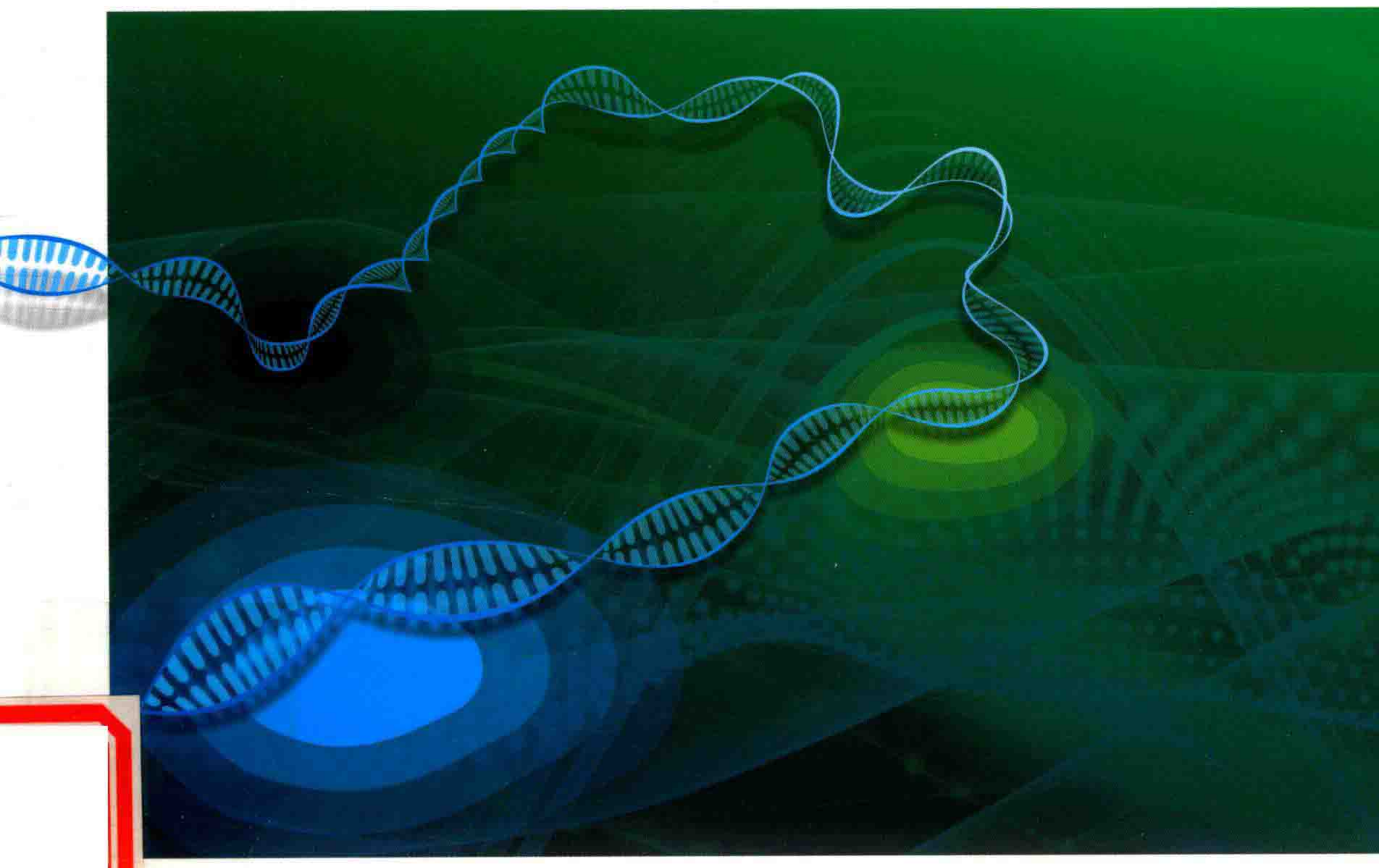
基因表达数据的特征选择及其识别算法研究

陆慧娟　严　珂　著



科学出版社

基因表达数据的特征选择及其识别算法研究

陆慧娟　严　珂　著

科学出版社
北京

内 容 简 介

为了有效处理基因表达数据，本书从数据集和分类器两个方向入手进行讨论。在数据集方面，采用不同算法进行特征选择，选择与分类目标密切相关的基因提高分类器模型的泛化性能。在分类器方面，构建训练集，利用集成方法提高旋转森林的分类精度和稳定性；利用改进后的粒子群算法优化核超限学习机的内权参数，提高分类器的分类精度；根据输出不一致测度，进行相异性集成，提高分类模型的分类精度和稳定性；通过在超限学习机模型中嵌入误分代价因素，实现对肿瘤的代价敏感分类。本书从机器学习的视角，提出了若干前沿的特征选择与分类算法，为后续基因表达数据识别的相关研究奠定了基础。

本书可作为从事计算机、自动控制、生物信息等领域的专家学者和应用人员的参考用书。

图书在版编目 (CIP) 数据

基因表达数据的特征选择及其识别算法研究 / 陆慧娟，严珂著.
—北京：科学出版社，2017.5
ISBN 978-7-03-051961-0

Ⅰ. ①基… Ⅱ. ①陆… ②严… Ⅲ. ①基因表达－模式识别－研究 Ⅳ. ①Q786

中国版本图书馆 CIP 数据核字(2017)第 040339 号

责任编辑：陈 静 赵微微 / 责任校对：郭瑞芝

责任印制：张 倩 / 封面设计：迷底书装

科学出版社出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

新科印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2017 年 5 月第 一 版 开本：720×1 000 1/16
2017 年 5 月第一次印刷 印张：9
字数：176 000

定价：48.00 元

序

癌症是对人类生命构成严重威胁的主要疾病之一，是由各种致癌因素导致的某些局部组织的细胞克隆性异常，在基因水平上失去对其生长的正常调控从而增生而成的新生物。癌症的早诊断是提高癌症患者成活率的关键。

