学校编码：10384　　分类号_____密级_____
学号：24320121152292　　UDC_____

厦门大学

硕 士 学 位 论 文

基于集成学习的多类基因微阵列数据应用研究

Study on Ensemble Algorithm for Multi-class Gene Microarray Datasets

曾志浩

指导教师姓名：刘昆宏 副教授
专 业 名 称：软 件 工 程

论文提交日期：2015 年 4 月
论文答辩日期：2015 年 5 月
学位授予日期：2015 年 月

指 导 教 师：_____

答辩委员会主席：_____

2015 年 5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果,均在文中以适当方式明确标明,并符合法律规范和《厦门大学研究生学术活动规范（试行)》。

另外,该学位论文为(　　)课题(组)的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。(请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。)

声明人（签名):

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘　要

集成学习是当前机器学习领域的一个研究热点，具体到多分类问题，旨在通过一组差异的分类器共同解决起初的多分类问题，然后经过大多数投票等策略将各个分类器的输出结果进行融合。集成多分类算法相比于单个的优秀分类器往往性能上更准确、更稳定，同时还具有更强的泛化能力。在解决多分类问题时，基于纠错输出编码算法（ECOC）。这是解决多分类问题的一种灵活、高效的算法框架，关键要点是将多分类转变为多个二分类问题。此外，遗传规划算法可用于解决二分类问题，通过进化计算得到准确的分类规则。本文在已有的研究基础上，对基因微阵列数据的集成多分类学习进行了理论探索和实践。

本文主要围绕着集成多分类算法，应用于基因微阵列数据的分析中，主要工作集中在以下方面：

1、提出了基于纠错输出编码的集成多分类学习算法。本文使用了三种 filter 特征选择算法对基因微阵列数据进行维度筛选，实现了将数据相关的纠错输出编码算法应用于基因微阵列数据的识别。通过集成多个纠错输出编码，提升了基因微阵列数据的分类准确率。

2、设计了用于计算纠错输出编码差异度的方法。在集成纠错输出编码时，成员之间差异度越大，集成学习越有效。本文提出了两种选择策略：局部差异度最大化和全局差异度最大化。局部差异度计算纠错输出编码两两之间的差异度，全局差异度计算某个纠错输出编码与剩余的候选编码矩阵差异度的总和。

3、提出了基于遗传规划的集成多分类算法。在解决多分类问题时采用纠错输出编码算法分解为二分类问题。设计遗传规划对于每一个二分类问题进行进化计算，然后集成最终的种群中适应度高、差异度大的个体。通过提升每一个二分类问题的准确率，基于遗传规划的集成多分类算法提升了整体的多分类性能。对于基因微阵列数据，该算法能够同步地筛选癌症关键基因。

关键词：多分类；纠错输出编码；遗传规划

Abstract

Ensemble learning is a current research focus in the field of machine learning. It applies a set of diverse classifiers together in order to solve the original task as specific to multi-class classification problems and fuses the output of each classifier through majority voting. Multi-class classification ensemble algorithm is more accurate and more stable than an excellent classifier, and has greater generalization ability. As to solve multi-class classification task, error-correcting output code algorithm (ECOC) is applied. The key point is transform the original task into multiple binary classification problems which is flexible and effective. Moreover, genetic programming can be used to solve binary classification problems and produces accurate classification rules through evolutionary computation. Based on previous analytical models, this dissertation analyzed multi-class classification ensemble algorithm applied to microarray datasets in theory and studied it in experiments.

This dissertation focuses on multi-class classification ensemble algorithms which are applied to the analysis of microarray datasets. All the work in this dissertation can be summarized as below:

(1) Multi-class classification ensemble algorithm based on ECOC has been proposed. This dissertation uses three different feature selection methods based on filter model and applies data dependent ECOC algorithms to the classification of microarray datasets. By fusing multiple ECOC coding matrices, the accuracy has been improved immensely.

(2) A new method to calculate the diversity among ECOC coding matrices has been proposed. Ensemble learning would be more effective considering the diversity optimization. This dissertation proposes two strategies named as local diversity maximize (MLD) and global diversity maximize (MGD). MLD method calculates the diversity among coding matrices pairwise and MGD method summaries the total diversity comparing with other coding matrices.

(3) Multi-class classification ensemble algorithm based on GP has been proposed. ECOC algorithm is used to transform the original task into multiple binary classification problems and GP algorithm is designed to tackle each binary classification problems. The individuals which have high fitness values and diversities are selected to ensemble. The classification accuracy is improved by enhance the binary learners using genetic programming. It can filter key genes which are related to cancers synchronously.

