转移函数，权值和激阀值调节和GCM模型相同。对于癌症样本$\bar{s}$的期望输出$\hat{O}$通过下式给出：

$$\hat{O} = \begin{cases} [1 \quad 0]^{\mathrm{T}} & if\ \bar{s} \in \tilde{C}_i \\ [0 \quad 1]^{\mathrm{T}} & else \end{cases} \tag{6.7}$$

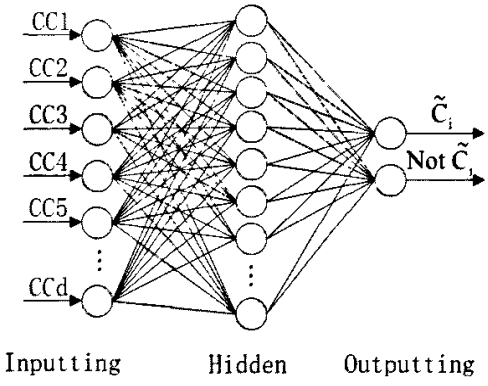

图6.3 癌症识别中的癌症组分量模型

6.3.3 基于组合GCM和CCM的癌症识别算法（EAGC）

GCM模型利用PCA提取样本的主分量$PC_j(1 \le j \le r)$作为输入变量，经过隐层神经元和权值的作用，在输出层判别输入样本的癌症类别$\tilde{C}_i(1 \le i \le k)$。CCM模型则利用癌症组内基因变量的相关性提取样本的癌症组分量，并输入CCM模型，在隐层神经元和权值的调节下，在输出层判别输入样本是否属于某种癌症类别（$\tilde{C}_i$ 或Not $\tilde{C}_i(1 \le i \le k)$）。GCM模型和CCM模型在基因特征抽取和癌症识别模型的构造上具有很强的互补性。如图6.4所示，假设数据集合中存在三个癌症模式ω_1、ω_2和ω_3，则Ω_1是不属于ω_1的样本构成的模式（称之为非ω_1模式）。相应地，Ω_2、Ω_3是非ω_2模式和非ω_3模式。不妨设$\hat{\varepsilon}_{\omega_1} = \{x_1, y_1\}$，$\hat{\varepsilon}_{\omega_2} = \{x_2, y_2\}$，$\hat{\varepsilon}_{\omega_3} = \{x_3, y_3\}$，$\hat{\varepsilon}_{\omega_1+\omega_2+\omega_3} = \{x, y\}$，在$x_1 - y_1$中可以区分$\omega_1$和$\Omega_1$，在$x_2 - y_2$中可以区分$\omega_2$和$\Omega_2$，在$x_3 - y_3$中可以区分$\omega_3$和$\Omega_3$，在$x - y$中可以区分$\omega_1$、$\omega_2$和$\omega_3$。CCM模型从癌症组内基因特征挖掘不同癌症模式$\tilde{C}_i$，是一种具有局部相关性的癌症识别模型。GCM模型从所有基因特征及特征组间相异性发现各种癌症模式$\tilde{C}_i$，是一种具有全局相关性的癌症识别模型。CCM模型可以识别单个癌症模式$\tilde{C}_i$，GCM模型则可以同时识别多个癌症模式$\tilde{C}_i$。从上述分析可知，GCM和CCM是两种具有互补性的癌症识别模型。

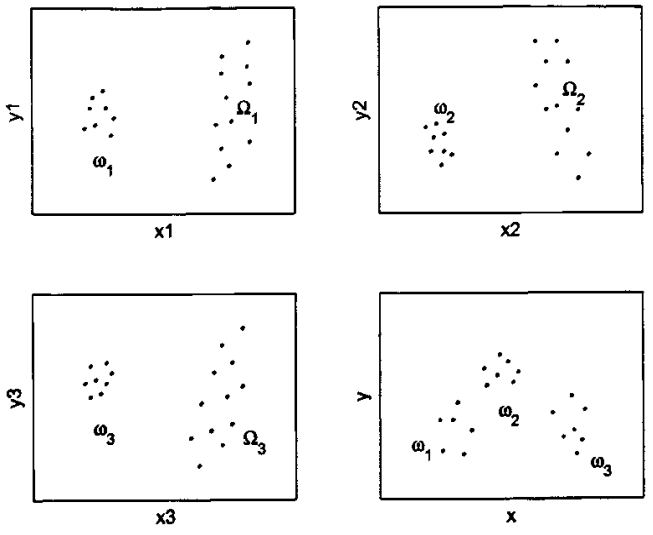

图6.4 CCM的癌症组分量和GCM的全局分量

本节提出一种基于组合GCM和CCM模型的癌症识别算法（图6.5）。首先在训练阶段利用基因数据中的训练子集建立GCM模型和$\widetilde{C}_i$的CCM模型，然后在测试阶段分别利用GCM模型和CCM模型识别测试样本，并利用基于权值的投票组合策略以识别样本的癌症类型。

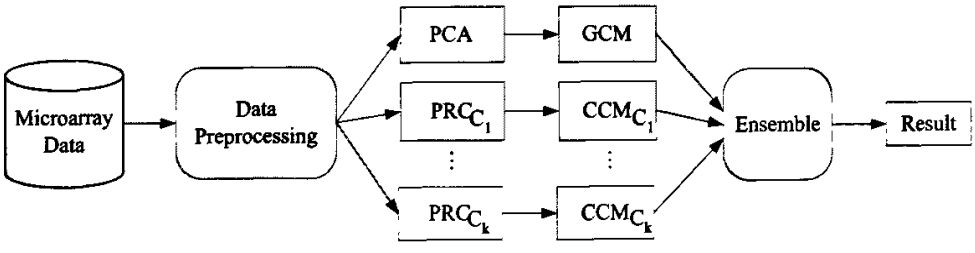

图6.5 基于组合GCM和CCM的癌症识别

对于测试样本$\bar{s}$，不妨设癌症$\widetilde{C}_i$的CCM模型的识别结果为$R(\widetilde{C}_i)^T = (r_{i1}, r_{i2})$，GCM模型的识别结果为$R(\widetilde{C})^T = (r_1, r_2, \ldots, r_k)$，基于权值的投票组合

策略描述如下：

$$\begin{cases} R(ensemble)^T = (r'_1, r'_2, ..., r'_k) \\ r'_i = \alpha r_{i1} + \beta r_i \\ \alpha + \beta = 1 \\ result = \widetilde{C}_{i'} \quad if\ r'_{i'} = max(R(ensemble)) \end{cases} \tag{6.8}$$

其中$R(ensemble)$为$R(\widetilde{C}_i)$ 和$R(\widetilde{C})$的组合结果，α、β分别为CCM和GCM模型的权值，$result$为测试样本$\bar{s}$的癌症类别。

EAGC有效综合GCM和CCM模型的癌症识别结果，消除基因数据中内在的噪声和冗余对单个分类器的影响，优化分类器的癌症识别结果，提高EAGC的泛化能力。基于组合GCM和CCM络模型的癌症识别算法具体描述如下：

算法 6.1 基于组合GCM和CCM模型的癌症识别算法(EAGC).

Require: 训练集（Training Set），测试集（Test Set），其中训练集中有k种不同类型的癌症，第i类癌症样本集合是$\widetilde{C}_i$，$\widetilde{C}_i$的表达矩阵C_i，$\widetilde{C} = \cup\widetilde{C}_i$，$q$=10，$CR \geq 85\%$，$d = 15$，$\eta = 0.5$，$\alpha = 0.4$，$\beta = 0.6$

Outputting: $\bar{t}$的癌症类别，$\bar{t} \in$Test Set

1: 对$\widetilde{C}$的表达矩阵X进行PCA分解，获取全局分量空间$\varepsilon_g = \{\bar{a}_1, \bar{a}_2, \cdots, \bar{a}_r\}$；
2: 给GCM模型的ω_{ij}和θ_i赋随机初值；
3: 训练GCM模型；
4: **for** $i = 1$ to k **do**
5: 获取癌症$\widetilde{C}$表达谱的协方差矩阵$Cov(\widetilde{C}_i^T)$；
6: 获取$Cov(\widetilde{C}_i^T)$的特征值$\lambda_1, \lambda_2, \ldots, \lambda_m$，特征向量$\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_m$；
7: 选取d个最小的$\lambda'_1, \lambda'_2, \ldots, \lambda'_d$对应的$\bar{p}'_1, \bar{p}'_2, \cdots, \bar{p}'_m$构成$\widetilde{C}_i$的最小扩展空间$\hat{\varepsilon}_i$；
8: 给CCM_{Ci}模型的ω_{ij}和θ_j赋随机初值；
9: 训练CCM_{Ci}模型；
10: **end for**
11: **for** each $\bar{t}$ in Test Set **do**
12: 获取$\bar{t}$ 的主分量$PC_j = \bar{t} \cdot \bar{a}_j$，并输入GCM，获得$\bar{t}$的识别结果为$R(\widetilde{C}) = (r_1, r_2, \ldots, r_k)^T$；
13: **for** $i = 1$ to k **do**
14: 获取$\bar{t}$在$\widetilde{C}_i$中的癌症组分量$CC_j = \bar{t} \cdot \bar{p}_j$，并输入$\mathrm{CCM}_{Ci}$，获得识别结果为$R(\widetilde{C}_i) = (r_{i1}, r_{i2})^T$
15: **end for**
16: 计算组合策略结果$R(ensemble) = (r'_1, r'_2, \ldots r'_k)^T$，其中$r'_i = \alpha r_{i1} + \beta r_i$；
17: 识别$\bar{t} \in \widetilde{C}_{i'}$，if $r'_{i'} = max(R(ensemble))$;
18: **end for**

6.3.4 讨论与分析

为了方便描述，设样本$\bar{s}$的主分量为$PC_j(\bar{s}_l)(1 \leq j \leq r)$，癌症组分量为$CC_j(\bar{s}_l)(1 \leq j \leq d)$。$M(\tilde{C}')$是一个虚拟样本，$M(\tilde{C}') = \frac{1}{|\tilde{C}'|}\sum_{\bar{s}_l \in \tilde{C}'} \bar{s}_l$，其中$\tilde{C}' \in \{\tilde{C}, \tilde{C}_i\}$，$|\tilde{C}'|$是$\tilde{C}'$的样本数量，$M(\tilde{C}')$表示$\tilde{C}'$的中心。

定义 6.7: （样本能量）对于癌症数据集$\tilde{C}$，$E(\tilde{C}, \varepsilon_g)$表示$\tilde{C}$中样本在全局分量空间$\varepsilon_g$的样本能量，$E(\tilde{C}, \varepsilon_g) = \frac{1}{n}\sum_{\bar{s}_l \in \tilde{C}} \{\sum_{j=1}^{r} (PC_j(\bar{s}_l) - PC_j(M(\tilde{C})))^2)\}^2$，其中$n$是$\tilde{C}$的样本数目。

定理 6.1: 对于癌症数据集$\tilde{C}$，假设ε_g是$\tilde{C}$的全局分量空间，$\varepsilon_g = \{\bar{a}_1, \bar{a}_2, \cdots, \bar{a}_r\}$，$\lambda_j(1 \leq j \leq r)$是$\bar{a}_i$对应的特征值，则$E(\tilde{C}, \varepsilon_g) = \sum_{j=1}^{r} \lambda_j$。

证明: 不妨设$\tilde{C} = \{\bar{s}_1, \bar{s}_2, \cdots, \bar{s}_n\}$，由PCA分析可知，$PC_j(\bar{s}_l) = \bar{s}_l \cdot \bar{a}_j$，$E(\tilde{C}, \varepsilon_g) = \frac{1}{n}\sum_{l=1}^{n}\sum_{j=1}^{r} (\bar{s}_l \cdot \bar{a}_j - M(\tilde{C}) \cdot \bar{a}_j)^2 = \frac{1}{n}\sum_{j=1}^{d}\sum_{l=1}^{n} (\bar{s}_l \cdot \bar{a}_j - M(\tilde{C}) \cdot \bar{a}_j)^2$，由式（1）可知，$\sum_{l=1}^{n} (\bar{s}_l \cdot \bar{a}_j - M(\tilde{C}) \cdot \bar{a}_j)^2 = n\lambda_j$，因此$E(\tilde{C}, \varepsilon_g) = \sum_{j=1}^{r} \lambda_j$。 □

从样本能量的定义可知，$E(\tilde{C}, \varepsilon_g)$表示癌症集$\tilde{C}$中样本和中心$M(\tilde{C})$在$\varepsilon_g$上的距离，是$\tilde{C}$中的癌症样本在$\varepsilon_g$上相似性的度量。$E(\tilde{C}, \varepsilon_g)$反映了$\tilde{C}$的子类$\tilde{C}_i$中样本的类间相异性。样本能量越大，$\tilde{C}$中样本具有越大的活跃性，则样本之间的相似程度越小，子类$\tilde{C}_i$的类间相异性越大。由于$min(\lambda_1, \lambda_2, \cdots, \lambda_r) \geq max(\lambda_{r+1}, \lambda_{r+2}, \cdots, \lambda_m)$，因此样本在$r$维的$\varepsilon_g$上具有最大的样本能量，则$\tilde{C}$中样本具有最大的活跃性。

定义 6.8: （组能量）对于癌症$\tilde{C}_i$，$E(\tilde{C}_i, \varepsilon)$表示$\tilde{C}_i$在关联空间$\varepsilon$的组能量，$E(\tilde{C}_i, \varepsilon) = \frac{1}{n_i}\sum_{\bar{s}_l \in \tilde{C}_i} \{\sum_{j=1}^{d} (CC_j(\bar{s}_l) - CC_j(M(\tilde{C})))^2)\}^2$，其中$n_i$是$\tilde{C}_i$的样本数目。

