DOI: 10.19650/j.cnki.cjsi.J1905227

混合整体趋势扩散的虚拟样本构建及其血液光谱分析应用*

高克铉[1],李志刚[2],徐长明[2],王巧云[1],李 博[1]
(1. 东北大学信息科学与工程学院 沈阳 110819; 2. 东北大学计算机科学与工程学院 沈阳 110819)

摘 要:准确的预测模型在光谱定量分析中起着非常重要的作用。针对小样本集空间信息匮乏、信息分布不均衡所造成的模型预测误差偏大的问题,基于传统多分布整体趋势扩散(MD-MTD)方法提出混合整体趋势扩散技术(Hybrid-MTD)构建虚拟样本空间,进一步扩充训练样本集,改善样本集空间的信息分布,从而显著降低模型的预测误差。分别利用全血样本的总胆固醇和甘油三酯光谱数据集进行对比实验验证。实验结果表明,基于添加虚拟样本后重构的数据集建立的偏最小二乘预测模型能够获得更低的平均预测均方差(MRmesp)。总胆固醇和甘油三酯的 MRmesp 分别为 0.41 和 0.45 mmol/L。对比 MD-MTD 方法,误差分别降低了 46.7% 和 22.4%。可见,所提出的 Hybrid-MTD 方法能够构建数量适宜的高质量的虚拟样本。填充后的样本集所对应的预测模型显著降低了预测误差,与现有的 MTD 方法相比具有更加优越的预测性能。混合整体趋势扩散技术在在血液光谱分析的应用有效提升了评估生理指标的质量,加快心血管疾病的筛查速度并降低其风险。

关键词: 混合整体趋势扩散;偏最小二乘;血液光谱分析;预测误差
中图分类号: TH741 O657.33 **文献标识码**: A **国家标准学科分类代码**: 150.2520

Virtual sample establishment of Hybrid-MTD and its application in blood spectrum analysis

Gao Kexuan[1], Li Zhigang[2], Xu Changming[2], Wang Qiaoyun[1], Li Bo[1]
(*1.College of Information Science and Engineering, Northeastern University, Shenyang 110819, China*;
2.School Computer Science and Engineering, Northeastern University, Shenyang 110819, China)

Abstract: An accurate prediction model plays a very important role in the quantitative spectrum analysis. Aiming at the problem of large model prediction error caused by information lacking and imbalanced information distribution in small sample set space, in this paper, based on traditional MD-MTD (multi-distribution mega trend diffusion) method, a Hybrid-Mega Trend Diffusion (Hybrid-MTD) technique is proposed to construct virtual sample space, which further expends the training sample set and improves the information distribution of the sample set space, and then obviously reduces model prediction error. The spectrum data sets of total cholesterol and triglyceride in whole blood samples were utilized to carry out comparison and experiment verification. The experiment results show that the PLS prediction models established based on the reconstructed data set with virtual samples added can provide lower mean prediction mean square error MRmesp (mean of RMSEP). The values of MRmesp of total cholesterol and triglyceride are 0.41 and 0.45 mmol/L, respectively. Compared with traditional MD-MTD method, the errors are reduced by 46.7% and 22.4%, respectively. The proposed Hybrid-MTD method can construct an adequate number of high-quality virtual samples; the prediction model corresponding to the sample set with the virtual samples filled obviously reduces the prediction error, and has superior prediction performance compared with the existing MTD method. The application of Hybrid-MTD technique in blood spectrum analysis effectively enhances the evaluation quality of physiological indicators, speeds up screening speed for cardiovascular disease and reduces its risk.

Keywords: hybrid-mega trend diffusion(Hybrid-MTD); partial least squares(PLS); blood spectrum analysis; prediction error

收稿日期:2019-06-04 Received Date:2019-06-04
*基金项目:国家自然科学基金(11404054)、河北省自然科学基金(F2016501138)项目资助

0 引 言

光谱技术已应用于临床诊断领域,以获得有关体液和组织成分的信息[1-3]。经研究发现,血液中的总胆固醇和甘油三酯含量的增高是心脏病、心肌梗塞和心血管疾病的主要发病诱因[4-6]。因此,探索基于光谱技术、免试剂的血液生理指标含量定量分析技术对快速筛查此类疾病风险具有重要意义。但在疾病相关的生理指标采集中往往由于样本采集操作的特殊性、样本自身的稀有性以及样本信息空间分布的不均衡性导致样本数量偏少或者样本信息不充分。样本量和特征信息的充分性影响预测模型效果[7-8]。同时,由于异常的样本数量较少也导致影响预测模型的泛化能力[9]。因此改造样本信息不均衡、不完备的小样本空间对于预测模型性能的提升具有深入研究价值[10-12]。

通常统计学认为,小样本集为获得样本的数量小于 50 的数据集[13-14]。由于样本数量较少且信息缺失无法提供建立预测模型所需的完整信息空间。如何通过现有观测信息空间的扩展对整体信息空间进行近似描述对于提升建模品质具有重大意义。

