相关性和相似度联合的癌症分类预测

张学扶 曾 攀 金 敏
(湖南大学信息科学与工程学院 长沙 410006)

摘 要 基于经验型组织病理学的癌症诊断往往误诊率很高。从基因层次对癌症进行分析和研究是现阶段提高癌症分类预测精度的重要途径之一。生物学研究表明,同种癌症的关联基因有着共同的功能特点。基于此,文中提出相关性和相似度联合的癌症分类预测集成方法。首先,一方面,从统计学角度分析基因的差异化表达,利用互信息方法对基因表达谱数据进行相关性计算;另一方面,从生物机理上进行基因间的相似性分析,结合拓扑相似性和语义相似性分别对蛋白质互作网络和GO数据进行基因间的功能相似度计算。以上两者结合,即通过同时最大化目标集合的相关性和相似度筛选出特征基因集。然后,通过Bootstrap方法对数据集进行多样性采样,在前面所选特征基因集的基础上利用多种机器学习算法训练得到多个差异化较大的分类预测模型。最后,利用得到的多模型对测试样本进行分类预测,通过决策模型得到最终的分类结果。对GEO中4种不同癌症数据集进行分类预测研究,并将所提方法与最近的研究方法进行综合对比,结果所提方法在各数据集上的分类预测精度均提高5%左右,相比IG/SGA方法最高能达到10%的精度提升。实验结果表明,相关性和相似度联合的方法有效提高了癌症的分类预测精度,选择得到的特征基因有利于揭示生物学意义,且将多种算法优势互补,可解决单个分类算法适用范围受限的问题。

关键词 癌症分类,相关性,语义相似性,拓扑相似性,多样性采样,多算法多模型

中图法分类号 TP391.9 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.07.046

Cancer Classification Prediction Model Based on Correlation and Similarity

ZHANG Xue-fu ZENG Pan JIN Min
(College of Computer Science and Electronic Engineering, Hunan University, Changsha 410006, China)

Abstract Cancer diagnosis based on empirical histopathology often has a high rate of misdiagnosis. Analyzing and studying cancer from the gene level is one of the important ways to improve the accuracy of cancer classification prediction at this stage. Biological studies have shown that the related genes of the same kind of cancer share common functional characteristics. Based on this, this paper proposes an integrated method of correlation and similarity for cancer classification prediction: First, on the one hand, statistical analysis of differential expression of genes The use of mutual information methods to perform correlation calculations on gene expression profiles. On the other hand, the similarity analysis between genes was performed on the basis of biological mechanisms, and the protein interaction network and GO data were genetically performed based on topological similarity and semantic similarity, respectively. The functional similarity calculation between the two, the combination of the two, that is, the feature set is selected by simultaneously maximizing the relevance and similarity of the target set; then the diversity of the data set is sampled by Bootstrap method, and the selected feature set in the front Based on the above, we use multiple different machine learning algorithms to train a number of differently differentiated prediction models. Finally, the multiple models are used to classify the test samples and obtain the final classification results through the decision model. The classification prediction of four different cancer datasets in GEO was compared with the latest research methods, and the classification accuracy on each dataset was improved by about 5%, which is up to 10% higher than that of IG/SGA methods. Increased accuracy. The experimental results show that the method of combining relevance and similarity can effectively improve the accuracy of cancer classification prediction. Selecting the obtained characteristic genes is beneficial for revealing biological significance, and the advantages of multiple algorithms can be complemented to solve the problem that the application scope of a single classification algorithm is limited. problem.

Keywords Cancer classification, Correlation, Semantic similarity, Topological similarity, Diversity sampling, Multiple algorithms and multiple models

到稿日期:2018-06-15 返修日期:2018-11-04 本文受国家自然科学基金项目(61773157)资助。

张学扶(1992—),男,硕士生,主要研究方向为数据分析,E-mail:xuefu_zhang@163.com;曾 攀(1990—),男,硕士生,主要研究方向为数据分析;金 敏(1973—),女,博士,教授,主要研究方向为数据分析,E-mail:jinmin@hnu.edu.cn(通信作者)。

1　引言

癌症一直是威胁人类生命健康的顽症。随着现代医疗水平的提升，早期癌症患者的治愈率约达到 80%～90%。因此，癌症的及早发现至关重要。传统的癌症诊断往往建立在病理形态学、组织学和免疫学等基础上，在很大程度上依赖于经验，并容易出现误诊。而分子生物学研究表明，不同阶段的癌细胞大多有着不一样的基因特性，从生物信息学的角度进行癌症分类研究已成为当今生物学领域的重要课题之一[1]。随着机器学习的发展和日渐成熟，机器学习为癌症分析和研究提供了有效的工具和全新的方法。现阶段，利用机器学习进行癌症分类研究主要包括 3 个方面：数据集的选取、特征选择和分类器的设计[2]。

在数据集方面，近年来随着生物高通量实验技术的快速发展，人们得到了越来越多的生物数据，包括基因表达谱数据、蛋白质互作网络以及功能注释信息数据等。Nguyen 等[3]利用基因表达谱数据以及 5 种不同的评价准则，通过分析基因表达水平的差异对层次分析法（AHP）进行改进，定量地对基因进行排序，从而提取出靠前的基因作为癌症功能基因，并使用马尔可夫模型作为分类器进行分类实验。这一方法在某些癌症数据集上相比其他分类器有一定的改进，但其分类精度在某些癌症数据集上并没有获得突破性的提高，这也进一步暗示了癌症的产生不仅仅与基因的表达差异有关，还可能涉及其他因素。因此为了更为精确地进行癌症分类，还需要进一步挖掘更多与癌症产生相关的因素，探究癌症产生的根源。在这个背景下，蛋白质互作网络（Protein Protein Interaction network，PPI）数据被用于各种研究，以确定癌症基因。Li 等[4]提出一种新的方法来鉴定候选疾病基因，这种方法使用了异质基因组和表型数据集，首先使用这两种数据集开发单独的基因网络，然后将这两个基因网络合并成一个网络并使用随机游走来识别疾病基因。这些数据集成方法都是通过引入 PPI 数据来进行特征选择，但是都没有进一步结合有效的分类模型进行癌症分类预测。笔者团队在 2015 年提出了一种新的结合拓扑相似性和语义相似性的疾病基因预测方法，其结合基于 GO（Gene Ontology）术语的语义相似性来弥补 PPI 数据的不完整性[5]。3 种生物数据都有各自的特点：基因表达谱数据是进行癌症诊断最为直接的数据来源，数据格式简单且统一；各个蛋白质互作网络数据库的数据量、存储格式、注释方式、查询方式不尽相同，阻碍了数据库之间的信息交换和数据整合，这导致了蛋白质互作网络数据往往不够完整，网络中的关系的精度也会大大下降[6]；对于功能注释信息数据，如 GO 数据，KEGG 数据库等都非常完善，是功能注释信息数据最常用的数据库。因此，如何将不同的生物数据综合应用到癌症诊断研究问题中是当前亟需解决的问题。