定理 6.2: 对于癌症样本集合$\tilde{C}_i$和表达矩阵C_i，假设ε_i是$\tilde{C}_i$的关联空间，ε_i的秩为d，$\lambda_j(1 \leq j \leq d)$是$\varepsilon_i$的方向扩展系数，则$E(\tilde{C}_i, \varepsilon_i) = \sum_{j=1}^{d} \lambda_j$。

证明: 对于癌症样本集合$\tilde{C}_i$，设$\tilde{C}_i = \{\bar{s}_1, \bar{s}_2, \cdots, \bar{s}_{n_i}\}$，因为$CC_j(\bar{s}_l) = \bar{s}_l \cdot \bar{p}_j$，所以$E(\tilde{C}_i, \varepsilon_i) = \frac{1}{n_i}\sum_{l=1}^{n_i}\sum_{j=1}^{d} (\bar{s}_l \cdot \bar{p}_j - M(\tilde{C}_i) \cdot \bar{p}_j)^2 = \frac{1}{n_i}\sum_{j=1}^{d}\sum_{l=1}^{n_i} (\bar{s}_l \cdot \bar{p}_j - M(\tilde{C}_i) \cdot \bar{p}_j)^2$，又设$p_j = \bar{s}_l \cdot \bar{p}_j = \bar{p}_j^T \bar{s}_l$，由于$Var(p_j) = \bar{p}_j^T Cov(C_i^T)\bar{p}_j = \bar{p}_j^T \lambda_j \bar{p}_j = \lambda_j$，所以$\sum_{l=1}^{n_i} (\bar{s}_l \cdot \bar{p}_j - M(\tilde{C}_i) \cdot \bar{p}_j)^2 = n_i\lambda_j$，因此$E(\tilde{C}_i, \varepsilon_i) = \sum_{j=1}^{d} \lambda_j$。 □

同理可知，组能量$E(\widetilde{C}_i,\varepsilon)$反映的是$\widetilde{C}_i$中的样本和$\widetilde{C}_i$的中心在$\varepsilon$上的距离，是$\widetilde{C}_i$中的癌症样本在$\varepsilon$上相似性的度量，组能量越小，组内样本的相似程度越大，反之组内样本的相异程度越大。从最小扩展空间的定义可知，$max(\lambda_1,\lambda_2,\cdots,\lambda_d) \leq min(\lambda_d,\lambda_{d+1},\cdots,\lambda_m)$，因此在$d$维最小扩展空间$\hat{\varepsilon}_i$上癌症$\widetilde{C}_i$中的样本具有最大的相似性。

从上述分析可知，GCM模型利用主分量PC_j作为基因特征来训练分类器，从癌症数据集的整体性来分析癌症数据以识别各类癌症$\widetilde{C}_i$。CCM模型则利用癌症组的癌症分量CC_j作为基因特征来训练分类器，从癌症组的局部性来分析癌症数据以识别癌症组$\widetilde{C}_i$。

这两类癌症识别模型从高维的癌症数据的不同特性来分析癌症数据，基于组合GCM和CCM的解决方案利用GCM模型和CCM模型的互补性，融合它们的识别结果以达到最优的识别结果。如图6.6所示，利用主分量建立的GCM模型和癌症组分量建立的CCM模型对患者样本的识别作为候选解决方案（Candidate solution），候选方案与患者的癌症类别（即最优解决方案，Optimal solution）存在较大的偏差，而基于组合GCM和CCM识别的解决方案（Estimated Solution）利用基于权值的投票组合策略综合了所有候选方案，有效地修正了候选解决方案的偏差。

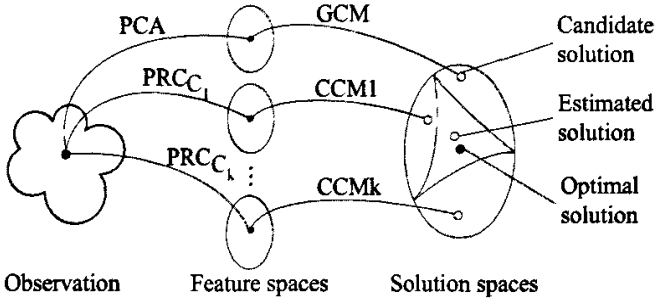

图6.6 基于组合GCM和CCM的解决方案

6.4 实验和分析

本节我们利用下面的六个基因表达谱数据集[43, 70, 106]（表6.1）来进行仿真实验，实验环境同3.5，并分析基于组合GCM和CCM分类算法的癌症识别性能。在此，我们将患者样本的癌症测试实验分为独立测试实验和交叉测试实验。在具有独立测试子集的前三个数据集上分别进行了独立测试实验和交叉测试实验，在没有独立测试子集的后三个数据集上只进行了交叉测试实验。

6.4.1 数据集

1. 急性白血病数据集（ALL-AML Leukemia）

急性白血病数据集[84]包含72 例急性白血病样本，每个样本均含7129个基因表达数据。其中47例样本被诊断为急性淋巴白血病（Acute Lymphoblastic Leukemia，ALL），25例样本被诊断为急性骨髓白血病（Acute Myeloid Leukemia，AML）。该数据集分为训练子集和测试子集，训练子集中包含38例训练样本（27例ALL+11例AML），测试子集中包含34例测试样本（20例ALL+14例AML）。

2. 乳腺癌数据集（Breast Cancer）

乳腺癌数据集[84]包含97 例乳腺癌样本，每个样本均含24481个基因表达数据。乳腺癌数据集记录了经过初次治疗超过5年时间后癌症患者的复发情况，在46例样本中癌症细胞发生转移(Metastases)，即癌症复发（Relapse），51例样本中癌症没有复发（Non-Relapse）。该数据集分为训练子集和测试子集，训练子集中包含78例训练样本（34例Relapse+44例Non-Relapse），测试子集中包含19例测试样本（12例Relapse+7例Non-Relapse）。

3. 前列腺癌数据集（Prostate Cancer）

前列腺癌数据集[84]共有136 例前列腺组织样本，每个样本均含12600个基因表达数据。其中75例为前列腺癌肿瘤样本（Prostate Tumor Sample，PTS），59例样本正常前列腺组织（Normal Prostate Sample，NPS）。该数据集分为训练子集和测试子集，训练子集中包含102例训练样本（52例PTS+50例NPS），测试子集中包含34例测试样本（25例PTS+9例NPS）。

4. 弥漫性大B细胞淋巴瘤数据集（DLBCL）

弥漫性大B细胞淋巴瘤数据集共有47例弥漫性大B细胞淋巴瘤样本，其中包括47例胚中心B细胞样（Germinal Center B-like，GCB）淋巴瘤样本和活性型周围B 细胞样（Activated Peripheral B-like，APB）淋巴瘤样本，每例样本均含4026个基因的表达数据。

5. 结肠癌数据集（Colon Tumor）

结肠癌数据集[71]共有62例结肠组织样本，其中包括40例结肠癌组织（Tumor Colon Tissue，TCT）和22例正常结肠组织（Normal Colon Tissue，NCT），每例样本均含2000个基因的表达数据。

6. 卵巢癌（Ovarian Cancer）

卵巢癌数据集共有253例卵巢组织样本，其中包括91例正常卵巢组织样本（Normal Ovarian Sample，NOS）和151例卵巢癌组织样本（Ovarian Cancer Sam-

ple，OCS），每例样本均含15154个基因的表达数据。

表6.1 基因表达谱数据集

Dataset	Genes	Training Samples	Test Samples
ALL-AML Leukemia	7129	38(27:11)	34(20:14)
Breast Cancer	24481	78(34:44)	19(12:7)
Prostate Cancer	12600	102(52:50)	34(25:9)
DLBCL	4026	47(24:23)	0
Colon Tumor	2000	62(40:22)	0
Ovarian Cancer	15154	253(91:162)	0

6.4.2 过滤噪声基因

利用Fayyad等[107]提出的基于启发式熵最小化的离散方法（Discretization）来过滤噪声基因。表6.2给出了六个数据集的基因过滤情况。

在样本集$\tilde{S}$有k个类别$\tilde{C}_1, \tilde{C}_2, \ldots, \tilde{C}_k$，假设$\tilde{S}$被分为两个子集$\tilde{S}_1$和$\tilde{S}_2$，$p(\tilde{C}_i, \tilde{S}_j)$为样本子集$\tilde{S}_j$中样本出现在$\tilde{C}_i$中的频率。则子集$\tilde{S}_j(j = 1,2)$的熵为：

$$Ent(\tilde{S}_j) = -\sum_{i=1}^{k} P(\tilde{C}_i, \tilde{S}_j)\log(P(\tilde{C}_i, \tilde{S}_j)) \tag{6.9}$$

假设由特征A在点T划分获取子集$\tilde{S}_1$和$\tilde{S}_2$，则用$E(A,T;\tilde{S})$描述此划分的类别信息熵，

$$E(A,T;\tilde{S}) = \frac{|\tilde{S}_1|}{|\tilde{S}|}Ent(\tilde{S}_1) + \frac{|\tilde{S}_2|}{|\tilde{S}|}Ent(\tilde{S}_2) \tag{6.10}$$

在特征A所有候选点中选取具有最小的类别信息熵作为划分点，然后在样本子集$\tilde{S}_j(j = 1,2)$上递归划分，并利用最小描述长度原则（Minimum Description Length Principle，MDL）作为停止条件，即满足：

$$Gain(A,T;\tilde{S}) < \frac{\log_2(N-1)}{N} + \frac{\delta(A,T;\tilde{S})}{N} \tag{6.11}$$

其中N是集合$\tilde{S}$中特征值的数目，$Gain(A,T;\tilde{S}) = Ent(\tilde{S}) - E(A,T;\tilde{S})$，$\delta(A,T;\tilde{S}) = \log_2(3^k - 2) - [k \cdot Ent(S) - k_1 \cdot Ent(S_1) - k_2 \cdot Ent(S_2)]$，$k_i$为$\tilde{S}_i$中类别的数目。

表6.2 噪声基因过滤

Dataset	Genes (Before filtering)	Genes (After filtering)
ALL-AML Leukemia	7129	866
Breast Cancer	24481	834
Prostate Cancer	12600	3071
DLBCL	4026	336
Colon Tumor	2000	135
Ovarian Cancer	15154	2945

6.4.3 性能评价

为了方便描述，将以上每个数据集划分为正例样本（Positives）和负例样本（Negatives）。正例样本分别为ALL、Relapse、PTS、GCB、TCT、NOS样本，负例样本分别为AML、Non-Relapse、NPS、APB、NCT、OCS样本。利用准确度（Accu）、灵敏度（Sn）和明确性（Sp）三个指标来进行性能评价。Accu、Sn、Sp定义如下所示：

$$Accu = \frac{TP + TN}{TP + FP + FN + TN} \tag{6.12}$$

$$Sn = \frac{TP}{TP + FN} \tag{6.13}$$

$$Sp = \frac{TN}{TN + FP} \tag{6.14}$$

其中TP、FP、TN、FN是由混乱矩阵所描述的样本数目，见表6.3。

表6.3 混乱矩阵

Observed	Predicted	
	Positives	Negatives
Positives	TP	FN
Negatives	FP	TN

6.4.4 独立测试实验

我们首先过滤Leukemia、Breast和Prostate数据集中的噪声基因（表6.2），将Leukemia、Breast过滤后的基因分别作为各数据集的特征基因，在Prostate数据集上挑选累积最大的前1000个作为Prostate的特征基因。然后利用每个数据集的训练子集和特征基因来训练Leukemia、Breast、Prostate的GCM模型，GCM模型的主分量设为15个，并分别训练正例样本子集和负例样本子集的CCM模型，CCM模型的癌症组分量也设为15个。最后利用EAGC算法给测试数据子集分类并计算性能评价指标Accu、Sn和Sp。上述分类实验重复10次，计算平均性能评价指标。并与Golub提出的加权投票法（Weighted Voting）[6]、SVM和KNN所获得的分类结果进行了比较，其中SVM采用径向基函数（RBF）作为核函数，KNN相似性度量函数采用Pearson相关系数，在加权投票法中利用50个特征基因，在SVM和KNN中利用SNR（信噪比）选取50个特征基因。加权投票法、SVM和KNN的分类实验同样重复10次，计算平均性能评价指标。

表6.4给出了独立测试实验的分类准确度Accu，从表6.4中可以看出在独立测试实验中相对于加权投票法和KNN，SVM取得了较好的分类准确度，在三个数据集上都优于其它两种分类器，并在Breast数据集上具有和EAGC相同的准确度。并且加权投票法、SVM和KNN在Prostate数据集上都没有取得较好的分类效果，缺乏泛化性。然而EAGC结合了具有互补性能的GCM模型和CCM模型，有效综合了GCM和CCM的解决方案，弥补了单个分类器的不足，在所有数据集上都取了较高的分类正确度。

表6.4 独立测试实验结果（Accu%）

	ALL-AML (Leukemia)	Breast Cancer	Prostate Cancer
Weighted Voting	85.3	78.9	67.6
SVM	94.1	94.7	75.6
KNN（K=5）	87.6	84.2	73.5
EAGC	97.1	94.7	91.4

图6.7给出了独立测试实验中灵敏度（Sn）和明确性（Sp）的情况。在Leukemia数据集中，Weighted Voting获得了较高的灵敏度，但明确性偏低。相对于KNN，Weighted Voting的灵敏度高于KNN，但明确性远低于KNN。在其它两个数据集中，SVM的明确性都高于Weighted Voting和KNN。在所有的数据数据集中，Weighted Voting方法的明确性偏低，KNN的灵敏度偏低，SVM则取得较好的