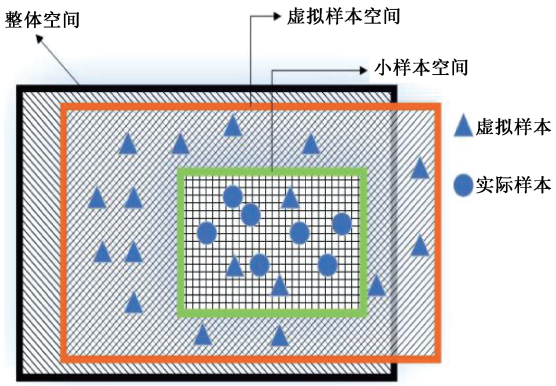

图 1 样本空间分布

Fig.1 Sample space distribution

研究表明,虽然小样本集包含大部分建模所需的特征信息,但传统的统计方法难以根据小样本集的贫乏信息为决策者提供可靠的结论[14]。目前,解决小样本问题的方法主要为欠采样法和过采样法。欠采样法通过筛选样本从而提升模型建立效果,如交叉验证法、Bootstrap 和 Kennard-Stone(KS)法等。欠采样有效地利用了原有样本再次进行数据整合保证了数据的真实性。但在小样本数据中样本数量过少易加剧局部信息缺失,因此无法使用欠采样法。过采样法则是在原样本的信息特征空间通过映射方式增加样本,其小样本空间、映射的虚拟样本空间整体空间以及虚拟样本和实际样本的包含关系如图 1 所示。目前主要有两种过采样方法:一种为综合过采样(synthetic ninority oversampling technique,SMOTE)及其衍生算法,另一种为基于高斯分布的虚拟样本生成技术 VSG 及其衍生法[10]。但 SMOTE 法仅在现有的样本信息下通过插值法在其间隔内部进行信息填充。模糊理论的提出后,Huang 等[15]基于模糊理论定义提出了正态扩散函数,实现了对于离散化的信息间隔进行填充。Huang 课题组认为随着样本间的信息间隔越小,其表达的样本信息越趋近于整体样本信息特性,样本的隶属度越高。随后,在 Huang 的正态扩散、信息填充技术的基础上,Li[16]进一步提出了整体趋势扩散技术(mega-trend-diffusion,MTD),定义了信息扩散的整体边界,得到整体趋势信息,实现在整体边界领域内对样本信息进行更均匀的扩充,提升了模型的泛化性[16]。在国内,陈忠圣和朱宝等针对虚拟样本边界建立和样本筛选的问题进一步优化,提出了多分布整体趋势扩散技术(multi-distribution mega-trend-diffusion,MD-MTD)技术,通过对样本分布区域进行多分割的方法,提高了生成虚拟样本的质量并克服 MTD 的缺陷[7,14,17-18]。在 MD-MTD 法中虽然能够通过对样本信息区域进行多分段分割进一步提升扩充的均匀性,但其在扩散边界附近样本数量过少、信息匮乏,仍然存在信息不均衡、不完备的缺陷,导致其影响各区域样本分布的均衡性以及建模的品质。

针对上述问题,本文首先提出基于 MTD 的改进型整体趋势扩散技术(advanced-MTD,AD-MTD)改善扩散区域的信息分布的均衡性。在此基础上进一步提出基于 MD-MTD 和 AD-MTD 生成的虚拟样本进行混合趋势扩散的技术 Hybrid-MTD,改善了信息扩散区边界点和原信息区的中心点 CL 附近信息扩散分布的均衡性,提高了光谱数据回归分析中的预测效果。本文针对人体血液中的总胆固醇和甘油三酯含量浓度进行基于全血光谱样本的定量分析,验证了利用 Hybrid-MTD 技术进行高维度信息空间虚拟样本的建立能够达到完善样本信息空间,减小预测的误差的效果。

1 数据集划分方式与虚拟样本生成技术

1.1 KS 算法

KS 算法在光谱应用分析领域中是一种广泛应用的数据集划分方法。KS 算法的核心理念是通过计算样本间的光谱信息欧氏距离,并依据样本均匀分布原则将数据集划分为相应数量的训练集,余下样本则为测试集[19-20]。

1.2 虚拟样本定义

对于虚拟样本,在不同的研究领域中有不同的定义,目前为止也未有关于虚拟样本的严格定义。常见的虚拟

样本定义如下[14]。

定义:令通过随机生成的训练样本集为,X,Y 为样本参数信息。基于先验知识下 $TD=(X,Y)$ 可以得到其对应的转换关系(T,K),生成新的样本 $TD'=(XT,KY)$,其中。这些新的样本称为虚拟样本。

Niyogi 等也从数学的角度证明了通过先验知识下构造出的虚拟样本能够与真实的样本信息一样提供有效训练样本[14]。在构建虚拟样本空间的过程中,样本间的特征信息和转换关系可以通过偏最小二乘回归 PLSR、支持向量机 SVM、神经网络 ANN 以及极限学习机 ELM 等机器学习法获得。因此通过有监督学习方式来建立整体样本空间的超平面 $H(x,\alpha)$- 即为理想预测模型(通常情况下无法获得) 和小样本集的推估平面 $H'(x,\alpha')$- 即为基于实际小样本集预测模型。α,α' 为其中的广义参数,超平面与推估平面的距离影响着理想信息空间与实际小样本信息空间的差异程度。整体理想超平面和实际小样集本推估平面的空间对应划分类别的效果如图 2 所示。

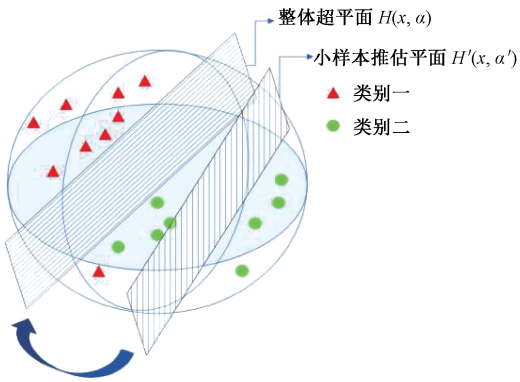

图 2　整体超平面与小样本推估平面

Fig.2　Mega-hyper plane and small sample estimation plane

加入含原有样本特征的虚拟样本集后,将缩短小样本集的推估平面 H' 与整体超平面 H 的距离,修正后的小样本预测模型与理想模型的相似度随之提升。因此含虚拟样本的小样本集的建立有效地提升预测模型的性能[4,13]。

1.3　基于整体趋势扩散技术的虚拟样本生成技术

MTD 技术是关于填补信息间隔的方式,其属于随机虚拟样本生成技术的一种改进型式。MTD 主要是通过隶属函数计算出其相应的虚拟样本信息的左边界 LB 和右边界 RB,从而在该范围内生成虚拟样本信息,其信息结构如图 3 所示。图 3 中,min 为样本参数最小边界,max 为样本参数最大边界,LB 为虚拟样本左边界,虚拟样本右边界 RB,样本数据中心点 CL,样本空间信息数据 X。LB 和 RB 的计算公式见式(1)~(4)。