在特征选择方面，由于基因表达数据具有分布不平衡、高维性、小样本、高噪声的特点，直接进行分类不仅耗时长，而且分类精度不高，因此首先需要对其进行降维，即去除无关基因，提取致癌基因[7]。特征选择方法分为过滤法（Filter）、包装法（Wrapper）和嵌入式方法（Embedded）等。过滤法简单快速，不依赖于具体的分类算法，直接从基因数据自身的特点出发，利用基因在不同癌症样本中表达水平的取值分布，比如互信息量、统计学检验等方法。这种方法很容易实现，但是忽略了基因与基因间的复杂关联，其选择的特征之间可能存在冗余。包装法的核心思想是用分类器的性能指标来评价基因或基因子集的重要性。当包装法与不同的分类器相结合时，能够筛选出相应的基因子集。包装法选择的基因子集能够与分类器的决策机制很好地吻合，在分类检验样本时可获得最高的准确率。但是包装法选择出的特征基因对分类算法有较大的依赖性，计算代价高，易出现过拟合的现象；而且该方法不是直接建立在基因表达差异上，所以筛选出的基因不具有明确的生物学意义。嵌入式法实际上是包装法的一种改进，它是在一个特定的分类器训练过程中进行特征基因选择的。虽然这样能得到一个理想的特征基因子集，而且对分类精度的提高也有很大的帮助，但是其时间复杂度太高。以上 3 种方法各有优劣，一种理想的特征选择方法不仅需要考虑到过滤法忽略的基因间的复杂关联问题，选择得到的特征基因还应具有明显的生物学意义。因此，需要结合所选择的数据集的数据特点和后续的分类模型做进一步研究。

在分类器设计方面，分类器是直接体现特征选择优劣的工具，好的分类器不但可以有效地验证所选择致癌基因的性能，还能大大缩短特征选择和分类所用的时间。目前有很多机器学习算法被用于分类的研究，常见的有支持向量机（Support Vector Machine，SVM）[8]、K 最近邻（K-Nearest Neighbor，KNN）[9]、随机森林（Random Forests）[10]等算法。SVM 是基于间隔最大化的一种监督分类学习方法，在解决小样本、非线性及高维模式识别时表现出许多特有的优势，但是核函数的选择和参数设置缺乏理论指导。K 最近邻分类算法是数据挖掘分类技术中最简单的方法之一，易于理解，且可用于非线性分类，在维度较低的数据中能获得较高的准确率，但这种算法不能很好地解决样本不平衡的问题。随机森林是利用多棵树对样本进行训练并预测的一种分类器，能够处理高维度数据，训练速度快，而且能平衡样本不平衡导致的误差，但是随机森林算法在某些噪声较大的数据上会出现过拟合的问题。而现有的很多集成算法，如多层感知神经网络 MLP 和集成学习算法 AdaBoost 等，分类时间由于参数的调优往往会很长，因此，任何一种分类模型都有其特定的适用范围。目前，在模型训练时有单模型、单算法多模型和多算法多模型 3 种方法[11]。单模型方法可能出现过拟合、泛化能力受限的问题。单算法多模型虽然能够很好地解决单模型可能出现的问题，但局限于单一算法，无法解决单个算法适用范围受限的问题。多算法多模型[12]针对多个算法，对每个算法生成一个模型，多个算法优势互补，克服了单个算法适用范围受限的缺点，但多算法多模型的算法选取上并没有理论的指导，需要通过理论分析和实验指导来确定选取何种算法[13]。

研究表明[14]，与同一种癌症相关联的基因有着共同的功能特点，这反映在它们编码的蛋白质有着相互作用，即这些基因参与相同的代谢通路。因此，癌症基因除了强烈的差异化表达外，其蛋白质产物与其他癌症基因的蛋白质产物有强烈的互作用。针对上述问题，本文基于同种癌症的关联基因有着共同功能的特点，提出相关性和相似度联合的癌症集成分

类预测模型。首先,一方面,从统计学角度分析基因的差异化表达,利用互信息方法对基因表达谱数据进行相关性计算;另一方面,从生物机理上进行基因间的相似性分析,利用拓扑相似性对蛋白质互作网络进行基因间的功能相似度计算,解决了过滤法忽略基因与基因间的复杂关联的问题,同时结合基于 GO 数据的语义相似性,克服了 PPI 数据不完整导致的网络关系弱化的问题,通过同时最大化所选特征基因集的相关度和相似度进行特征选择。然后,利用 Bootstrap 方法进行多样性采样,以解决单分类算法过拟合的问题,使泛化能力得到增强;在前面所选特征基因集的基础上使用 SVM,KNN 和 RandomForest 算法训练得到多个差异化较大的分类模型,将多个算法优势互补,以解决单个分类算法适用范围受限的问题。最后,在这些模型的基础上投票表决,得到最终的分类结果。本文将该方法应用在 GEO(Gene Expression Omnibus)数据库中 4 种不同癌症的数据集上,通过实验分析比较不同特征选择方法以及不同的分类器模型的分类精度。