性能。EAGC由于综合了不同的解决方案以弥补了单个分类器的不足，取得了明显优于Weighted Voting，SVM和KNN的效果。

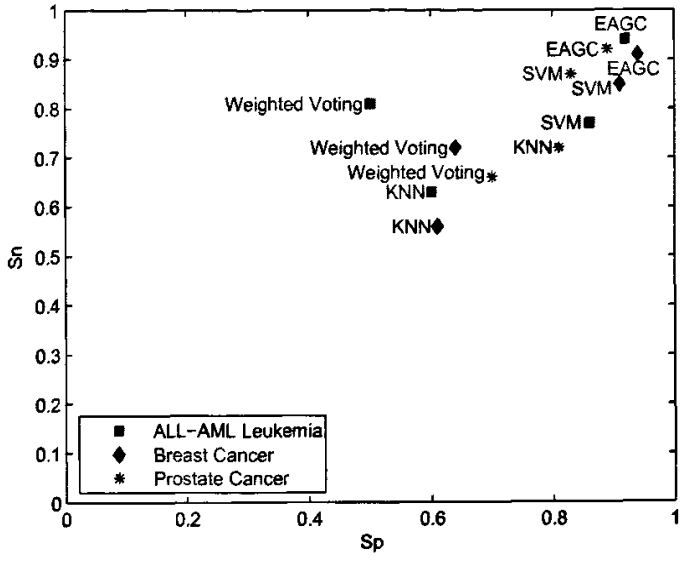

图6.7 独立测试实验结果（Sn和Sp）

6.4.5 交叉测试实验

在所有的六个数据集上进行交叉测试实验，包括“留一交叉检验”（Leave-One-Out Cross Validation，LOOCV）和“5折交叉检验”（Five-Fold Cross Validation，FFCV）。

在LOOCV中，每次从数据集中挑选一个不同的样本作为测试样本，其余样本作为训练数据集训练GCM模型和CCM模型，其中GCM模型和CCM模型的输入变量与独立测试实验相同，然后利用EAGC识别测试样本。重复该过程，直到每一个样本作为测试样本时为止。统计所有被正确识别的样本，并计算性能评价指标Accu、Sn和Sp。上述分类实验重复10次，计算平均性能评价指标。

在FFCV中，将数据集平均分为成五部分，每次挑选不同的一部分作为测试样本，其余样本作为训练数据集训练GCM模型和CCM模型，然后利用EAGC识别测试样本。重复分类过程五次，直到每部分样本作为测试样本时为止。统计所有被正确识别的样本，并计算性能评价指标Accu、Sn和Sp。上述分类实验重复10次，计算平均性能评价指标。并与加权投票法、SVM和KNN进行比较，其中

加权投票法、SVM和KNN的分类参数设置与独立测试实验相同，分类实验同样重复10次，并计算平均性能评价指标。

表6.5和表6.6 给出了交叉测试实验的分类准确度Accu，从表6.5 可以看出在LOOCV中，相对于加权投票法和KNN，SVM同样取得了较好的分类准确度，并且在DLBCL上取得最好的分类准确度97.8%，高于EAGC。然而在Prostate上Accu低于其它分类器。在所有数据集中KNN则相对表现了较好的泛化能力。EAGC除了在Prostate上分类准确度略低于SVM之处，在其它数据集上都高于加权投票法、KNN和SVM。同样从表6.6可以看出在FFCV中，SVM同样取得了较好的分类准确度，加权投票法次之。EAGC除了在DLBCL上分类准确度略低于SVM之处，在其它数据集上都高于加权投票法、KNN和SVM。

表6.5 LOOCV测试实验结果（Accu%）

	ALL-AML (Leukemia)	Breast Cancer	Prostate Cancer	DLBCL	Colon Tumor	Ovarian Cancer
Weighted Voting	90.3	77.3	70.4	88.6	93.5	63.5
SVM	95.3	84.6	68.9	97.8	91.9	82.4
KNN（K=5）	86.1	84.2	80.1	92.8	85.6	73.3
EAGC	99.3	97.9	91.4	93.4	96.8	92.6

表6.6 FFCV测试实验结果（Accu%）

	ALL-AML (Leukemia)	Breast Cancer	Prostate Cancer	DLBCL	Colon Tumor	Ovarian Cancer
Weighted Voting	88.2	76.5	74.5	88.3	90.1	73.2
SVM	95.1	83.7	63.5	93.2	90.1	83.1
KNN（K=5）	85.2	82.2	73.6	86.2	83.2	68.6
EAGC	98.4	95.2	90.2	91.0	94.3	90.4

图6.8和图6.9分别给出了留一交叉测试实验（LOOCV）和五折交叉测试实验（FFCV）中灵敏度（Sn）和明确性（Sp）的情况。在LOOCV实验中Weighted Voting在Colon数据集上取得了较好的分类效果，但在别的实验中的效率低于其它方法。除了在FFCV实验中的Ovarian数据集上，KNN的分类效果明显不及SVM之外，在其它的实验中则与SVM相当，并在Prostate的LOOCV和FFCV都优于SVM。说明在不同的数据集上的分类效果由于方法不同存在较大差别，在癌症样本检测中分类方法缺乏泛化性。从图6.8和图6.9可以看出，EAGC则表现的较好的泛化能力，在所有的实验中都表现了很好的分类性能。

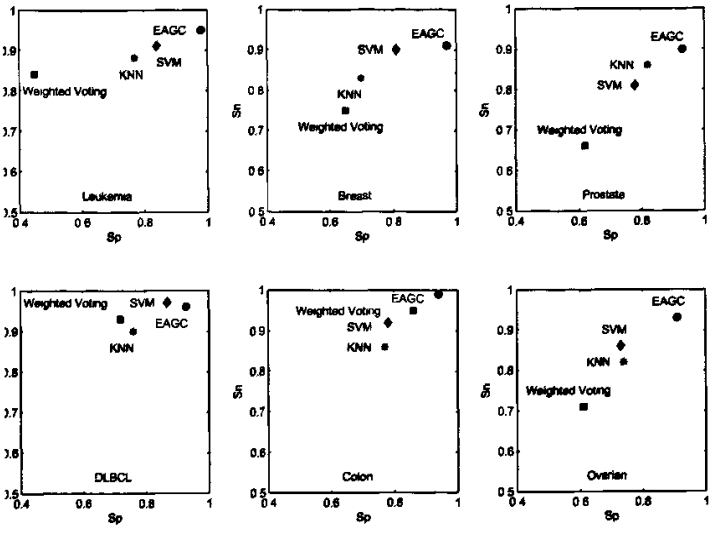

图6.8 LOOCV交叉测试实验结果（Sn和Sp）

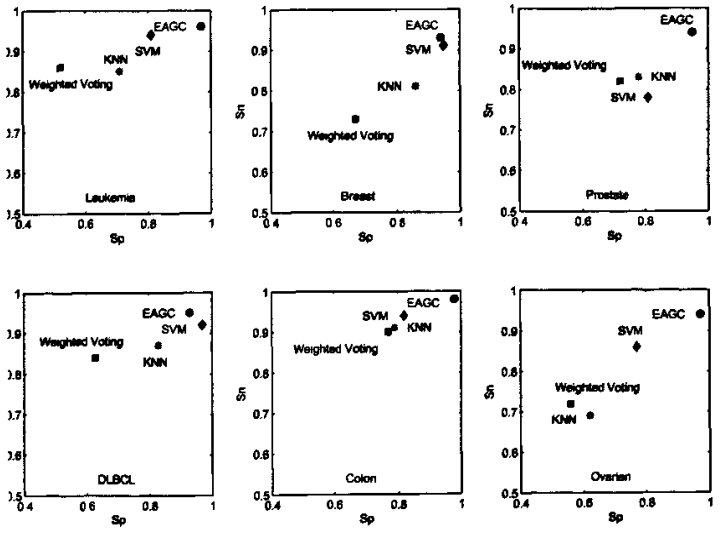

图6.9 FFCV交叉测试实验结果（Sn和Sp）

6.5 本章小结

在利用基因表达谱对患者样本进行癌症识别中，针对基因表达谱数据特点提出了两种癌症识别模型（GCM模型和CCM模型），并结合GCM模型和CCM模型的互补性，利用基于权值的投票组合策略提出一种组合分类算法（EAGC）。然后在Leukemia、Breast、Prostate、DLBCL、Colon、Ovarian等六个数据集上分别进行了独立测试实验和交叉测试实验。相对于传统算法过分依赖特征基因，缺乏泛化性的不足，EAGC有效综合了GCM和CCM识别模型的解决方案，弥补了单个分类器的不足，扩展了EAGC的解决方案，在所有数据集上都取了很好的分类性能。

第7章 基于显现模式的癌症分类算法研究

基于基因表达谱的分类技术对于疾病检测具有十分重要的研究意义。利用显现模式（Emerging Pattern，EP）的基因分类方法不仅可以识别癌症样本，同时可以挖掘出可解释的具有生物意义的基因模式，从基因协同调控的角度揭示癌症病理。本章首先讨论癌症检测中如何挖掘高效的显现模式，针对提取显现模式时在小样本情况下概率近似计算问题和显现模式的基因分割点的选择问题分别提出增强的显现模式（EPI）和高级显现模式（EPA），并在EPI和EPA两种基因模式上提出了两种癌症检测策略（CCEPI/CCEPA和KEPI/KEPA）。最后在急性白血病数据集上进行实验，结果表明EPI和EPA有效地提高了EP的癌症检测性能，两种检测策略都取得了较好的癌症识别率。

7.1 概述

基于微阵列的癌症基因表达谱记录了一组基因在多个癌症样本下的表达值，为研究癌症与基因之间存在的关联提供了新的途径。这种基因表达谱数据的一个显著特点是样本维数非常高，每个样本都记录了组织细胞中所有测试基因的表达水平，但实际上只有少数基因才真正同样本类别相关，这些包含了样本分类信息的基因被称为分类特征基因。如何有效分析癌症基因表达谱，并利用分类特征基因找出决定样本类别的一系列基因表达规则，对于癌症的诊断与治疗以及药物发现都具有重要意义，也是当前生物信息学研究的重点课题[6, 55, 99]。

聚类是基因分类中广泛使用的工具，常见的聚类算法有K-means、SOM等[108]。聚类是一种无监督的学习方法，而且没有特征选择机制，更好的识别癌症的途径是使用有监督的分类算法。Khan等将人工神经网络应用于疾病检测[15]。Guyon等采用SVM在白血病和结肠癌两种数据集上取得了较好的分类性能[14]。但是这些分类算法不能挖掘有生物学意义的基因表达规则，不利于生物学研究人员更好的理解疾病与基因间的本质联系。

最近Li等提出一种新的基于显现模式（Emerging Pattern，EP）的PCL基因分类算法[109]。通过显现模式捕获两种类别之间（例如：食用与有毒蘑菇之间、正常组织与病变组织之间等）基因表达模式的显著变化[110]。在不同类别的癌症数据集上频率变化越显著的模式具有越强的癌症识别能力。显现模式可以挖掘出具有生物意义的基因表达规则，能更好地研究癌症病理和基因的协同调控。PCL基因分类算法的基本思想是：基于熵的分离方法分离出分类特征基因及其基因表达的分割点，在此基础上挖掘出显现模式并构造PCL分类器。但该算法存在一些缺点：

首先，在计算基因含有的分类信息量即熵的过程中，通过样本集S_j中类别C_i出现的频率来估计S_j属于类别C_i的概率$P(C_i, S_j)$，这种估计在样本容量很小的癌症基因表达数据中存在偏差，此时的熵不能真实反映基因的分类能力；其次，通过相邻基因的平均值作为候选分割点的方法也存在不足；再次，PCL分类算法利用测试样本中所满足EPs的频率与训练样本集中EPs频率计算测试样本归属癌症类型的似然度（Likelihood）[109]时，没有考虑在似然度相同以及接近的情况下测试样本的所属类，而是简单地识别为DN类。

本章首先在显现模式的抽取中，通过增加虚拟样本来扩展总体的样本容量，利用贝叶斯方法来估计概率$P(C_i, S_j)$，并抽取增强的显现模式（EPI）。然后在显现模式中候选分割点的选择中，通过假设分割点满足高斯分布来增强候选分割点的可靠性，抽取高级的显现模式（EPA）。最后，在EPI和EPA基因模式上分别提出了两种癌症检测策略（CCEPI/CCEPA和KEPI/KEPA）。并在急性白血病数据集上进行实验，结果表明EPI和EPA有效地提高了EP的癌症检测性能，两种检测策略都取得了较好的癌症识别率。

7.2 预备知识

定义 7.1: （基因表达规则）$gene_i@[c,d]$ 是一个基因表达规则，表示$gene_i$ 的表达值属于[c,d]。

定义 7.2: （基因表达模式）基因表达模式X是一组基因表达的集合，X定义为：

$$X = \{gene_{i_1}@[a_{i_1}, b_{i_1}], gene_{i_2}@[a_{i_2}, b_{i_2}], \ldots, gene_{i_k}@[a_{i_k}, b_{i_k}]\} \tag{7.1}$$

其中$i_t \neq i_s, 1 \leq t, s \leq k$，如果$k \geq 1$，则称模式X的表达长度为k。

定义 7.3: （支持度）对于基因数据集D，则定义$supp_D(X)$为基因表达模式X在数据集D中的支持度，

$$supp_D(X) = count_D(X)/|D| \tag{7.2}$$

其中$count_D(X)$ 为D中满足模式X的数据元素个数，$|D|$为数据集D中元素的个数。

定义 7.4: （增长率）对于基因数据集D1和D2，利用GrowthRate(X)表示基因

表达模式X从数据集D1到数据集D2的增长率，

$$GrowthRate(X) = \begin{cases} 0, & \text{if } supp_{D_1}(X) = 0 \text{ and } supp_{D_2}(X) = 0 \\ \infty, & \text{if } supp_{D_1}(X) = 0 \text{ and } supp_{D_2}(X) \neq 0 \\ \frac{supp_{D_2}(X)}{supp_{D_1}(X)}, & \text{otherwise} \end{cases} \quad (7.3)$$