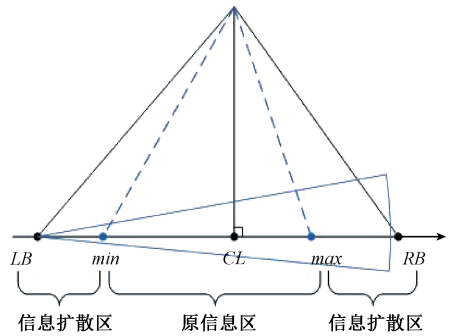

图 3　MTD 示意图

Fig.3　Schematic diagram of MTD

$$LB=\begin{cases} CL-\dfrac{N_L}{N_L+N_R}\times\sqrt{-2\times S^2/N_L\times\ln(10^{-20})}, \\ \qquad LB<min \\ min, \qquad LB>min \end{cases} \tag{1}$$

$$RB=\begin{cases} CL+\dfrac{N_R}{N_L+N_R}\times\sqrt{-2\times S^2/N_R\times\ln(10^{-20})}, \\ \qquad RB>max \\ max, \qquad RB<max \end{cases} \tag{2}$$

$$CL=(max+min)/2 \tag{3}$$

$$\hat{S}_x^2=\frac{\sum_i^k(x_i-\bar{x})^2}{n-1} \tag{4}$$

左边界 LB 和右边界 RB 决定了信息扩散的整体范围,虚拟样本信息应在其区间范围内。其中,N_L 和 N_R 分别代表位于中心点数据 CL 左边和右边样本数量,样本方差为 S_x^2,k 为小样本集样本的数量。对于虚拟样本信息的生成方法如式(5)所示。

$$X_{VLi}=LB+s(RB-LB),i=1,2\cdots,n \tag{5}$$

式中:X_{VLi}为虚拟样本信息;n 为生成虚拟样本数量;s 为正态分布随机数。通过式(5)就可以在两边界内生成服从正态分布的虚拟样本信息。MTD 算法中,基于隶属函数的虚拟样本信息分布对于信息的扩散具有一定的提升效果[18]。但在虚拟样本的信息生成法对于样本的分布未能充分考虑,回归模型的选定中 BP 神经网络对于高维信息样本预测的效果较差。MD-MTD 是基于 MTD 的一种改进算法,其优点在于其能在不同的区域扩充虚拟样本,从而能够保证在信息扩充区域生成样本的同时也能在原样本信息区域扩充虚拟样本,避免了样本分布的不平衡[13-14]。

MD-MTD 能够进一步提升虚拟样本信息的均匀分布,并且能够进一步确定在不同区域中的虚拟样本生成方式。但是其虚拟样本的扩散区域过多,虚拟信息的生

成中在叠加情况下易超出区域范围,并且在高维度信息样本空间下无法解决其信息的有效分散程度,因此在高维度信息下可能会在各扩散区域中产生局部密集的虚拟样本信息,从而无法保证区域内的样本均匀分布。

1.4 混合整体趋势扩散虚拟样本构建技术

针对上述问题,基于 MTD 算法,首先提出了 AD-MTD。AD-MTD 的区间划分如图 4 所示。

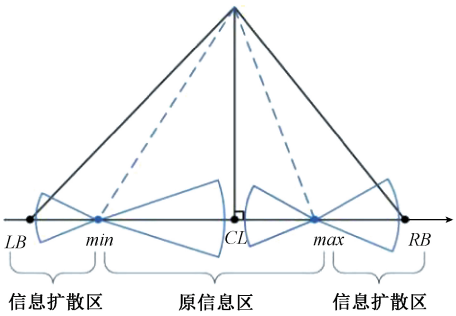

图 4 AD-MTD 示意图

Fig.4 Schematic diagram of AD-MTD

其中,N_L 和 N_R 分别代表位于中心点数据 CL 左边和右边样本数量。max 和 min 代表真实样本簇特征光谱吸收度的最大和最小值。通过对小样本中的分析物浓度含量 Y 进行排序得到光谱信息空间相应的 max 和 min。假设要求生成的虚拟样本总数为 N_{Vir},在[LB,CL]边界范围内所产生的虚拟样本数量为 N_{VirL},在[CL,RB]边界范围内所产生的虚拟样本数量为 N_{VirR},其相对应的关系如式(6)所示。其中左半扩展部分虚拟样本数量 N_{VirL}和右半扩展部分虚拟样本数量 N_{VirR}的计算公式如下式(7)和(8)所示。

$$N_{Vir} = N_{virL} + N_{virR} \tag{6}$$

$$N_{VirL} = \frac{\| CL - LB \|_2}{\| RB - LB \|_2} \times N_{Vir} \tag{7}$$

$$N_{VirR} = \frac{\| RB - CL \|_2}{\| RB - LB \|_2} \times N_{Vir} \tag{8}$$

式(6)表示虚拟样本总数和左右两信息扩散区域内的虚拟样本数间的关系式。式(7)和(8)为计算其划分的[LB,CL]和[CL,RB]区间长度占总长度的比例值,通过该比例值得到相应该区间范围内的虚拟样数量。在虚拟样本的生成方法中,主要是通过插值的思想实现对虚拟样本信息在信息区域范围内的扩散。通过 AD-MTD 方法,依据式(9)和(10)所描述的原则生成虚拟样本。

$$X_{VLi} = \begin{cases} min - \theta_i \times \left(\dfrac{CL - LB}{2}\right), s_i < 0.5 \\ min + \theta_i \times \left(\dfrac{CL - LB}{2}\right), s_i > 0.5 \end{cases},$$