目前的癌症诊断方法，主要是通过观察显微镜下细胞的大小、颜色和形状来确定肿瘤的类型。这种诊疗方法建立在形态学之上，存在很大的缺陷，同一类型的肿瘤可能会出现临床上的差异，对治疗的敏感性不够。癌症的发生是一个多阶段逐步演变的过程，在这一过程中，常伴随着多种基因的改变。从分子生物学水平发现、识别与癌症相关的重要基因是生物信息学研究的一个重要课题，对癌症患者早期诊断和进行个性化治疗具有重要意义。它不仅能提高患者的生存率，而且能提高患者的生存质量，从而引起人们的广泛关注。

该书是著者近 5 年来在国家自然科学基金和浙江省自然科学基金项目资助下，获得的一系列关于基因表达数据的特征选择及其识别算法研究成果的结晶。书中介绍了基因表达数据高维、小样本、高噪声、样本不平衡的特点，以及针对这些特点所设计的分类流程，包括特征选择和分类器构建两大步骤。从机器学习的视角，提出了若干前沿的特征选择与分类算法，为后续基因表达数据识别的相关研究奠定了基础。此外，书中还介绍了神经网络、支持向量机、超限学习机等分类器的原理以及在基因数据分类中的应用，并利用实际基因数据进行了验证。根据基因数据的特点，提出使用集成学习、代价敏感学习等机器学习技术。

该书内容丰富、结构清晰、创新明显，是一本融合理论研究和工程应用的学术专著，是从事相关研究的科技人员很好的参考用书。该书的出版将对基因表达数据的特征选择及其识别理论研究和技术应用起到积极的推动作用。

何新贵

中国工程院院士、北京大学教授

2016 年 12 月

前　言

肿瘤组织无论在细胞形态上还是组织结构上，都与其发源的正常组织有不同程度的差异。肿瘤在本质上是基因病。各种环境的和遗传的致癌因素以协同或序贯的方式引起 DNA 损害，从而激活原癌基因和灭活肿瘤抑制基因，加上凋亡调节基因和 DNA 修复基因的改变，继而引起表达水平的异常，使靶细胞发生转化。被转化的细胞先多呈克隆性的增生，经过一个漫长的多阶段的演进过程，其中一个克隆相对无限制的扩增，通过附加突变，选择性地形成具有不同特点的亚克隆（异质化），从而获得浸润和转移的能力（恶性转化），形成恶性肿瘤。肿瘤在所占据的组织中形成肿块，其大小、外形、界限、硬度、表面情况、与邻近组织关系等可作为检查与诊断肿瘤的依据。