2 相关性和相似度联合的特征选择方法

由于基因表达数据具有分布不平衡、高维性、小样本、高噪声的特点,直接进行分类不仅耗时长,而且分类精度不高,因此首先对其进行降维;同时为减少降维的时间消耗,本文采用过滤法作为基因选择的方法,通过整合基因表达数据和功能注释信息数据进行特征基因的选取。首先,利用互信息[15]和 Jaccard 相似系数[16]分别进行基因相关性分析以及基因间功能拓扑相似度计算,解决了过滤法忽略了基因间的复杂关联的问题,并结合语义相似性分析,克服了 PPI 数据不完整导致网络关系弱化的问题。然后,挑选出若干特征基因,使该特征基因集的相关度和相似度同时最大化。此类问题本质上是图论中的图覆盖问题[17]。

2.1 功能相似性计算

一般而言,与同一种癌症相关联的基因往往具有共同的功能特征;这些基因的蛋白质产物也具有相互作用。因此,癌症基因的一个重要特征就是其蛋白质产物与其他癌症基因蛋白质产物紧密相关。同时有研究发现:蛋白质互作网络(PPI)中距离越短的蛋白质往往涉及相同的生物功能,并且相邻的蛋白质比非相邻的蛋白质更有可能具有相同的生物功能。这是因为同种癌症的致癌基因都有着相似的功能特点,这反映为基因编码的蛋白质有着互相作用,即这些致癌基因参与相同的代谢通路。因此需要定量地计算两个基因之间的功能相似度,本文采用 Jaccard 相似系数和 HRSS(Hybrid Relative Specificity Similarity)方法来分别对蛋白质互作网络和 GO 注释信息进行拓扑相似性计算和语义相似性计算[18]。HRSS 方法是由吴晓梅等[19]提出的,其通过融合基于边和点方法的优势,在计算语义相似领域取得了不错的效果。

(1)拓扑相似性计算

蛋白质互作网络[20]通常表示为无向图,其中图的结点对应为蛋白质,而图的边表示蛋白质之间的相互作用关系。这种相互作用既包括蛋白质之间直接的物理相互作用,也包括蛋白质之间间接的功能相关性;除了包含实验数据、从文献中进行文本挖掘的结果和综合其他数据库数据外,还包含利用生物信息学方法预测的结果。蛋白质互作网络图中边的权重就是对这些不同方法得到的结果给予一定的权重,最终给出一个综合的置信度。

在蛋白质互作网络图中,本文采用式(1)的相似系数进行拓扑相似度计算,Jaccard 相似系数值越大,基因功能相似度越高。对于给定基因 A_i,其相邻基因集用 N_i 表示,$w_{ij}\in[0,1]$为基因 A_i 和基因 $A_j\in N_i$ 之间边的置信度。置信度 w_{ij} 和相邻基因集 N_i 均可以从 PPI 网络图中获得。N_{ik} 表示基因 A_i 和基因 A_k 的公共相邻基因集,即 $N_{ik}=N_i\cap N_k$。$N_{i\backslash k}$($N_{i\backslash k}=N_i-N_{ik}$)基因集表示基因 A_i 的相邻基因集,而不是基因 A_k 的相邻基因集。

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|} \tag{1}$$

基因 A_i 和基因 A_k 的拓扑相似度可表示为:

$$J(A_i,A_k)=\frac{\sum\limits_{A_j\in N_{ik}}\min\{w_{ij},w_{kj}\}}{\sum\limits_{A_j\in N_{i\backslash k}}w_{ij}+\sum\limits_{A_j\in N_{ik}}\max\{w_{ij},w_{kj}\}+\sum\limits_{A_j\in N_{k\backslash i}}w_{kj}} \tag{2}$$

从式(2)可以看出,如果两个基因的相邻基因集和对应的置信度越接近,则这两个基因的功能相似度就越高。另一方面,如果两个基因没有公共的相邻基因,则它们在功能上是不相似的。同样,式(2)满足 Jaccard 相似系数公式的性质:

1)$J(A_i,A_k)\leqslant 1$;

2)$J(A_i,A_k)=1$,当且仅当 $N_{ik}=N_i=N_k$,$w_{ij}=w_{kj}$,$\forall A_j\in N_{ik}$;

3)$J(A_i,A_k)=0$,当且仅当 $N_{ik}=\varnothing$;

4)$J(A_i,A_k)=J(A_k,A_i)$。

(2)语义相似性计算

本文采用 GO 注释信息作为基因的语义信息。衡量基因间的语义相似性即衡量基因本体术语在 GO 的有向无环图中的位置关系以及每个术语所包含的信息量差别。而术语描述了基因的分子功能、生物过程、细胞组成等方面的信息。因此,语义相似性程度较大的基因之间,其功能也较为接近。对于两个基因 g_1 和 g_2,其语义相似度由式(3)计算:

$$HRSS^{GO}_{MAX}(g_1,g_2)=\max_{\substack{go_i\in tg_1\\go_j\in tg_2}}(HRSS(go_i,go_j)) \tag{3}$$

其中,tg_1 和 tg_2 分别是 g_1 和 g_2 两个基因的术语集合。$HRSS(go_i,go_j)$则为两个 GO 术语之间的相似性公式:

$$HRSS(go_i,go_j)=\frac{1}{1+\gamma}\frac{\alpha_{IC}}{\alpha_{IC}+\beta_{IC}} \tag{4}$$

其中,α_{IC}代表了最大信息量共同祖先(*MICA*)的特异性,β_{IC}表示术语 i 和术语 j 的普遍性,γ 衡量了术语 i 和术语 j 在经过共同祖先时的距离。同理,式(3)也满足 Jaccard 相似系数公式的性质。