$$i = 1,2,3,\cdots,N_{VirL} \tag{9}$$

$$X_{VRj} = \begin{cases} max - \theta_j \times \left(\dfrac{RB - CL}{2}\right), s_i < 0.5 \\ max + \theta_j \times \left(\dfrac{RB - CL}{2}\right), s_i > 0.5 \end{cases},$$

$$j = 1,2,3,\cdots,N_{VirR} \tag{10}$$

式中:θ_i,$\theta_i \in (0,1)$、s_i,$s_i \in (0,1)$ 分别为第 i 虚拟样本所对应的正态随机数和均匀分布随机数;θ_j,$\theta_j \in (0,1)$、s_j,$s_j \in (0,1)$ 分别为第 j 个虚拟样本所对应的正态随机数和均匀分布随机数。通过偏最小二乘回归(PLSR)对虚拟样本空间 X_{Vir} 进行回归分析,从而得到对应的虚拟样本的浓度含量值。首先基于原有样本信息建立回归模型 $H(X)$;其次对虚拟样本信息进行回归计算得到预测数据[14]。

AD-MTD 虽然有效地增加了信息扩散区域的虚拟信息,但在原信息区内生成的虚拟信息仍距离中心点 CL 较远,导致在原信息区虚拟样本较少。

为此结合 AD-MTD 与 MD-MTD 最终提出了混合整体趋势扩散技术 Hybrid-MTD。其核心是利用两种虚拟样本建立的方法的各自优势分别在提升信息扩散区域和中心点 CL 附近原信息区域进行虚拟样本的构建,使区域内的样本均匀分布。

2 模型的验证和评估

虚拟样本构建的有效性以及重构样本集的预测模型的预测精度是否有效提升,是检验虚拟样本空间信息覆盖有效性的重要指标。本文中主要基于两方面考虑:

1)虚拟样本集与原样本集的同分布判别;

2)融合虚拟样本后模型的预测效果评估。

针对于虚拟样本生成的有效性探讨,目前的方法仍未有一个合理而统一的评价。本文基于统计学的知识,通过 TSKS (two samples kolmogorov-smironv)检验法[21]判别虚拟样本与小样本集是否具有同分布,从而在置信区间内接受虚拟样本的有效性。步骤如下:

1)对信息集 Y_{comb} 与小样本集 Y 进行合并成 $Y_1Y_2Y_3\cdots Y_{n+m}$(n、m 为两个样本集的样本总数),依照升序进行排序 $Y_{(1)} \leqslant Y_{(2)} \cdots \leqslant Y_{(n+m)}$;

2)计算两样本集的经验分布函数 $F_{1n}(Y)$、$F_{2m}(Y)$,构造统计量 $D_{m,n}$ 得到如式(11)所示;

$$D_{m,n} = \max\{| F_{1n}(Y_{(i)}) - F_{2m}(Y_{(i)}) |\} \tag{11}$$

3)给定显著性水平 $\alpha = 0.025$,确定检验 p 值;

4)若 $p > \alpha$,则认定信息集 Y_{comb} 与小样本集 Y 属于同分布,反之则不属于同分布。

在式(11)中,为求信息集经验分布函数 $F_{1n}(Y)$ 与小样本集经验分布函数 $F_{2m}(Y)$ 的每个划分区段的差值,并找到差值最大,将其最大差值赋予构造统计量 $D_{m,n}$。式中的 i 为划分经验分布函数划分的第 i 区间。显著性水平

α 的数值大小决定了接受该 $D_{m,n}$ 离差程度的置信区间的大小范围,其数值越小, 置信区间越大,则接受离差的程度越大。为了验证新建立的样本集,使用式(12)~(15)给出的预测模型误差性能评价指标。

$$RMSEP = \sqrt{\frac{\sum_{i}^{n}(y_{xi} - \hat{y}_{xi})^2}{n - 1}} \tag{12}$$

$$RMSEP_{Vir} = \sqrt{\frac{\sum_{i}^{n}(y_i - \hat{y}_{xi})^2}{n - 1}} \tag{13}$$

$$MRmsep = \frac{\sum_{j=1}^{k}(RMSEP_{Vir})_j}{k} \tag{14}$$

$$OTR = \frac{RMSEP - MRmsep}{RMESP} \times 100\% \tag{15}$$

式中: n 表示测试集数量;k 为对含虚拟样本的训练模型的预测次数,通常 $k = 10$;y_{xi} 表示为测试集参考值;y_i 和 $\hat{y}_{xi}$ 分别表示含虚拟样本建模和不含虚拟样本建模所对应的预测值。$RMSEP$ 表示预测均方差、$RMSEP_{vir}$ 表示填充虚拟样本优化后的训练集所对应的预测模型的预测均方差;$MRmsep$ 表示优化后的训练集所对应的预测模型的平均预测均方差,数值越低则预测模型的预测准确性越高;OTR 则为优化率,主要用来衡量误差减少的程度。OTR 数值越高则模型的性能越好。

基于样本分布判别以及$MRmsep$和OTR两个指标,既能通过基于统计学习法衡量虚拟样本的生成与原样本间是否具有同分布,保证了虚拟样本与原样本间的相关性,也可对于加入虚拟样本后对模型的优化程度进行合理的评价。

3　虚拟样本建立和预测验证分析

3.1　虚拟样本建立

具体构建虚拟样本的步骤如下:

1)确定信息扩散边界与虚拟样本数量;

2)生成虚拟样本信息 X_{Vir};

3)通过 PLSR 法对原有训练集样本生成小样本推估平面 $H'(X)$ 并结合虚拟样本信息的达到虚拟样本的预测回归值 Y_{Vir};

4)将虚拟样本集 $D(X_{Vir},Y_{Vir})$ 与原训练集组 $D(X_{train},Y_{train})$ 合成新的信息集 $D(X_{comb},Y_{comb})$;

5)判断检验 p 值是否满足大于 α,不满足则返回第 2 步;

6)基于新的信息集 $D(X_{comb},Y_{comb})$生成新的回归模型;