目前的肿瘤诊断方法，主要是通过观察显微镜下细胞的大小、颜色和形状来确定肿瘤的类型。这种诊疗方法建立在形态学之上，存在很大的缺陷，同一类型的肿瘤可能会出现临床上的差异，对治疗的敏感性不够。本书主要研究从基因层面诊断癌症，期望通过对基因数据的处理来达到更早、更准确地发现癌症。为了有效处理基因表达数据，主要采用神经网络、支持向量机以及决策树来设计分类器。为了提高分类系统的稳定性，拟利用分类器集成技术，来改进相关算法。

鉴于此，作者在国家自然科学基金和浙江省自然科学基金项目资助下，近 5 年来，一直从事基因表达数据的特征选择及其识别算法的研究，从机器学习的视角，提出若干前沿的特征选择与分类算法，为后续基因表达数据识别的相关研究奠定了基础。此外，还研究神经网络、支持向量机、超限学习机等分类器的原理以及在基因数据分类中的应用，并利用实际基因数据进行验证。根据基因数据的特点，使用集成学习、代价敏感学习等机器学习技术，为机器学习在其他领域的进一步应用奠定技术基础，具有重要的理论意义和实际的应用价值。

本书是作者在国内外本领域权威期刊以及有影响的国际会议论文集上，发表的 10 多篇学术论文的基础上进一步加工、深化而成的，是对已有成果的全面总结。具体研究内容从数据集和分类器两个方向入手。在数据集方面，利用适当的方法进行特征选择，选择与分类目标密切相关的基因以提高分类器模型的泛化性能；创造性地结合两种不同的特征选择算法对基因数据集进行特征选择，能够有效地克服传统特征选择算法的弊端。在分类器方面，构建训练集，利用集成方法提高

旋转森林的分类精度和稳定性；利用改进后的粒子群算法优化核超限学习机内权参数，达到提高分类器的分类精度的效果；根据输出不一致测度，进行相异性集成，提高分类模型的分类精度和稳定性；通过在 ELM 模型中嵌入误分代价因素，实现对肿瘤的代价敏感分类。

在本书的编写过程中，作者得到了中国工程院院士、北京大学何新贵教授的悉心指导，而且何院士欣然为本书作序，令作者深受鼓舞，在此向何院士表示衷心的感谢！天津大学邹权研究员，杭州电子科技大学高志刚副教授，中国计量大学潘晨教授、叶敏超博士等为书稿提出了许多宝贵的建议；全书在编写过程中也得到了孟亚琼、陈俊颖、杨磊、马露露、高慧云、王黎、于海燕、魏莎莎、刘亚卿、王石磊、刘金勇、杜帮俊、陈晓青等的帮助；“计算机应用技术”浙江省重点学科（中国计量大学）也为本书的出版提供了大力的帮助。在此一并表示衷心的感谢!

在本书的编写过程中，参考了国内外有关研究成果，在此对所涉及的专家和研究人员表示衷心的感谢。挂一漏万，可能书中所列出的参考文献不够全面，在此也对可能被遗漏的专家和研究人员表示衷心的感谢。此外，本书的出版得到了国家自然科学基金项目(项目编号：61272315，60842009，61602431)和浙江省自然科学基金项目(项目编号：Y1110342)的资助，在此一起表示鸣谢！

本书可作为从事计算机、自动控制、生物信息等领域的专家学者和应用人员的参考书。

由于学术水平所限，书中难免有疏漏之处，敬请读者指正。作者的联系方式：hjlu@cjlu.edu.cn。

作　者

2016 年 12 月于中国计量大学

目　　录

第1章 绪 论

肿瘤是由各种致癌因素导致的某些局部组织的细胞克隆性异常，在基因水平上失去对其生长的正常调控从而增生而成的新生物。肿瘤一般可分为良性和恶性两大类。恶性肿瘤又称为癌症。在2012年，全球约有1410万新发癌症病例，820万患者死于癌症。其中57%的癌症患者以及65%的癌症死亡患者来自于发展中国家[1]。作为人类健康的第一杀手，恶性肿瘤已经成为我国主要的公共卫生问题之一。所以，对肿瘤的预防和治疗是全世界关注的焦点。