综上所述,本文综合了两种相似性的计算方法,将两个基因 A_i 和 A_j 的综合功能相似度表示为:

$$S(A_i,A_k)=\max(J(A_i,A_k),HRSS^{GO}_{MAX}(g_1,g_2)) \tag{5}$$

2.2 互信息的相关性计算

互信息是一种信息度量,表示一个随机变量包含另一个随机变量的信息量,用于度量两个变量之间的相关性[21]。本文采用互信息来计算基因表达水平与类别标签之间的相关性。基因 A_i 与标签 D 之间的相关性 $\gamma(A_i,D)$为:

$$\gamma(A_i,D)=I(A_i,D) \tag{6}$$

其中，$I(A_i,D)$表示基因 A_i 与标签 D 之间的互信息量，计算如下：

$$I(A_i,D)=H(A_i)+H(D)-H(A_i,D) \tag{7}$$

其中，$H(A_i)$和 $H(D)$分别表示基因 A_i 与标签 D 的信息熵；$H(A_i,D)$表示基因 A_i 与标签 D 的联合信息熵。信息熵用来衡量随机变量的不确定性。给定一离散随机变量 X 和其概率密度函数 $p(x)=\Pr\{X=x\}$，$x\in X$，离散随机变量 X 的信息熵定义为：

$$H(X)=-\sum_{x\in X}p(x)\log p(x) \tag{8}$$

同理，对于两个随机变量 X 和 Y，其联合信息熵为：

$$H(X,Y)=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log p(x,y) \tag{9}$$

其中，$p(x,y)$为联合概率密度函数。

在基因表达数据集中，样本的类别标签由离散的标识符表示，而基因表达值是连续的。因此，为了用互信息来衡量基因与标签之间的相关性，需要将连续的基因表达值划分成若干个离散的部分，然后计算先验概率和联合概率，以得到基因与标签之间的相关程度。本文采用文献[22]中的离散化方法来离散化连续的基因表达值。对于 n 个基因表达值，基于其均值 μ 和标准方差值 σ 进行离散化：若表达值大于 $\mu+\sigma$，则置为 1；若表达值位于 $\mu-\sigma$ 和 $\mu+\sigma$ 两者之间，则置为 0；若表达值小于 $\mu-\sigma$，则置为 -1。这 3 个状态值分别对应为基因水平的高表达、基准表达和低表达。

2.3　特征选择流程

本文综合基因表达数据和蛋白质网络数据，以进行癌症基因的提取。因此，基于基因表达数据集中的所有基因 C，通过同时最大化所选基因集 S 的相关性和功能相似度，得到最终的特征基因集 S。假设集合 $C=\{A_1,\cdots,A_i,\cdots,A_j,\cdots,A_m\}$为基因表达数据集中的 m 个基因，S 为所选特征基因。定义 $\gamma(A_i,D)$为基因 A_i 与标签 D 之间的相关度，$S(A_i,A_j)$为基因 A_i 和基因 A_j 的功能相似度。因此，所选特征基因集 S 的总相关度为：

$$\tau_{\text{relevance}}=\sum_{A_i\in S}\gamma(A_i,D) \tag{10}$$

所选特征基因集 S 的总功能相似度为：

$$\tau_{\text{similarity}}=\sum_{A_i\neq A_j\in S}S(A_i,A_j) \tag{11}$$

因此，算法的关键在于选择特征基因集 S，使得该基因集的总相关度 $\tau_{\text{relevance}}$ 和总功能相似度 $\tau_{\text{similarity}}$ 最大，即最大化目标函数：

$$\tau=\alpha\tau_{\text{relevance}}+(1-\alpha)\tau_{\text{similarity}} \tag{12}$$

其中，α 为平衡因子。

本文采用贪婪法解决这一最大化问题，具体步骤如算法 1 所示。

算法 1

输入：原基因集合 $C=\{A_1,\cdots,A_i,\cdots,A_j,\cdots,A_m\}$，m 为基因表达谱数据集中的基因个数

输出：特征基因集合 S

Step 1　初始化 $S=\varnothing$。

Step 2　对于每一个基因 $A_i\in C$，计算 $\gamma(A_i,D)$。

Step 3　选择相关度最高的基因 A_i，将其从集合 C 中删除并添加至集合 S 中。

Step 4　重复 Step 5 和 Step 6，直到集合 S 中的个数满足要求。

Step 5　计算集合 C 中每一个基因相对集合 S 中的总功能相似度，并删除集合 C 中相似度为 0 的基因。

Step 6　在集合 C 中剩下的基因中，选择基因 A_j，使得式(13)最大：

$$\alpha\gamma(A_i,D)+\frac{1-\alpha}{|S|}\sum_{A_i\in S}S(A_i,A_j) \tag{13}$$

将基因 A_j 从集合 C 中删除并添加至集合 S 中。

Step 7　停止。

3　癌症分类集成模型

标准的支持向量机学习算法问题可以归结为求解一个受约束的二次规划(Quadratic Programing，QP)问题，SVM 的原理是寻找满足分类要求的最优分类超平面，使得该超平面在保证分类精度的同时，能够使其两侧的空白区域最大化。理论上，支持向量机能够实现对线性可分数据的最优分类。

KNN 算法也称 K 近邻算法，是一种基本分类与回归方法，其主要思想是：如果一个样本在特征空间中的 K 个最相邻(即特征空间中最邻近)的样本中的大多数都同属于某一个类别，则该样本也属于这个类别。

随机森林学习算法是基于决策树的一种集成学习算法。决策树是被广泛应用的一种树状分类器，在树的每个节点通过选择最优的分裂特征不停地进行分类，直到达到建树的停止条件。当输入待分类样本时，决策树确定一条由根节点到叶节点的唯一路径，该路径上叶节点的类别就是待分类样本所属的类别。随机森林解决了决策树性能瓶颈的问题，对噪声和异常值有较好的容忍性，对高维数据分类问题具有良好的可扩展性和并行性。此外，随机森林是由数据驱动的一种非参数化方法，只需通过给定样本来学习训练分类规则，并不需要先验知识。

由于基因表达数据具有样本小且分布不平衡的特点，本文采用 Bootstrap 方法[23]进行多样性采样。Bootstrap 方法也称为自助法，它是一种有放回的抽样方法，是一种用小样本估计总体值的非参数统计方法。已经证明，在初始样本足够大的情况下，Bootstrap 抽样能够无偏差地接近总体的分布。其基本步骤是：1)设定抽样比例，即从原始样本中按照指定的比例抽取样本；2)设定样本集个数 N，即重复上述 N 次，得到 N 个样本集。