7)判断重复次数 f 是否等于 k,不等则 f 累加 1 后返回 2),相等则进入下一步;

8)评估验证。

为了对比和验证本文提出的虚拟样本构建方法的有效性以及解决光谱生理液体分析技术中样本空间信息不足的瓶颈,本文基于全血光谱的实际样本空间以及经过 MD-MTD、AD-MTD 和 Hybrid-MTD 方法填充了虚拟样本后的修正样本空间构建 PLS 预测回归模型进行了总胆固醇定量分析验证。将 51 个未进行预处理的总胆固醇中红外光谱样本进行样本划分,本试验通过 KS 划分法选取 37 个(70%)样本作为训练集,其余 14 个样本作为测试集,生成虚拟样本数为 40 个(多次试验确定)。分别通过 MD-MTD、AD-MTD 和 hybrid-MTD 生成虚拟样本。总胆固醇相关虚拟样本信息如图 5 所示,生成虚拟样本和原训练样本分布如图 6 所示。

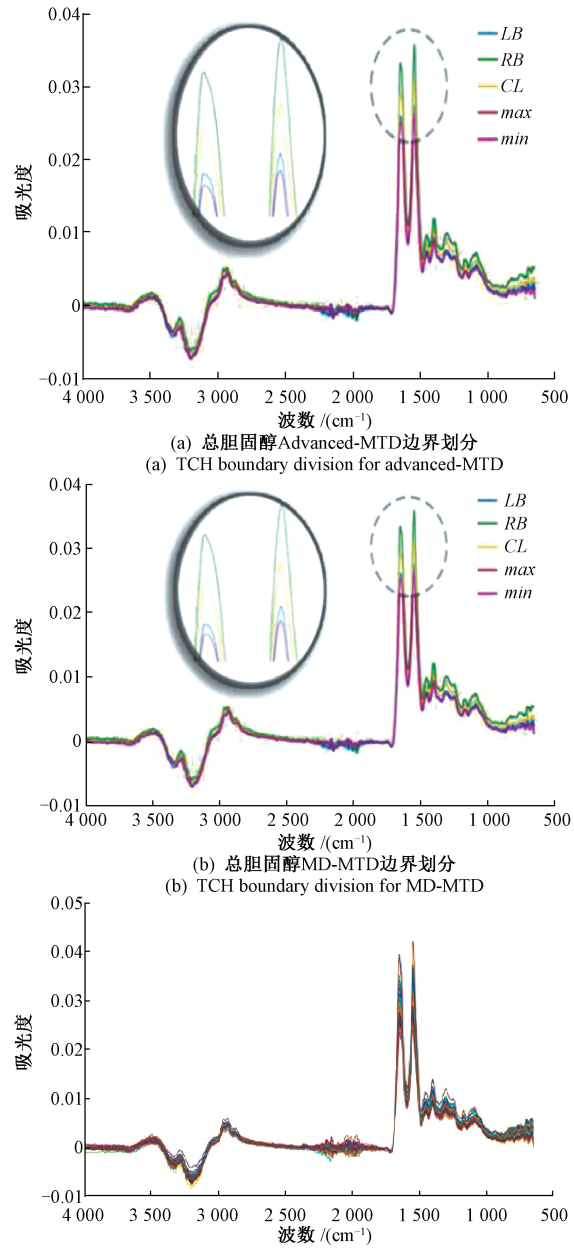

图 5　总胆固醇相关谱图信息

Fig.5　TCH related spectrogram information

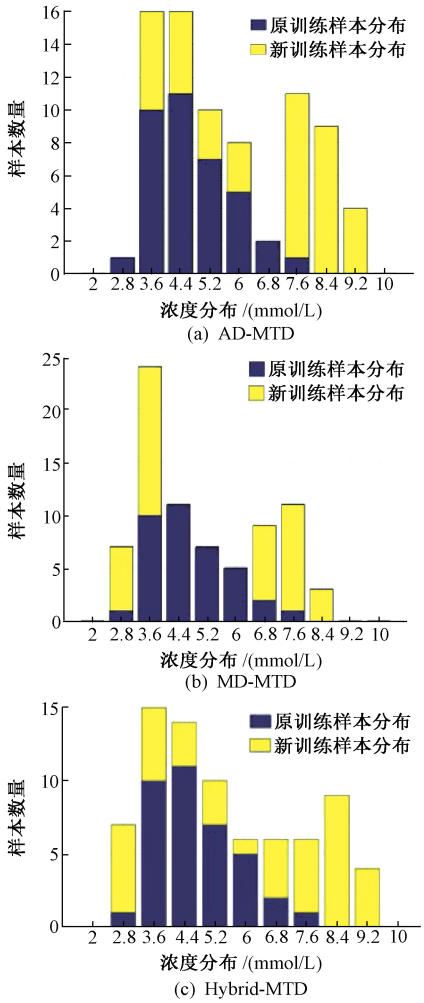

图 6 各虚拟样本建立 TCH 的样本分布

Fig.6 Sample distribution constructed from various virtual samples for TCH

图 6 中,基于 Hybrid-MTD(PLS)法的虚拟样本在实际总胆固醇浓度范围内的分布具有均匀性,而 MD-MTD 法生成的虚拟样本集中在边界,因此影响的样本的分布均匀性。在填充虚拟样本后的样本空间建立 PLS 回归预测模型进行胆固醇的定量分析。基于 Hybrid-MTD 填充虚拟样本后样本集对应的 PLS 回归预测模型的预测值与实际参考值的关系如图 7 所示。

图 7 中,Test、Train、AD-vir 和 MD-vir 分别表示测试集、原训练集、AD-MTD 和 MD-MTD 生成的虚拟样本集。其中,原训练集、AD-MTD 和 MD-MTD 生成的虚拟样本集混合后形成新的训练集,新训练集用于训练预测模型。从图中可知 AD-MTD 和 MD-MTD 的虚拟样本分布情况,两者虚拟样本混合后的 Hybrid-MTD 在整体信息分布中更具均匀性,因此有效提升了预测模型的回归效果。其最终总胆固醇预测结果如表 1 所示。

