基于网格划分局部线性嵌入算法的近红外光谱相似性度量方法

徐宝鼎[1]，丁香乾[1]，秦玉华[2*]，侯瑞春[1]，张磊[3]

[1]中国海洋大学信息科学与工程学院，山东 青岛 266100；

[2]青岛科技大学信息科学技术学院，山东 青岛 266061；

[3]山东烟草研究院有限公司，山东 济南 250101

摘要 近红外光谱数据的高维、高冗余、高噪声和非线性的特性严重影响了光谱相似性度量的准确性，针对该问题，提出了一种基于网格划分局部线性嵌入(GGLLE)算法的近红外光谱相似性度量方法。首先，根据关键化学成分在光谱中的表达，将高维光谱数据划分为多个网格子空间。其次，对局部线性嵌入(LLE)算法做了两方面改进，并采用改进的 LLE 算法依次实现每个子空间从高维空间向低维空间的特征映射，计算生成子空间的相似度矩阵。最后，将子空间相似度矩阵归一化处理并求解所累加和生成光谱样本集的相似度矩阵，实现光谱的相似性度量。实验选取两组某烟草企业提供的烟叶光谱构建了光谱的相似性度量模型，以相似性度量的准确率作为算法优劣的衡量标准。实验结果表明，GGLLE 算法构建的相似性度量模型的准确率为 93.3%，明显优于主成分分析、栈式自编码器和 LLE 算法的 64.2%、67.5%和 82.5%，从而证明了 GGLLE 算法的有效性。

关键词 光谱学；近红外光谱；相似性度量；改进局部线性嵌入算法；网格子空间；测地线距离；高维数据

中图分类号 O433.4 **文献标识码** A **doi**：10.3788/LOP56.033001

Similarity Measurement Method of Near Infrared Spectrum Based on Grid Division Local Linear Embedding Algorithm

Xu Baoding[1], Ding Xiangqian[1], Qin Yuhua[2*], Hou Ruichun[1], Zhang Lei[3]

[1] *College of Information Science and Engineering, Ocean University of China, Qingdao, Shandong* 266100, *China*;

[2] *College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, Shandong* 266061, *China*;

[3] *Shandong Tobacco Research Institute Co. Ltd., Jinan, Shandong* 250101, *China*

Abstract The high-dimension, high-redundancy, high-noise and nonlinear characteristics of near-infrared spectroscopy data seriously affect the accuracy of spectral similarity measurement. Aiming at this problem, a similarity measurement method of the near infrared spectrum based on the grid division local linear embedding (GGLLE) algorithm is proposed. First, the high-dimensional spectral data is divided into multiple grid subspaces according to the expression of key chemical components in the spectrum. Second, two aspects for the local linear embedding (LLE) algorithm are improved, and the improved LLE algorithm is used to sequentially map the feature of each subspace from high- to low-dimensional space and calculate the similarity matrix of the generated subspace. Finally, the subspace similarity matrix is normalized, and the similarity matrix of the accumulated and generated spectral sample set is to be solved to realize a similarity measurement of the spectrum. Two sets of tobacco leaf spectral data provided by a tobacco company are selected to construct a model of the spectral similarity measurement. The accuracy of the similarity measurement is a criterion of the pros and cons of the algorithm. The experimental results show that the accuracy of the similarity measurement model constructed by the GGLLE algorithm is 93.3%, which is obviously better than the accuracies achieved by principal component analysis, stacked

收稿日期：2018-07-06；**修回日期**：2018-08-08；**录用日期**：2018-08-17

基金项目：国家重点研发计划项目(2017YFB1400903)

* **E-mail**：yuu71@163.com

auto encoders, and LLE algorithms, which are 64.2%, 67.5%, and 82.5%, respectively. Thus, the effectiveness of the GGLLE algorithm is proved.

Key words spectroscopy; near-infrared spectrum; similarity measurement; improved local linear embedding algorithm; grid subspace; geodesic distance; high-dimensional data