为了有效减少单分类算法可能出现的过拟合问题，增强模型泛化能力，增加模型多样性，提高分类精度并使多个算法优势互补，利用 Bootstrap 方法对基因表达数据集进行多次可重复采样，然后利用 SVM 算法、KNN 算法和随机森林算法分别对采样后的训练数据训练出多个差异化较大的模型。

4　完整框架流程

本文提出的分类模型基于多算法多模型的分类方法，通过 Bootstrap 方法进行多样性采样，在前面所选特征基因集的基础上利用 SVM，KNN 和 RandomForest 算法训练得到多个差异化较大的分类模型，并在这些模型的基础上投票表决，得到最终的分类结果。

该方法的具体步骤如图 1 所示,主要分为以下 3 步。

第 1 步 数据集选取及特征选择。影响癌症分类准确度的因素主要涉及基因表达差异程度和基因与基因之间的关联程度。为充分利用这两个因素,本文不仅选用基因表达的数据,还选用蛋白质互作网络数据,并且采用互信息衡量基因表达的差异程度,使用 Jaccard 相似系数衡量基因与基因之间的功能相似度。在后续步骤开始前,本文实验采用十折交叉验证方法将基因表达数据划分为训练样本集和测试样本集。对训练样本集进行特征基因选择,具体方法见第 1 节。

第 2 步 构建多模型。在 k 折交叉验证法的基础上,通过 Bootstrap 方法,使用采样率 μ 对基因表达数据集进行 m_1,m_2 和 m_3(均为奇数)次可重复采样。利用 SVM 算法训练 m_1 个模型 $S_1,S_2,\cdots,S_{m_1}$;利用 KNN 算法训练 m_2 个模型 $K_1,K_2,\cdots,K_{m_2}$;利用 RandomForest 算法训练 m_3 个模型 $R_1,R_2,\cdots,R_{m_3}$。得到多模型后,利用这些模型对训练样本集进行训练。

第 3 步 对测试样本进行分类。利用第 1 步得到的训练样本集和测试样本集、第 2 步特征选择得到的特征基因集,以及第 3 步训练得到的多模型,对测试样本集进行测试,并计算分类精度。

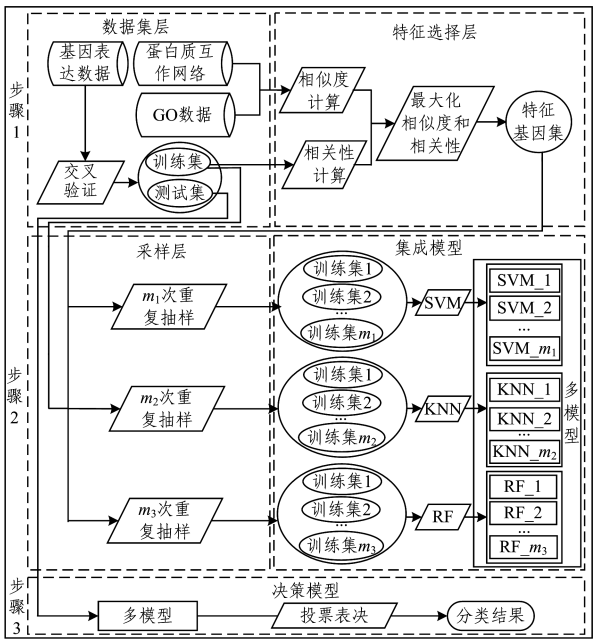

图 1 相关性和相似度联合的癌症分类预测模型框架图

Fig. 1 Correlation and similarity combined cancer classification prediction model framework

5 实验研究

本文实验的环境为:Inter Core i5-4210M 处理器,8GB 内存,2.60GHz,Windows 10 操作系统。

5.1 实验数据集

实验所用数据集包括基因表达谱数据、蛋白质互作网络数据和 GO 数据库。其中,基因表达数据是从 GEO[1)] 数据库[24]中下载得到,共 4 种癌症数据集,详细情况如表 1 所列。蛋白质互作网络数据来源于 STRING(Search Tool for the Retrieval of Interacting Genes)数据库[25],STRING 数据库是一个记录已知蛋白质并预测其相互作用关系的综合数据库。另外,本文还使用 Uniprot 数据库来对应蛋白质及编码蛋白质的基因。

表 1 基因表达谱数据集的总体情况

Table 1 Overall situation of gene expression profile dataset

基因表达谱数据集	癌症类型	总样本数	正常样本数	癌症样本数	探针数
GSE9476	AML	64	38	26	22283
GSE10797	乳腺癌	66	10	56	22277
GSE25070	结肠癌	52	26	26	24526
GSE19804	肺癌	120	60	60	54675

5.2 模型评估指标

对于二分类问题的性能评估方法,分类结果的“混淆矩阵”(见表 2)是各项性能评估的重要标准。根据样本的真实类别与学习器预测类别的组合,将样本划分为 TP(True Positive)、FN(False Negative)、FP(False Positive)和 TN(True Negative)4 种情形。令 TP,FN,FP,TN 分别表示其对应的样本数,则显然有 $TP+FN+FP+TN=$样本总数。本文所用到的性能评估指标如下:

$$accuracy=\frac{TP+TN}{TP+FN+FP+TN} \tag{14}$$

表 2 分类结果的混淆矩阵

Table 2 Confusion matrix of classification result

真实情况	预测结果	
	癌症样本	正常样本
癌症样本	TP	FN
正常样本	FP	TN

5.3 实验结果分析

在本文实验研究中,采用 10 折交叉验证法,采样率 $\mu=0.7$,重复采样次数 $m_1=5,m_2=5,m_3=5,m_4=5$。本文通过简单的实验来设置各算法中相应的参数,例如在 SVM 算法的核函数选取上,本文分别采用 3 个核函数(线性核函数、多项式核函数和径向基函数)进行实验,结果显示线性核函数和径向基函数得到的分类精度相差不超过 1%,故本文使用径向基函数作为 SVM 算法的核函数;同样,RandomForest 算法中树数目为 10;KNN 算法中最近邻数目为 5,采用欧几里得距离作为距离度量方式。