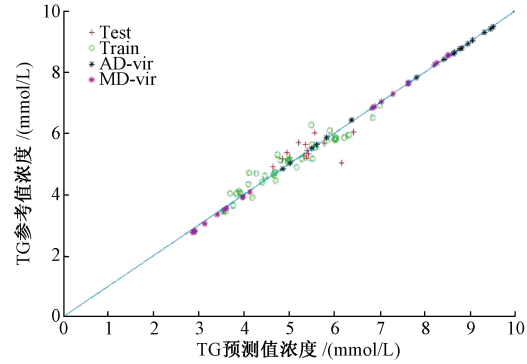

图 7 总胆固醇 Hybrid-MTD 样本集 PLS 模型的预测值与参考值之间的关系

Fig.7 The actual reference TCH concentration versus the predicted TCH concentration obtained from the PLS model based on Hybrid-MTD sample set

表 1 总胆固醇浓度预测结果分析

Table 1 Prediction result analysis of TCH concentration

类型	*Nv*	*H′*	*H*	*hN/pv*	*MRmsep*	*OTR*/%
–	–	–	PLS	5	0.76	–
MD-MTD	40	PLS	PLS	5	0.50	38.4
AD-MTD	40	PLS	PLS	5	0.43	43.4
Hybrid-MTD	40	PLS	PLS	5	0.41	46.7

表 1 中,*Nv* 为生成虚拟样本数量; *H′* 列对应为构建小样本推估超平面所使用的模型;*H* 列对应为构建整体推估平面所使用的模型,*hN/pv* 为隐含层数 / 潜变量数,*MRmsep*(mmol/L)为 10 次定量分析预测试验后 *Rmsep*(mmol/L)的均值。表 1 数据表明经过填充虚拟样本的方式改善后的样本集所对应建立的 PLS 定量分析预测模型的性能均有所提升。性能提升幅度最大的是基于本文提出的 Hybrid-MTD 方法填充虚拟样本后的数据集所建立的 PLS 预测模型,其 *MRmsep* 值为 0.41 mmol/L。同时,*OTR* 指标也显示预测误差降低了 46.7%,对于总胆固醇浓度的预测效果有着显著的提高。因此,基于 Hybrid-MTD 生成虚拟样本填充后样本空间对应的 PLS 预测模型性能明显优于经过其它虚拟样本生成方法填充的样本空间所对应预测模型的性能。

3.2 全血中的甘油三酯光谱样本集预测验证

为了更客观全面的验证本文提出的虚拟样本构建方法,再以全血中的甘油三酯样本集进行定量分析验证。将 52 个未进行预处理的甘油三酯样本进行样本划分,本试验通过 KS 划分法选取 32 个样本(60%)作为训练集,选取 20 个样本(40%)作为测试集进行验证测试,虚拟样本 40 个(多次试验确定)。通过上述方法得到虚拟样本

并建立回归模型,新训练集共 72 个样本。同样,分别通过 MD-MTD、AD-MTD 和 hybrid-MTD 生成虚拟样本。虚拟样本相关信息如图 8 所示。

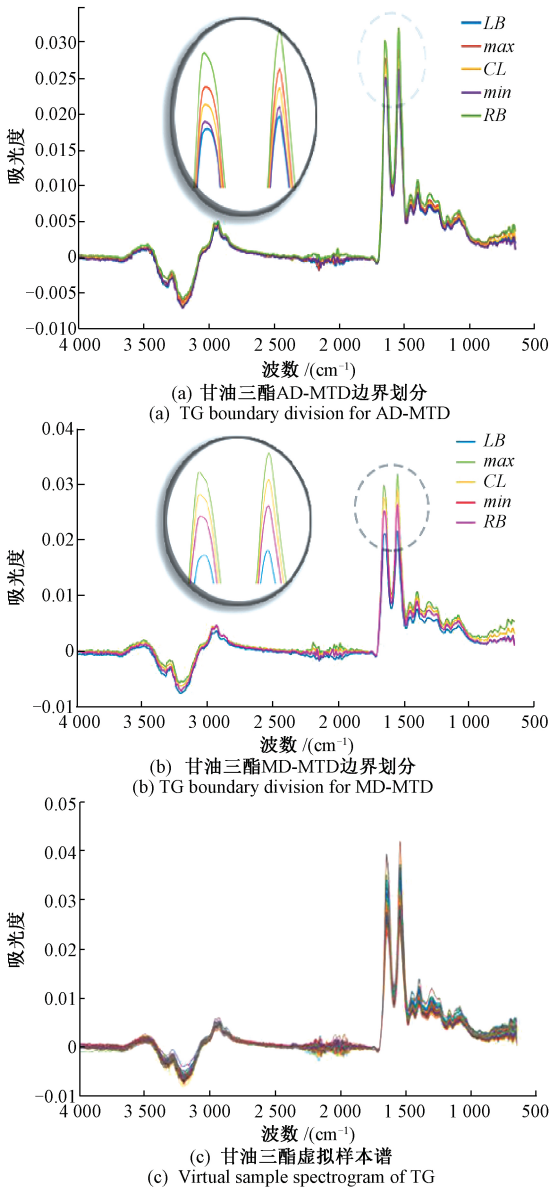

图 8　甘油三酯相关谱图信息

Fig.8　TG related spectrogram information

生成虚拟样本和原训练样本分布如图 9 所示。图 9(b)中,MD-MTD 法生成的虚拟样本集中在中部,影响的样本的分布均匀性。基于 Advanced-MTD 生成虚拟样本则在样本扩散边界部分,有效地扩充了样本的整体信息。从图 9(c)中两者的融合体现了 AD-MTD 对于 MD-MTD 的边界样本有所补充,使整体信息分布得到改善。

