一种面向组学数据的中级融合分类方法

李明达 郑浩然

摘要 目的 对组学数据进行深入分析有助于推动医疗诊断等方面的研究。利用单一种类组学数据的分析方法无法满足解决某些复杂生物医学问题的需要。为了利用多种组学信息来解决复杂的生物医疗问题，本文提出一种中级融合分类方法。**方法** 引入偏最小二乘法(partial least squares, PLS)分别对各种组学数据进行降维，然后利用支持向量机（Support Vector Machine, SVM）对融合后的数据进行分类。**结果** “非小细胞肺癌与肾癌”和“结肠直肠癌与结肠直肠腺瘤”这两个组学数据集被用于测试本文方法的有效性。在这两个癌症组学数据集上的应用体现出该方法，不但能够有效降低高维组学数据的维数，而且具有较高的分类准确率（接受者操作特征曲线下的面积高达 0.95 以上）。**结论** 本文提出的中级融合方法能够利用多种组学数据对癌症样本进行分类，可以有效提高疾病诊断的准确率。

关键词 组学数据，降维，中级融合，偏最小二乘法，支持向量机

中图分类号 R318.04 文献标志码 A 文章编号 1002-3208

A mid-level fusion method for omics dataset

LI Mingda ZHENG Haoran

Department of Computer Science and Technology, University of Science and Technology of China, Hefei, 230027

【Abstract】Objective The analysis of omics data is of great importance to medical diagnosis. Methods to analyze only one type of omic dataset can’t meet the need of solving some complex biomedical problems. In order to solve the complex biomedical problems by using different kinds of omics datasets, a mid-level fusion method was proposed. **Method** Partial least squares (PLS) is used to reduce the dimension, then support vector machine (SVM) is used for classification. **Results** “Non-small cell lung cancer vs renal cancer” and “colorectal cancer vs colorectal adenomas” datasets are used for testing the method’s effectiveness. The results of the experiment demonstrated that the mid-level fusion method could not only reduces the dimension of omics data but also obtains a high classification accuracy (The area under the receiver operating characteristic curve is higher than 0.95). **Conclusions** The mid-level fusion method takes advantage of different kinds of omics datasets for classification and improves the accuracy of diagnosis.

【Keywords】 omics data, dimensionality reduction, mid-level fusion, partial least squares, support vector machine,

0 引言

对生命体来说，各种组学数据能够从不同的层面反映组织器官的功能和代谢的状态。因此，组学数据分析在对生物系统进行广泛监控的过程中具有非常重要的作用[1]。对组学数据的分析可以提供对生物体更加综合的理解，从而达到对疾病进行区分和诊断的目的。然而，组学数据处理面临着这样两个困难：首先，仅仅利用单一种类组学数据进行的分析技术不足以揭示复杂的生物体系中蕴含的与疾病相关的信息，如何对多种组学数据信息进行融合是组学数据研究能否成功的关键[2]。其次，在对组学数据进行的研究当中，从不同分析设备上产生的各种组学数据往往是海量的，而这些数据的海量性通常体现在高维度上。组学数据研究还需要对组学数据进行降维，从中提取出同疾病密切相关的知识和信息[2]。

为了从海量组学数据中挖掘出与疾病相关的信息，一些组学数据融合分析方法被提出。L.E.Wangen等[3]提出一种叫做多模块偏最小二乘法(multiblock partial least squares, MBPLS)的算法，该算法是一种有监督多模块建模的分类方法，曾被应用于基于多发硬化的代谢数据融合中[4]，并且适用于利用代谢和蛋白组学数据融合的分类问题[5]；Julien Boccard等[2]提出了一个通过建立数据之间联系的融合分析策略，该策略基于正交偏最小二乘判别分析法(orthogonal partial least squares discriminant analysis, OPLS-DA)[6]，可以对多个组学数据模块进行核融合。

基金项目：973 计划(2011CB910200)资助

作者单位：中国科学技术大学计算机科学与技术学院(合肥 230027)