按现在的医疗水平，对早期癌症患者的治疗有80%以上的治愈率；但是，晚期的癌症患者在治疗后很少能生存5年以上。因此，早发现、早预防、早治疗是挽救患者的重要手段[2]。目前的肿瘤诊断方法，主要是通过观察显微镜下细胞的大小、颜色和形状来确定肿瘤的类型。这种诊疗方法建立在形态学之上，存在很大的缺陷，如同一类型的肿瘤可能会出现临床上的差异，对治疗的敏感性不够[3]。癌症的发生是一个多阶段逐步演变的过程，在这一过程中，常伴随着多种基因的改变。从分子生物学水平发现、识别与癌症相关的重要基因是生物信息学研究的一个重要课题，对癌症患者早期诊断和进行个性化治疗具有重要意义。它不仅能提高患者的生存率，而且能提高患者的生存质量。

目前，民众迫切追求高质量医疗服务，但是医疗成本处于单调递增状态，因此提高质量和降低成本已成为医疗服务业关注的焦点。一方面，我国医院长期以来将重点放在质量管理方面，并取得了较大进步；另一方面，医院还需要加强对疾病诊断和治疗过程的科学管理，尽量避免治疗的随意性、用药的盲目性、过度检查等现象。如今，许多医院已经意识到该问题，并不断寻找解决之道。许多严重的遗传病、绝症等无法用药物进行有效的治疗，唯有探索人类基因的秘密，从基因入口进行研究，才可能从根本上进行解决。

由于人类基因组(测序)计划的稳步实施以及分子生物学等相关学科的迅速发展，基因序列数据快速增长，更多的微生物与动植物的基因组序列能够得到测定。所以，如何研究不同基因在生命过程中所担负的各种功能就成了全球生命科学工作者共同关注的课题。

基因芯片的出现使同时检测成千上万个基因在生物体内活性的梦想成为现实。目前，DNA微阵列技术已广泛应用于医学、生物学和信息学研究的各个领域，成为生命科学研究的基本工具，如基因序列分析[4]、癌症诊断[5-7]及新药研发[8]等。

1999 年，Golub 等[9]在 *Science* 上发表了关于采用基因芯片技术研究癌症分类问题的文章之后，该研究方向逐渐成为生物信息学领域的研究热点之一，医学、计算机科学、控制科学、生物医学等领域的很多研究人员都在该方向做了大量研究，并根据各自的领域知识提出了大量有效的技术与方法。

基因芯片技术[10,11]为解决肿瘤分类问题拓展了新的思路。通过基因芯片获取肿瘤相关基因表达数据，对肿瘤进行分类，是肿瘤诊断的一个全新手段，也是计算机科学、生物信息学、生物医学等的一个重要交叉研究领域[12]，其可以正确分类组织形态相似的肿瘤亚型，不仅能发现肿瘤的致病基因，还能够挖掘肿瘤发生的本质[13]。

基因表达数据具有高维、小样本、分布不平衡和高噪声等特点。如何对此类数据进行模式学习和数据挖掘，是当前模式识别和机器学习领域内的一个研究热点和亟待解决的问题。

基因表达数据的模式识别过程为：首先进行原始数据预处理，然后进行特征选择-提取，最后基于特征进行分类[14]。然而在实际环境下，训练样本集的分布通常是不平衡的，即在含有若干个类别的训练样本集中每个类别的样本数量不相等，甚至相差很多。这种不平衡会使分类器训练、预测偏向于大类样本的类别，从而对决策产生不良影响。因此，在实际应用中必须考虑样本集分布对分类器训练、预测产生的偏向性。

为了有效处理基因表达数据，主要采用神经网络(Neural Networks，NN)、支持向量机(Support Vector Machine，SVM)、超限学习机(Extreme Learning Machine，ELM)以及决策树(Decision Tree，DT)来设计分类器。为了提高分类系统的稳定性，拟利用分类器集成技术，来改进相关算法。分类器集成可以显著地提高分类器系统的泛化能力和输出稳定性，且已经成功地应用到了很多领域，如地震波分类、光学字符识别、人脸识别等。该技术在计算机辅助医疗诊断方面也具有很好的应用前景。