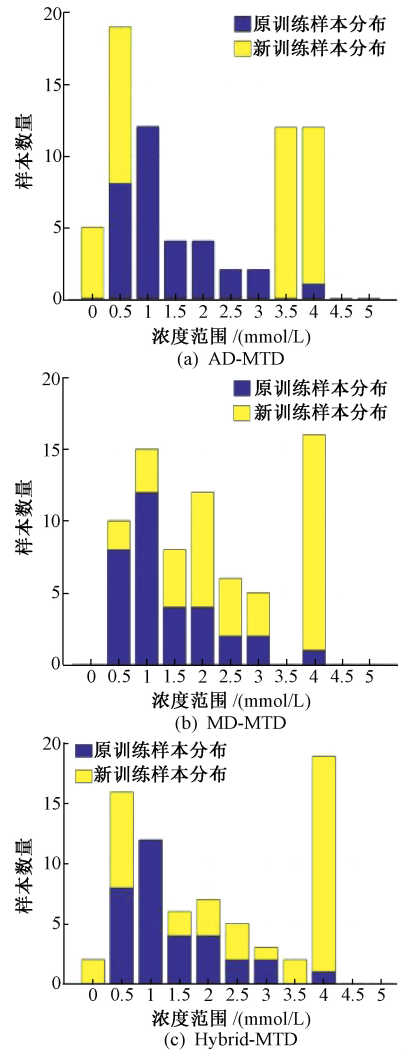

图 9　各虚拟样本建立的 TG 样本分布

Fig.9　The sample distribution constructed from various virtual samples for TG

图 9 中的原训练样本和不同方法建立的新训练样本的分布情况可进一步看出 MD-MTD 方法中心周围的样本扩充较多,边界样本较少。而图 9(c)中则可看出在 AD-MTD 的基础上结合 MD-MTD 增加的虚拟样本后得到的 Hybrid-MTD 可改善信息分布的完整性。同样,在填充虚拟样本后的样本空间建立 PLS 回归预测模型进行甘油三酯的定量分析。基于 Hybrid-MTD 填充虚拟样本后样本集对应的 PLS 回归预测模型的预测值与实际参考值的关系如图 10 所示。其最终甘油三酯浓度预测结果如表 2 所示。

由表 2 可知,与总胆固醇类似,对应于虚拟样本填充后的样本集对应的 PLS 的模型,MRmsep 有所降低,但与胆固醇不同,甘油三酯预测误差降低的效果不显著 OTR 指标也提升有限。原因在 MD-MTD 侧重增加内部的

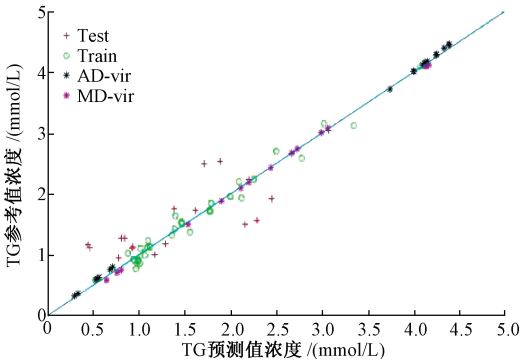

图 10 甘油三酯 Hybrid-MTD 样本集 PLS 模型的预测值与参考值之间的关系

Fig.10 The actual reference TG concentration versus the predicted TG concentration obtained from the PLS model based on Hybrid-MTD sample set

表 2 甘油三酯浓度预测结果分析

Table 2 Prediction result analysis of TG concentration

类型	*Nv*	*H′*	*H*	*hN/Pv*	*MRmsep*	*OTR*/%
–	–	–	PLS	5	0.58	–
MD-MTD	40	PLS	PLS	5	0.51	10.8
AD-MTD	40	PLS	PLS	5	0.52	10.3
Hybrid-MTD	40	PLS	PLS	5	0.45	22.4

虚拟样本,AD-MTD 侧重增加边界虚拟样本,均导致的虚拟信息不完整性对于甘油三酯的预测模型影响较严重。Hybrid-MTD 为对 MD-MTD 和 AD-MTD 各取 20 个的虚拟样本,与原训练样本融合为新的训练样本集,进行 PLS 模型预测,结果表明,基于 Hybrid-MTD 填充虚拟样本后的样本集建立的模型甘油三酯(TG)的定量分析平均预测均方差 $MRmesp = 0.45$ mmol/L。同时,*OTR* 指标也显示预测误差降低了 22.4%。因此,对于 Hybrid-MTD 来说,甘油三酯的预测效果同样得到了显著的提升。

4 结 论

本文针对样本信息空间分布的不均衡性和小样本空间信息的不完备性,提出 Hybrid-MTD 方法生成虚拟样本对样本空间进行填充,完善样本信息空间。对比 MD-MTD、Hybrid-MTD 与 AD-MTD 法可得知,MD-MTD 算法在虚拟样本的扩充区域主要为样本的中心区,而 AD-MTD 则主要在扩散边界附近增加样本,Hybrid-MTD 法基于前两者的混合,兼顾了扩散区和原信息区的分布,同时考虑增加边界扩散区域的虚拟样本和内部的虚拟样本,更好的改善实际样本空间信息分布。同时,通过对总胆固醇和甘油三酯血液光谱样本集进行虚拟样本构建与定量分析验证。总胆固醇和甘油三酯的全血光谱样本集定量分析实验表明,相对于 MD-MTD 和 AD-MTD 来说基于 Hybrid-MTD 生成的虚拟样本填充后的样本空间建立的总胆固醇和甘油三酯含量的 PLS 预测模型的预测性能最优。综上所述,本文提出的 Hybrid-MTD 虚拟样本生成方法能够有效解决样本空间信息不均衡性对所建立预测模型性能的影响,达到了提升小样本集及信息不均衡样本集的预测模型性能、提高其在血液光谱分析应用中的质量的目的。

参考文献

[1] 陈文亮,徐可欣,杜振辉, 等.人体无创血糖检测技术[J].仪器仪表学报, 2003,24(Z1): 258-261.
CHEN W L, XU K X, DU ZH H, et al. Techniques of human body non-invasive blood glucose detection [J]. Chinese Journal of Scientific Instrument 2003, 24(Z1): 258-261.