作者简介：李明达(1988-)，男，硕士研究生

通讯作者：郑浩然。E-mail: hrzheng@ustc.edu.cn

数据融合方式通常分为三类[7, 8]：第一类是低级融合，即将所有不同数据连接在一起，所有特征相邻；第二类是中级融合，该策略在对不同的数据分别进行预处理和特征选择之后再进行连接；第三类是高级融合，即将不同的数据分别进行建模，将各自的预测结果汇总起来进行投票。

已有的组学数据融合分析方法虽然可以完成对组学数据的融合和降维，但它们对组学数据的种类较为敏感，而且分类准确率存在一定的提升空间。为克服传统融合分析方法对组学数据种类较为敏感的局限，达到高效利用组学数据来解决分类问题的目的，本文提出了一种中级融合方法。该方法将两种对组学数据具有普适性的分析方法——偏最小二乘法(partial least squares, PLS)[9]和支持向量机(support vector machine, SVM)[10]通过中级融合的方式进行结合，以克服对组学数据种类的敏感性。通过实验验证发现该方法不但能够融合不同种类的组学数据，而且显示出很高的分类准确率，对于解决融合不同种类组学数据的分类问题有着十分重要的意义。

1 中级融合算法

1.1 融合算法流程

假设对 n 个样本进行检测得到两种组学数据集X_1和X_2，其中X_1的特征数目为p_1，X_2的特征数目为p_2。下面通过对组学数据集X_1和X_2进行的分析为例阐述本文提出的中级融合算法流程：

(1)数据筛选。对获得的组学数据进行初步筛选，删除没有表达值的特征。经筛选，X_1的剩余特征数为p_1'，X_2的剩余特征数为p_2'。

(2)预处理。对已完成筛选的数据，针对可能出现的生物特征数值差异较大的情况，对所有数据通过 z-score 方法进行数据标准化处理。标准化之后的X_1和X_2数据集分别为X_1'和X_2'。

(3)确定训练样本。经过预处理之后，为保证数据样本的平衡性，训练集中每类样本的数量要相同或相近。于是，令X_1'和X_2'均有n_1个样本用于训练，其余n_2个样本用于测试，X_1'中划分的训练集为X_train_1'，测试集为X_test_1'；X_2'中划分的训练集为X_train_2'，测试集为X_test_2'。

(4)提取潜在变量。利用训练集中的X_train_1'和X_train_2'，通过 PLS 模型进行训练。在训练时，设定训练样本的潜在变量集X_train_1''中的特征数为h_1，X_train_2''中的特征数为h_2。利用得到的模型再分别对测试集的X_test_1'和X_test_2'进行降维，提取出测试样本的潜在变量矩阵X_test_1''和X_test_2''。在对训练集中的样本使用 PLS 时，为避免过拟合现象的发生，将可提取的潜在变量数目上线设定为 10(h_1，$h_2 \le 10$)。

(5)融合分类。融合测试集的各种组学数据潜在变量，即将X_test_1''和X_test_2''合并，将这一融合矩阵输入到 SVM 分类器中进行分类，最终给出分类结果$Y_{predict}$。该算法流程如图 1 所示。

1.2 融合算法分析

1.2.1 对高维特征的降维

组学数据分析中的研究对象是样本数少、特征多的小样本数据，直接利用分类器进行分析往往无法得到预期效果，甚至会产生过拟合的现象。因此，为了高效、准确地分析小样本组学数据，需要在分类前对组学数据进行降维，即将高维特征（经过预处理的高维组学特征）转化为低维特征（潜在变量）。与 PCA 不同，PLS 是一种有监督的降维方法，该方法被广泛应用在对高维小样本数据的降维处理中。其所用的分析手段是将观测变量 X 和预测变量 Y 投影到一个全新的空间之中，从而寻找一个线性的回归模型。利用 PLS 实现降维的过程如下：