定义 7.5: （显现模式）支持度在基因数据集D1和基因数据集D2中具有显著变化的基因表达模式，被称之为显现模式（Emerging Pattern, EP）。

定义 7.6: （ρ-显现模式）设X是从数据集D1到D2的显现模式，对于增长率阀值$\rho > 1$，如果$GrouthRateX \geq \rho$则称模式X是从D1到D2的ρ-显现模式。

7.3 显现模式

7.3.1 基因数据集

急性白血病的基因表达谱（Leukemia Dataset）[1] 包含72个急性白血病样本，其中有47个急性淋巴性白血病(ALL)、25个急性骨髓性白血病(AML)。每个样本都包含了7129个基因的表达数据。

7.3.2 表达规则

对于分类特征基因的选择，文献[110]采用熵最小化的离散方法来产生基因的表达规则。

在样本集$\widetilde{S}$有k个类别$\widetilde{C}_1, \widetilde{C}_2, \ldots, \widetilde{C}_k$，假设$\widetilde{S}$被分为两个子集$\widetilde{S}_1$和$\widetilde{S}_2$，$p(\widetilde{C}_i, \widetilde{S}_j)$为样本子集$\widetilde{S}_j$中样本出现在$\widetilde{C}_i$中的概率。则子集$\widetilde{S}_j (j = 1,2)$的熵为：

$$Ent(\widetilde{S}_j) = -\sum_{i=1}^{k} P(\widetilde{C}_i, \widetilde{S}_j) \log(P(\widetilde{C}_i, \widetilde{S}_j)) \quad (7.4)$$

假设由特征A在点T划分获取子集$\widetilde{S}_1$和$\widetilde{S}_2$，则用$E(A, T; \widetilde{S})$描述此划分的类别信息熵（Class Information Entropy，CIE），

$$E(A, T; \widetilde{S}) = \frac{\left|\widetilde{S}_1\right|}{\left|\widetilde{S}\right|} Ent(\widetilde{S}_1) + \frac{\left|\widetilde{S}_2\right|}{\left|\widetilde{S}\right|} Ent(\widetilde{S}_2) \quad (7.5)$$

在特征A所有候选点中选取具有最小的类别信息熵作为分割点，然后在样本

[1] http://sdmc.lit.org.sg/GEDatasets/Datasets.html#ALL-AML_Leukemia

子集$\widetilde{S}_j(j=1,2)$上递归划分，并利用最小描述长度原则（Minimum Description Length Principle, MDL）作为停止条件，即满足：

$$Gain(A,T;\widetilde{S}) < \frac{\log_2(N-1)}{N} + \frac{\delta(A,T;\widetilde{S})}{N} \tag{7.6}$$

其中，

$$\begin{aligned} Gain(A,T;\widetilde{S}) &= Ent(\widetilde{S}) - E(A,T;\widetilde{S}) \\ \delta(A,T;\widetilde{S}) &= \log_2(3^k-2) - [k\cdot Ent(S) - k_1\cdot Ent(S_1) - k_2\cdot Ent(S_2)] \end{aligned} \tag{7.7}$$

式7.6中N为集合$\widetilde{S}$中特征值的数目，式7.7中k_i为$\widetilde{S}_i$中类别的数目。

通过该方法在白血病数据集中分离出866个分类特征基因[43, 103]，在PCL中，主要考虑前25个特征基因，表7.1列出了类别信息熵最小的前25个特征基因及其分割点。根据基因表达的分割点可以生成基因的表达规则，第1个基因的两个表达规则编号为1和2，分别为$gene_1@[0,994]$和$gene_1@[994,\infty]$，则第i个基因的两个表达规则编号为（$i\times 2-1$）和（$i\times 2$）。

7.3.3 PCL

在利用显现模式对基因进行分类中，Li等提出了一种基于集合似然的预测算法（Prediction by Collective Likelihood, PCL）[109]。首先从训练样本集中的ALL和AML子集中挖掘出对应的EPs，并根据EP的出现频率进行降序排列，分别记为$EPs(ALL)$和$EPs(AML)$，

$$\begin{aligned} EPs(ALL) = &\ \{EP_1^{(ALL)}, EP_2^{(ALL)}, \ldots, EP_i^{(ALL)}\} \\ EPs(AML) = &\ \{EP_1^{(AML)}, EP_2^{(AML)}, \ldots, EP_i^{(AML)}\} \end{aligned} \tag{7.8}$$

假设测试样本T中分别包含以下的$EPs_{ALL}(T)$和$EPs_{AML}(T)$，

$$\begin{aligned} EPs_{ALL}(T) = &\ \{EP_{i_1}^{(ALL)}, EP_{i_2}^{(ALL)}, \ldots, EP_{i_x}^{(ALL)}\} \\ EPs_{AML}(T) = &\ \{EP_{j_1}^{(AML)}, EP_{j_2}^{(AML)}, \ldots, EP_{j_y}^{(AML)}\} \end{aligned} \tag{7.9}$$

其中$i_1 < i_2 < \cdots < i_x \le i$，$j_1 < j_2 < \cdots < j_y \le j$。

然后依据$EPs(ALL)$、$EPs(AML)$、$EPs_{ALL}(T)$和$EPs_{ALL}(T)$分别计算测试

表7.1 离散方法分离出的前25个特征基因及分割点

Gene Index	Gene	Cut Point	Description
1	X95735	994.0	Zyxin
2	M55150	1346.0	FAH Fumarylacetoacetate
3	M31166	83.5	PTX3 Pentaxin-related gene, rapidly induced by IL-1 beta (PTX3)
4	M27891	1419.5	CST3 Cystatin C (amyloid angiopa- thy and cerebral hemorrhage)
5	X70297	339.0	CHRNA7 Cholinergic receptor, ni- cotinic, alpha polypeptide 7
6	P31483	80.5	Nucleolysin TIA-1
7	L09209	992.5	APLP2 Amyloid beta (A4) precursor -like protein 2
8	U46499	156.5	Glutathione S-transrerase, microso- mal
9	M16038	651.5	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog
10	M92287	1869.5	CCND3 Cyclin D3
11	D14874	185.0	ADM Adrenomedullin
12	U50136	1341.0	Leukotriene C4 synthase (LTC4S) gene
13	U22376	1423.0	C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds
14	M27783	197.5	ELA2 Elastatse 2, neutrophil
15	D88422	658.0	CYSTATIN A
16	M21551	398.5	Neuromedin B mRNA
17	M23197	401.5	CD33 CD33 antigen (differentiation antigen)
18	U46751	2909.5	Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA
19	Y12670	911.0	LEPR Leptin receptor
20	M83652	541.5	PFC Properdin P factor, complement
21	M98399	245.0	CD36 CD36 antigen (collagen type I receptor, thrombospondin receptor)
22	M54995	156.5	PPBP Connective tissue activation peptide III
23	U02020	381.5	Pre-B cell enhancing factor (PBEF) mRNA
24	M31523	415.5	TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)
25	M81933	128.0	CDC25A Cell division cycle 25A

样本T属于癌症亚型ALL和AML的预测似然度$score_{ALL}(T)$和$score_{AML}(T)$，

$$\begin{aligned} score_{ALL}(T) &= \sum_{m=1}^{k} \frac{frequency(EP_{jm}^{ALL})}{frequency(EP_{m}^{ALL})} \\ score_{AML}(T) &= \sum_{m=1}^{k} \frac{frequency(EP_{jm}^{AML})}{frequency(EP_{m}^{AML})} \end{aligned} \tag{7.10}$$

其中k是用于计算预测似然度选取的EP的数目。

最后，利用预测似然度$score_{ALL}(T)$和$score_{AML}(T)$判定样本T的癌症类型。如果$score_{ALL}(T) > score_{AML}(T)$，则样本$T \in ALL$；否则$T \in AML$。

7.4 基于增强显现模式的癌症检测

7.4.1 增强显现模式

由式7.4和7.5可知，生成基因表达规则的分割所对应的类别信息熵越小，则该基因分割具有越强的类别识别能力。在理想情况下，如果基因表达规则可以完全将两类癌症类型分隔开，则$E(A,T;\widetilde{S}) = 0$。

Li等[109, 111, 112]通过式7.4来计算熵$Ent(\widetilde{S}_j)$时，利用样本集S_j中类别C_i出现的频率替代概率$P(C_i,S_j)$。由伯努利大数定理可知，样本集S_j中类别C_i出现的频率在数据集中癌症样本容量趋于无穷大时则收敛于概率$P(C_i,S_j)$。而在现实情况下，由于基因数据集中的样本一般都很小，仅为$10^1 \sim 10^2$，并且经过研究人员手工挑选，存在较大的非客观因素影响。导致样本集S_j中类别C_i出现的频率与概率$P(C_i,S_j)$存在较大偏差。在此，引入贝叶斯估计（m—估计）：

$$P(C_i,S_j) = \frac{n_i + mp}{n + m} \tag{7.11}$$

其中，

$$\begin{aligned} n_i &= |\{s\}| \mid s \in S_j \wedge s \in C_i \\ n &= |\{s\}| \mid s \in S_j \end{aligned} \tag{7.12}$$

式7.11中m是一个虚拟样本集的容量，m是一个常量。在理想情况下，p应为真实概率，但在缺少其他信息时选择p的常用方法是假定其为均匀分布概率。式7.11利用m个均匀分布的虚拟样本来修正概率$P(C_i,S_j)$。上式可以理解为在观察的n个癌症样本中，添加m个按p分布的虚拟样本以扩展实际的观察样本。一般情况下，m不宜过大，否则导致假定概率p在式7.11中占主导，使得估计的概率$P(C_i,S_j)$收敛于p。本章中令$m = |S_j|/4$，$p = 1/2$。

根据式7.4-7.12，本节从Leukemia的7129个基因中筛选分类特征基因以及基因表达的分割点，依据最小描述长度原则选择出857个特征基因以及对应这些基因表达的分割点。然后通过定义7.1，我们生成基因的表达规则，称之为增强的基因表达规则。尽管现在只有857个特征基因，但产生基因表达模式的计算量仍然巨大。在此，我们主要考虑前25个特征基因，表7.2列出了分类信息熵最小的前25个基因。基因表达的分割点将这25个特征基因分割成50个基因表达区

表7.2 基于m-估计的离散方法分离出的前25个特征基因及分割点

Gene Index	Gene	Cut Point	Description
1	X95735	994.0	Zyxin
2	M55150	1346.0	FAH Fumarylacetoacetate
3	M31166	83.5	PTX3 Pentaxin-related gene, rapidly induced by IL-1 beta (PTX3)
4	M27891	1419.5	CST3 Cystatin C (amyloid angiopa-thy and cerebral hemorrhage)
5	X70297	339.0	CHRNA7 Cholinergic receptor, ni- cotinic, alpha polypeptide 7
6	P31483	80.5	Nucleolysin TIA-1
7	L09209	992.5	APLP2 Amyloid beta (A4) precursor -like protein 2
8	U46499	156.5	Glutathione S-transrerase, microso- mal
9	D88422	658.0	CYSTATIN A
10	M21551	398.5	Neuromedin B mRNA
11	M23197	401.5	CD33 CD33 antigen (differentiation antigen)
12	U46751	2909.5	Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA
13	U50136	1987.0	Leukotriene C4 synthase (LTC4S) gene
14	Y12670	911.0	LEPR Leptin receptor
15	M27783	357.5	ELA2 Elastatse 2, neutrophil
16	M83652	541.5	PFC Properdin P factor, complement
17	M98399	245.0	CD36 CD36 antigen (collagen type I receptor, thrombospondin receptor)
18	M54995	156.5	PPBP Connective tissue activation peptide III
19	U02020	381.5	Pre-B cell enhancing factor (PBEF) mRNA
20	M31523	415.5	TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)
21	D14874	185.0	ADM Adrenomedullin
22	M16038	651.5	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog
23	M92287	1869.5	CCND3 Cyclin D3
24	U22376	1423.0	C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds
25	J05243	245.0	SPTAN1 Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)

间，由定义7.1，增强的基因表达规则由基因表达区间产生。为了区分这50个规则项以及后续算法的简化，我们将第i个基因的两个增强的基因表达规则编号

为$(i \times 2-1)$和$(i \times 2)$。这样的编号便于我们标识基因表达模式，例如，基因表达模式{2 11}就代表$\{geneX95735@[994,+\infty], geneP31483@(-\infty,80.5)\}$。

我们以未采用贝叶斯估计所得的25个分类特征基因及其分割点作为参照，对比表7.1和表7.2，发现引入贝叶斯估计后88%的分类特征基因及其基因表达规则没有变化。但是引入贝叶斯估计后利用虚拟样本对概率进行近似估计，使得这88%的分类特征基因的排序相对发生了一些微小变化。不过对于辨别能力最强，即基因分割点划分信息熵的前8个分类特征基因及其表达规则，两种方法达到了100%的吻合。另有12%的分类特征基因及其表达规则发生了较大变化，表7.3和表7.4则分别列出了这部分基因在测试集中的辨别能力。

表7.3 在测试集上显现模式抽取中的三个特征基因在分割点的类别信息熵

Gene	Cut Point	CIE
U50136	1341.0	0.57621
M27783	197.5	0.58595
M81933	128.0	0.59916

表7.4 在测试集上增强显现模式抽取中的三个特征基因在分割点的类别信息熵

Gene	Cut Point	CIE
U50136	1987.0	0.45446
M27783	357.5	0.62831
J05243	245.0	0.45684

由表7.3和表7.4可知，测试集中除了基因M27783在分割点197.5比在分割点357.5的类别信息熵要小6.7%外，基因U50136在分割点1341.0的类别信息熵比在分割点1987.0的类别信息熵高26.8%。基因M81933与基因J05243分别为表7.1和表7.2 的特征基因的第25位，但基因M81933比基因J05243的分类信息熵高31.2%。同样由表7.5和表7.6可以发现，增强的基因表达规则U50136@(-∞,1987.0)可以完全正确的区分ALL和AML，同时J05243@[245.0,+∞)也很好地识别了ALL和AML。而与之对应的基因表达规则U50136@(-∞,1341.0)和M81933@[128.0,+∞)的辨别能力则相对较差。

