核映射和 Rank-Order 距离的局部保持投影相似性度量方法

秦玉华[1]，张　萌[1*]，杨　宁[2]，单秋甫[3]

1. 青岛科技大学信息科学技术学院，山东 青岛　266061
2. 青岛蓝智现代服务业数字工程技术研究中心，山东 青岛　266071
3. 云南中烟工业有限责任公司技术中心，云南 昆明　650024

摘　要　针对近红外光谱高维、高冗余、非线性和小样本等特点导致光谱相似性度量时出现的“维度灾难”，提出一种基于核映射和 rank-order 距离的局部保持投影(KRLPP)算法。首先将光谱数据经过核变换映射到更高维空间，有效保证了流形结构的非线性特征。然后改进局部保持投影(LPP)算法对数据进行降维操作，将 rank-order 距离替代传统的欧氏距离或测地线距离，通过共享邻近点的信息，得到更加准确的局部邻域关系。最后在低维空间通过距离的计算实现光谱的度量。该方法不仅有效解决了高维空间存在的“距离失效”问题，同时还提高了相似性度量结果的精度。为了验证 KRLPP 算法的有效性，首先根据降维前后数据集信息残差的变化确定了最佳参数近邻点的个数 k 和降维后的维数 d。其次，从光谱降维投影效果和模型分类效果两个角度与 PCA，LPP 和 INLPP 算法进行了对比，结果表明 KRLPP 算法对于烟叶的部位有较好的区分能力，降维效果以及对于不同部位的正确识别率明显优于 PCA，LPP 和 INLPP。最后，从某品牌卷烟叶组配方中选取了 5 个代表性烟叶作为目标烟叶，分别采用 PCA，LPP 和 KRLPP 方法从 300 个用于配方维护的烟叶样品中为每个目标烟叶寻找相似烟叶，并从化学成分和感官评价两方面对替换前后的烟叶及叶组配方进行了评价分析。其中 LPP 和 KRLPP 用于降维的参数选择保持一致，PCA 选择前 6 个主成分。结果表明，由 KRLPP 选出的替换烟叶与替换配方在总糖、还原糖、总烟碱、总氮等化学成分以及香气、烟气、口感等感官指标上较 PCA、LPP 方法差异最小，相似性度量准确度最高。该方法可应用于配方产品替换原料的查找，辅助企业实现产品质量的维护。

关键词　近红外光谱；局部保持投影算法；核映射；rank-order 距离；相似性度量

中图分类号：O657.33　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2021)10-3117-06

引　言

近年来，近红外光谱分析技术(NIR)因其简便、环保、速度快、不损坏样品等优点，已经在农业、石化、食品、烟草等众多领域占有重要地位[1]。产品的近红外光谱中含有超90%的结构信息，能够较全面的表征产品的质量信息。而相似性度量作为一种衡量数据间差异的重要方法，广泛应用于机器学习和数据挖掘领域[2]。对近红外光谱进行相似性度量，可实现产品之间质量相似性的评价；该方法也可用于食品、卷烟等各类配方产品相近原料的查找和替换。但近红外光谱数据具有高维、非线性、高噪声、高冗余的特点，同时存在数据分布稀疏和空空间现象[3]，导致相似性度量在低维空间常用的距离度量方式失效，因此需要研究一种高效的适用于高维数据的相似性度量方法，解决高维空间存在的“维度灾难”问题。

贺玲等[4]对高维空间进行基于网格划分的子空间相似性度量，但只能避免噪声对高维数据的影响。谢明霞等[5]提出了一种高维数据的相似性度量函数，可以有效缓解高维的影响，但此函数的提出基于聚类算法，并不具有普遍性。曹鹏云等[6]提出一种基于核方法和测地线距离的高维空间相似性度量方法，解决了传统度量中低维保距映射的问题，但不适用于稀疏的光谱样本。徐宝鼎等[7]改进局部线性嵌入算法中的距离度量公式并在子空间进行降维，但此方法需要通过特征筛选实现对子空间的划分，算法复杂且计算量较大。由此可见，由于直接对高维数据进行相似性度量较为困难，因此