OCIS codes 300.6340; 070.4790; 070.5010

1 引 言

近红外(NIR)光谱分析技术作为一项“绿色”分析技术,近几年来被广泛地用于烟草、石油、制药、乳品、农业等领域[1-2]。而近红外光谱的相似性度量方法[3]作为此项技术的一个重要部分,有着诸多应用场景。如在烟草企业中辅助配方设计工作,由于原材料的生产量有限,当配方中某烟叶出现库存短缺或价格、质量方面的波动时,需要用另一种品质特征近似的烟叶来替换。目前,寻找品质特征相近的烟叶主要依靠配方专家对烟叶常规化学成分、外观质量等指标的分析[4]。但是该方法费时、费力、步骤复杂且带有配方人员的主观性,已无法满足企业对产品均匀性的需求。所以,本文致力于寻找一种近红外光谱的相似性度量方法来代替传统的烟叶品质之间的相似性度量。相似性度量模式一般包括基于距离、夹角余弦等,其中距离(欧氏距离)是最常用的相似性度量方式。但是在处理维数高达上千维的近红外光谱数据时,多个距离之间的差值变得越来越小,“距离失效”问题尤为突出。而借助主成分分析(PCA)进行降维后,又会导致原始样本间距离结构、拓扑结构发生变化。高维空间中距离度量困难,使研究者们转向对降维方法的研究,期望通过一种能尽量保持样本集原有“距离结构”和信息的降维方法来实现高维向低维空间的转换,从而在低维空间实现相似性度量。曹鹏云等[5]提出了一种基于核变换和测地线距离的局部线性嵌入(LLE)算法的相似性度量方法来度量烟叶相近程度,但 LLE 算法要求样本集稠密均匀,对稀疏的光谱样本效果并不理想;丁玲等[6]利用等距特征映射(ISOMAP)算法对高光谱遥感数据进行非线性降维,但 ISOMAP 算法要求数据所在的流形等距欧氏空间的子集是凸集,但光谱数据空间难以保证满足凸集条件;并且上述两种方法都没有考虑到光谱数据高冗余、高噪声的特点,光谱之间的相似信息通常会被淹没在少数的噪声维中,最终导致度量结果不准确。

针对上述问题,本文提出了一种基于网格划分局部线性嵌入算法(GGLLE)的近红外光谱相似性度量方法。首先,将高维光谱数据根据关键化学成分在光谱中的主要吸收谱段划分为多个网格子空间。从高维数据空间的子空间出发,在某些主要维度上探讨数据间的相似性可以避免光谱中冗余和噪声维[7]的影响。然后,在 LLE 算法中,引入测地线距离[8]代替欧氏距离,解决了欧氏距离在度量高维数据时出现的“距离失效”问题,并改进了距离计算公式,使高维空间下的光谱数据集分布更均匀,避免因光谱数据样本稀疏导致的不确定性。采用 GGLLE 算法依次实现每个子空间从高维空间向低维空间的特征映射,并在低维空间中计算子空间相似度矩阵。最终生成光谱样本集的相似度矩阵,从而可以找出相似度最高的光谱。

2 原理与方法

2.1 LLE 算法

LLE 算法是 2000 年由 Roweis[9]提出的一种基于流行的非线性降维方法。它能够在保留数据原有的几何结构的基础上,有效实现数据从高维空间到低维空间的映射。LLE 算法在降维时能保持样本的局部不变性特征且具有待定参数少、适用于非线性数据处理等,已被广泛用于图像识别、高维数据可视化等领域[10-11]。

LLE 算法的思想为,假设数据在较小的局部是线性的,即某一个数据可以由它邻域中的几个样本来线性表示。例如样本 $\boldsymbol{x}_1$ 在它的原始高维邻域里用 K 近邻思想找到和离它最近的 k 个(如 3 个)样本 $\boldsymbol{x}_2$、$\boldsymbol{x}_3$、$\boldsymbol{x}_4$,然后假设 $\boldsymbol{x}_1$ 可以由 $\boldsymbol{x}_2$、$\boldsymbol{x}_3$、$\boldsymbol{x}_4$ 线性表示,即

$$\boldsymbol{x}_1 = w_{12}\boldsymbol{x}_2 + w_{13}\boldsymbol{x}_3 + w_{14}\boldsymbol{x}_4, \tag{1}$$

式中 w_{12}、w_{13}、w_{14} 为权重系数。在通过 LLE 降维后,希望 $\boldsymbol{x}_1$ 在低维空间对应的投影 $\boldsymbol{x}_1'$ 和 $\boldsymbol{x}_2$、$\boldsymbol{x}_3$、$\boldsymbol{x}_4$ 对应的投影 $\boldsymbol{x}_2'$、$\boldsymbol{x}_3'$、$\boldsymbol{x}_4'$ 也尽量保持同样的线性关系,即

$$\boldsymbol{x}_1' \approx w_{12}\boldsymbol{x}_2' + w_{13}\boldsymbol{x}_3' + w_{14}\boldsymbol{x}_4'。 \tag{2}$$

也就是说,投影前后线性关系的权重系数 w_{12}、w_{13}、w_{14} 是尽量不变或者最小改变的。由此可以看出,线性关系只在样本的附近起作用,距离远的样本对局部的线性关系没有影响,因此降低了降维的复杂度。