在100个基因表达规则的基础上，所有潜在的增强的显现模式（Improved Expression Pattern，EPI）有$2^{100}-1$个，即使只选择表达长度小于9的EPIs，也存在$C_{7129}^{9} \times 2^{9}$个EPIs。显然，采用朴素的穷举法是不可行的。在此，我们采用一种基于边界LargeBorder[111,112]的EPI挖掘算法。LargeBorder算法详细描述参见文献[111]。首先利用Max-Miner算法挖掘基因数据集D1和D2上的基因表达规则最大

表7.5 测试集样本在基因表达规则上的分布情况

Gene	Expression Rule	ALL	AML	Total
U50136	U50136@(-∞,1341.0)	2	7	9
U50136	U50136@[1341.0,+∞)	18	7	25
M27783	M27783@(-∞,197.5)	9	12	21
M27783	M27783@[197.5,+∞)	11	2	13
M81933	M81933@(-∞,128.0)	15	5	20
M81933	M81933@[128.0,+∞)	5	9	14

表7.6 测试集样本在增强基因表达规则上的分布情况

Gene	Expression Rule	ALL	AML	Total
U50136	U50136@(-∞,1987.0)	0	7	7
U50136	U50136@[1987.0,+∞)	20	7	27
M27783	M27783@(-∞,357.5)	8	10	18
M27783	M27783@[357.5,+∞)	12	4	16
J05243	J05243@(-∞,245.0)	19	5	24
J05243	J05243@[245.0,+∞)	1	9	10

边界$LargeBorder_{\theta}(D1)$和$LargeBorder_{\delta}(D2)$，

$$\begin{aligned} LargeBorder_{\theta}(D1) = & < \{\phi\}, \{D_1, D_2, \ldots, D_m\} > \\ LargeBorder_{\delta}(D2) = & < \{\phi\}, \{E_1, E_2, \ldots, E_n\} > \end{aligned} \tag{7.13}$$

换言之，$LargeBorder_{\theta}(D1)$是数据集D1中支持度为θ的基因表达规则的最大频繁项集，$LargeBorder_{\delta}(D2)$是数据集D2中支持度为δ的基因表达规则的最大频繁项集，

$$\begin{aligned} LargeBorder_{\theta}(D1) &= \{X | X\ is\ EPI, and\ supp_{D1}(X) \geq \theta\} \\ LargeBorder_{\delta}(D2) &= \{X | X\ is\ EPI, and\ supp_{D2}(X) \geq \delta\} \end{aligned} \tag{7.14}$$

并且满足$\theta = \rho \times \delta$。

然后根据每个D_j的边界差（BorderDiff）产生增强的基因表达模式(EPIs)。

$$\begin{aligned} EPIs &= \bigcup_j (EPIBorder(D_j)) \\ EPIBorder(D_j) &= BorderDiff(< \{\phi\}, \{D_j\} >, < \{\phi\}, \{E'_1, E'_2, \ldots, E'_k\} >) \end{aligned} \tag{7.15}$$

式中$E'_i = E_i \cap D_j$

在急性白血病的两种亚型（急性淋巴性白血病（ALL）和急性骨髓性白血病（AML））中，根据上述方法获取了50个分类特征基因的100个基因表达规

则，并采用基于边界LargeBorder的EPI挖掘算法分别抽取ALL和AML中增长率大于等于1的EPI，在ALL中挖掘了12043个EPI，在AML中挖掘了9654个EPI。这里在表7.7、表7.8中分别列出ALL和AML中增长率最大的前25个EPI。

表7.7 ALL样本中增长率最大的前25个EPIs

EPI Index	EPIs	Frequency in ALL	Frequency in AML
1	{11 15}	0.97872	0
2	{11 13 15}	0.97872	0
3	{11 15 61}	0.97872	0
4	{11 13 15 61}	0.97872	0
5	{13 15}	0.97872	0
6	{13 15 61}	0.97872	0
7	{1 73}	0.95745	0
8	{1 87}	0.95745	0
9	{1 11 73}	0.95745	0
10	{1 11 87}	0.95745	0
11	{1 13 21}	0.95745	0
12	{1 13 73}	0.95745	0
13	{1 13 87}	0.95745	0
14	{1 61 73}	0.95745	0
15	{1 61 87}	0.95745	0
16	{1 73 87}	0.95745	0
17	{1 11 13 21}	0.95745	0
18	{1 11 13 73}	0.95745	0
19	{1 11 13 87}	0.95745	0
20	{1 11 61 73}	0.95745	0
21	{1 11 13 87}	0.95745	0
22	{1 11 73 87}	0.95745	0
23	{1 13 21 61}	0.95745	0
24	{1 13 21 89}	0.95745	0
25	{1 13 61 73}	0.95745	0

由表7.7和表7.8，可以发现：

1. 利用基于边界LargeBorder的EPI挖掘算法抽取的增强的基因表达规则具有很强的癌症辨别能力，具有很高的支持度。

2. 在两种癌症亚型中增长率最大的前25个EPI均至少包含有一个表7.2中的特征基因，从而说明表7.2中的特征基因具有很强的癌症辨别能力。

3. 一些增强的基因表达模式具有很强的癌症辨别能力，例如模式{11 13 15 61}，其中包含有4个特征基因和4个增强的基因表达规则，该模式在ALL样本集中达到0.97872的支持度，意味着几乎每一个ALL样本的表达谱都满足由这4个基

表7.8 AML样本中增长率最大的前25个EPIs

EPI Index	EPIs	Frequency in ALL	Frequency in AML
1	{6 8}	0	0.92
2	{6 9}	0	0.92
3	{6 8 9}	0	0.92
4	{6 8 17}	0	0.92
5	{6 8 64}	0	0.92
6	{6 9 17}	0	0.92
7	{6 9 64}	0	0.92
8	{6 8 9 17}	0	0.92
9	{6 8 9 64}	0	0.92
10	{6 8 17 64}	0	0.92
11	{6 9 17 64}	0	0.92
12	{6 8 9 17 64}	0	0.92
13	{8 9}	0	0.92
14	{8 9 17}	0	0.92
15	{8 9 64}	0	0.92
16	{8 9 17 64}	0	0.92
17	{9 64}	0	0.92
18	{9 17 64}	0	0.92
19	{4 6}	0	0.92
20	{4 17}	0	0.92
21	{4 20}	0	0.92
22	{4 6 64}	0	0.92
23	{4 17 64}	0	0.92
24	{4 20 64}	0	0.92
25	{23 64}	0	0.92

因构成的模式，并且没有一个AML样本的表达谱满足该模式。

4. EPI支持度不受其表达长度影响，表达长度越小的模式的支持度不一定大于表达长度越大的模式的支持度，反之也成立。例如模式{12}在AML样本中的支持度为0.84，而模式{6 8 9 17 64}的支持度为0.92。

7.4.2 基于EPI的癌症识别算法

利用集合似然的预测算法（PCL）可以利用EP对患者样本进行有效地癌症类型预测。但存在以下的不足：首先PCL方法没有充分考虑作为参照的训练样本集EPs自身频率对似然度的影响；其次没有充分考虑$score_{ALL}(T) = score_{AML}(T)$以及$score_{ALL}(T)$与$score_{AML}(T)$接近的情况下如何识别样本T。在PCL算法中，当$score_{ALL}(T) = score_{AML}(T)$时，PCL识别T为AML亚型；

当$score_{ALL}(T)$与$score_{AML}(T)$接近时，PCL对T类别判断不足。

例如，在表7.7、表7.8中，我们发现训练样本中两类子集的EPIs的频率并不相同，但是同一子集则连续存在若干个频率相同的EPIs，那么容易出现类似如下的情况。设k=1，假设$frequency(EP_1^{(ALL)}) = 0.97872$，$frequency(EP_1^{(AML)}) = 0.92$，如果样本T中同时满足$EP_1^{(ALL)}$和$EP_1^{(AML)}$，那么$frequency(EP_{i_1}^{(ALL)}) = 0.97872$，$frequency(EP_{j_1}^{(AML)}) = 0.92$。根据公式7.10可知，$score_{ALL}(T) = score_{AML}(T)$，但是显然$EP_1^{ALL} > EP_1^{AML}$，似然度应偏向频率大的ALL类。同样，对于似然度不同的情况，在利用样本T中EPIs与训练样本集EPIs计算似然度时，考虑对应的训练样本集中的EPIs自身的频率对似然度的影响也是合理的。于是基于增强显现模式的癌症分类算法中采用如下的似然度计算方法：

$$\begin{aligned} score_{ALL}(T) &= \sum_{m=1}^{k}\left(e^{\frac{frequency(EPI_{im}^{(ALL)})}{frequency(EPI_m^{(ALL)})}} + \frac{frequency(EPI_m(ALL))}{2k}\right) \\ score_{AML}(T) &= \sum_{m=1}^{k}\left(e^{\frac{frequency(EPI_{jm}^{(AML)})}{frequency(EPI_m^{(AML)})}} + \frac{frequency(EPI_m(AML))}{2k}\right) \end{aligned} \tag{7.16}$$

既在似然度计算中保持PCL中原来似然度的主导作用，又增加训练样本集EPIs自身频率对似然度的影响。针对以上分析，本节提出了一种基于增强显现模式的癌症分类算法（Cancer Classification with EPI，CCEPI），CCEPI算法的详细描述见算法7.1。

从上面的分析，可以知道基于集合似然的预测算法（PCL）没有较好地处理$score_{ALL}(T) = score_{AML}(T)$时，以及$score_{ALL}(T)$和$score_{AML}(T)$相近时如何预测样本T的癌症亚型的问题，由于KNN算法在分类中分析和综合了距离邻近的K个对象对测试样本的识别，是一种协同的分类方法，从而避免在似然度相同和相近的情况下过于轻率的判别。在此，借鉴KNN的思想，在基于EPI的癌症识别中提出一种K-EPI近邻算法（K-Emerging Pattern nearest neighbors，KEPI）。其基本思想是利用基于LargeBorder的EPI挖掘算法从数据集Leukemia中抽取出ALL和AML的EPIs，EPIs(ALL)和EPIs(AML)，然后对于Leukemina中的训练样本S，假设S的癌症类型是SD，$SD \in \{ALL, AML\}$，抽取满足EPIs(SD)的$EPIs_{SD}(S)$，

$$EPIs_{SD}(S) = \{EPI_{i_1}^{(SD)}, EPI_{i_2}^{(SD)}, \ldots, EPI_{i_x}^{(SD)}\} \tag{7.17}$$

同样，对于测试样本T抽取满足EPIs(SD)的$EPIs_{SD}(T)$，

$$EPIs_{SD}(T) = \{EPI_{i'_1}^{(SD)}, EPI_{i'_2}^{(SD)}, \ldots, EPI_{i'_x}^{(SD)}\} \tag{7.18}$$

算法 7.1 增强显现模式的癌症分类算法（CCEPI）

Require: 癌症亚型训练数据集ALL和AML，测试样本T

Outputting: T的癌症亚型

1: 利用$D_{Leukemia}$表示ALL和AML的所有样本集合，$D_{Leukemia} = ALL \cup AML$；
2: 利用7.4.1节中的方法分割基因的表达水平，根据基因分割点的CIE对基因排序，并挖掘增强的基因表达规则；
3: 利用Max-Miner算法分别挖掘基因数据集ALL和AML中基因表达规则的最大边界$LargeBorder_{\theta}(ALL) =< \{\phi\}, \{D_1^{ALL}, D_2^{ALL}, \ldots, D_m^{ALL}\} >$以及最大边界$LargeBorder_{\delta}(AML) =< \{\phi\}, \{D_1^{AML}, D_2^{AML}, \ldots, D_n^{AML}\} >$；
4: **for** Each of {ALL, AML} **do**
5: SD是处理数据的癌症亚型，$AD = \{ALL, AML\} - SD$，k是最大边界中右边界中模式的数目；
6: **for** j=1 to k **do**
7: 根据边界差$BorderDiff(< \{\phi\}, \{D_j^{SD}\} >, < \{\phi\}, \{D_1', D_2', \ldots, D_k'\} >)$生成EPI的表达边界$EPIBorder(D_j^{SD})$，其中$D_i' = C_i^{AD} \cap D_j^{SD}$；
8: $EPIs^{(SD)} = \bigcup_j (EPIBorder(D_j^{SD}))$；
9: **end for**
10: **end for**
11: 抽取T中存在于$EPIs^{(ALL)}$和$EPIs^{(AML)}$中的EPIs，并根据在ALL和AML中的频率排序，$EPIs_{ALL}(T) = \{EPI_{i_1}^{(ALL)}, EPI_{i_2}^{(ALL)}, \ldots, EPI_{i_x}^{(ALL)}\}$，$EPIs_{AML}(T) = \{EPI_{j_1}^{(AML)}, EPI_{j_2}^{(AML)}, \ldots, EPI_{j_y}^{(AML)}\}$；
12: 根据式7.16计算样本T属于ALL和AML的似然度$score_{ALL}(T)$和$score_{AML}(T)$；
13: 预测样本T的癌症类型，如果$score_{ALL}(T) > score_{AML}(T)$，则样本$T \in ALL$；否则$T \in AML$

然后，根据$EPIs_{SD}(S)$和$EPIs_{SD}(T)$计算样本S和样本T之间的似然度score(S,T)，

$$score(S,T) = \sum_{m=1}^{k} \frac{frequency(EPI_{i_m}^{SD})}{frequency(EPI_m^{SD})} \tag{7.19}$$

其中k是用于计算预测似然度选取的EPI的数目。

最后将Leukemia中每一个S的Score(S,T)按降序排列，并挑选K个具有最大Score(S,T)训练样本S'识别样本T的癌症类型，

$$T \in \begin{cases} ALL & if\ |\{S'\}| > \lfloor K/2 \rfloor \wedge S' \in ALL \\ AML & else \end{cases} \tag{7.20}$$