收稿日期：2020-09-24，**修订日期**：2021-01-30

基金项目：国家重点研发计划项目(2018YFB1701704)，云南中烟工业有限责任公司项目(2019XX02)资助

作者简介：秦玉华，1971 年生，青岛科技大学信息科学技术学院教授　　e-mail：yuu71@163.com

*通讯作者　　e-mail：1427193350@qq.com

往往先采用降维的方法进行特征提取，消除高维数据中的噪声和冗余，在低维空间中进行数据的度量。但是目前存在的相似性度量方法都选择以测地线距离作为距离的度量方式，并没有真正映射到准确的邻域信息，因此得到的度量结果会出现不同程度的偏差。

针对上述问题，提出一种基于核变换和 rank-order 距离局部保持投影相似性度量方法，首先，将光谱数据映射到更高维的数据空间，采用改进的局部保持投影算法对数据进行降维，引入 rank-order 距离替代欧氏距离，可以更有效地保证映射到低维空间局部邻域信息的准确性，同时使得在降维之后的低维空间得到的相似光谱更加准确。将该方法应用于卷烟配方替换烟叶的寻找并取得了较好的效果，实现了卷烟配方的辅助维护。

1 算法与原理

1.1 经典局部保持投影算法

局部保持投影(LPP)算法作为一种经典的无监督的特征提取算法，由 He[8] 等首次提出。LPP 算法综合了 PCA 算法和 LE 算法的优点，有较强的泛化能力，在模式识别、数据挖掘等领域取得了显著成效[9]。假设在高维欧氏空间 R^D 中有 n 个 D 维数据集 $X=\{x_1, x_2, x_3, \cdots, x_n\}$，$x_i \in R^D$，($i=1, 2, 3, \cdots, n$)，LPP 算法的核心思想是选取一个最佳的变换矩阵 U 将高维数据集 X 映射到低维空间 R^d ($d \ll D$)[10]，在低维空间重构局部邻域信息，获得低维特征矩阵 $Y=\{y_1, y_2, y_3, \cdots, y_n\}$，使得降维之后的特征空间仍保持高维空间局部邻域信息不变。

LPP 算法的基本步骤如下：

Step1：通过欧氏距离的计算为样本点 x_i($i=1, 2, 3\cdots, n$)选出 k 个距离最近的点作为近邻点并构建邻接图 $G=(V, E)$。

Step2：利用热核函数 S_{ij}[11] 度量样本点 x_i 和邻近点 x_j 的相似性，即两点之间边的权重值。计算公式为

$$S_{ij}=\begin{cases}\exp\left(-\dfrac{\| x_i-x_j \|^2}{t}\right), & \text{if } x_i \in N_k(x_j) \,\forall\, x_j \in N_k(x_i)\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

式(1)中，t 为无关参数，$N_k(x_j)$可表示 x_i 所有的邻近点。

Step3：利用 $y_i=U^{\mathrm{T}}x_i$ 获得由高维数据 X 映射到低维空间的矩阵 Y。定义目标函数 $O(u)$为

$$O(u)=\sum_{i,j}^{n} \| u^{\mathrm{T}}x_i-u^{\mathrm{T}}y_j \|^2 S_{ij} \tag{2}$$

为了得到最佳变换矩阵 U，变换目标函数并将其最小化。

然后根据矩阵论通过式(3)计算广义特征方程的特征值。

$$XLX^{\mathrm{T}}u=\lambda XDX^{\mathrm{T}}u \tag{3}$$

其中，对角矩阵 $D_{jj}=\sum_i S_{ij}$，$L=U-S$ 为拉普拉斯矩阵。假定所求广义特征特征方程前 d 个特征值为 $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 \cdots \geqslant \lambda_d$，那么对应的特征向量解为 $u_1, u_2, \cdots, u_d$，由此得出最佳变换矩阵为 $U=[u_1, u_2, \cdots, u_d]$。

1.2 rank-order 距离

rank-order 距离作为一种新的距离度量方式，由 Zhu[12] 在 2011 年提出。rank-order 距离是利用数据点间共同的邻近点信息来计算样本间的距离，在高维空间中两点之间的直线距离不一定准确，但是加上邻近点的共享信息会极大的提高距离度量的准确性[13]。rank-order 距离计算步骤如下：

