2013

Micro NIR1700 对 15 类固体化学药品判定分析的研究报告

QKX

KYSS

2013-5-7

JDSU MicroNIR™ 便携式近红外光谱仪对 15 类固体化学药品判定分析的研究报告

1 近红外光谱分析技术简述

1.1 近红外的定义

按照美国实验和材料协会（ASTM）的定义，近红外区域是指波长在 **780~2526nm** 范围内的电磁波，近红外谱区于 1800 年被天文学家 william Herschel 所发现，是人类最早发现的非可见光区域。在电磁波谱中，靠近可见光红光区域的电磁波称为红外光，红外包括三部分：近红外（NIR）、中红外（MIR）、远红外（FIR），近红外是红外的一部分，因最靠近红光，所以叫近红外。在整个电磁波谱中，从近红外开始往长波方向，依次为中红外、远红外、微波、中波、长波；往短波方向，依次为紫外、x-射线、γ-射线、宇宙射线。每一波长范围的电磁波都有其特性和应用。

1.2 近红外光谱的原理和特性

近红外分析技术的基本原理是利用物质中 C-H、N-H、O-H 、S-H 及 C=O、C=C 等基团对近红外光吸收较强的特点，根据有机物质的近红外光谱信息，对相应的物理、化学特征进行定性分析和定量测量。近红外光谱是以吸收波长为横坐标（标示吸收峰的位置），以吸光度为纵坐标（标示吸收峰的强度），构成的一幅图形。所有近红外光谱的吸收谱带都是中红外吸收基频的一级、二级、三级倍频及合频。即有机物质的分子振动，吸收近红外区域的光，产生了分子振动能级从基频到第二、三、四能级的跃迁，产生一级、二级、三级倍频。其吸收形成的光谱图就是近红外光谱图。

近红外光谱的特点：

1）C-H、N-H、O-H、S-H 及 C=O、C=C 等官能团的倍频峰、合频峰叠加在一起，形成一个宽峰，谱峰重叠严重。不像在中红外区域的比较尖锐的谱峰，容易分辨每个峰所代表特定基团的吸收。近红外区域的宽峰包含了很多的吸收信息，单从谱图上无法直接分辨是哪一种物质的吸收峰。

2)近红外区域的倍频或合频吸收系数很小，通常比中红外区域的基频吸收弱 20~50 倍。这一特点使得被测样品无需用溶剂稀释即可直接测定，便于生产过程的实时测定。另外，吸收系数小会妨碍样品中微量杂质的测定，但也保证了微量杂质或在近红外吸收弱的组分不至

于干扰测定。这也是近红外不能进行痕量分析的原因。

正是近红外光谱具有谱峰重叠严重、吸收系数低这两个特点，所以必须依赖计算机从谱图中提取信息，然后通过化学计量学软件对复杂的光谱信息进行解析，得到我们需要的分析数据。现代近红外分析技术突飞猛进的发展和成为一门独立的分析技术是与计算机的快速发展和化学计量学的深入研究密不可分。

2 研究的主要内容

本报告主要研究了 JDSU MicroNIR™ 便携式近红外光谱分析仪对不同类化学药品进行模式识别的思路和方法，选取了不同厂家、不同批次的 15 类化学药品，每一类化学药品分别扫描得到近红外光谱 10 条，分析了各模式识别算法在实际鉴定中的应用。结果表明 JDSU MicroNIR™ 便携式近红外光谱分析仪在实际应用中的可行性，为近红外模型的建立和优化以及新项目的研究和开发提供了一定的技术支持。

3 材料与方法

3.1 试验仪器与软件

仪器：JDSU MicroNIR™ 便携式近红外光谱仪（北京凯元盛世科技发展有限责任公司），主要部件包括：光学部分、控制部分、USB 接口数据线、笔记本电脑。仪器波长范围为 950 – 1650nm，扫描次数 50 次，单次积分时间 5000us。挪威 CAMO 公司 The Unscrambler 分析软件。