对于 n 个样本，p 个特征的某种组学数据矩阵 X(n×p)及其类别矩阵 y(n×1)，可以通过转化成具有 h 个潜在变量的得分矩阵 T(n×h)，如公式(1)(2)所示。其中，P(p×h)为符合矩阵；b(h×1)为解释系数矩阵；E(n×p)为观测残差矩阵；f(n×1)为响应残差矩阵。

$$X = TP^t + E \tag{1}$$

$$y = Tb + f \tag{2}$$

令残差平方和 J 最小，J 的表达式如下：

$$\mathrm{J} = \min(\|E\|^2 + \|f\|^2) \tag{3}$$

通过化简，拉格朗日函数的构建，以及偏最小二乘的迭代算法可以依次求出每个样本的潜在变量。由于 PLS 被广泛应用于各类组学数据的降维处理中，因此，通过 PLS 降维得到的潜在变量可以最大化地概括各种原始高维组学数据中对分类起到决定性作用的重要信息。

1.2.2 利用潜在变量的分类

虽然各种组学数据的潜在变量与类别信息相关程度很大，然而它们的组合不一定构造出一个能够对样本进行线性区分的空间。而 SVM 的目标就是要构造一个目标函数对两类模式进行最大程度地区分，其处理方法包括针对“线性可分”和“线性不可分”这两种情况的策略：