给定Leukemia中的两个训练子集ALL和AML以及测试样本T，KEPI算法的详细描述见算法7.2。KEPI以训练样本S与测试样本T的似然度score(S,T)衡量训练样本S与测试样本T的距离，通过距离T最近的前K个训练样本的投票预测T的癌症类别。KEPI算法综合似然度score(S,T)最高的前K个训练样本对T的识别，克

服PCL分类算法的不足。在KEPI算法中，K设为奇数。

算法 7.2 基于增强显现模式的癌症识别算法（KEPI算法）

Require: 癌症亚型训练数据集ALL和AML，K，测试样本T
Outputting: T的癌症亚型

1: 利用$D_{Leukemia}$表示ALL和AML的所有样本集合，$D_{Leukemia} = ALL \cup AML$;
2: 利用7.4.1节中的方法分割基因的表达水平，根据基因分割点的CIE对基因排序，并挖掘增强的基因表达规则;
3: 利用Max-Miner算法分别挖掘基因数据集ALL和AML中基因表达规则的最大边界$LargeBorder_\theta(ALL)$以和最大边界$LargeBorder_\delta(AML)$;
4: 利用基于$LargeBorder$的EPIs的挖掘算法分别抽取$EPIs^{(ALL)}$和$EPIs^{(AML)}$;
5: 抽取样本T中符合$EPIs^{(ALL)}$ $EPIs^{(AML)}$的EPIs，$EPIs_{ALL}(T)$和$EPIs_{AML}(T)$，$EPIs_{TD}(T) = \{EPI_{i'_1}^{(TD)}, EPI_{i'_2}^{(TD)}, \ldots . EPI_{i'_x}^{(TD)}\}$，$TD \in \{ALL, AML\}$;
6: **for** Each S $\in$ {ALL, AML} **do**
7: 设SD是样本S的类别标识，k是最大边界中右边界中模式的数目;
8: 抽取样本S中符合$EPIs^{(SD)}$的EPIs，$EPIs_{SD}(S) = \{EPI_{i'_1}^{(SD)}, EPI_{i'_2}^{(SD)} \ldots . , EPI_{i'_x}^{(SD)}\}$，$SD \in \{ALL. AML\}$;
9: 根据式7.19计算样本S和T的似然度，并选择K个与T具有最高似然度的样本{S'};
10: **end for**
11: **for** Each $S' \in \{S'\}$ **do**
12: **if** $S' \in ALL$ **then**
13: Count(ALL)++;
14: **else**
15: Count(AML)++;
16: **end if**
17: **end for**
18: **if** Count(ALL)>Count(AML) **then**
19: 样本S识别为ALL;
20: **else**
21: 样本S识别为AML;
22: **end if**

7.4.3 实验结果与分析

采用“留一交叉检验法”（Leave One Out Cross Validation，LOOCV）[103]进行样本类型的识别实验。实验环境同3.5。在Leukemia数据集的72个样本中选择一个样本作为测试样本T，其余71个样本作为训练数据集，利用CCEPI和KEPI预测测试样本T的癌症类型。然后又从Leukemia中重新选择一个没有经过测试的样本作为测试样本，重复上述过程，直到所有样本都经过CCEPI和KEPI 测试为止。统计所有识别结果正确的样本数，并计算算法的分类正确率。在似然度$score_{ALL}(T)$，$score_{AML}(T)$和$score(S,T)$ 的计算

中，k设为20。在KEPI中，$K=5$。并利用PCL，SVM和KNN在Leukemia数据集上进行LOOCV测试实验。在PCL中似然度$score_{ALL}(T)$，$score_{AML}(T)$的计算中k设为20。在SVM和KNN训练中，SVM采用径向基函数（RBF）作为核函数，KNN相似性度量函数采用Pearson相关系数，$K=5$，选择7.4.1节选取的50基因作为特征基因。上述实验重复10遍，并计算平均正确率，表7.9列出了实验结果。

表7.9 试验结果比较

Classification	Gene Patterns including Genes, EPs, EPIs	Results(%)
CCEPI	20	96.2
KEPI(K=5)	20	97.5
PCL	20	93.1
SVM	50	94.5
KNN(K=5)	50	85.3

从表7.9可以看出，CCEPI和KEPI利用20个EPIs来计算似然度，并取得了很好的分类精度，比PCL分别提高了3.1%、4.4%。原因在于，在抽取EPI时，引入贝叶斯估计（m一估计），增加虚拟样本空间，扩展了样本容量，使得抽取的特征基因和基因表达规则更具有辨别性，有效消除了噪声对基因表达模式的影响。同时，由于在ALL和AML中存在较多的EPs的支持度相同，而PCL没有考虑似然度接近和相同时如何识别测试样本类别的问题。在CCEPI中，通过引入EPI的自身频率来修改似然度；在KEPI中，则通过K个样本的投票来决定测试样本的类别。相对于KNN，KEPI大大提高了样本识别的正确率，将近11%。

7.5 基于高级显现模式的癌症检测

7.5.1 高级显现模式

在EP的挖掘过程中，产生基因表达规则的基因表达分割点的方法是将基因表达排序后，利用相邻基因表达的平均值作为候选的基因表达分割点。然后利用7.3.2节的方法挖掘基因表达规则。然而并不能保证两个相邻表达的平均值一定就是最好的分割点，也不一定满足最小描述长度的原则。在此，我们做出如下的假设：如果某个分类特征基因排序后的表达谱为$(gene_{sort}(1), gene_{sort}(2), \ldots, gene_{sort}(n))$，满足$gene_{sort}(1) < gene_{sort}(2) < \ldots < gene_{sort}(n)$，假设产生基因表达规则的分割点$cutpoint'$满足$genesort(i-1) < cutpoint' < genesort(i+2)$，并且$cutpoint'$是$(genesort(i-1)genesort(i+2))$上的正态分布。并定义$cutpoint'$为随机分割点（Radomn Cut Point，RCP）。根据这个设

想，我们同样利用7.4.1节的方法抽取25个分类特征基因和基因表达的分割点，并称分割点生成的基因表达规则为高级基因表达规则。图7.1 给出了25个基因在基因表达分割点的CIE，并对比了生成EP时的25个特征基因在分割点（称为基因表达原分割点，Original Cut Point）的CIE。

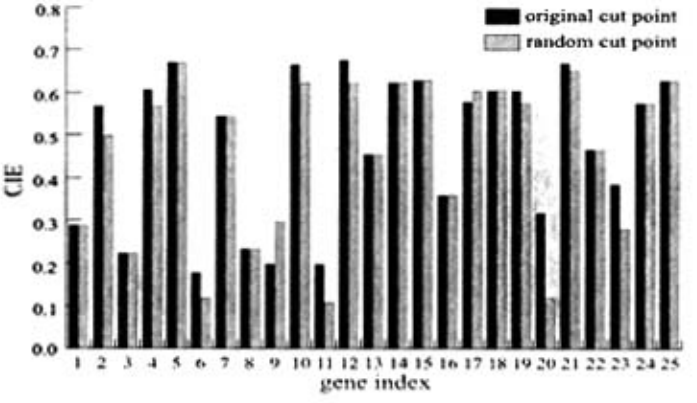

图7.1　随机割点与原割点在测试集中分类性能比较

随机分割点使得分割点的选取具有了局部随机性，原分割点可视为是随机割点的一种特殊情况。在实验中我们发现40%的分类特征基因在采用随机割点后对测试集的分类能力有不同程度的提高，48%的随机割点的能力与原割点持平，另有12% 的随机割点的分类能力比原始割点有小幅下降，因此随机分割点有效地提高了基因表达规则的辨别能力。

然后我们利用基于最大边界的显现模式挖掘算法从高级基因表达规则中挖掘高级显现模式（Advanced Emerging Pattern，EPA）。对于急性白血病，分别抽取了ALL，AML中增长率大于等于1的EPA，$EPAs(ALL)$和$EPAs(AML)$，其中分别有17065、5654个EPA。表7.10和表7.11 分别列出ALL和AML中增长率最大的前20个EPAs。从表中可以发现EPAs比EPIs具有更高的支持度。

7.5.2 基于高级显现模式的癌症识别算法

相对于基于EP的集合似然的癌症预测算法，本节也提出了两种癌症识别算法：基于高级显现模式的癌症分类算法（CCEPA）和基于高级显现模式的癌症识别算法（KEPA）。CCEPA、KEPA的基本思想分别与CCEPI和KEPI的基本思想是一致的，只是在算法中利用EPA来代替EPI，因此算法的详细描述在此不再重复。

表7.10 ALL中增长率最大的前20个EPAs

EPA Index	EPAs	ALL Support	AML Support
1	{1}	1	0
2	{1 5}	1	0
3	{1 7}	1	0
4	{1 17}	1	0
5	{1 19}	1	0
6	{1 21}	1	0
7	{1 23}	1	0
8	{1 25}	1	0
9	{1 27}	1	0
10	{1 29}	1	0
11	{1 31}	1	0
12	{1 33}	1	0
13	{1 5 7}	1	0
14	{1 5 17}	1	0
15	{1 5 19}	1	0
16	{1 5 21}	1	0
17	{1 5 23}	1	0
18	{1 5 25}	1	0
19	{1 5 27}	1	0
20	{1 5 29}	1	0

7.5.3 实验结果与分析

采用“留一交叉检验法”（LOOCV）进行样本类型的识别实验[103]。实验环境同3.5。本次实验分为两部分：第一部分，在CCEPA和KEPA中，计算似然度$score_{ALL}(T)$，$score_{AML}(T)$和$score(S,T)$时取$k = 20$，令KEPA中$K = 5$。并利用PCL，SVM和KNN在Leukemia数据集上进行LOOCV测试实验。在PCL中，计算似然度$score_{ALL}(T)$，$score_{AML}(T)$时设$k = 20$。在SVM和KNN训练中，SVM采用径向基函数（RBF）作为核函数，KNN相似性度量函数采用Pearson相关系数，$K = 5$，选择7.4.1节选取的50个基因作为特征基因。上述实验重复10遍，并计算平均正确率。第二部分，在CCEPA和KEPA中，计算似然度$score_{ALL}(T)$，$score_{AML}(T)$和$score(S,T)$时取$k = 10$，令KEPA中$K = 5$。同样利用PCL，SVM和KNN在Leukemia数据集上进行LOOCV测试实验。在PCL中，计算似然度$score_{ALL}(T)$，$score_{AML}(T)$时$k = 10$。在SVM和KNN训练中，SVM采用径向基函数（RBF）作为核函数，KNN相似性度量函数采用Pearson相关系数，$K = 5$，选择7.4.1节选取的25个基因作为特征基因，并且设KNN中的K为5。上述实验重复10遍，并计算平均正确率，表7.12列出了实验结果。

表7.11 AML中增长率最大的前20个EPAs

EPA Index	EPAs	ALL Support	AML Support
1	{2}	0	1
2	{2 4}	0	1
3	{2 9}	0	1
4	{2 12}	0	1
5	{2 14}	0	1
6	{2 16}	0	1
7	{2 42}	0	1
8	{2 44}	0	1
9	{2 45}	0	1
10	{2 47}	0	1
11	{2 4 9}	0	1
12	{2 4 12}	0	1
13	{2 4 14}	0	1
14	{2 4 16}	0	1
15	{2 4 42}	0	1
16	{2 4 44}	0	1
17	{2 4 45}	0	1
18	{2 4 47}	0	1
19	{2 9 12}	0	1
20	{2 9 14}	0	1

表7.12 试验结果比较

Classification	Gene Patterns including Genes, EPs, EPAs	Results(%)
CCEPA	20	97.2
KEPA(K=5)	20	98.6
PCL	20	93.5
SVM	50	95.2
KNN(K=5)	50	84.5
CCEPA	10	96.5
KEPA(K=5)	10	97.1
PCL	10	90.0
SVM	25	83.2
KNN(K=5)	25	80.5

从表7.12可以看出，CCEPA和KEPA利用20个EPAs来计算似然度，并取得了很好的分类精度，并且比CCEPI和KEPI的分类精度高出1%。比PCL分别提高了3.7%、5.1%。比50个特征基因的SVM和KNN算法样本识别的正确率高出10%。

同时，EPAs具有很强的辨别能力，在较小的高级基因显现模式中同样有很高的识别率。减少各种分类器中用于训练的基因模式，发现CCEPA和KEPA中的EPAs由20个减少到10个时，样本的正确识别率几乎没有影响；而PCL中的EPs由20个减少到10个时，样本的正确识别率存在较大的变化；SVM和KNN中的特征基因数目降为25个时，正确识别率则影响很大。

7.6 本章小结

研究利用显现模式的基因分类方法问题。抽取了基因表达中的增强的显现模式（EPI）和高级显现模式（EPA），缓解了在样本容量很小情况下概率估计的偏差，提高候选分割点的可靠性，有效地增强了显现模式的癌症辨别性能。并在EPI和EPA基因模式上分别提出了两种癌症检测策略（CCEPI/CCEPA和KEPI/KEPA）。最后在急性白血病数据集上进行实验，结果显示，两种检测策略都取得了较好的癌症识别率。

结 论

本文主要论述博士学位课题“基于DNA微阵列基因表达谱数据的癌症检测研究”的研究成果。通过微阵列基因表达谱数据的分析来检测癌症不仅能够预测患者样本的癌症类型，针对患者制定有效的治疗方案帮助患者进行癌症治疗，还能够辅助研究人员开发与研制新的抗癌药物。因此利用基因表达谱数据进行癌症检测研究具有非常重要的实际意义和应用价值。有许多研究人员对其进行了广泛深入的研究。由于微阵列基因表达谱数据具有数据量巨大、高维性、高噪声、高冗余和数据分布不均衡的特点，DNA微阵列数据分析又给癌症检测带来了巨大挑战。本文研究高维基因表达数据中最具辨别能力的特征基因的选择方法；基因组和癌症组中隐含的潜在的基因表达模式；癌症基因组之间的相互调控关系，具有可解释性的显现模式；利用选取的特征基因和抽取的基因表达模式以及显现模式建立合适的癌症分类模型，预测患者样本的癌症模型；癌症预测中的多分类模型的组合。该课题在现有的研究成果和进一步深入研究将为基因的功能性研究、癌症的基因表达分析、疾病的临床诊断、抗癌药物的研制提供重要的科学依据。