对 N 个样本点的 D 维数据集采用 LLE 算法降维到 d 维($d \ll D$),步骤如下:

1）对每个样本点求 K 个最近邻点。K 是预先设定的一个数值。对每个样本点 $\boldsymbol{x}_i$（$i=1,2,\cdots,N$）通过度量与其他样本点的欧氏距离来选取 K 个距离最小的点作为它的近邻点。

2）计算局部重构权值矩阵 $\boldsymbol{W}$。将每个样本点由 K 个近邻点近似表示的均方差定义损失函数，即

$$J(\boldsymbol{W})=\sum_{i=1}^{N}\left\|\boldsymbol{x}_i-\sum_{j=1}^{K}w_{ij}\boldsymbol{x}_j\right\|^2, \tag{3}$$

式中 $\boldsymbol{x}_j$（$j=1,2,\cdots,K$）为 $\boldsymbol{x}_i$ 的 K 个近邻点，w_{ij} 为样本点 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 之间的权重系数。对权重系数 w_{ij} 做归一化的限制，即 $\sum_{j=1}^{K}w_{ij}=1$。将损失函数矩阵化为

$$J(\boldsymbol{W})=\sum_{i=1}^{N}\boldsymbol{W}_i^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{x}_j)\boldsymbol{W}_i, \tag{4}$$

式中 $\boldsymbol{W}_i=(w_{i1},w_{i2},\cdots,w_{ik})^{\mathrm{T}}$。构造矩阵 $\boldsymbol{Z}_i=(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{x}_j)$ 并与 $\sum_{j=1}^{K}w_{ij}=1$ 结合，采用拉格朗日乘数法求出最优权重系数 $\boldsymbol{W}_i$ 为

$$\boldsymbol{W}_i=\frac{\boldsymbol{Z}_i^{-1}\boldsymbol{1}_k}{\boldsymbol{1}_k^{\mathrm{T}}\boldsymbol{Z}_i^{-1}\boldsymbol{1}_k}, \tag{5}$$

式中 $\boldsymbol{1}_k$ 为 k 维全 1 向量。

3）将数据集所有样本点映射到低维空间。算法希望降维前后数据保持局部不变性线性，即最小化损失函数 $J(\boldsymbol{Y})$，表示为

$$J(\boldsymbol{Y})=\sum_{i=1}^{N}\left\|\boldsymbol{y}_i-\sum_{j=1}^{K}\boldsymbol{W}_{ij}\boldsymbol{y}_j\right\|^2, \tag{6}$$

式中 $\boldsymbol{y}_i$ 是 $\boldsymbol{x}_i$ 的输出向量，$\boldsymbol{y}_j$（$j=1,2,\cdots,K$）为 $\boldsymbol{y}_i$ 的 K 个近邻点。在求解过程中，为了保证输出 $\boldsymbol{Y}$ 是惟一的，要满足如下约束条件，即

$$\sum_{j=1}^{N}\boldsymbol{y}_i=0,\frac{1}{N}\sum_{j=1}^{N}\boldsymbol{y}_i\boldsymbol{y}_i^{\mathrm{T}}=\boldsymbol{I}_{d\times d}, \tag{7}$$

用 $\boldsymbol{W}_i$ 表示 $\boldsymbol{W}$ 矩阵的第 i 列，$\boldsymbol{I}_i$ 表示 N 维单位矩阵的第 i 列。那么(6)式可矩阵化为

$$J(\boldsymbol{Y})=\sum_{i=1}^{N}\left\|\boldsymbol{Y}\boldsymbol{I}_i-\boldsymbol{Y}\boldsymbol{W}_i\right\|^2=$$
$$\mathrm{tr}[\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{W})^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{Y}]=\mathrm{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{Y}) \tag{8}$$

式中 $\boldsymbol{M}=(\boldsymbol{I}-\boldsymbol{W})^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{W})$。要使损失函数值最小，则取 $\boldsymbol{Y}$ 为 $\boldsymbol{M}$ 的 d 个最小非零特征值所对应的特征向量。将特征值从小到大排列，第一个特征值几乎为零，可舍去。通常取第 2～($d+1$)间的特征值所对应的特征向量作为输出结果。