Step1：计算每个样本点 x_i 和其他样本点的欧氏距离，并根据距离的远近排序得到 x_i 邻近点的顺序表。

Step2：计算每两个样本点 x_i 和 x_j 间的不对称 rank-order 距离 $d(x_i, x_j)$。分别定义 $f_{x_i}(m)$为样本点 x_i 在顺序表中第 m 个邻近点，$R_{x_i}(x_j)$表示 x_j 是 x_i 在邻近顺序表中第几个邻近点，即样本点 x_j 在 x_i 邻近顺序表中的序号。则 $R_{x_j}(f_{x_i}(m))$表示的是样本点 $f_{x_i}(m)$在 x_j 邻近点顺序表中的序号。因此样本点 x_j 和 x_i 的非对称 rank-order 距离公式如式(4)

$$d(x_i, x_j)=\sum_{k=0}^{R_{x_i}(x_j)} R_{x_j}(f_{x_i}(m)) \tag{4}$$

由式(4)可知，$d(x_i, x_j)$表示的是样本点 x_i 的几个最近邻点在 x_j 的邻近点顺序表中所在位置序号的总和，并且 $d(x_i, x_j)$的值越小，空间中的样本点的局部邻域信息越准确。

如图 1 所示，样本点 x_i 和 x_j 之间的非对称 rank-order 距离为

$$\begin{aligned}d(x_i, x_j)&=\sum_{k=0}^{4} R_{x_j}(f_{x_i}(m))=R_{x_j}(f_{x_i}(0))+R_{x_j}(f_{x_i}(1))\\&+R_{x_j}(f_{x_i}(2))+R_{x_j}(f_{x_i}(3))+R_{x_j}(f_{x_i}(4))\\&=6+2+0+3+5=16\end{aligned} \tag{5}$$

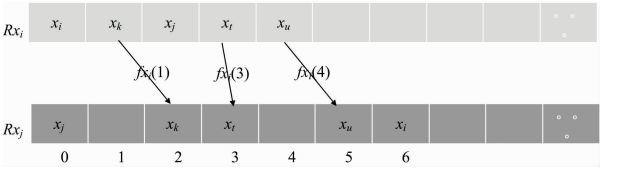

图 1 样本点 x_i 和 x_j 共享邻近点信息

Fig. 1 Shared neighbor information of sample points x_i and x_j

Step3：将 Step2 中计算得出的不对称 rank-order 距离进行归一化，可得对称的 rank-order 距离为

$$d_R(x_i, x_j)=\frac{d(x_i, x_j)+d(x_j, x_i)}{\min(R_{x_j}(x_j)+R_{x_j}(x_i))} \tag{6}$$

1.3 基于核映射 rank-order 距离的局部保持投影算法(KRLPP)

为了能准确地找出样本的邻近点，提出了基于核映射和 rank-order 距离的局部投影(KRLPP)算法，先通过核变换将数据集映射到更高维的空间，同时引入 rank-order 距离替换欧氏距离，通过共享局部邻近点的信息来重新度量样本点的相似关系，以此提高低维空间中相似性度量结果的精度。

KRLPP 算法步骤如下：

Step1：将近红外光谱数据矩阵 X 中每个样本通过高斯核函数 $K(m, n)=\exp\left(\dfrac{\| m-n \|^2}{2\sigma^2}\right)$进行核变换 $x \rightarrow \Phi(x)$[14]，非线性映射到希尔伯特空间 H 中。由核函数 $K=\Phi(X)^T\Phi(X)$ 可得 $K(x_i, x_j)=\langle \Phi(x_i), \Phi(x_j) \rangle=\Phi(x_i)^T\Phi(x_j)$。

Step2：根据 1.2 中算法的描述计算得出矩阵中任意两

点间的 rank-order 距离 $d_R\langle\Phi(x_i),\Phi(x_j)\rangle$，并以 $d_R\langle\Phi(x_i),\Phi(x_j)\rangle$ 寻找样本点的邻近点。