3.2 试验方法

3.2.1测试原理

当样品受到频率连续变化的近红外光照射时，将在其表面和内部产生漫反射。由于分子吸收了某些频率的辐射，并由其振动或转动运动引起偶极矩的净变化，产生分子振动和转动能级从基态到激发态的跃迁，使相应于这些吸收区域的反射光强度减弱。记录近红外光的漫反射光的强度与波数或波长关系曲线，得到近红外光谱。通过化学计量学软件，可使不同生产厂家，不同批次的同一类药品得到鉴别，并且使不同类别的样品得到较好地分离。

3.2.2测试方法：

3.2.2.1 样品采集与处理：

不同厂家、不同批次的药品共 15 类，具体生产厂家及批号如下：

（1）聚乙二醇 12000：① Fluka AG,Chemische fabric CH-9470 Buchs② Fluka 81285

（2）磷酸三苯酯：国药集团化学试剂有限公司，F20101109

（3）甲壳素：上海源叶生物科技有限公司，2011/12

（4）硼酸：北京朝阳区金盏化工厂，820707

（5）过氧化苯甲酰：山东邹平恒泰化工有限公司，20091101

（6）3，5-二硝基水杨酸：北京化学试剂公司，20051130

（7）琼脂糖：上海东海制药厂，850723

（8）D-果糖：北京欣经科生物技术有限公司，2934B28

（9）马铃薯淀粉：北京红星化工厂，781226

（10）十二烷基硫酸钠：北京化学试剂公司，20060613

（11）L-(+)-酒石酸：长城生物化学工程有限公司，20040518

（12）焦性没食子酸：国药集团化学试剂有限公司，20061102

（13）纤维素粉：国药集团化学试剂有限公司，20071105

（14）透明质酸：烟台三丰生化制品有限公司，20080906

（15）海藻糖：上海熔岩精细化工有限公司，040801

3.2.2.2近红外光谱采集：

使用JDSU MicroNIR™ 近红外光谱仪采集样品的光谱数据。将样品自然放置于样品盒中，样品盒放置在光谱仪的探头上，不同厂家和批次的每一类样品共扫描10次获得10条光谱，15类样品总共得到150条近红外光谱。

3.2.3**定性方法**

3.2.3.1建立PCA模型

将扫描得到的光谱导入挪威 CAMO 公司 The Unscrambler 化学计量学分析软件，采用线性建模方法主成分分析（principal component analysis, PCA）提取信息，对样品进行分类。主成分分析（principal component analysis, PCA）是对多变量数据进行统计处理的一种数据线性投影方法，它在尽可能保留原有信息的基础上将高维空间中的样本映射到较低维的主成分空间中。其基本思路是以一种最优化方法浓缩量测数据（用 Y 表示）信息，使数据矩阵简化，降低维数，寻找少数几个由原始变量线性组合的主成分，以提示数据 Y 结构特征，

提取基本信息。

3.2.3.2聚类分析

聚类分析（clusters）是一种无管理模式识别方法，常用于目标观测对象的分类，即利用表征观测对象的一组变量对目标进行分类。聚类分析不需要训练集，是无管理模式识别方法的典型代表，有很大的实用价值，特别适用于样品归属不清楚的情况。聚类分析所讨论的对象是n个样本组成的样本测量数据集合或样本数据矩阵。假如样本集合是不连续的，则这个样本集合可以看作是包含未知的若干个性质不同的子集，即不同的类。聚类分析的目的就是寻找这些子集，其基本思想是在多维模式空间中，任何一个子集内部样本之间的相似性即同类内相似性大于不同子集样本之间的相似性，即类与类之间的相似性。