案例 1 本实验对特征选择方法中式(12)的不同平衡因子 α 的分类精度进行对比研究。因为平衡因子 $\alpha\in(0,1)$,本文以 0.1 为步调,取该区间中所有可能的 α 值,α 有 9 种取值情况,即 $\alpha=\{0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9\}$。表 3 列出 4 种癌症数据集在不同平衡因子 α 下的分类精度情况。实验对每一 α 取值利用本文特征选择方法进行特征选择计算,最后都采用本文分类预测集成模型进行分类。图 2 是 4 种癌症数据集随着平衡因子的变化得到的分类精度预测曲线图。本实验在进行特征选择时,均以 20 个特征基因为目标。

1) http://www.ncbi.nlm.nih.gov/geo/

表 3　不同平衡因子 α 在各数据集上实验的分类精度

Table 3　Classification accuracy of experiments on different datasets with different balance factors

（单位：%）

数据集	不同的平衡因子								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
GSE9476	92.3	93.5	95.2	95.6	96.3	98.0	100.0	97.6	97.8
GSE10797	96.7	95.9	90.9	91.4	95.6	96.1	98.5	97.2	97.0
GSE25070	98.6	97.8	96.5	98.6	97.2	98.2	99.0	99.2	99.0
GSE19804	82.3	83.5	84.7	85.9	86.6	86.2	87.5	86.7	88.0

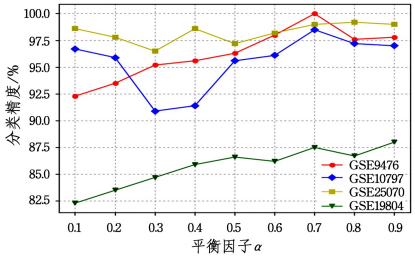

图 2　不同平衡因子 α 在各数据集上的分类预测精度曲线图

Fig. 2　Curve of classification and prediction precision of different equilibrium factors in different data sets

表 3 的分类预测结果显示，对于不同的平衡因子 α，本文分类预测集成模型的分类预测精度差异较大。平衡因子 α 和分类预测精度并没有明显的线性关系，但是从表 3 的数据可以知道当 $\alpha=0.7$ 时，本文分类预测集成模型在各数据集上实验的分类预测精度均比其他的 α 取值要高。图 2 的分类预测精度曲线显示，GSE9476 数据集的分类预测精度曲线存在一个凸峰点，即当 $\alpha=0.7$ 时，精度达到最高，为 100.0%。分析其他数据集的精度曲线发现，在 $\alpha=0.7$ 时均达到凸峰或是接近凸峰。通过深入分析发现，α 在 0.1～0.6 之间以及在 0.8 以上时，各数据集上的分类预测精度都不高；结合式(12)可知，α 越小，特征选择方法就越偏向于相关性分析，α 越大，就越偏向于相似性分析。这一实验结果表明，α 的取值不是越大越好，也并非越小越好，而是需要深入权衡相关性分析与相似性分析在特征选择时的权重。下文所有实验中 α 均取 0.7。

案例 2　本实验对不同特征基因数目进行研究，分别对 5 种不同特征基因数(基因数为 5，10，20，50 和 100)进行实验。表 4 给出不同特征基因个数在各数据集上实验的分类精度情况。实验对每一特征基因数目使用本文特征选择方法进行特征选择计算，最后都采用本文分类预测集成模型进行分类。图 3 是 4 种癌症数据集随着特征基因数的变化得到的分类精度预测曲线图。

表 4　不同特征基因个数在各数据集上实验的分类精度

Table 4　Classification accuracy of experiments with different number of genes on each dataset

（单位：%）

数据集	不同特征基因个数				
	5	10	20	50	100
GSE9476	99.5	100.0	100.0	99.5	99.5
GSE10797	96.7	97.8	98.8	97.3	99.5
GSE25070	98.5	100.0	99.0	98.6	99.1
GSE19804	87.5	86.5	87.4	85.3	86.6

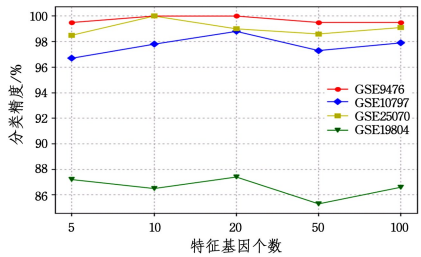

图 3　不同特征基因数在各数据集上的分类预测精度曲线图

Fig. 3　Curve diagram of classification and prediction accuracy of different characteristic genes in each data set

表 4 的分类预测结果表明，特征基因数对分类预测精度没有明显的影响，例如 GSE9476 数据集的精度波动范围小于 0.5%。分类预测精度曲线图显示，特征基因数与分类精度之间并没有一定的规律和趋势。这进一步表明，特征基因数对分类预测精度影响不大，因此，本文所有实验中特征基因数都取 20。

为进一步分析选择出来的特征基因的表达差异水平，本实验对 4 种癌症数据集的 20 个特征基因的表达差异进行分析。图 4 是 GSE25070 数据集的特征基因表达热图。热图中左侧部分是各特征基因在正常样本中的表达水平分布，右侧部分则是相应癌症样本的特征基因表达水平分布。白色表示低表达，浅灰色表示基准表达，而深灰色表示基因的高表达。对于每一个基因的表达，从热图中可以看到，基因在正常样本与癌症样本中的表达颜色有显著的差异。这进一步表明了本文特征选择方法的有效性，且其对后续的癌症分类有很大的帮助。

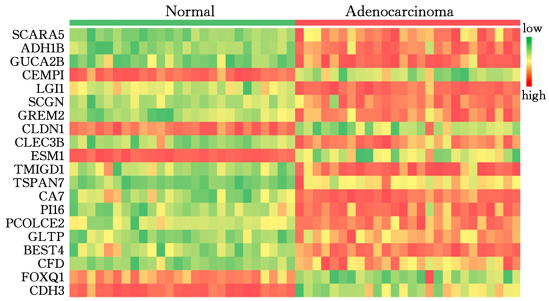

图 4　GSE25070 数据集特征基因表达热图

Fig. 4　Gene expression heat map of GSE25070 dataset

案例 3　本实验对本文特征选择方法进行对比研究，对比方法包括未进行特征选择、只利用基因表达谱数据的互信息方法。表 5 列出了不同特征选择方法在各数据集上实验的分类精度情况。图 5 是 4 种癌症数据集利用不同特征选择方法得到的分类精度预测曲线图。