Step3：在 H 空间中，最小化目标函数，则式(3)可写为

$$\Phi(X)L\Phi(X)^{\mathrm{T}}u=\lambda\Phi(X)D\Phi(X)^{\mathrm{T}}u \tag{7}$$

用核变换后的核矩阵 $K=(K_{ij})$ 表示式(8)为

$$KLKu=\lambda KDKu \tag{8}$$

核矩阵 K 为半正定矩阵。

求式(8)的前 d 个广义特征值及特征向量得到最佳变换矩阵，并求出低维映射矩阵 Y。

Step4：对于降维得到低维特征矩阵 Y，通过欧氏距离进行相似样本点的寻找，样本个数自行设定。相似点的距离度量公式作为相似度的度量标准定义如式(9)

$$W(x_i)=\min_{1,2,\cdots,l} d(x_i,x_j)=\|x_i-x_j\|_{L_2} \tag{9}$$

2　实验部分

2.1　样品制备

选取 2017 年—2019 年 300 个用于调配卷烟配方的单料烟和 1 个需要维护的某品牌卷烟叶组配方(叶组配方是专家根据各种烟叶的主要化学成分、物理特征及感官等品质因素，将不同的单料烟按照一定原则和比例配制而成具有特定吸味风格和品质要求的卷烟产品)。将烟叶样品放置于 60 ℃的烘箱中干燥 4 h，用旋风磨磨碎过 40 目筛，密封平衡后进行光谱数据的采集。

2.2　光谱采集和预处理

选用尼高力公司的 Antaris Ⅱ 近红外光谱仪，扫描范围为 4 000～10 000 cm^{-1}，分辨率为 8 cm^{-1}。每个实验样品称重 15 g，放于在样品杯中用压样器压实进行光谱采集，室温保持在 18～25 ℃，为减少不确定性，每个样品扫描 3 次，取平均值作为该样品的最终光谱。

为消除高频噪声和基线漂移等对光谱造成的影响，选取二阶导数加 Savitzky Golay 平滑对光谱进行预处理。

2.3　评价方法

烟叶中总烟碱、总糖、还原糖、总氮等化学成分[15]及感官质量对烟叶的品质有重要影响，本研究主要以化学成分及感官评吸打分两种方式对替换前后的烟叶和叶组配方进行了对比。其中化学成分通过近红外方法检测三次取平均值得出，感官评吸由 10 位配方专家组成感官质量评价小组，依据 YC/T 497—2014《卷烟中式卷烟感官评价方法》，对烟叶的香气、烟气、口感特性(各占比 40%，40%和 20%)分别打分，总分为百分制。同时为了更直观的展示替换前后烟叶及叶组配方的总体质量差异，以 0.5 为梯度进行质量特征差异评价打分，评价标准如表 1 所示。

3　结果与讨论

3.1　参数的选择

近邻点个数 k 和降维后的维数 d 为 LPP 算法中两个重要的参数。k 取值过大会使部分重要的邻域结构信息被忽略，取值过小得到的邻域信息会比较局限；d 选取过大则可能会包含较多的噪声信息，在以往的算法中，往往都是根据经验选择，因此不同参数的选取对降维结果的影响颇大。本工作根据降维前后数据集信息残差的变化来确定参数。残差的计算公式如式(10)

$$R=1-\rho(D_X,D_Y) \tag{10}$$

式(10)中，D_X 和 D_Y 分别为降维前后数据的距离矩阵，ρ 为两者的线性相关系数，ρ 越大，代表高维数据降维之后得到的 D_Y 的信息量越大。图 2 为选取不同 k 值和 d 值的残差图。

表 1　总体质量评价标准

Table 1　Evaluation criteria of overall quality

质量偏差档次	质量偏差描述
0	无
1	轻微
2	略有
3	有
4	稍大
5	较大

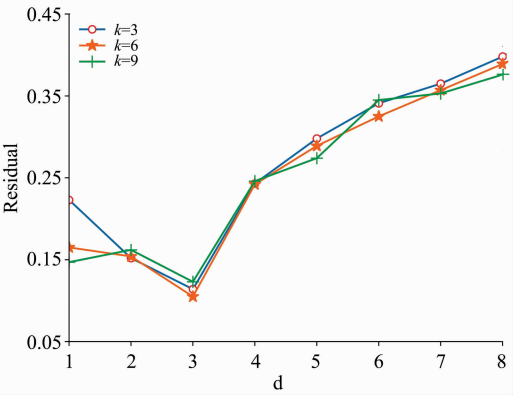

图 2　不同参数取值残差图

Fig. 2　Residual diagram with different parameter values

可以看出，当 k 值为 6，d 为 3 时，残差最小，表示此时映射到低维空间的特征矩阵获得最大的信息量。

3.2　投影结果对比分析

针对烟叶光谱数据的内在规律提取和相似性度量，有学者提出了改进邻域的局部保持投影方法 INLPP，将类别信息参数加入到距离计算中，对于不同香型风格的烟叶有较好的区分效果。图 3 为分别采用 PCA，LPP，INLPP 和 KRLPP 对烟叶光谱数据进行降维的投影效果对比。