具体研究内容从数据集和分类器两个方面入手。在数据集方面，利用适当的方法进行特征选择，选择与分类目标密切相关的基因提高分类器模型的泛化性能；创造性地结合两种不同的特征选择算法对基因数据集进行特征选择，能够有效地克服传统特征选择算法的弊端。在分类器方面，构建训练集，利用集成方法提高旋转森林(Rotation Forest，RoF)算法的分类精度和稳定性；利用改进后的粒子群算法优化核超限学习机的内权参数，提高分类器的分类精度；根据输出不一致测度，进行相异性集成，提高分类模型的分类精度和稳定性；通过在超限学习机模型中嵌入误分代价因素，实现对肿瘤的代价敏感分类(Cost-Sensitive Classification，CSC)等。

上述研究内容，构建了一种适用于基因表达数据分类问题的算法框架，如图 1-1 所示，提高了肿瘤基因表达数据的分类精度，一定程度上解决了该研究领域的难

点问题，对推进高维、不平衡数据的研究具有重要理论意义和实用价值。另外，可将研究成果应用于临床肿瘤分类诊断，深入研究肿瘤的发生发展机理及相关致癌基因的表达与调控，促进肿瘤的预测和预防工作，提高人类健康水平。更进一步，可以将不平衡数据挖掘技术推广到信用卡欺诈检测、网络入侵检测、故障诊断等众多应用领域，这将对社会经济的发展产生重要的推动作用。