<1>对于“线性可分”的情况，假设存在一个超平面可

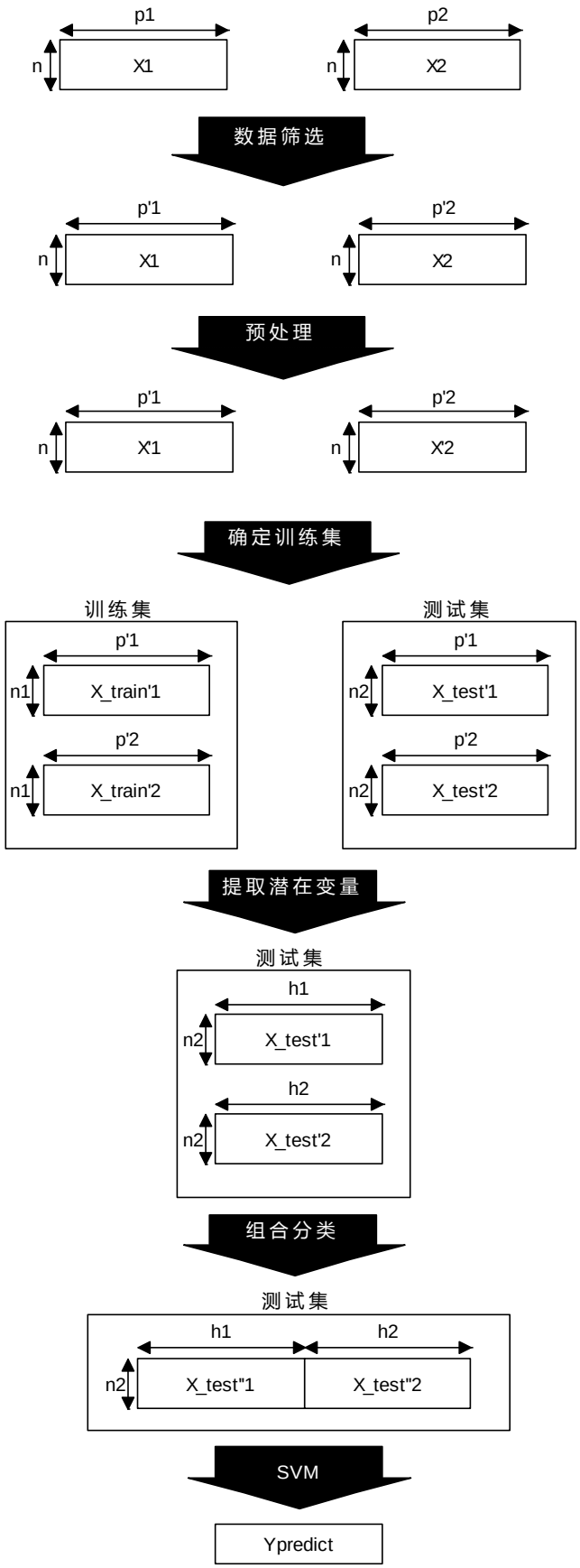

图 1 中级融合算法流程

Figure 1 the workflow of the mid-level fusion method

以把组学训练样本全部分开，将该超平面表述为：

$$\omega \cdot x + b = 0 \quad (4)$$

其中，“·”是点积，ω为 n 维向量，b 是偏移量。这里的最优超平面就是能让两类样本当中与超平面的距离最近的向量与超平面的间距最大的平面。这里将求最优超平面的问题转化为下面的二次优化问题：

$$\min\Phi(\omega) = \frac{1}{2}\|\omega\|^2 \quad (5)$$

其需要满足：

$$y_i(\omega \cdot x_i + b) \geq 1 \quad i = 1,2,\ldots\ldots n \quad (6)$$

最优解可以通过求解拉格朗日函数的鞍点得到。<2>对于“线性不可分”的情况，可以把潜在变量组合矩阵X_{mix}映射到一个高维的特征空间 H 中，在这一空间当中利用原空间函数去实现内积运算。根据泛函中的理论，需要让一种核函数满足 Mercer 条件，其就能对应某一空间的内积。于是，只需在最优的分类面上采用恰当的内积函数便可以解决此类“线性不可分”的问题。此时的目标函数为：

$$\max W(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i * x_j) \quad (7)$$

其对应的分类函数为：

$$f(x) = \sum_{i=1}^{n} y_i\alpha_i K(x * x_i) + b^* \quad (8)$$

对于“线性可分”和“非线性可分”两种情况，SVM 均是能够给出最优的分类效果。因此，本文的中级融合方法在处理各种组学数据时均能够产生优异的分类效果。

2 结果及分析

2.1 低级融合法

在利用低级融合法[11]进行对照实验时，首先将各种组学数据合并在一起，然后利用偏最小二乘判别分析法(partial least squares discriminant analysis, PLS-DA)进行分类，将得到低级融合法的分类结果。

2.2 高级融合法

在利用高级融合法[11]进行对照实验时，首先利用 PLS-DA 或 SVM 分类器分别对各种组学数据进行分类，将各种组学数据的分类结果汇总后进行投票，将投票结果作为最终的高级融合分类结果。

2.3 分类评价指标

对分类结果进行评估的时候采用的评价指标是敏感性(Sensitivity)、特异性(Specificity)、准确率(Accuracy)以及 ROC 曲线下的面积(AUC)。其中，Sensitivity、Specificity 和 Accuracy 的计算公式如下

所示：

$$Sensitivity = TP/(TP + FN) \quad (9)$$

$$Specificity = TN/(TN + FP) \quad (10)$$

$$Accuracy = (TP + TN)/(TP + FN + TN + FP) \quad (11)$$

非小细胞肺癌与肾癌数据集的混淆矩阵如表 1 所示。结肠直肠癌与结肠直肠腺瘤数据集的混淆矩阵如表 2 所示。本文的 ROC 曲线通过渐进的方式计算出来的。

表 1 非小细胞肺癌与肾癌数据集的混淆矩阵

Table 1 Confusion matrix for “non-small cell lung cancer vs renal cancer” dataset

	非小细胞肺癌样本	肾癌样本
被预测为非小细胞肺癌样本	TP	FN
被预测为肾癌样本	FP	TN

表 2 结肠直肠癌与结肠直肠腺瘤数据集的混淆矩阵

Table 2 Confusion matrix for “colorectal cancer vs colorectal adenomas” dataset

	结肠直肠癌样本	结肠直肠腺瘤样本
被预测为结肠直肠癌样本	TP	FN
被预测为结肠直肠腺瘤样本	FP	TN

表 3 非小细胞肺癌与肾癌数据集的分类结果对比

Table 3 Performance comparison on “non-small cell lung cancer vs renal cancer” dataset