表 5　不同特征选择方法在各数据集上的分类精度

Table 5　Classification accuracy of experiments on different datasets by differentfeature selection methods

（单位：%）

数据集	未特征选择	互信息	本文方法
GSE9476	90.5	95.2	100.0
GSE10797	88.3	93.4	98.8
GSE25070	90.6	95.1	100.0
GSE19804	70.8	80.2	87.4

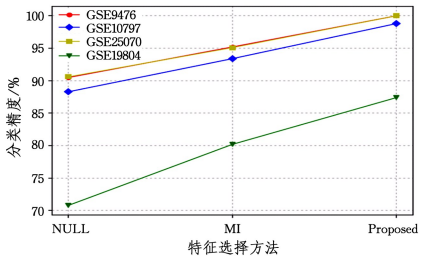

图 5 不同特征选择方法在各数据集上实验的分类预测精度曲线图

Fig. 5 Curve diagram of classification and prediction accuracy of different feature selection methods on different data sets

表 5 的分类预测结果表明，未进行特征选择得到的精度远远低于本文相关性分析和相似度分析结合的特征选择方法得到的精度，而利用互信息的相关性分析方法得到的精度同样比本文方法的精度低 5%左右。图 5 显示，在各数据集上，各方法的精度都呈现线性递增的变化趋势。这一实验结果表明，相似性分析对于特征选择有很大的改善作用，这也更好地揭示了基因间的复杂关联。

案例 4 本实验对 SVM、KNN、多层感知神经网络 MLP、集成学习算法 AdaBoost 以及本文分类预测集成模型进行研究。表 6 列出了不同分类模型算法在各数据集上实验的分类精度情况。图 6 是不同分类模型算法在各数据集上的分类精度曲线图。

表 6 不同分类模型算法在各数据集上的分类精度

Table 6 Classification accuracy of experiments on different datasets by different classification model algorithms

（单位：%）

数据集	SVM	KNN	MLP	AdaBoost	本文模型
GSE9476	98.0	96.5	97.6	97.8	100.0
GSE10797	96.8	94.1	94.4	95.2	98.8
GSE25070	98.2	95.8	94.7	97.2	100.0
GSE19804	83.6	80.5	82.9	82.6	87.4

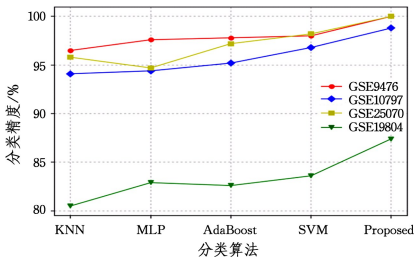

图 6 不同分类模型算法在各数据集上实验的分类预测精度曲线图

Fig. 6 Curve diagram of classification prediction accuracy of different classification model algorithms in different data sets

表 6 的分类预测结果表明，相比单分类算法 KNN，SVM 和 MLP，本文模型在精度上有很大的提高，在 GSE9476 数据集中，本文方法比 KNN 的精度高 4.5%，比 SVM 的精度高 2.0%，比 MLP 的精度高 2.4%。而相比集成学习算法 AdaBoost，本文模型的精度也高出 2.2%。这说明，本文模型比单分类算法和集成学习算法 AdaBoost 的泛化能力更强。图 6 的分类预测精度曲线表明，在各数据集上，各分类算法的精度都呈现近似线性递增的变化趋势，且本文模型的分类精度都较其他算法高。这一实验结果表明，本文模型综合了多个分类算法的优点，解决了单个分类算法适用范围有限的问题，泛化能力更强。

案例 5 本实验对最近的研究方法进行综合对比研究，对比研究方法包括文献[6]中提出的马尔可夫模型（HMMs）、文献[26]中提出的基于信息增益和遗传算法的分类模型（IG/SGA）和文献[27]中提出的基于 C4.5 算法的改进模型（PSOC4.5）。表 7 列出了不同研究方法在各数据集上实验的分类精度情况。图 7 是不同研究方法在各数据集上的分类精度曲线图。

表 7 不同研究方法在各数据集上实验的分类精度

Table 7 Classification accuracy of different research methods on each data set

（单位：%）

数据集	HMMs	IG/SGA	PSOC4.5	Proposed
GSE9476	95.7	97.1	94.8	100.0
GSE10797	93.5	94.1	93.0	98.8
GSE25070	100.0	96.8	97.1	100.0
GSE19804	80.3	78.2	79.9	87.4

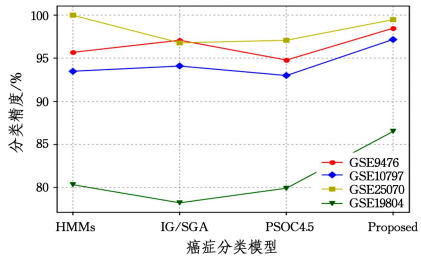

图 7 不同研究方法在各数据集上的分类预测精度曲线图

Fig. 7 Curve diagram of classification accuracy of different research methods on different data sets

从表 7 中可以看出，虽然 HMMs 模型在 GSE25070 数据集上的分类精度比本文模型高 0.5%，但在其他数据集上本文模型都要高 4%左右。另外，其他两种算法 DBN-NN 和 PSOC4.5 在各数据集上的分类精度都比本文模型低，这进一步验证了本文特征选择方法和分类预测集成模型的高效性。图 7 的分类预测精度曲线同样表明，在各数据集上，本文模型的分类精度都较其他研究方法高。这一实验结果表明，相关性分析和相似度分析结合的分类预测模型相比现有的研究方法，泛化能力更强，有更高的分类精度，并且相似度分析的加入不仅分析了基因的差异化表达，还揭示了基因与基因间的关联关系。

结束语 本文提出基于基因表达数据和功能注释信息的癌症分类集成模型。对 GEO 数据库中的 4 种癌症进行实验研究，结果表明，本文特征基因选择算法对于癌症分类精度有不错的改进。相比其他分类算法，本文癌症分类集成模型有较高的分类精度。通过整合基因表达数据和功能注释信息，联合相关性和相似度，不仅分析了基因的表达差异，还考虑到同种癌症基因有着共同功能的特点；对训练集进行多样性采样，克服了单分类算法过拟合的问题，泛化能力增强，使多个算法优势互补，解决了单个分类算法适用范围受限的问题。而本文中实验数据未对癌症类型进行细分，只包含癌症样本和正常样本，后续我们将尝试将本文方法应用于多分类的数

据集上并验证模型的有效性。

参 考 文 献

[1] SONG N F. Design and Analysis of Ensemble Classifier for Gene Expression Data of Cancer[J]. Wireless Internet Technology, 2016(7):71-72. (in Chinese)
宋年丰. 癌症基因表达数据的集成分类器设计与分析[J]. 无线互联科技，2016(7):71-72.