不同部位的烟叶在化学成分、质量方面存在较多的差异，烟叶能提供的香味与部位间存在直接的相关性，因此配方设计和维护中配方人员要充分考虑烟叶部位的差异。由图 3 投影图可以看出，PCA 算法无法有效区分上、中、下不同部位的烟叶，LPP 算法对于区分不同部位烟叶边界仍明显存在交叉现象，INLPP 算法对于部位的区分效果明显优于 PCA 和 LPP 算法，但是中部和下部还是有少部分重叠的区域，而 KRLPP 算法对于烟叶上部、中部、下部三部分区分边界较为明显，降维效果优于 INLPP 方法。

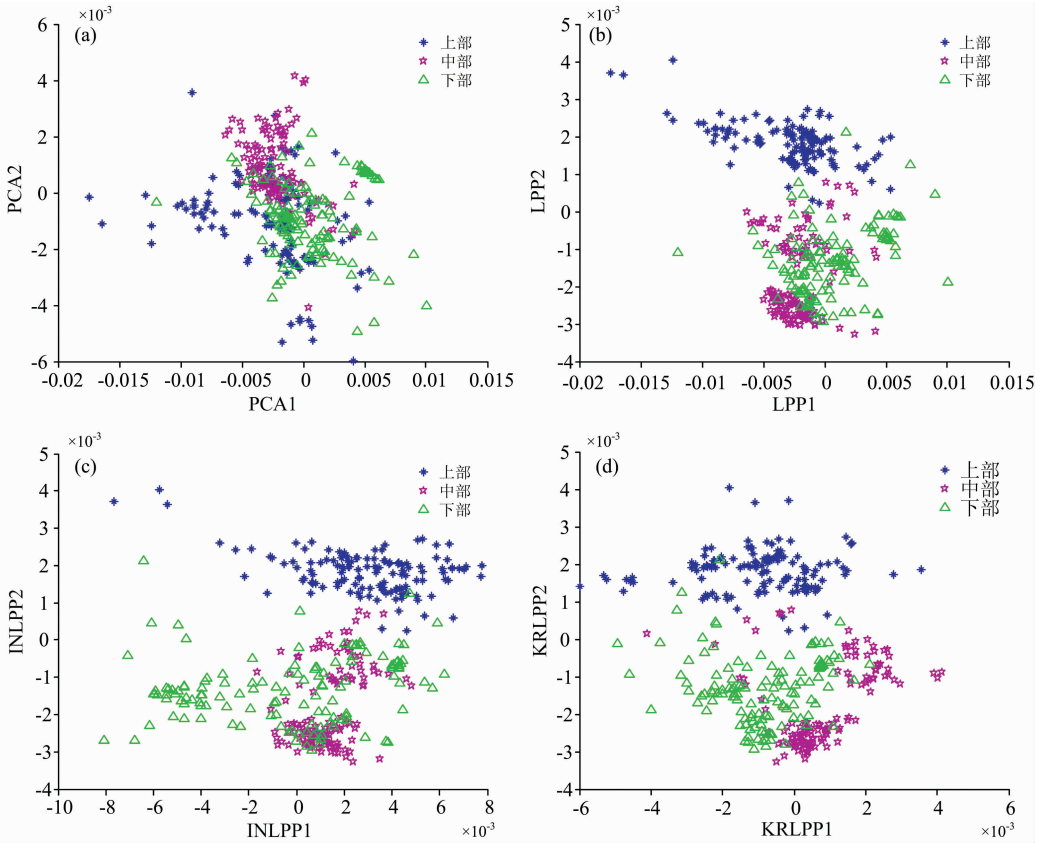

图 3 PCA，LPP，INLPP 和 KRLPP 算法降维投影图

Fig. 3 Dimensionality reduction projection of PCA，LPP，INLPP，KRLPP algorithms

表 2 不同降维算法烟叶部位分类结果对比

Table 2 Comparison of the classification results of tobacco stalk position based on different dimensional reduction algorithms