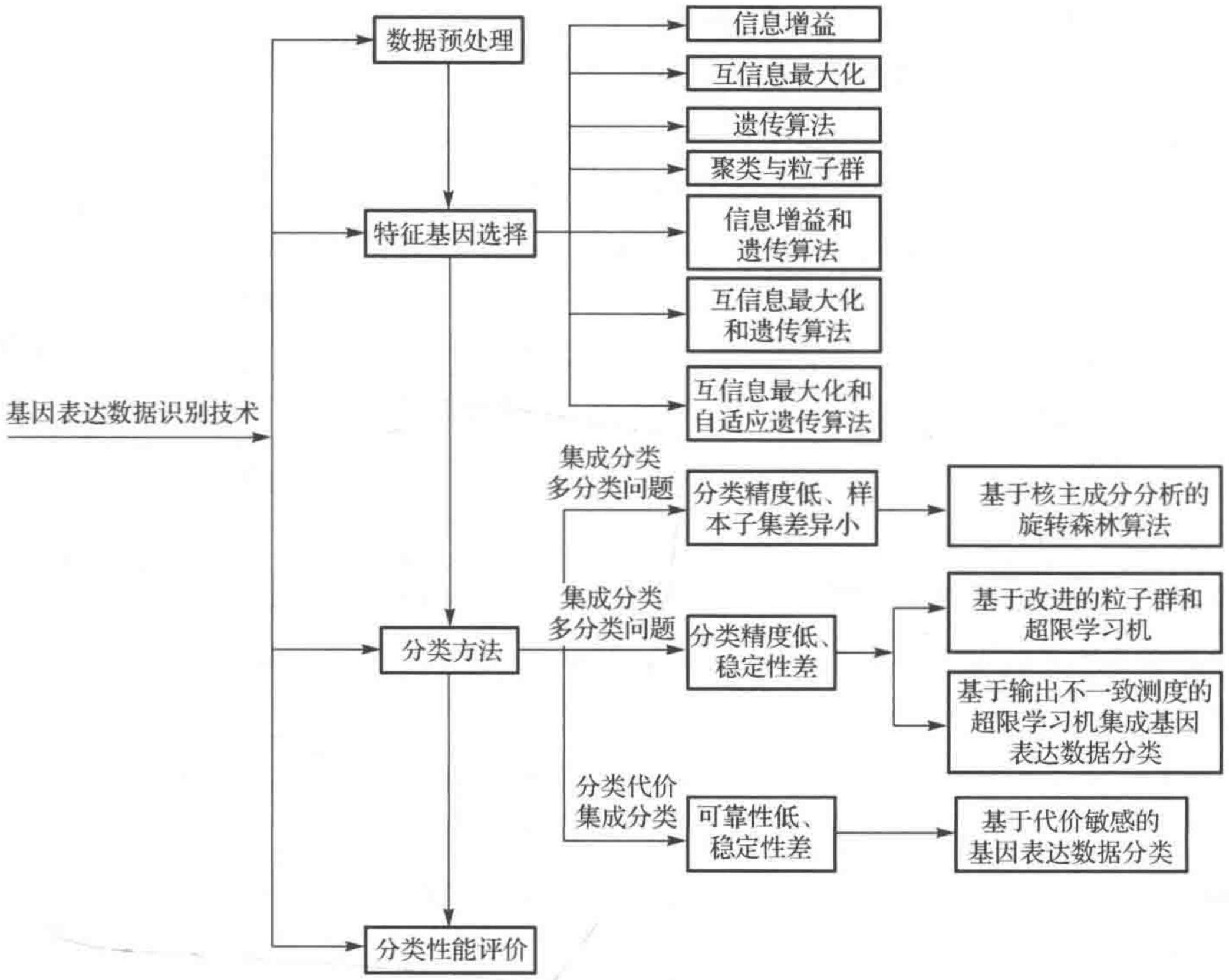


图 1-1 研究内容框架图

参 考 文 献

[1] 林森. 大数据解读癌症[J]. 百科知识,2016,(8):4-7.

[2] Fizazi K. Biennial report on genitourinary cancers[J]. European Journal of Cancer, 2016(66): 125-130.

[3] 施京华. 基于数据挖掘的癌症诊疗决策优化研究[D].上海: 上海交通大学, 2011.

[4] Ao S I, Palade V. Ensemble of Elman neural networks and support vector machines for reverse engineering of gene regulatory networks[J]. Applied Soft Computing, 2011, 11(2): 1718-1726.

[5] Javed K, Wei S, Markus R, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks[J]. Nature Medicine, 2001, 7(6): 673-679.

[6] Pomeroy L, Pablo T, Michelle G, et al. Prediction of central nervous system embryonal tumor outcome based on gene expression[J]. Nature, 2002, 415(6870): 436-442.

[7] Ross D T, Scherf U, Eisen B, et al. Systematic variation in gene expression patterns in human cancer cell lines[J]. Nature Genetics, 2000, 24(3): 227-234.

[8] Valafar F. Pattern recognition techniques in microarray data analysis a survey[J]. Annals of the New York Academy of Sciences, 2002, 980 (1): 41-64.

[9] Golub T R, Slonim D K, Tamayo P, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286(15): 531-537.

[10] Roobol M J. Contemporary role of prostate cancer gene 3 in the management of prostate cancer[J]. Current Opinion in Urology, 2011, 21(3): 225-229.

[11] Evans W E, Guy R K. Gene expression as a drug discovery tool[J].Science Translational Medicine, 2011, 3(107): 107-109.

[12] 黄德双. 基因表达谱数据挖掘方法研究[M]. 北京: 科学出版社, 2009.

[13] Detterbeck F C, Bolejack V, Arenberg D A, et al. The IASLC lung cancer staging project: Background data and proposals for the classification of lung cancer with separate tumor nodules in the forthcoming eighth edition of the TNM classification for lung cancer[J]. Journal of Thoracic Oncology, 2016, 11(5): 681-692.

[14] Roukos D H. Current status and future perspectives in gastric cancer management[J]. Cancer Treatment Reviews, 2000, 26(4): 243-255.

第 2 章　理论基础与相关工作

2.1　基因表达数据特征选择方法

基因表达数据分析作为微阵列技术[1]，蕴涵了巨大的科学价值。它不仅联系了人类基因组序列与临床医学，为人类疾病的诊断和防治开辟了全新的途径，还能够帮助人们探索生物体内基因调控及其相互作用的机理。微阵列可用来检测在不同组织类型中的基因表达差异，如正常细胞和癌细胞，或不同阶段的癌症，以便进行基因表达数据的分类从而实现疾病的识别和诊断，这就要求研究者建立正确反映这些关系的癌症分类模型。