Key Words: Multi-class Classification; Error-correcting Output Code; Genetic Programming

目 录

Contents

第一章 绪 论

1.1 基因微阵列简介

1.1.1 基因微阵列技术

基因（遗传因子）是支撑遗传、变异的主要物质，储存着生命的诞生、成长、凋亡过程的全部内在信息。基因的表达强度受到环境因素和基因自身两方面的影响，因此生物个体之间基因的表达具有极大的差异性。量化基因的表达强度具有重要的生物意义。一种方法是基因组测序，成本高、速度慢，另一种方法是基因微阵列技术[1]，也被称为 DNA 芯片，具有成本低、速度快的优点，但是精准度相比于前者更低。基因在生物体内的经由 DNA 到 RNA，最后到蛋白质的表达过程。通过将细胞中的 RNA 分子逆转录为 DNA 分子并使用染色杂交技术进行检测，可以得到不同基因片段的表达强度。基因微阵列技术（DNA microarray）就是根据这一原理，由物理学、微电子学和生物信息学等多学科交叉，以及现代芯片制造工艺形成的一种高尖端技术[2]。

基因微阵列技术具有多种重要的用途，包括度量基因表达强度的变化，检测单核苷酸多态性（Single Nucleotide Polymorphisms，简称 SNPs），以及基因分型或定向重测序。基因微阵列技术在癌症等重大遗传性疾病的病理和临床诊断，以及药物的研究与开发发挥着越来越重要的作用。同时，在预测和理解基因的功能及其所参与的生物过程方面也具有广泛的应用前景。与传统的基因测序分析技术相比，微阵列在制造、运作、准确度、效率和成本等多个层面具有较大差异，首先效率更高、成本更低，其次制造、运作更容易，缺点在于准确性更低。在当前的集成技术的辅助下，一块基因微阵列芯片（DNA microarray chip）的大小仅为平方厘米级，如图 1-1 所示。在基因微阵列芯片的表面，密集有序地排列了成千上万个探针点，可以并行地探测大量目标基因的表达强度或者对基因组的多个区域进行基因分型（Genotype）。

基因微阵列生物实验的基本流程包含 5 个主要步骤，如图 1-2 所示。首先，收集测试样本和参考样本。其次，从样本中提取 mRNA（messenger RNA）。使用反转录酶将 mRNA 转录为 cDNA（complementary DNA），并使用荧光分子进

行荧光标记。然后，将 cDNA 置于微阵列芯片上，并与微阵列互补探针杂交。探针的类型包括 cDNA，寡核苷酸及聚合酶链反应生成的基因片段。

图 1-1 Affymetrix 公司生产的基因芯片（资料来源：Wikipedia - DNA Microarray）

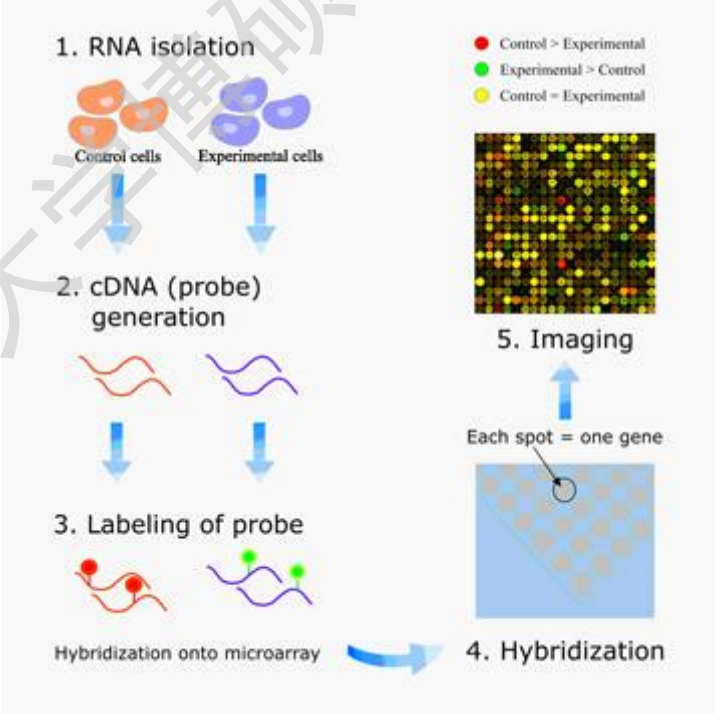

图 1-2 基于 cDNA 微阵列的基因表达检测实验基本流程图（资料来源：Wiki spaces - Southern Blotting and DNA Microarray）