2.2 基于网格划分局部线性嵌入算法

LLE 算法适用于非线性数据降维且待定参数少的情况，但是面对上千维、高噪高冗余、样本稀疏的光谱数据时仍有不足之处，针对上述问题，做出如下改进。

1）将光谱数据矩阵划分成多个网格子空间。光谱数据中大量冗余信息和噪声维对样本间的相似性度量产生了很大的干扰，样本间的相似信息也会被淹没在噪声维中，导致度量结果的不准确性。而且在高维数据中，维数越高，“维度灾难”问题越严重。因此本文依据文献[12]给出的特征选择方法（互信息结合遗传算法的特征选择方法），从光谱矩阵中分别筛选出对样品 N 种主要化学成分有较强相关性的特征，并分别生成 N 个网格子空间。从子空间出发度量光谱样本之间的相似性，削弱了“维度灾难”和噪声维对高维光谱相似性度量的影响，提高了计算精度。

2）引入 ISOMAP 中的测地线距离代替欧氏距离。在高维空间中，欧氏距离未必能够反映两个样本点的真实距离，在距离计算过程中可能会因两个曲面距离较小而导致不同曲面上的点进入同一个局部邻域，从而影响 LLE 在高维流形数据上的降维效果。而测地线距离是空间中两点的最短路径，所以空间中两点的测地线距离比欧氏距离更能反映空间点的拓扑结构。测地线距离与欧氏距离对比如图 1 所示。

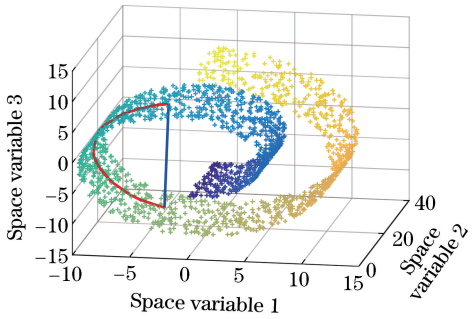

图 1 欧氏距离与地测线距离的比较

Fig. 1 Comparison between Euclidean distance and geodesic distance

图 1 中，红色线为空间中两个样本点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的测地线距离，蓝色线为欧氏距离。样本点 $\boldsymbol{x}_i$ 到 $\boldsymbol{x}_j$ 最短路径的权值之和 $d_{\mathrm{G}}(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 即为两点间的地测线距离，仿真数据采用 Dijkstra 算法计算得到。

改进 LLE 算法的距离计算公式。光谱数据样本集难免分布不均匀，LLE 算法在样本点稠密区域较小的 K 值就可以取得良好的效果，而在样本点稀疏区域为了保持各样本点之间的相对位置关系，往往需要较大的 K 值。在 LLE 计算过程中 K 值是唯一的，所以只有改善样本集的分布，使其分布均匀

才能避免上述问题。本文改进的距离计算公式为

$$D_{ij}=\frac{2d_{\mathrm{G}}(\boldsymbol{x}_i,\boldsymbol{x}_j)}{M_i+M_j},\tag{9}$$

式中：$d_{\mathrm{G}}(\boldsymbol{x}_i,\boldsymbol{x}_j)$表示$\boldsymbol{x}_i$，$\boldsymbol{x}_j$两个样本点之间的测地线距离；$M_i$、$M_j$分别表示$\boldsymbol{x}_i$、$\boldsymbol{x}_j$与其他点的平均距离。改进后的距离公式使整个光谱样本集分布更均匀，提高了算法的精度和效率。

具体方法和步骤如下：

1）网格子空间划分。通过文献[12]中给出的互信息结合遗传算法的特征选择方法，分别筛选出对N种化学成分相关性强的特征谱段$\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_N$，根据获得的特征谱段，将需要相似性度量的光谱矩阵划分成N个网格子空间。

2）将步骤1)中获得的N个网格子空间分别采用改进后的LLE算法映射到低维空间，并在低维空间中根据欧氏距离计算生成N个子空间的距离矩阵分别用$\boldsymbol{W}_1,\boldsymbol{W}_2,\cdots,\boldsymbol{W}_N$来表示。

3）将$\boldsymbol{W}_1,\boldsymbol{W}_2,\cdots,\boldsymbol{W}_N$分别通过$d_{ij}=(d_{ij}-d_{\min})/(d_{\max}-d_{\min})$归一化处理，若$d_{ij}>0.3$则令$d_{ij}=+\infty$。生成最终的相似度矩阵$\boldsymbol{W}=\boldsymbol{W}_1+\boldsymbol{W}_2+\cdots+\boldsymbol{W}_N$。寻找和样本$\boldsymbol{x}_i$相似度最高的样本时，只需查询相似度矩阵的$\boldsymbol{W}$在第$i$行的最小值对应的列值$j$，$\boldsymbol{x}_j$即为相似度最高的样本光谱。