算法	正确识别率/%
PCA	66.2
LPP	77.6
INLPP	87.8
KRLPP	91.2

为进一步验证投影结果的有效性，表 2 为分别采用 PCA，LPP，INLPP 和 KRLPP 四种算法进行特征提取，使用 SVM 分类器建立不同部位烟叶的分类模型正确识别率的对比。

由表 2 可以得出，由 KRLPP 算法进行降维操作后的烟叶光谱不同部位的识别率为 91.2%，明显高于其他算法，说明该方法对于烟叶部位分类信息特征提取更为有效。

3.3 相似性度量结果对比分析

从卷烟叶组配方中选取 5 个代表性烟叶作为目标替换烟叶，然后分别采用 PCA，LPP 和 KRLPP 方法从 300 个用于配方维护的烟叶样品中为每个目标烟叶寻找相似烟叶，用于叶组配方中原料的替换。其中 LPP 和 KRLPP 用于降维的参数选择保持一致，PCA 选择前 6 个主成分。

为了验证实验结果的准确性，本文采用化学成分和感官评吸打分两种评价方式，分别从替换前后的单料烟和叶组配方两个角度进行了评价，从而保证了配方维护结果的可靠性。

3.3.1 单料烟替换前后评价结果对比

表 3 列出了 1 个目标烟叶分别采用 PCA，LPP 和 KRLPP 三种方法，根据相似度计算标准，从单料烟度量角度选出的 3 个替换烟叶与目标烟叶的化学成分和感官评吸结果对比。其他 4 个目标烟叶的替换推荐结果与该表所列的结果类似，不再详细列出。

由表 3 可得，三种方法所选出的替换烟叶与目标烟叶从化学成分和感官特征两方面皆有较小的偏差。其中由 PCA 算法选出的替换烟叶较目标烟叶偏差相对略大，LPP 次之，KRLPP 偏差最小。特别是 KRLPP 算法选出的 3 个替换烟叶，在总糖、还原糖、总烟碱、总氮等化学成分指标以及香气、烟气、口感等感官特征上与目标烟叶非常接近。说明该方法在卷烟配方维护中寻找相似烟叶的效果最好。

3.3.2 叶组配方替换前后评价结果对比

烟叶有效替换是配方维护的重要环节，但由于烟叶的种植受气候、土壤、栽培甚至是年份的影响，为保证产品质量的稳定性，通常要根据原料库存等实际需要对配方进行调整，寻找相似度高的替换烟叶决定了最终配方维护的稳定性和一致性，因此从叶组配方整体角度对比替换前后的配方产品评价更能体现配方的维护效果。

表 4 为采用 PCA，LPP 和 KRLPP 方法选出的 3 个替换烟叶(表 3 结果)，从叶组配方整体角度使用上述替换烟叶对目标烟叶进行替换，从而调配生成 3 个不同的替换配方与原配方的化学成分和感官评吸结果对比。

由表 4 可得，KRLPP 算法所得的替换配方在化学成分和感官指标上较 PCA 和 LPP 最接近于原配方，尤其是替换配方 1，各种指标几乎相同，配方质量差异最小，说明该方法得到的度量结果准确度最高。主要原因是该方法经过核变换和 rank-order 距离改进，使得高维数据在降维之后更能有效的保持局部邻域信息，因此相似性度量结果的稳定性和准确性更好，该方法能更有效地指导烟叶的配方设计与维护工作。