[2] MITCHELL A L, GAJJAR K B, HEOPHILOU G, et al. Vibrational spectroscopy of biofluids for disease screening or diagnosis: translation from the laboratory to a clinical setting[J]. Journal of Biophotonics, 2014, 7(3-4): 153-165.

[3] 张小青,孙小亮,潘庆华, 等.衰减全反射傅里叶变换红外光谱技术的临床应用研究进展[J].光谱学与光谱分析, 2017,37(2): 408-411.
ZHANG X Q, SUN P L, PAN Q H, et al. The advancement of attenuated total reflection Fourier transform infrared technology in clinical application [J]. Spectroscopy and Spectral Analysis, 2017, 37(2): 408-411.

[4] NORDESTGARRD B, BORGE G, Anette V. Triglycerides and cardiovascular disease [J]. 2014, The Lancet, 384 (9943): 626-635.

[5] 赵水平.甘油三酯对心血管疾病的影响[J].岭南心血管病杂志, 2013, 19(1): 1-3.
ZHAO SH P. Effects of triglycerides on cardiovascular disease [J]. Lingnan journal of cardiovascular disease, 2013, 19(1): 1-3.

[6] ANDERSON J, KEAVEN M. Cardiovascular disease risk profiles. American heart journal, 1991, 121 (1): 293-298.

[7] HE Y L, WANG P J, ZHANG M Q, et al. A novel and effective nonlinear interpolation virtual sample generation method for enhancing energy prediction and analysis on small data problem: A case study of Ethylene industry [J]. Energy, 2018, 147: 418-427.

[8] CHANG C J, LI D C, HUANG Y H, et al. A novel gray

forecasting model based on the box plot for small manufacturing data sets [J]. Applied mathematics and computation, 2015, 265: 400-408.

[9] CHANG C J, LI DC, DAI W L, et al. A latent information function to extend domain attributes to improve the accuracy of small-data-set forecasting [J]. Neuro computing, 2014, 129: 343-349.

[10] JULIAN J F, NICOLE H A. When small data beats big data [J]. Statistics & Probability Letters, 2018, 136: 142-145.

[11] PIERCESARE S. On the role of statistics in the era of big data: a call for a debate [J]. Statistics & Probability Letters, 2018, 136: 10-14.

[12] CHANG C J, LI D C, CHEN C C, et al. A forecasting model for small non-equigapdata sets considering data weights and occurrence possibilities [J]. Computers & Industrial Engineering, 2014, 67: 139-145.

[13] WANG Y Q, WANG Z Y, SUN J Y, et al. Gray bootstrap method for estimating frequency-varying random vibration signals with small samples [J]. Chinese Journal of Aeronautics, 2014, 27(2): 383-389.

[14] 朱宝. 虚拟样本生成技术及建模应用研究[D].北京:北京化工大学, 2017.
ZHU B. Research on virtual sample generation technologies and their modeling application [D]. Beijing: Beijing University of Chemical Technology, 2017.

[15] HUANG C F, CLAUDIO M B. A diffusion-neural-network for learning from small samples[J]. International Journal of Approximate Reasoning, 2003, 35(2): 137-161.

[16] LI D C, WU C S, TSAI T I, et al. Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge[J]. Computers & Operations Research, 2007, 34(4):966-982.

[17] CHEN ZH SH, ZHU B, HE Y L. A PSO based virtual sample generation method for small sample sets: Applications to regression datasets [J]. Engineering Applications of Artificial Intelligence, 2017, 59: 236-243.

[18] 朱宝,陈忠圣,余乐安.一种新颖的小样本整体趋势扩散技术[J].化工学报,2016, 67(3): 820-826.
ZHU B, CHEN ZH S, YU L A. A novel mega-trend-diffusion for small sample. CIESC Journal, 2016, 67(3): 820-826.

[19] LI ZH G, LI T H, LV H, et al. Quantitative analysis of biofluids based on hybrid spectra space [J]. Chemometrics and Intelligent Laboratory Systems, 2017, 165: 22-28.

[20] 李志刚,彭思龙,杨妮, 等.基于导数光谱融合建模的红外光谱定量分析方法[J].分析化学, 2016, 44(3): 437-443.
LI ZH G, PENG S L, YANG N, et al. Quantitative analysis of infrared spectra based on derivative spectra ensemble modeling method [J]. Analytica Chimica Acta, 2016, 44(3): 437-443.

[21] HUBERT L. On the Kolmogorov-Smirnov test for normality with mean and variance unknown [J]. Journal of the American Statistical Association, 1967, 62(318): 399-402.

作者简介

高克铉,2017 年于沈阳工程学院获得学士学位,现为东北大学在读硕士研究生,主要研究方向为中红外光谱分析与信号处理。
E-mail:gaokexuan01@163.com

Gao Kexuan received his B. Sc. degree in 2017 from Shenyang Institute of Engineering. Now, he is a master student in Northeastern University. His main research interest includes mid-infrared (MIR) spectrum analysis and signal processing.

李志刚(通信作者),1999 年于燕山大学获得学士学位,2002 年于燕山大学获得硕士学位,2005 年于天津大学获得博士学位,现为东北大学副教授,主要研究方向为中红外光谱分析与信号处理。
E-mail:lizhigang@neuq.edu.cn

Li Zhigang (Corresponding author) received his B. Sc. degree in 1999 and M. Sc. degree in 2002 both from Yanshan University, received his Ph. D. degree in 2005 from Tianjin University. Now, he is an associate professor in Northeastern University. His main research interest includes mid-infrared (MIR) spectrum analysis and signal processing.