在基因表达数据分类过程中，由于数据维数高达上万，直接分类不仅时间消耗很大，而且分类精度不高，因此首先需要对其进行降维。通常有两种方法进行降维：一是特征选择，二是特征提取。前者从高维数据中选择部分特征，保持原始数据属性；后者通过空间变换，原始数据属性会被破坏。本章主要介绍采用特征选择技术进行数据降维。

特征选择分为过滤法、缠绕法和嵌入法等。过滤法简单、快速，不依赖于具体的分类算法，基因选择结果可以用于不同类型的分类器。缠绕法与特定的分类器结合，由分类器的分类指标来确定选择哪些基因，在迭代过程中逐步优化特征子集，使得分类精度最大化。嵌入法是对缠绕法的改进，通过在一个特定的分类器训练的过程中进行特征基因选择。

对于过滤法，Golub 等[2]利用信噪比准则对每个特征基因所含分类信息进行排序，定义分类信息指标 $G(k)=\left|\mu_{k+}-\mu_{k-}\right|/(\mu_{k+}+\mu_{k-})$，其中 μ_{k+} 和 μ_{k-} 分别是第 k 个基因在正负样本中表达量的均值。将 $G(k)$升序排列取对应的前 d 个基因，在 Leukemia 数据集上选择 50 个基因，取得了 94.74%的正确率。

Fisher 准则[3]利用基因表达水平的均值和方差定义了基因的分类能力：$J(k)=(\mu_{k+}-\mu_{k-})^2/(\mu_{k+}+\mu_{k-})$。Fabian 等[4]采用 Fisher 准则的方法在 Leukemia（白血病）数据集上选择出 $J(k)$值最大的 5 个基因，采用 8 折交叉验证，使用 SVM 分类器取得了 95%的分类正确率。

李颖新等[5,6]认为即使基因表达均值不同，当方差差异很大时，从生物学角度分析，该基因很可能是与急性淋巴细胞性白血病（Acute Lymphoblastic Leukemia，

ALL)致病机理紧密相关的特征基因，并通过对 Golub 等提出的信噪比方法进行改进，加入了方差因素，取得了更好的结果。

信噪比、t-test 等[7]基于统计量对基因进行打分排序的方法有一个假设：数据集服从正态分布。事实上这种假设通常是不成立的。邓林等[8]证明了多数肿瘤数据集不服从正态分布，并提出基于秩和统计的特征选择方法，利用 SVM 对相关基因表达数据进行训练，建立肿瘤诊断模型，在结肠癌数据和白血病数据上取得了较好效果。信息增益[9]也是一种重要的过滤方法，它通过统计某一个基因在分类系统中提供的信息量，来确定对于该基因在分类系统中的重要程度。过滤法是一种简单有效的方法，但没有考虑各个特征基因之间的相关性，没有对各个特征基因进行优化组合，通常无法获得最优特征子集，而缠绕法则是针对这个缺点提出的。

缠绕法是将特征选择和分类器相结合的方法，如贝叶斯分类器、SVM、神经网络、近邻法等，将分类精度作为评价特征子集的标准。

Chris 和 Peng[10]提出了最小冗余特征选择方法，通过定义基因与分类的相关度和基因之间的冗余度，获得了与分类相关度高、冗余度小的特征子集。Li 等[11]利用遗传算法(Genetic Algorithm，GA)和最大似然分类法对多分类基因表达数据集进行特征选择，降低了特征子集冗余度，在多分类问题中获得了更高的分类精度。Shah 和 Kusiak[12]、Maldonado 和 Weber[13]与 Nguyen 和 Torrea[14]采用遗传算法，在保持较高分类精度的前提下，最小化特征冗余，但这种方法时间复杂度较高。王树林等[15]提出一种以 SVM 分类精度为评价标准的选择特征基因子集的启发式宽度优先搜索算法，其优点是能够有效减少特征基因个数，同时使特征子集包含尽可能多的分类相关信息。肿瘤样本集中各个类别的样本个数通常差异较大，针对此问题，李建中等[16]提出了一种解决样本不平衡问题的与数据分布无关的特征基因选择方法，在最小化类内差异和最大化类间差异的策略下，选择敏感的度量函数提高算法的鉴别能力，利用类内差异和类间差异的一致性来增加算法的稳定性与适用性。