表 3　替换烟叶与目标烟叶评价结果对比

Table 3　Evaluation comparison of replacement tobacco and target tobacco

方法	单料烟	烟叶相似性	化学成分/%				感官特征				质量差异
			总烟碱	总糖	还原糖	总氮	香气特征	烟气特征	口感特征	总分	
	目标烟叶	—	2.24	25.51	21.83	1.97	35.1	34.4	15.0	84.5	—
	替换烟叶 1	10.8	2.42	25.62	21.94	1.92	34.8	34.0	14.7	83.5	1.5
PCA	替换烟叶 2	11.2	2.44	25.65	21.92	1.91	34.7	34.1	14.6	83.4	2.0
	替换烟叶 3	11.5	2.44	25.64	21.89	2.03	34.8	33.9	14.7	83.9	2.5
	替换烟叶 1	9.27	2.32	25.54	21.86	1.94	34.9	34.2	14.8	83.9	1.0
LPP	替换烟叶 2	9.56	2.34	25.56	21.86	1.93	34.8	34.1	14.8	83.7	1.5
	替换烟叶 3	9.73	2.36	25.57	21.87	1.93	34.9	34.1	14.7	83.7	1.5
	替换烟叶 1	4.54	2.22	25.52	21.83	1.96	35.1	34.4	15.0	84.5	0
KRLPP	替换烟叶 2	7.90	2.20	25.53	21.84	1.95	35.0	34.4	15.0	84.4	0
	替换烟叶 3	7.78	2.18	25.52	21.82	1.95	35.0	34.4	14.0	84.4	0.5

表 4　替换配方与原配方评价结果对比

Table 4　Evaluation comparison of replacement formula and original formula

方法	叶组配方	烟叶相似性	化学成分/%				感官特征				质量差异
			总烟碱	总糖	还原糖	总氮	香气特征	烟气特征	口感特征	总分	
	原配方	—	2.23	25.53	21.87	2.00	36.4	34.7	16.0	87.1	—
	替换配方 1	10.8	2.42	25.61	21.98	1.98	36.1	34.5	15.7	86.4	1.5
PCA	替换配方 2	11.2	2.43	25.66	21.95	1.96	36.2	34.4	15.8	86.4	1.5
	替换配方 3	11.5	2.45	25.65	21.93	2.06	36.1	34.4	15.7	86.2	2.0
	替换配方 1	9.27	2.31	25.55	21.89	1.98	36.3	34.7	15.9	86.9	1.0
LPP	替换配方 2	9.56	2.34	25.56	21.91	1.97	36.2	34.7	15.7	86.6	1.5
	替换配方 3	9.73	2.37	25.58	21.90	1.97	36.3	34.6	15.8	86.7	1.0
	替换配方 1	4.54	2.22	25.53	21.87	2.00	36.4	34.7	16.0	87.1	0
KRLPP	替换配方 2	7.90	2.20	25.54	21.86	2.00	36.4	34.6	16.0	87.0	0.5
	替换配方 3	7.78	2.20	25.52	21.87	1.99	36.3	34.7	16.0	87.0	0.5

4　结　论

基于核变换和 rank-order 距离的相似性度量方法 KRLPP 有效的提高了相似性度量的准确性，将光谱数据经过核变换之后，更能保持数据的空间结构，改进距离度量公式，则保证映射后的局部邻域信息更准确，使得高维空间中存在“距离失效”导致的维度灾难问题得到有效的解决。通过对替换前后烟叶和叶组配方两个角度进行化学成分和感官质量评吸打分可得，本文提出的相似性度量方法更有效的寻找替换烟叶和叶组配方的维护，该方法可有效推进配方产品辅助设计与维护工作，保持产品质量的稳定性。

References

[1] CHU Xiao-li, XU Yu-peng, LU Wan-zhen(褚小立，许育鹏，陆婉珍). Chinese Journal of Analytical Chemistry(分析化学), 2008, 23(5): 702.

[2] SONG Chun-jing, DING Xiang-qian, XU Peng-min, et al(宋春静，丁香乾，徐鹏民，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2017, 37(7): 2032.

[3] Li W, Wang G, Li K, et al. Chinese High Technology Letters, 2017, 65(2): 1764.

[4] HE Ling, CAI Yi-chao, YANG Zheng(贺　玲，蔡益朝，杨　征). Computer Science(计算机科学), 2010, 37(5): 155.

[5] XIE Ming-xia, GUO Jian-zhong, ZHANG Hai-bo, et al(谢明霞，郭建忠，张海波，等). Computer Engineering and Science(计算机工程