1 工作总结

本课题在综述了基于基因表达数据研究癌症检测的研究现状的基础上，主要对癌症检测研究中的癌症类型识别 基因表达模式的挖掘和基因特征的抽取问题内容进行了系统研究 主要贡献如下：

第三章中，提出了在癌症检测中一种自适应的基因选择方法：针对基因表达数据高噪声 高冗余的特点，提出了一种分步的特征基因选择方法 在不同的特征分析步骤中分别去除基因数据中的高噪声和高冗余，达到特征基因选择降维的目的；然后，引入 Gap Statistic 理论，提出了一种自适应的特征基因的选择方法，弥补目前特征基因选择算法中缺乏较好的基因数目预置机制的不足。

第四章中，提出了一种基于隐含变量模型的基因分类算法：利用主分量分析方法（Primary Component Analysis，PCA）和独立分量分析方法（Independent Component Analysis，ICA）挖掘基因表达谱中的隐含的基因调控因子，以揭示基因表达之间存在的相互影响以及调控机制，在癌症研究中获取了PCAE和ICAE两种基因表达模型；利用抽样的办法减少噪声对PCAE和ICAE的影响，并利用提出的癌症检测算法CDHV 对患者样本进行癌症预测，CDHV有效地挖掘了基因表达数据的隐含表达变量，同时提高了预测效率。

第五章中，提出了基于癌症组关联空间的基因表达模式抽取与癌症识别算法：利用癌症组基因表达存在的局部特征相关性的生物病理特点，抽取不同癌症组的特

征模式和基因表达模式，讨论与癌症组相关联的基因表达模式在癌症组中的表达以及调控，并提出了癌症组相关联的基因表达模式的癌症预测算法。

第六章中，提出了基于组合分类算法的癌症识别算法：在高维的基因表达谱中，由于不同特征选择法采用不同的搜索机制和评价策略，挑选出的特征基因明显不同，导致分类器的癌症识别结果不稳定。针对癌症组基因数据和基因组数据提出一组具有互补性分类器，然后利用组合分类算法提高癌症分类效果，并增强了癌症检测算法的泛化性能。

第七章中，研究基于显现模式的癌症分类算法。癌症的产生和发展过程是由部分基因共同调控和表达的结果，需要我们从基因协同表达来分析基因数据。针对提取显现模式时在小样本情况下概率近似计算问题和显现模式的基因分割点的选择问题分别提出增强的显现模式（EPI）和高级显现模式（EPA），并在EPI和EPA两种基因模式上提出了两种癌症检测策略最后在急性白血病数据集上进行实验，结果表明EPI和EPA有效地提高了EP的癌症检测性能，两种检测策略都取得了较好的癌症识别率。

2 研究展望

基于微阵列基因表达谱数据的癌症检测研究是一个十分宽广的研究领域，具有非常重要的研究意义，将促进疾病诊断和基因治疗的发展。本文针对基因表达数据高维、高噪声和高冗余的特点，提出了自适应的特征基因的抽取算法，隐含的潜在的基因表达模式挖掘算法；针对“维数灾难”问题，提出了癌症基因组的基因表达模式的挖掘算法，提示其基因表达；并在抽取或挖掘的基因模式中，提出了有效的癌症检测算法。而且针对复杂的基因调控系统，挖掘了具有可解释性的显现模式，并提出了相应的癌症识别算法。但下面的问题仍有待进一步的研究：

1. 生物学解释和应用问题

生物信息学研究是以生物学问题作为驱动力的，任何脱离实际的生物学问题而沉溺于算法的研究都是不可取的。因此，将来的研究工作首先应考虑它的生物学意义，脱离了生物学问题，再漂亮的数学公式都一文不值。

2. 异类表达数据的融合问题

目前基于基因表达数据的癌症检测研究中，利用的仅仅是反应mRNA丰度的基因表达数据。在很多情况下，这一数据本身提供的信息量是有限的，提供的信息不能不是非常精确。因此，将基因芯片数据和其他数据，比如序列数据、结构数据、蛋白质芯片、组织芯片等数据结合起来研究癌症识别和基因治疗将成为将来的研究重点。

3. 复杂基因调控系统的挖掘问题

在人类基因组中，基因表达之间存在相互影响，并且具有复杂的基因调控原理和机制。本文通过PCA/ICA挖掘了简单的基因协同表达模式，但是整个基因组的调

控机制，以及复杂的基因调控系统的挖掘还需要很多的研究工作。

4. 癌症类型发现问题

本文主要研究了基于基因表达数据的有监督的癌症识别问题，没有考虑癌症样本的无监督分析，从而发现癌症的新类型。癌症类型的发现问题对癌症的研究同样非常有用。

另外，癌症识别算法的泛化性问题，也将成为我们后续工作的重要方向。

参考文献

[1] Roberts L., Davenport R. J., Pennisi E., et al. A history of the Human Genome Project. Science. 2001, 291(5507):1195

[2] Roberts L. The human genome. Controversial from the start. Science. 2001, 291(5507):1182–1188

[3] Lander E. S., Linton L. M., Birren B., et al. Initial sequencing and analysis of the human genome. Nature. 2001, 409(6822):860–921

[4] Chee M., Yang R., Hubbell E., et al. Accessing genetic information with high-density DNA arrays. Science. 1996, 274(5287):610–614

[5] Papadopoulos N., Nicolaides N. C., Wei Y. F., et al. Mutation of a mutL homolog in hereditary colon cancer. Science. 1994, 263(5153):1625–1629

[6] Golub T. R., Slonim D. K., Tamayo P., et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999, 286(5439):531–537

[7] Fodor S. P., Read J. L., Pirrung M. C., et al. Light-directed, spatially addressable parallel chemical synthesis. Science. 1991, 251(4995):767–773

[8] Marshall A., Hodgson J. DNA chips: an array of possibilities. Nat Biotechnol. 1998, 16(1):27–31

[9] Service R. F. Microchip arrays put DNA on the spot. Science. 1998, 282(5388):396–399

[10] Kricka L. J. Revolution on a square centimeter. Nature Biotechnology. 1998, 16:513 – 514

[11] Goffeau A. DNA technology: Molecular fish on chips. Nature. 1997, 385(6613):202–203

[12] Shah N., Lepre J., Tu Y., et al. Can we identify cellular pathways implicated in cancer using gene expression data? Proc IEEE Comput Soc Bioinform Conf. 2003, 2:94–103

[13] Furey T. S., Cristianini N., Duffy N., et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics. 2000, 16(10):906–914

[14] Guyon I., Weston J., Barnhill S., et al. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning. 2002, 46(1-3):389–422

[15] Khan J., Wei J. S., Ringnér M., et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med. 2001, 7(6):673–679

[16] Young R. A. Biomedical discovery with DNA arrays. Cell. 2000, 102(1):9–15

[17] 张东晖, 黄颖, 蔡军等译. 生物信息学. 中信出版社, 2003

[18] Zhang M. Q. Large-scale gene expression data analysis: a new challenge to computational biologists. Genome Res. 1999, 9(8):681–688

[19] Stipp D. Gene chip breakthrough. Fortune. 1997, 135:56–66

[20] News and Staffs . Breakthrough of the year. The runners-up. Science. 1998, 282(5397):2157–2161

[21] Hacia J. G. Resequencing and mutational analysis using oligonucleotide microarrays. Nat Genet. 1999, 21(1 Suppl):42–47

[22] Schena M., Shalon D., Davis R. W., et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995, 270(5235):467–470

[23] DeRisi J. L., Iyer V. R., Brown P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science. 1997, 278(5338):680–686

[24] Brown P. O., Botstein D. Exploring the new world of the genome with DNA microarrays. Nat Genet. 1999, 21(1 Suppl):33–37

[25] Tanaka T. S., Jaradat S. A., Lim M. K., et al. Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray. Proc Natl Acad Sci U S A. 2000, 97(16):9127–9132

[26] Golub T. R. Genome-wide views of cancer. N Engl J Med. 2001, 344(8):601–602

[27] Bull J. H., Ellison G., Patel A., et al. Identification of potential diagnostic markers of prostate cancer and prostatic intraepithelial neoplasia using cDNA microarray. Br J Cancer. 2001, 84(11):1512–1519

[28] Heller R. A., Schena M., Chai A., et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proc Natl Acad Sci U S A. 1997, 94(6):2150–2155

[29] Buates S., Matlashewski G. Identification of genes induced by a macrophage activator, S-28463, using gene expression array analysis. Antimicrob Agents Chemother. 2001, 45(4):1137–1142. URL http://dx.doi.org/10.1128/AAC.45.4.1137-1142.2001

[30] Schena M., Shalon D., Heller R., et al. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci U S A. 1996, 93(20):10614–10619

[31] Kauraniemi P., Hedenfalk I., Persson K., et al. MYB oncogene amplification in hereditary BRCA1 breast cancer. Cancer Res. 2000, 60(19):5323–5328

[32] Okutsu iJ., Tsunoda T., Kaneta Y., et al. Prediction of chemosensitivity for patients with acute myeloid leukemia, according to expression levels of 28 genes selected by genome-wide complementary DNA microarray analysis. Mol Cancer Ther. 2002, 1(12):1035–1042

[33] Kan T., Shimada Y., Sato F., et al. Gene expression profiling in human esophageal cancers using cDNA microarray. Biochem Biophys Res Commun. 2001, 286(4):792–801

[34] Lossos I. S., Alizadeh A. A., Eisen M. B., et al. Ongoing immunoglobulin somatic mutation in germinal center B cell-like but not in activated B cell-like diffuse large cell lymphomas. Proc Natl Acad Sci U S A. 2000, 97(18):10209–10213

[35] Brown V., Jin P., Ceman S., et al. Microarray identification of FMRP-associated brain mRNAs and altered mRNA translational profiles in fragile X syndrome. Cell. 2001, 107(4):477–487

[36] Oestreicher N., Ramsey S. D., Linden H. M., et al. Gene expression profiling and breast cancer care: what are the potential benefits and policy implications? Genet Med. 2005, 7(6):380–389

[37] Kumar-Sinha C., Ignatoski K. W., Lippman M. E., et al. Transcriptome analysis of HER2 reveals a molecular connection to fatty acid synthesis. Cancer Res. 2003, 63(1):132–139

[38] Rogers P. D., Barker K. S. Evaluation of differential gene expression in fluconazole-susceptible and -resistant isolates of Candida albicans by cDNA microarray analysis. Antimicrob Agents Chemother. 2002, 46(11):3412–3417

[39] Marx J. Medicine. DNA arrays reveal cancer in its many forms. Science. 2000, 289(5485):1670–1672

[40] Brazma A., Vilo J. Gene expression data analysis. FEBS Lett. 2000, 480(1):17–24

[41] Quackenbush J. Microarray analysis and tumor classification. N Engl J Med. 2006, 354(23):2463–2472

[42] Han J., Kamber M. Data Mining: Concepts and Techniques,The Morgan Kaufmann Series in Data Management Systems. Jim Gray, Series Editor Morgan Kaufmann Publishers, 2006

[43] Tan A.-H., Pan H. Predictive neural networks for gene expression data analysis. Neural Netw. 2005, 18(3):297–306

[44] Tavazoie S., Hughes J. D., Campbell M. J., et al. Systematic determination of genetic network architecture. Nat Genet. 1999, 22(3):281–285

[45] Eisen M. B., Spellman P. T., Brown P. O., et al. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998, 95(25):14863–14868

[46] Tamayo P., Slonim D., Mesirov J., et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A. 1999, 96(6):2907–2912

[47] Toronen P., Kolehmainen M., Wong G., et al. Analysis of gene expression data using self-organizing maps. FEBS Lett. 1999, 451(2):142–146

[48] Herzel H., Beule D., Kielbasa S., et al. Extracting information from cDNA arrays. Chaos. 2001, 11(1):98–107

[49] Alizadeh A. A., Eisen M. B., Davis R. E., et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000, 403(6769):503–511

[50] Alon U., Barkai N., Notterman D. A., et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A. 1999, 96(12):6745–6750

[51] Herrero J., Valencia A., Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics. 2001, 17(1):126–136

[52] Kohonen T. Self-Organizing Maps. Springer Verlag,Berlin,Heidelberg, 2001

[53] Wang J., Delabie J., Aasheim H., et al. Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. BMC Bioinformatics. 2002, 3:36

[54] Veer v. tL. J., Dai H., Vijver v. dM. J., et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002, 415(6871):530–536

[55] Kuramochi M., Karypis G. Gene classification using expression profiles: a feasibility study. International Journal on Artificial Intelligence Tools. 2005, 14(4):641–660

[56] Mao S., Dong G. Discovery of Highly Differentiative Gene Groups from Microarray Gene Expression Data Using the Gene Club Approach. J. Bioinformatics and Computational Biology. 2005, 3(6):1263–1280

[57] Hu Z., Fan C., Oh D. S., et al. The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics. 2006, 7:96

[58] Slonim D. K., Tamayo P., Mesirov J. P., et al. Class prediction and discovery using gene expression data. In Proc. of RECOMB. 2000, 263–272

[59] Xiong H., Chen wX. Kernel-based distance metric learning for microarray data classification. BMC Bioinformatics. 2006, 7:299

[60] Yao Z., Ruzzo W. L. A regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. BMC Bioinformatics. 2006, 7 Suppl 1:S11