3.2.3.3 SIMCA分类法

SIMCA分类法（soft independent modeling of class analogy）又称相似分析法，是在化学中得到广泛应用的模式识别方法。SIMCA算法用于模式识别分类的基本思路是对训练集中每一类样本测量数据矩阵分别进行主成分分析，建立每一个类的主成分回归数学模型，然后在此基础上对未知样本进行判别分类，即分别试探该未知样本与各类样本数学模型进行拟合，以确定其属于哪一类或不属于任何一类。由此可见，SIMCA不仅适用于两类的分类问题，而且也适用于两类以上或某样本同时属于两类或多类的问题。

SIMCA分类法是建立在主成分分析基础上的一种模式识别方法，其基本思路是先利用主成分分析的显示结果得到一个样本分类基本印象，然后分别对各类样本建立相应的模型，继而应用这些模型来对未知样本进行判别分析，以确定其属于哪一类，或不属于哪一类。

3.3 试验目的

对 15 种化学药品的近红外光谱进行采集，并利用化学计量学软件对其进行模式识别分析。

4 结果与分析

4.1 样品光谱的采集

对上述15种化学药品进行近红外光谱采集，测定的原始吸收光谱图见图1，图中横坐标为波长，纵坐标为反射率。从图1可看出，这15类化学样品的光谱交错重叠，十分相似，不易辨认。只从光谱特征上，难以区分各类药品，因此需要用化学计量学模式识别方法对光谱数据进行处理。

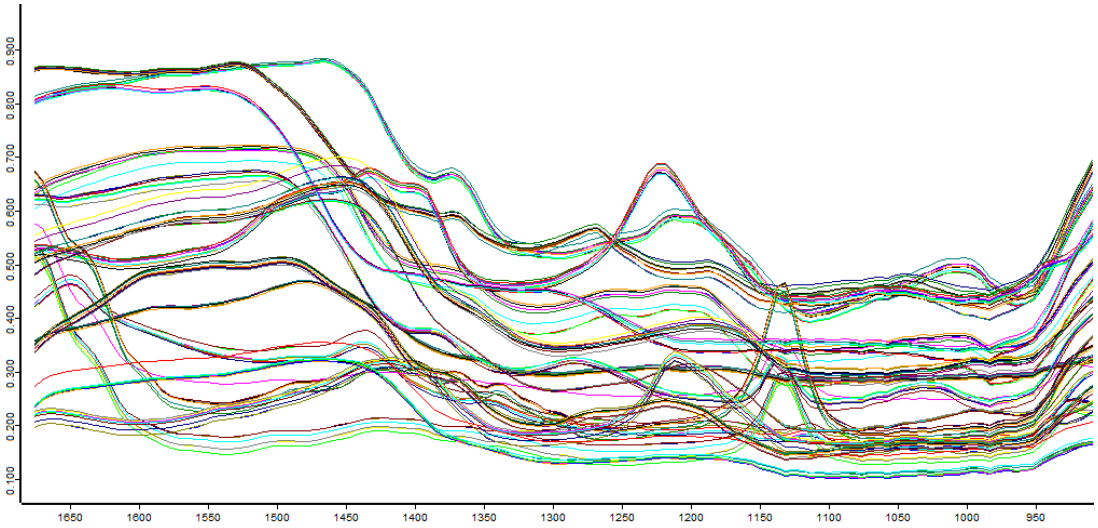
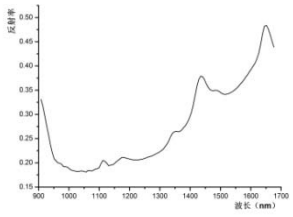
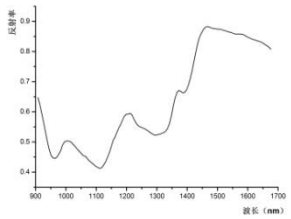