3 实验方法

3.1 数据来源

实验选用两组某烟草企业提供的烟叶数据：第一组为268个库存烟叶（主要化学成分含量已知），用于实验参数的训练，以及烟叶产地、部位、等级的相似性度量验证。第二组为174对配方专家历史配方调整的烟叶，共348个烟叶样本，用于检验本文算法相似性度量与专家推荐结果的一致性。

每一对配方调整的烟叶都包含替换与被替换两个烟叶，当某配方中某种烟叶短缺时，配方专家通过化学成分检测、感官评吸等综合评定，找出品质特征最相似的烟叶作为替换。因此被替换与替换烟叶的光谱相似度很高。

3.2 光谱采集与预处理

光谱数据采集选用尼高力公司的Antaris II近红外光谱仪，光谱扫描范围为4000～10000 $\mathrm{cm^{-1}}$。样品在60℃的烘箱中烘4 h，磨粉后过40目筛，常温下避光密封储存。每个样品称重20 g，放置在直径为5 cm的样品杯中并用压样器压实，放入近红外光谱仪中扫描，实验室温度控制在18～25 ℃、湿度＜60%，采用漫反射方式，对扫描样品杯底部7个不同位置光谱取平均值。为避免样品的均匀性不一致，每个样品均重复装样扫描三次，计算三次扫描的平均值作为该样品光谱。

本次实验采用二阶导数加Norris(11)点平滑预处理方法，来消除背景噪声、仪器随机误差和基线漂移等干扰项对光谱数据的影响。光谱的预处理方法和建模均使用MATLAB R2015b软件完成。

4 结果与讨论

4.1 网格子空间划分

利用文献[12]中给出的互信息结合遗传算法的特征选择方法，在第一组库存烟叶样本集光谱中筛选出与烟碱有较强相关性的特征47个，如图2所示。

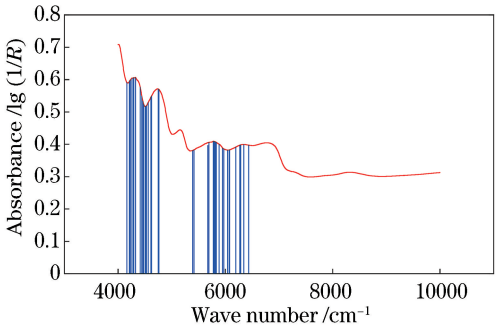

图2 烟碱相关性强的特征表达

Fig. 2 Feature expression of strong correlation for nicotine

由图2可知，蓝色直线与红色光谱相交的这47个特征可以很好地表达烟碱。接下来将它们从配方调整的光谱矩阵$\boldsymbol{M}$中挑选出来，整理合成一个特征子集矩阵，即为第一子空间矩阵$\boldsymbol{S}_1(47)$。同样的计算方法得到与总糖强相关的特征81个，整合成第二子空间矩阵$\boldsymbol{S}_2(81)$；以此类推，共得到烟碱、总糖、还原糖、总氮、淀粉的5个子空间矩阵：$\boldsymbol{S}_1(47)$、$\boldsymbol{S}_2(81)$、$\boldsymbol{S}_3(64)$、$\boldsymbol{S}_4(39)$、$\boldsymbol{S}_5(52)$。至此，从光谱矩阵划分出5个网格子空间，避免了水分吸收谱段和噪声谱段造成的影响。

4.2 确定参数K和本征维数

对于局部邻域k值的选取和本征维数d大小的确定，目前并没有足够好的理论方法，但是它们选取是否恰当对LLE算法的结果至关重要。K值选取过大，导致整个流形过于平滑，忽略了某些重要信息，同时也增加了计算量。K选取过小，则不能反映全局特征。本征维数取值过小，不能充分反映高维空间中数据的结构，过大则容易混入噪声信息。

本文提出了一种有监督的训练 K 值和本征维数值的方法，在第一组样本光谱数据集上，通过协同调整 K 值和 d 值并逐一计算残差来确定参数值。残差的计算公式为

$$R=1-\rho(\boldsymbol{D}_y,\boldsymbol{D}_n), \tag{10}$$

式中：$\boldsymbol{D}_y$ 子空间低维映射数据集的距离矩阵；$\boldsymbol{D}_n$ 为对应的化学成分含量差值矩阵；ρ 为线性相关系数。图 3 为第一子空间 $\boldsymbol{S}_1$ 在 K 值和本征维数取值不同时的残差图。