[2] CHEN J, ZHANG M, SHAO X G. Gene selection and cancer classification based on Monte Carlo and non-negative matrix factorization:CN 104462817 B[P]. 2017. (in Chinese)
陈晶，张苗，邵学广. 基于蒙特卡洛和非负矩阵因子分解的基因选择和癌症分类方法:CN 104462817 B[P]. 2017.

[3] NGUYEN T, KHOSRAVI A, CREIGHTON D, et al. Hidden Markov models for cancer classification using gene expression profiles[J]. Information Sciences, 2015, 316(C):293-307.

[4] LI Y, LI J. Disease gene identification by random walk on multigraphs mergingheterogeneous genomic and phenotype data[J]. Bmc Genomics, 2012, 13(7):1-12.

[5] LIU B, JIN M, PAN Z. Prioritization of candidate disease genes by combining topological similarity and semantic similarity[J]. Journal of Biomedical Informatics, 2015, 57(C):1-5.

[6] LIU G, WONG L, CHUA H N. Complex discovery from weighted PPI networks[J]. Bioinformatics, 2009, 25(15):1891.

[7] WANG H, JING X, NIU B. A discrete bacterial algorithm for feature selection in classification of microarray gene expression cancer data[J]. Knowledge-Based Systems, 2017, 126(C):8-19.

[8] GEORGE V S, RAJ C. Review On Feature Selection Techniques And The Impact Of Svm For Cancer Classification Using Gene Expression Profile[J]. International Journal of Computer Science & Engineering Survey, 2011, 2(3):16-27.

[9] BOUAZZA S H, HAMDI N, ZEROUAL A, et al. Gene-expression-based cancer classification through feature selection with KNN and SVM classifiers[C]// Intelligent Systems and Computer Vision. IEEE, 2015:1-6.

[10] NIKUMBH S, GHOSH S, JAYARAMAN V K. Biogeography-based informative gene selection and cancer classification using SVM and Random Forests[C]// Evolutionary Computation. IEEE, 2012:1-6.

[11] LI J, ZHAO Z, LIU Y, et al. A Comparative Study on Machine Classification Model in Lung Cancer Cases Analysis[C]// International Conference on Frontier Computing. Singapore: Springer, 2016:343-357.

[12] NAGARAJAN R, UPRETI M. An ensemble predictive modeling framework for breast cancer classification[J]. Methods, 2017, 131.

[13] ZHOU M, JIN M. Holographic Ensemble Forecasting Method for Short-Term Power Load[J]. IEEE Transactions on Smart Grid, 2017, PP(99):1-1.

[14] GOH K I, CUSICK M E, VALLE D, et al. The human disease network[J]. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(21):8685-8690.

[15] ALZUBAIDI A, COSMA G, BROWN D, et al. Breast Cancer Diagnosis Using a Hybrid Genetic Algorithm for Feature Selection Based on Mutual Information[C]// International Conference on Interactive Technologies and Games. IEEE, 2016.

[16] REAL R, VARGAS J M. The Probabilistic Basis of Jaccard's Index of Similarity[J]. Systematic Biology, 1996, 45(3):380-385.

[17] KOMM D, KRÁLOVIČ R, MÖMKE T. On the Advice Complexity of the Set Cover Problem[C]// International Computer Science Symposium in Russia. Berlin: Springer, 2012:241-252.

[18] WANG X, GULBAHCE N, YU H. Network-based methods for human disease gene prediction[J]. Briefings in Functional Genomics, 2011, 10(5):280-293.

[19] WU X, PANG E, LIN K, et al. Improving the Measurement of Semantic Similarity between Gene Ontology Terms and Gene Products: Insights from an Edge- and IC-Based Hybrid Method [J]. Plos One, 2013, 8(5):e66745.

[20] SZKLARCZYK D, FRANCESCHINI A, WYDER S, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life[J]. Nucleic Acids Research, 2015, 43:D447.

[21] VANITHA C D A, DEVARAJ D, VENKATESULU M. Multi-class cancer diagnosis in microarray gene expression profile using mutual information and Support Vector Machine[J]. Intelligent Data Analysis, 2016, 20(6):1425-1439.

[22] DING C, PENG H. Minimum Redundancy Feature Selection from Microarray Gene Expression Data[J]. Journal of Bioinformatics & Computational Biology, 2005, 3(2):185-205.

[23] JOHNSON R W. An introduction to the bootstrap[J]. Teaching Statistics, 2001, 23(2):49-54.

[24] BARRETT T, SUZEK T O, TROUP D B, et al. NCBI GEO: mining millions of expression profiles—database and tools[J]. Nucleic Acids Research, 2005, 33(Database Issue):D562.

[25] TIMALSINA P, CHARLES K, MONDAL A M. STRING PPI Score to Characterize Protein Subnetwork Biomarkers for Human Diseases and Pathways[C]// IEEE International Conference on Bioinformatics and Bioengineering. IEEE, 2014: 251-256.

[26] SALEM H, ATTIYA G, EL-FISHAWY N. Classification of human cancer diseases by gene expression profiles[J]. Applied Soft Computing, 2017, 50:124-134.

[27] CHEN K H, WANG K J, WANG K M, et al. Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data[J]. Applied Soft Computing, 2014, 24(C):773-780.