与科学)，2010，32(5)：92.
[6] CAO Peng-yun，FU Qiu-juan，GONG Hui-li，et al(曹鹏云，付秋娟，宫会丽，等). Chinese Tobacco Science(中国烟草科学)，2013，34(3)：84.
[7] XU Bao-ding，DING Xiang-qian，QIN Yu-hua，et al(徐宝鼎，丁香乾，秦玉华，等). Laser & Optoelectronics Progress(激光与光电子学进展)，2019，56(3)：251.
[8] Lu K，He X F. Pattern Recognition，2005，38(11)：2047.
[9] ZHANG Zhi-wei，YANG Fan，XIA Ke-wen，et al(张志伟，杨 帆，夏克文，等). Journal of Electronics and Information Technology(电子与信息学报)，2008，45(3)：539.
[10] Gu X H，Gong W G，Yang L P. Neurocomputing，2011，74(17)：1452.
[11] HUANG Dong-mei，ZHANG Xiao-tong，ZHANG Ming，et al(黄冬梅，张晓桐，张 明，等). Laser & Optoelectronics Progress(激光与光电子学进展)，2019，56(2)：63.
[12] Zhu C，Wen F，Sun J. A·Rank-Order Distance Based Clustering Algorithm for Face Tagging，CVPR 2011，2011，481. doi：10.1109/CVPR.2011.5995680.
[13] ZHAO Chun-hui，TIAN Ming-hua，LI Jia-wei(赵春晖，田明华，李佳伟). Journal of Harbin Engineering University(哈尔滨工程大学学报)，2017，38(8)：1179.
[14] Agelet L E，Ellis D D，Duvick S. J. Cereal. Sci.，2012，55(4)：160.
[15] Meesa C，Souard F，Delported C. Talanta，2018，177(9)：4.

Local Preserving Projection Similarity Measure Method Based on Kernel Mapping and Rank-Order Distance

QIN Yu-hua[1]，ZHANG Meng[1*]，YANG Ning[2]，SHAN Qiu-fu[3]
1. College of Information Science and Technology，Qingdao University of Science and Technology，Qingdao 266061，China
2. Qingdao Lanzhi Modern Service Industry Digital Engineering Research Center，Qingdao 266071，China
3. China Tobacco Yunnan Industrial Co.，Ltd.，Technical Research Center，Kunming 650024，China

Abstract Aiming at the curse of dimensionality problem in measuring spectral similarity caused by the high dimensionality，high redundancy，non-linearity and small samples of the near-infrared spectrum，a local preserving projection algorithm based on kernel mapping and rank-order distance (KRLPP) is proposed in this paper. First，the spectral data is mapped to a higher-dimensional space through a kernel transformation，which effectively ensures the manifold structure's nonlinear characteristics. Then，the dimensionality of the data is reduced by the locality preserving projections (LPP) algorithm，the rank-order distance is introduced instead of the traditional Euclidean distance or geodesic distance，and a more accurate local neighborhood relationship can be obtained by sharing the information of neighboring points. Finally，the measurement of the spectrum is realized by calculating the distance in low-dimensional space. This method solves the problem of distance failure in high-dimensional space and improves the accuracy of similarity measurement results. In order to verify the effectiveness of the KRLPP algorithm，firstly，the best parameters including the number k of the nearest neighbors and the dimensionality d of the reduced space were determined according to the residuals variation of the dataset before and after dimension reduction. Secondly，it compared with PCA，LPP，and INLPP algorithms from the perspectives of the projection effect of the spectra dimension reduction and the model classification ability. The results show that the KRLPP algorithm has a better ability to distinguish tobacco positions，and the effects of dimension reduction and correct identification of different tobacco positions are significantly better than PCA，LPP and INLPP methods. Finally，five representative tobacco were selected as target tobacco from a certain brand of cigarette formula. At the same time，PCA，LPP and KRLPP methods were used to find similar tobacco for each target tobacco from 300 tobacco samples used for formula maintenance，and the tobacco and cigarette formulas before and after replacement were evaluated from the aspects of chemical composition and sensory. Among them，the parameter selection of LPP and KRLPP for dimensionality reduction is consistent，and 6 principal components were selected for PCA. The results showed that，compared with PCA and LPP methods，the chemical components of total sugar，reducing sugar，total nicotine，total nitrogen and sensory indexes such as aroma，smoke and taste of the replacement tobacco and the replacement formula selected by the KRLPP algorithm had the least difference，and the accuracy of similarity measurement was the highest. This method can be applied to search for alternative raw materials for formula products and assist enterprises in maintaining product quality.

Keywords Near infrared spectroscopy；Local preservation projection algorithm；Kernel mapping；Rank-order distance；Similarity measure

* Corresponding author

(Received Sep. 24，2020；accepted Jan. 30，2021)