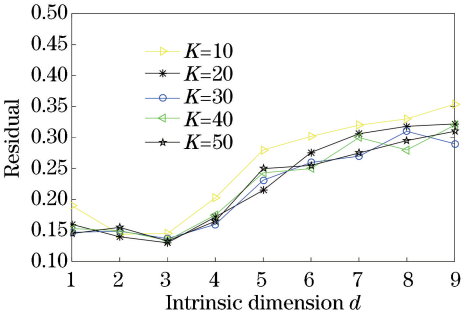

图 3 第一子空间在不同 K 和 d 值时的残差图

Fig. 3 Residuals of the first subspace at different K and d values

由图 3 可知，当调整局部邻域 K 的取值，本征维度 d 为 3 时，残差总是最小，并且当 d 取值为 3 时，K 的取值从 20 之后继续增大对残差影响不大，且 K 越大则计算量越大。所以本文对第一子空间通过改进的 LLE 降维，取 $K=20$，$d=3$。利用上述参数估计方法可得其余子空间的最佳参数，在 $\boldsymbol{S}_2,\boldsymbol{S}_3,\cdots,\boldsymbol{S}_5$ 中，除 $\boldsymbol{S}_2$ 的本征维度取值为 4 以外，其余最佳取值均为 3，且 $K=20$ 在所有子空间中都比较合理。

4.3 相似度矩阵

将子空间矩阵 $\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_5$ 通过改进的 LLE 进行降维(LLE 的参数设定如 3.2 节所示)，降维后采用欧氏距离计算样本点之间的距离，归一化后生每个成子空间的距离矩阵。例如将第一子空间 $\boldsymbol{S}_1$ 降维后，计算得到 $\boldsymbol{S}_1$ 的距离矩阵 $\boldsymbol{W}_1$，即

$$\boldsymbol{W}_1=\begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1m} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2m} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3m} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \cdots & x_{mm} \end{bmatrix}=$$

$$\begin{bmatrix} 0 & 0.074 & 0.237 & \cdots & 0.233 \\ 0.074 & 0 & 0.316 & \cdots & 0.282 \\ 0.237 & 0.316 & 0 & \cdots & 0.108 \\ \vdots & \vdots & \vdots & & \vdots \\ 0.233 & 0.282 & 0.108 & \cdots & 0 \end{bmatrix}, \tag{11}$$

式中：m 为样本数；$\boldsymbol{x}_{ij}$ 为第 i 个样本与第 j 个样本之间的距离，$\boldsymbol{x}_{ij}=\boldsymbol{x}_{ji}$；$\boldsymbol{W}_1$ 是一个 $m\times m$ 的斜对角对称矩阵。同样的方法求得剩余子空间的距离矩阵 $\boldsymbol{W}_2,\boldsymbol{W}_3,\cdots,\boldsymbol{W}_5$。将 $\boldsymbol{W}_1,\boldsymbol{W}_2,\cdots,\boldsymbol{W}_5$ 这 5 个距离矩阵相加即得为光谱的相似度矩阵 $\boldsymbol{W}$，即

$$\boldsymbol{W}=\begin{bmatrix} 0 & 0.468 & 1.741 & \cdots & 1.920 \\ 0.468 & 0 & 2.014 & \cdots & 1.895 \\ 1.741 & 2.014 & 0 & \cdots & 1.523 \\ \vdots & \vdots & \vdots & & \vdots \\ 1.920 & 1.895 & 1.523 & \cdots & 0 \end{bmatrix}。 \tag{12}$$

相似度度量方法认为，空间中的两个样本点距离越近，则相似度越高，即相似度与距离成反比。所以，度量两个样本的相似度时，只需从相似度矩阵 $\boldsymbol{W}$ 中查询这两个样本点的距离值，距离越小则相似度越高。如需查找某个样本点 i 的相似样本，则要查找 $\boldsymbol{W}$ 中的第 i 行的除 0 以外的最小值所对应的列 j，或查找 $\boldsymbol{W}$ 中的第 i 列的除 0 以外的最小值所对应的行 j，第 j 个样本即第 i 个样本的相似样本。

4.4 投影对比分析

为验证本文方法的有效性，将第一组 268 个库存烟叶光谱投影到二维空间中，从空间表达上比较 PCA、LLE、GGLLE 几种光谱相似性度量方法的性能差异，结果如图 4 所示。

领域专家认为同产地的烟叶相似度高，因此相似性度量模型应使同产地的烟叶尽可能靠近，不同产地的烟叶尽可能分开。由图 4 可知，本文算法可以更好地区分不同产地的烟叶，性能明显优于 PCA 和 LLE。