	Sensitivity	Specificity	Accuracy	AUC
低级融合	1.0000	0.0000	0.5294	0.3381
高级融合	1.0000	0.2500	0.6471	0.8144
文献[3]方法	0.6667	0.8750	0.7647	0.8184
文献[2]方法	0.8889	0.7500	0.8235	0.8780
本文方法	1.0000	1.0000	1.0000	1.0000

表 4 结肠直肠癌与结肠直肠腺瘤数据集的分类结果对比

Table 4 Performance comparison on “colorectal cancer vs colorectal adenomas” dataset

	Sensitivity	Specificity	Accuracy	AUC
低级融合	0.8750	0.6250	0.7500	0.8247
高级融合	0.6250	0.8750	0.7500	0.7999
文献[3]方法	0.6250	0.8750	0.7500	0.7999
文献[2]方法	0.7500	1.0000	0.8750	0.9060
本文方法	0.8750	1.0000	0.9375	0.9541

2.4 非小细胞肺癌与肾癌数据集

非小细胞肺癌与肾癌的数据来自美国国家癌症研究所的 60 人癌症细胞系数据库(The US National Cancer Institute 60 human tumour cell line, NCI-60)。该数据库中的组学数据来自于代谢、蛋白以及基因等组学实验。实验样本取自 60 个人的癌症细胞系[12]，该细胞系来自包括结肠、肺等 9 种人体组织器官。分类实验的样本分为两类：非小细胞肺癌样本(9 个)和肾癌样本(8 个)。所利用的两种组学数据类型是 Ross 基因集[13](5643 维)、Staunton 基因集[14](243 维)。

由于可以用于实验的样本总数较少，实验验证采用留一法交叉验证。在进行对照比较实验采用的是 2.1 中的低级融合法和 2.2 中的高级融合法，一种中级融合方法[3]以及一种基于核融合的方法[2]。本文

算法在利用 PLS 提取潜在变量时，分别对 Ross 和 Staunton 数据集提取的潜在变量数均为 4，SVM 使用的是线性核函数。该参数是使分类效果达到最优的前提下能够设定的潜在变量数最小值。初级融合的潜在变量数设定为 1，高级融合所使用的分类器为 PLS-DA，且潜在变量数为 1。

分类结果如表 3 所示。本文提出的中级融合方法的分类效果在所有对比方法中表现最优，文献[2]、文献[3]方法以及高级融合方法表现其次，而低级融合方法表现最差。

2.5 结肠直肠癌与结肠直肠腺瘤数据集

结肠直肠癌数据集源于文献[15]。利用 94 个人的血清样本，通过相关技术提取出了他们的生物标记物(2 维)、荧光反应记录(19 维)以及核磁共振记录(其中 Carr-Purcell-Meiboom-Gill, 201 维；Nuclear Overhauser Enhancement Spectroscopy, 254 维)。这 94 个样本分为两类：47 个确诊患有结肠直肠癌，其余为结肠直肠腺瘤样本。

为了避免过拟合现象的发生，本文根据文献[15]中的描述选取了 78 个样本作为训练集，用于建立分类模型；16 个样本作为测试集，用于对结果模型进行最终的验证。由于生物标记物和荧光反应记录的维数较低，因此仅对两种核磁共振记录分别利用 PLS 进行降维，然后将四种组学数据进行融合分类。本文算法对两种核磁共振记录的潜在变量数的设定，经试验得出的最优潜在变量数均为 6，SVM 使用的是线性核函数。初级融合的潜在变量数设定为 1，高级融合所使用的分类器为 SVM。

分类结果如表 4 所示。本文方法的分类效果依然最优，在敏感性、特异性、准确率以及 ROC 曲线下面积等评判准则下均高于低级融合、高级融合、文献[2]和文献[3]。