缠绕法通过不断迭代的方式逐步约简特征子集，可以最大限度地减少冗余基因，并保持较高的分类精度。然而由于它需要结合具体的分类器在特征选择过程中进行分类，时间复杂度通常很高。

嵌入法实际上是缠绕法的一个改进，它是通过在一个特定的分类器训练的过程中进行特征基因选择的。一个典型方法是利用 SVM 进行特征基因的递归筛选。其中，SVM 作为一个分类器，首先作用在整个训练样本集上，然后对每个基因，计算剔除该基因时 SVM 分类性能的变化。选择分类函数中关联权重绝对值最小的特征基因，并将其从特征基因集合中剔除，重复此过程直至训练集数据为空，最后一起删除的特征基因子集就是最优分类子集。虽然这样能得到一个理想的特征基因子集，但是其时间复杂度太高。

综上所述，目前的基因表达数据分类的特征选择方法主要是基于特征重要程度排序的过滤法、依赖具体分类器的缠绕法和通过缠绕法改进的嵌入法[17-19]。

2.2　神 经 网 络

人工神经网络[20](Artificial Neural Networks，ANN)是对生物神经系统的简单模拟。一组人工神经元按照一定的规则紧密联系在一起，构成神经网络的各层，其中每一个单元有相应的输入，并产生单一的输出。随着计算机技术和生物学的发展，人们对人工神经系统的研究越来越深入。由于实际生物神经系统的复杂性，人工神经系统还只能模拟简单计算、存储记忆等功能。20 世纪 80 年代，Rumelhart[21]提出了感知器模型，首次把神经网络研究应用于工程实践。1986 年，Lecun 等[22]学者提出的多层感知器反向传播(Back Propagation，BP)算法是神经网络领域的重大突破，克服了感知器模型发展的主要困难。目前，ANN 的应用已经渗透到各个领域，如智能控制、模式识别、信号处理、优化计算、生物医学工程等[23]。

人工神经元是构成神经网络的最小单位，一个简单的具有输入、输出、计算功能的人工神经元结构如图 2-1 所示。

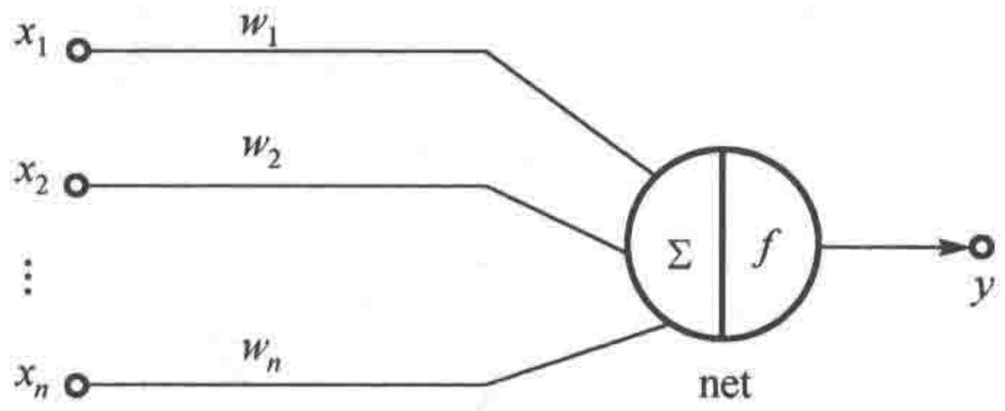


图 2-1　人工神经元结构图

图中 $(x_1,x_2,\cdots,x_n)$ 是输入向量，$(w_1,w_2,\cdots,w_n)$ 是输入向量神经元之间的连接权值，f 为非线性激活函数，神经元计算如下：

$$\text{net}=\sum_{i=1}^{n} w_i x_i \tag{2-1}$$

$$y=f(\text{net}) \tag{2-2}$$

当 f 为阈值函数时，设阈值为 θ，则其输出为

$$y=\text{sgn}(\text{net}-\theta) \tag{2-3}$$

对于要求激活函数 f 可微的情况，一般选取 f 为 Sigmoid 函数(其中 e 为自然常数，约为 2.7)：

$$y=\frac{2}{1+\text{e}^{-2x}}-1 \tag{2-4}$$