4.5 相似性度量结果对比分析

选取第一组和第二组烟叶样本近红外光谱构建光谱相似性度量模型，将本文方法的相似性度量准确率分别与 PCA、栈式自编码器(SAE)[13]和传统的 LLE 算法进行对比。PCA 选取前 6 个主成份(累积贡献率 90%)，SAE 为一个 1557-150-3 的网络结构，即总共有两个隐藏层，第一个隐藏层包含 150 个节点，第二个隐藏层包含 3 个节点。LLE 的局部邻域 K 设为 20，本征维度 d 为 3。为了避免实验结果的偶然性，光谱相似性度量实验从两个不同角度进行。

实验 1 是从第一组 268 个库存烟叶中随机选出 90 个烟叶，利用相似性度量模型寻找其相似烟叶，并记录准确率。判断依据是两个烟叶是否为相同产区、相同部位、相近等级。领域专家认为相同产区、相同部位、相近等级的烟叶相似度高。

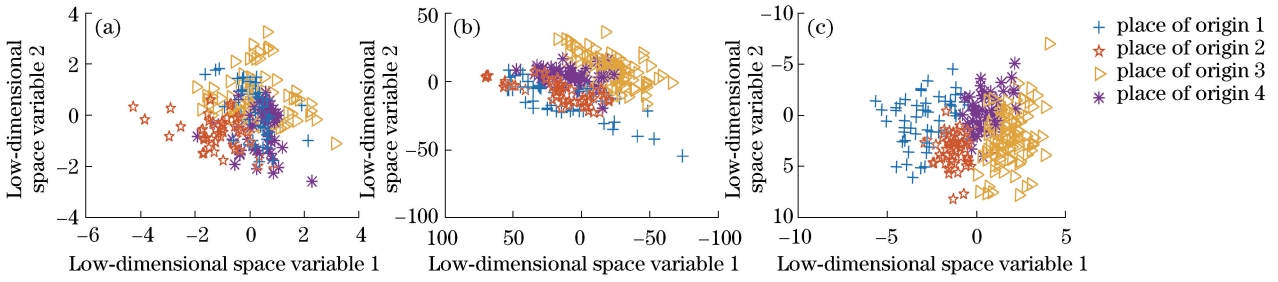

图 4 (a) PCA、(b) LLE 和(c) GGLLE 投影对比图

Fig. 4 Projection comparison of (a) PCA, (b) LLE, and (c) GGLLE

实验 2 是从 348 个配方烟叶中随机选出 120 个烟叶，利用相似性度量模型分别找出与它们最相似的前 3 个烟叶，与专家推荐的替换烟叶进行对比，并记录模型推荐烟叶与专家推荐烟叶一致性的概率。

两次实验的结果如表 1 所示。

表 1 不同算法性能对比

Table 1 Comparison of performance with different algorithms %

Algorithm	Accuracy of experiment 1	Accuracy of experiment 2
PCA	64.2	55.8
SAE	67.5	48.3
LLE	82.5	70.8
GGLLE	93.3	81.7

由表 1 可以看出，本文算法构建的烟叶近红外光谱相似性度量模型，寻找相似烟叶和替换烟叶的准确率分别为 93.3%和 81.7%，都明显高于其他算法，说明本文算法构建的光谱相似性度量模型稳定性更强、预测能力更好。PCA 算法为线性降维分析方法，当处理非线性的光谱数据集时，降维结果不能有效反映数据的本质特征，因此度量效果最差。SAE 是最近热点研究方向深度学习中的模型，在处理大规模数据的特征提取方面表现优异，但是 SAE 模型需要大规模的训练数据集来支撑，在处理小样本的光谱数据时，表现并不理想。LLE 算法能够在降维时保持样本位置的局部不变，取得较好的效果，但是容易受光谱中噪声维的影响以及本身的距离计算方法存在缺陷，所以仍具有可提高的空间。本文算法在 LLE 基础上进行了改进，实验结果证明，其相似性度量效果明显优于 PCA、SAE 和 LLE 算法。

5 结 论

基于网格划分的局部线性嵌入算法的近红外光谱相似性度量方法有效提高了近红外光谱相似性度量的稳健性和准确性。该算法将近红外光谱划分成多个网格子空间，有效剔除了噪声维并降低了单次 LLE 降维维度。同时，又对 LLE 算法进行了改进：引入测地线距离代替欧氏距离，解决了欧氏距离在度量高维数据时出现的“距离失效”问题；改进了距离计算公式，使高维空间下的光谱数据集分布更均匀，避免光谱数据样本稀疏导致的不确定性。实验结果表明，本文算法所建模型待定参数少、稳健性好、精度高，可辅助完成卷烟配方的维护与设计，同时对高维数据的相似性度量有普遍的参考意义。下一步研究重点将放在如何降低算法复杂度、提高效率上。