3 讨论与结论

随着生物科技的发展，越来越多与疾病相关的海量组学数据得以被检测。面对各种各样的高维组学数据，如何将生物信息综合起来进行分析以达到对疾病进行判断的目的是一个十分棘手的问题。现有的融合分析方法对组学数据的种类较为敏感，而且分类准确率有待提升。本文提出了一种面向组学数据的中级融合分类方法。该算法选用对各种组学数据具有普适性的 PLS 方法进行降维，提取出原始信息中利于分类的潜在变量，从而解决对组学数据种类较敏感的问题；采用的中级融合策略能够将不同种类的组学数据信息进行充分融合；利用可以处理“线性可分”和“线性不可分”两种情况的 SVM 分类器以达到对生物样本进行准确分类的目的。在两个公共组学数据集进行的验证中，通过与低级融合算法、高级融合算法以及文献[2]、[3]方法的分类结果进行比较，本文算法显示出了更高的分类准确率。在分子生物学研究中的面向疾病的组学数据融合分析方面，本文方法具有积极的指导意义。

参考文献

[1] JoyceAR, Palsson BO. The model organism as a system:integrating ‘omics’ data sets[J]. Nat Rev Mol Cell Biol, 2006, 7 (3):198-210.

[2]Boccard J, Rutledge D N. A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion[J]. Analytica chimica acta, 2013, 769:30-39.

[3] L.E. Wangen, B.R. Kowalski, A multiblock partial least squares algorithm forinvestigating complex chemical systems [J]. Chemom, 3 (1989) 3–20.

[4]A.K. Smilde, M.J. van der Werf, S. Bijlsma et al. Fusion of mass spectrometry-based metabolomics data [J]. Anal. Chem,2005, 77, 6729–6736.

[5]T. Moyon, F. Le Marec, E.M. Qannari et al. Alexandre-Gouabau, Statistical strategies for relating metabolomics and proteomics data: a real case study in nutrition research area [J]. Metabolomics, 2012, 1–12.

[6] Bylesjö M, Rantalainen M, Cloarec O, et al. OPLS discriminant analysis: combining the strengths of PLS‐DA and SIMCA classification[J]. Journal of Chemometrics, 2006, 20:341-351.

[7] Roussel S, Roger JM, Bellon-Maurel V, Grenier P. Fusion of Aroma, FT-IR and UV Sensor Data Based on the Bayesian Inference. Application to the Discrimination of White Grape Varieties [J]. Chemom. Intell. Lab. Syst. 2003, vol. 65 (2), pp 209 - 219.

[8] Smolinska A, Blanchet L, Buydens L M C, et al. NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review[J]. Analytica chimica acta, 2012, 750: 82-97.

[9] Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics[J]. Chemometrics and intelligent laboratory systems, 2001, 58(2):109-130.

[10] Cortes C, Vapnik V. Support-vector networks [J]. Machine

learning, 1995, 20(3):273-297.

[11] Alessandra Biancolillo, Remo Bucci, Antonio L. Magrì, Andrea D. Magrì, Federico Marini, Data-fusion for multiplatform characterization of an italian craft beer aimed at its authentication [J]. Analytica Chimica Acta, 2014, 820:23-31.

[12] Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen[J]. Nat Rev Cancer, 2006, 6:813–823.

[13]Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO: Systematic variation in gene expression patterns in human cancer cell lines[J]. Nat Genet, 2000, 24:227-235.

[14]Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J,Scherf U, Lee JK, Reinhold WO, Weinstein JN, Mesirov JP, Lander ES,Golub TR: Chemosensitivity prediction by transcriptionalprofiling[J]. Proc Natl Acad Sci USA, 2001, 98:10787-10792.

[15] Bro R, Nielsen H J, Savorani F, et al. Data fusion in metabolomic cancer diagnostics[J]. Metabolomics, 2013, 9(1):3-8.