使用激光扫描仪读取充分杂交后的微阵列数据，微阵列上每一个点将呈现不同的颜色，其中红色指示表达，绿色代表抑制，黄色则是中性，黑色对应非活跃的基因。使用图像处理软件识别每个探针点上的真实颜色强度并转换为实数值，得到基因微阵列数据。

1.1.2 基因微阵列数据

基因微阵列数据可以视为一个二维数值矩阵，其中行向量代表一个样本所有基因的表达强度，而列向量代表一个基因在所有样本上的基因表达强度。由于探针点数量（即测量的基因数量）庞大，列向量维度通常在 5,000 - 15,000 的量级。由于实验复杂、费用昂贵，样本数量却稀少珍贵，通常小于 100。这种维度与样本之间极度不平衡的特点被概括为“高维度、小样本”。一方面，DNA 微阵列技术为遗传疾病的诊断、药物研发和生物学研究带来前所未有的机遇，另一方面，其所产生的基因微阵列数据具有高维度、小样本等特点，传统的数据分析方法往往无法获得理想的结果。值得注意的是维度之间不是完全独立的，潜在的相互关系异常复杂，导致两种趋势：计算复杂度急剧增大，或者大量有用的变量被隐藏。此外，基因微阵列数据还天生具有高噪声和高变异等难点，进一步加大了基因微阵列数据的分析难度。

目前，应用于基因微阵列数据分析的方法主要包括：

1．数据处理：评估数据的可靠性和重复性，量化基因表达强度，数据规格标准化，能够使研究人员更有效地利用基因微阵列数据。数据处理是从基因微阵列生物实验到基因微阵列数据分析的关键步骤，对于降低系统误差、消除人为失误，以及保证基因表达水平的可比性、重复性具有重要的意义。许多研究人员致力于探索和改进基因微阵列数据的处理方法，形成了领域内一项重要的研究分支。常见的预处理方法包括：缺省值（Missing Value）修补、奇异值（Outlier）修正，以及数据归一化（Normalization）等。

2．图像分析：一方面是指基因微阵列原始数据可视化，或者分析结果可视化。另一方面，作为一个独立的思路，对基因微阵列生物实验获得的图像数据，采用网格化、热点识别（Spot Recognition）等图像算法[3]，达到移除冗余特征或者标识有效基因等目的。

3．维度规约：在对基因微阵列数据进行分析之前，通常需要降低维度。维

度规约也被称为特征提取。线性方法例如 PCA（principal components analysis）[4]，以及非线性流形学习，Laplacian eigenmaps[5], local linear embedding[6], locally preserving projections[7], 以及 Sammon's mapping[8]等。

4．假设驱动的统计分析：针对基因微阵列数据，采用 t-test, ANOVA, 贝叶斯方法，Mann-Whitney test 方法，识别基因表达强度在样本间具有统计学意义的差异[9]。此类方法通常基于数据满足某种分布的假设，计算数据呈现的差异的统计意义，进而降低后续分析中的“弃真”和“存伪”两类错误。同时这种统计显著性的高低，可用于特征选择。

5．无监督学习：将样本或者基因进行聚类学习，常用的算法包括 k-means，层次聚类算法，以及 Self-Organizing Maps（SOM）[10]等，在构建类簇（cluster）的时候，通过计算基因对之间的距离。这种距离的涵义是线性的，然而基因微阵列数据本身具有丰富多样的结构，甚至无法简单地通过类簇的概念进行描述[11]。因此，一些研究人员探索了能够挖掘出基因相互之间非线性关系的聚类算法，考虑到时间开销，提出了一种折中方案：两个基因已经足以显著地减低线性相似度衡量指标的局限。相关的改进算法包括 FLAME[12]，GeneClust[13]，biclustering[14]和 CLIFF[15]等。

6．监督式学习：基于微阵列数据和分类学习算法构建分类预测模型，对于标签未知的样本，可以预测其类别归属。常用的分类算法包括线性回归、k-近邻算法、决策树、随机森林，以及神经网络等。此外，研究人员探索了基于不同种类的进化算法，例如遗传算法、粒子群算法，以及蚁群算法等。在监督式学习中，基因微阵列数据经过基于不同指标的特征归约和筛选，例如信息增益法、Gini impurity 准则等[16]。

总而言之，基因微阵列技术不仅涉及生物信息学的研究，而且还包含微阵列数据的分析处理方法。