[61] Mewes H. W., Amid C., Arnold R., et al. MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res. 2004, 32(Database issue):D41–D44

[62] Mewes H. W., Heumann K., Kaps A., et al. MIPS: a database for genomes and protein sequences. Nucleic Acids Res. 1999, 27(1):44–48

[63] Brown M. P., Grundy W. N., Lin D., et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A. 2000, 97(1):262–267

[64] Semolini R., Von Zuben F. Transductive support vector machines for classification of microarray gene expression data. Neural Networks, 2003. Proceedings of the International Joint Conference on. 2003, vol. 4, 2946–2951vol.4

[65] Liu Y. Active learning with support vector machine applied to gene expression data for cancer classification. J Chem Inf Comput Sci. 2004, 44(6):1936–1941

[66] Gerald L. B., Tang S., Bruce F., et al. A decision tree for tuberculosis contact investigation. Am J Respir Crit Care Med. 2002, 166(8):1122–1127

[67] Feinglass J., Yarnold P. R., McCarthy W. J., et al. A classification tree analysis of selection for discretionary treatment. Med Care. 1998, 36(5):740–747

[68] Gaudart J., Poudiougou B., Ranque S., et al. Oblique decision trees for spatial pattern detection: optimal algorithm and application to malaria risk. BMC Med Res Methodol. 2005, 5:22

[69] Dettling M., Bühlmann P. Boosting for tumor classification with gene expression data. Bioinformatics. 2003, 19(9):1061–1069

[70] Tan A. C., Gilbert D. Ensemble machine learning on gene expression data for cancer classification. Appl Bioinformatics. 2003, 2(3 Suppl):S75–S83

[71] Cho S.-B., Won H.-H. Machine Learning in DNA Microarray Analysis for Cancer Classification. In the Proc. the First Asia-Pacific Bioinformatics Conference (APBC 2003). Australian Computer Society, 2003, 189–198

[72] Keller A. D., Schummer M., Hood L., et al. Bayesian Classification of DNA Array Expression Data. Tech. Rep. UW-CSE-2000-08-01, University of Washington, 2000

[73] Lee K. E., Sha N., Dougherty E. R., et al. Gene selection: a Bayesian variable selection approach. Bioinformatics. 2003, 19(1):90–97

[74] Jaeger J., Sengupta R., Ruzzo W. L. Improved gene selection for classification of microarrays. Pac Symp Biocomput. 2003:53–64

[75] Goh L., Song Q., Kasabov N. K. A Novel Feature Selection Method to Improve Classification of Gene Expression Data. In the Proc. of the Second Asia-Pacific Bioinformatics Conference (APBC 2004). Australian Computer Society, 2004, 161–166

[76] 李颖新, 李建更, 阮晓钢. 基于支持向量机的肿瘤分类特征基因选取. 计算机学报. 2006, 26(2):324–330

[77] 李颖新, 阮晓钢. 基于基因表达谱的肿瘤亚型识别与分类特征基因选取研究. 电子学报. 2005, 33(4):651–655

[78] Conde L., Mateos A., Herrero J., et al. Unsupervised reduction of the dimensionality followed by supervised learning with a perceptron improves the classification of conditions in DNA microarray gene expression data. IEEE Press (New York), 2002, 77–86

[79] Raychaudhuri S., Stuart J. M., Altman R. B. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput. 2000:455–466

[80] Chu S., DeRisi J., Eisen M., et al. The transcriptional program of sporulation in budding yeast. Science. 1998, 282(5389):699–705

[81] Liebermeister W. Linear modes of gene expression determined by independent component analysis. Bioinformatics. 2002, 18(1):51–60

[82] Hori G., Inoue M., ichi S., et al. Blind gene classification based on ICA of microarray data. In the Proc. of the 3rd International Conference on Independent Component Analysis and Signal Separation(ICA2001). San Diego, San Diego, San Diego, San Diego, California, U. S. A., 2001, vol. 3, 332–336

[83] Hori G. IM., Nishimura S. NH. Blind Gene Classification- An ICA-based Gene Classification./Clustering Method. RIKEN BSI BSIS Technical Report. 2002, (02-5)

[84] Shipp M. A., Ross K. N., Tamayo P., et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med. 2002, 8(1):68–74

[85] Hu X., Yoo I. Cluster Ensemble and Its Applications in Gene Expression Analysis. In the Proc. the Second Asia-Pacific Bioinformatics Conference (APBC 2004). Australian Computer Society, 2004, 297–302

[86] Lukashin A. V., Fuchs R. Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. Bioinformatics. 2001, 17(5):405–414

[87] Mateos A., Herrero J., Tamames J., et al. Supervised Neural Networks For Clustering Conditions In DNA Array Data After Reducing Noise By Clustering Gene Expression Profiles. Kluwer Academic, 2002, 91–103

[88] Tibshirani R., Walther G., Hastie T. Estimating the number of clusters in a dataset via the gap statistic. Tech. Rep. 208, Dept. of Statistics, Stanford University. 2000

[89] Pierga J.-Y., Reis-Filho J. S., Cleator S. J., et al. Microarray-based comparative genomic hybridisation of breast cancer patients receiving neoadjuvant chemotherapy. Br J Cancer. 2007, 96(2):341–351

[90] Wall M. E., Rechtsteiner A., Rocha L. M. A Practical Approach to Microarray Data Analysis. Kluwer: Norwell, MA, 2003, 91–109

[91] Hyvarinen A. Fast and robust fixed-point algorithms for independent component analysis. Neural Networks, IEEE Transactions on. 1999, 10(3):626–634

[92] Hyvarinen A., Karhunen J., Oja E. Independent Component Analysis. Wiley, New York, 2001

[93] Hyvarinen A., Oja E. Independent component analysis: algorithms and applications. Neural Netw. 2000, 13(4-5):411–430

[94] 杨竹青, 李勇, 胡德文. 独立成分分析方法综述. 自动化学报. 2002, 28(5):762–773

[95] Pearson K. On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine. 1901, 2:559–572

[96] Hotelling H. Analysis of a Complex of Statistical Variables into Principal Components. Journal of Educational Psychology. 1933, 24:417–441

[97] Comon P. Independent component analysis, a new concept? Signal Process. 1994, 36(3):287–314

[98] Amari S. Super Efficiency in Blind Source Separation. IEEE Trans on Signal Processing. 1999, 47(4):936–944

[99] 李霞, 张田文, 郭政等. 一种基于递归分类树的集成特征基因选择方法. 计算机学报. 2004, 27(5):675–682

[100] Parsons L., Haque E., Liu H. Subspace clustering for high dimensional data: a review. SIGKDD Explorations. 2004, 6(1):90–105

[101] Kasabov N. Evolving Connectionist Systems: Methods and Applications in Bioinformatics, Brain Study and Intelligent Machines. Springer Verlag, 2002

[102] 李颖新, 阮晓钢. 基于支持向量机的肿瘤分类特征基因选取. 计算机研究与发展. 2005, 42(10):1796–1801

[103] Li J., Wong L. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. Bioinformatics. 2002, 18(5):725–734

[104] 王正群, 陈世福, 陈兆乾. 优化分类型神经网络线性集成. 软件学报. 2005, 16(11):1902–1908

[105] O'Neill M. C., Song L. Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect. BMC Bioinformatics. 2003, 4:13

[106] Liu B., Cui Q., Jiang T., et al. A combinational feature selection and ensemble neural network method for classification of gene expression data. BMC Bioinformatics. 2004, 5:136

[107] Fayyad U. M., Irani K. B. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In the Proc. of the 13th International Joint Conference on Artificial Intelligence(IJCAI). Morgan Kaufmann, San Francisco, CA, 1993, 1022–1029

[108] Jiang D., Tang C., Zhang A. Cluster analysis for gene expression data: a survey. Knowledge and Data Engineering, IEEE Transactions on. 2004, 16(11):1370–1386

[109] Li J., Liu H., Downing J. R., et al. Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. Bioinformatics. 2003, 19(1):71–78

[110] Li J., Wong L. Emerging patterns and gene expression data. Genome Inform. 2001, 12:3–13

[111] Dong G., Li J. Mining border descriptions of emerging patterns from dataset pairs. Knowl. Inf. Syst. 2005, 8(2):178–202

[112] Dong G., Li J. Efficient Mining of Emerging Patterns: Discovering Trends and Differences. KDD. 1999, 43–52

攻读博士学位期间所发表和投稿的论文

[1] 卢新国, 林亚平, 王海军, 李小龙. 基于微阵列基因表达谱的一种关联空间的癌症分类算法. 电子学报. (已接收)

[2] **Xinguo Lu**, Yaping Lin, Haijun Wang, Siwang Zhou, Xiaolong Li. A Novel Relative Space Based Gene Feature Extraction and Cancer Recognition. the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2007), Nanjing, China, LNAI 4426: 712-719

[3] **Xinguo Lu**, Yaping Lin, Xiaolin Yang, Lijun Cai, Haijun Wang, Gustaph Sanga. Using Most Similarity Tree Based Clustering to Select the Top Most Discriminating Genes for Cancer Detection. The Eighth International Conference on Artificial Intelligence and Soft Computing (ICAISC 2006). Zakopane, Poland, June 2006, LNAI 4029: 931-940. [SCI, EI, ISTP]

[4] **Xinguo Lu**, Yaping Lin, Xiaolong Li, Yeqing Yi, Lijun Cai, Haijun Wang. Gene Cluster Algorithm Based on Most Similarity Tree. In the Proc. of the 8th International Conference on High Performance Computing in Asia Pacific Region (HPC Asia2005), Beijing, China, 2005, p652-656. [EI, ISTP]

[5] **Xinguo Lu**, Yaping Lin, Wen Yue, Haijun Wang, Siwang Zhou. ICA Based Supervised Gene Classification of Microarray Data in Yeast Functional Genome. In the Proc. of the 8th International Conference on High Performance Computing in Asia Pacific Region (HPC Asia2005), Beijing, China, 2005, p633-638. [EI, ISTP]

[6] 卢新国, 林亚平, 陈治平. 基于互信息的改进特征选择预处理算法. 湖南大学学报. 2005, 32(1): 104-107. [EI]

[7] 卢新国, 林亚平. 癌症识别中一种基于组合神经网络的分类算法. 电子与信息学报. (已投)

[8] 蔡立军, 林亚平, 卢新国, 易叶青, 李小龙. 基于遗传算法的基因分类. 电子学报. 2006, 34(11): 2115-2119

[9] Haijun Wang, Yaping Lin, **Xinguo Lu**, Yalin Nie. A Novel EPA-KNN Gene Classification Algorithm. the Fourth International Symposium on Neural Networks (ISNN2007), LNCS 4492:1254-1263

[10] 周四望, 林亚平, 张建明, 欧阳竞成, 卢新国. 传感器网络中基于环模型的小波数据压缩算法,软件学报, 2007, 18(3): 679-690

[11] 刘娟, 林亚平, 易叶青, 卢新国, 吴巧敏. Ad hoc网络中一种视频编码的多径路由算法. 计算机工程与应用. 2006, 42(21): 114-117

[12] Xiaolong Li, Yaping Lin, Siqing Yang, Yeqing Yi, Jianping Yu, **Xinguo Lu**. A Key Distribution Scheme Based on Public Key Cryptography for Sensor Networks. In the Proc. of the 2006 Advances in Computational Intelligence and Security (CIS 2006), LNCS. (已接收)

[13] Wen Yue, Zhiping Chen, **Xinguo Lu**, Feng Lin, and Juan Liu. Using Query Expansion and Classification for Information Retrieval. In the Proc. Of the 2005 International Conference on Semantics, Knowledge and Grid. 2005, 359-367

[14] 李小龙, 林亚平, 易叶青, 卢新国, 羊四青. 传感器网络中不依赖节点位置信息的节点调度算法. 2007年中国计算机大会. （已接收）

[15] 李小龙, 林亚平, 易叶青, 余建平, 卢新国. 传感器网络中基于虚拟坐标的节点调度算法. 软件学报. （已接收）

致 谢

值此论文完成之际，谨向给予我无私帮助的老师和同学们致以诚挚的谢意!

首先感谢我敬爱的导师林亚平教授和师母苏爱珍老师！多年来，在生活处世和学习科研上我得到了林老师的悉心关怀和精心指导，使我克服了许多困难，圆满地完成了学业。本论文的研究工作正是在林老师最初的建议下展开的。从选题、攻关到得出一系列结果，以至于最后成文，都凝聚着恩师的大量心血。林老师渊博的知识、严谨的治学态度、深邃的洞察力、持之以恒的探索精神、朴实无华的工作与生活作风，将永远激励我刻苦学习努力工作，让我受益终生。同时也特别感谢师母在我学习期间给予我的耐心帮助和热情关怀。学生对导师和师母的教育和培养之恩在此再次表示深深的感谢！

衷心感谢湖南大学计算机与通信学院的各位领导和老师对我的帮助和指导，感谢陈治平、王雷等博士师兄在学术上的讨论和交流给予我的启发与帮助。北京工业大学李颖新博士，新加坡南洋理工大学Li博士（J.Y. Li），芬兰赫尔辛基科技大学Oja教授（E. Oja）都对作者提供过许多有益的帮助，在此，作者对他们表示衷心的感谢！

感谢博士生欧阳竞成、周四望、张建明、谭义红，李小龙，易叶青等给我的无私帮助和支持。感谢计算机网络与机器学习实验室的兄弟姐妹们，陪伴我度过了这长久的学习、研究阶段，帮助我解决问题，开拓思想，特别感谢王海军对我研究和实验工作的大力支持！

特别感谢在本人的博士论文预答辩及盲审期间认真审阅我的论文并提出宝贵意见的各位专家教授！

衷心感谢我慈祥的父亲对我的学习和工作的关心和支持，感谢我的亲人们给予我的关怀、支持和理解！

衷心感谢答辩委员会的各位专家的评议和指导！