致谢 感谢中国烟草总公司山东省公司科技项目《基于近红外光谱的山东烟叶自动分选及化学成分协调性快速评价研究》资助。

参 考 文 献

[1] Wang L J, Yang Y Y. Purification and noise elimination of near-infrared spectrum in rapid detection of milk components concentration by using principal component weight resetting[J]. Acta Optica Sinica, 2017, 37(10): 1030003.
王丽杰, 杨羽翼. 利用主成分权重重置实现牛奶成分浓度快速检测中近红外光谱的净化去噪[J]. 光学学报, 2017, 37(10): 1030003.

[2] Kong Q Q, Ding X Q, Gong H L. Application of improved random forest pruning algorithm in tobacco origin identification of near infrared spectrum [J]. Laser & Optoelectronics Progress, 2018, 55(1): 013006.
孔清清, 丁香乾, 宫会丽. 改进的修剪随机森林算法在烟叶近红外光谱产地识别中的应用研究[J]. 激光与光电子学进展, 2018, 55(1): 013006.

[3] Zhao C H, Tian M H, Li J W. Research progress on spectral similarity metrics [J]. Journal of Harbin Engineering University, 2017, 38(8): 1179-1189.

赵春晖, 田明华, 李佳伟. 光谱相似性度量方法研究进展[J]. 哈尔滨工程大学学报, 2017, 38(8): 1179-1189.

[4] Du W, Tan X L, Yi J H, *et al*. Evaluation of leaf tobacco quality using chemical composition data[J]. Acta Tabacaria Sinica, 2007, 13(3): 25-31.
杜文, 谭新良, 易建华, 等. 用烟叶化学成分进行烟叶质量评价[J]. 中国烟草学报, 2007, 13(3): 25-31.

[5] Cao P Y, Fu Q J, Gong H L, *et al*. Similarity measurement method of tobacco leaves in high dimensional space [J]. Chinese Tobacco Science, 2013, 34(3): 84-88.
曹鹏云, 付秋娟, 宫会丽, 等. 高维空间下烟叶质量相似性度量方法研究[J]. 中国烟草科学, 2013, 34(3): 84-88.

[6] Ding L, Tang P, Li H Y. Dimensionality reduction and classification for hyperspectral remote sensing data using ISOMAP [J]. Infrared and Laser Engineering, 2013, 42(10): 2707-2711.
丁玲, 唐娉, 李宏益. 基于 ISOMAP 的高光谱遥感数据的降维与分类[J]. 红外与激光工程, 2013, 42(10): 2707-2711.

[7] He L, Cai Y C, Yang Z. Researches on similarity measurement of high dimensional data[J]. Computer Science, 2010, 37(5): 155-156,227.
贺玲, 蔡益朝, 杨征. 高维数据的相似性度量研究[J]. 计算机科学, 2010, 37(5): 155-156,227.

[8] Tenenbaum J B. A global geometric framework for nonlinear dimensionality reduction [J]. Science, 2000, 290(5500): 2319-2323.

[9] Roweis S T. Nonlinear dimensionality reduction by locally linear embedding [J]. Science, 2000, 290(5500): 2323-2326.

[10] Gou H Y, Zhou Y, Zhu C C, *et al*. Semi-supervised LLE algorithm of face recognition [J]. Computer Engineering and Design, 2011, 32(8): 2825-2828, 2908.
勾红云, 周勇, 朱长成, 等. 半监督 LLE 人脸识别算法[J]. 计算机工程与设计, 2011, 32(8): 2825-2828,2908.

[11] Liu J M, Zhou X L, Zhu S J, *et al*. Ear recognition based on locally linear embedding and its improved algorithm [J]. Opto-Electronic Engineering, 2012, 39(12): 132-137.
刘嘉敏, 周晓莉, 朱晟君, 等. 基于 LLE 及其改进算法的人耳识别[J]. 光电工程, 2012, 39(12): 132-137.

[12] Kong Q Q, Gong H L, Ding X Q, *et al*. Research on genetic algorithm based on mutual information in the spectrum selection[J]. Spectroscopy and Spectral Analysis, 2018, 38(1): 31-35.
孔清清, 宫会丽, 丁香乾, 等. 基于互信息的遗传算法在光谱谱段选择中应用[J]. 光谱学与光谱分析, 2018, 38(1): 31-35.

[13] Wang Y S, Yao H X, Zhao S C. Auto-encoder based dimensionality reduction[J]. Neurocomputing, 2016, 184: 232-242.