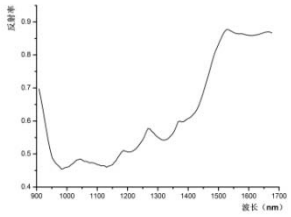
图 1　15 类样品的 150 条近红外光谱



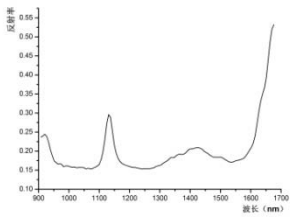
3,5-二硝基水杨酸



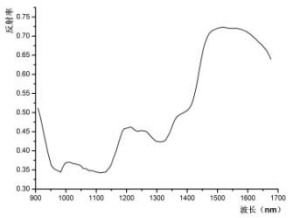
D-果糖



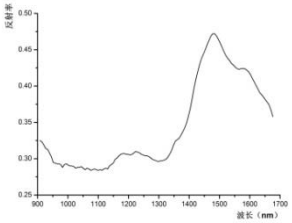
L-(+)-酒石酸



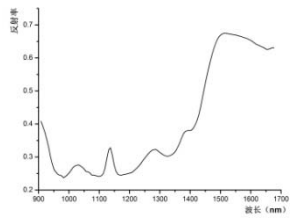
过氧化苯甲酰



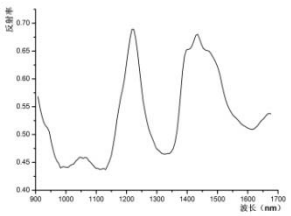
海藻糖



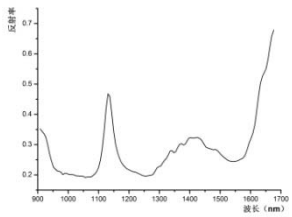
甲壳素



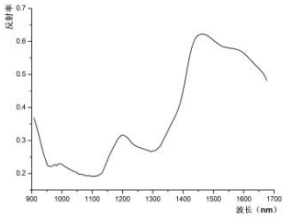
焦性没食子酸



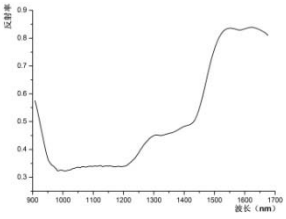
聚乙二醇 12000



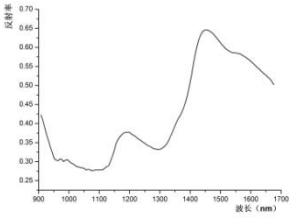
磷酸三苯酯



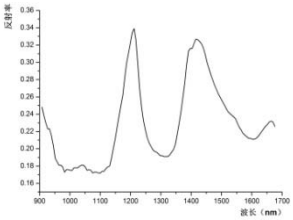
马铃薯淀粉



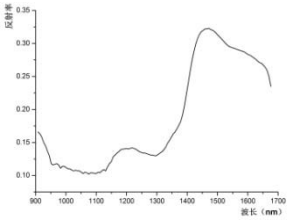
硼酸



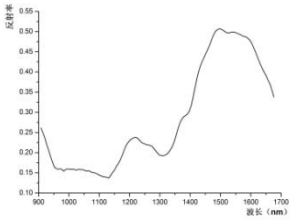
琼脂糖



十二烷基硫酸钠



透明质酸



纤维素粉

4.2 模式识别分析

将扫描得到的所有近红外光谱导入 The Unscrambler 化学计量学分析软件，然后分别采用主成分分析，聚类分析和 SIMCA 分析法对样品进行模式识别分析。

4. 2.1 主成分分析（PCA）

把 150 个样品合并，导入 The Unscrambler 化学计量学分析软件，采用 PCA 提取信息，对样品进行分类。分类结果如图 2（PC1 vs PC2 得分图），从图 2 中可以看出来，同一类样品分别被聚集在了一起，并且 15 类样品均得到了较好的分离；其中，有些样品类间距离较近，如 L-(+)-酒石酸和 D-果糖，透明质酸和 3，5-二硝基水杨酸。但总体来说，这几类样品还是可以得到较好的分离，分类结果较为满意。

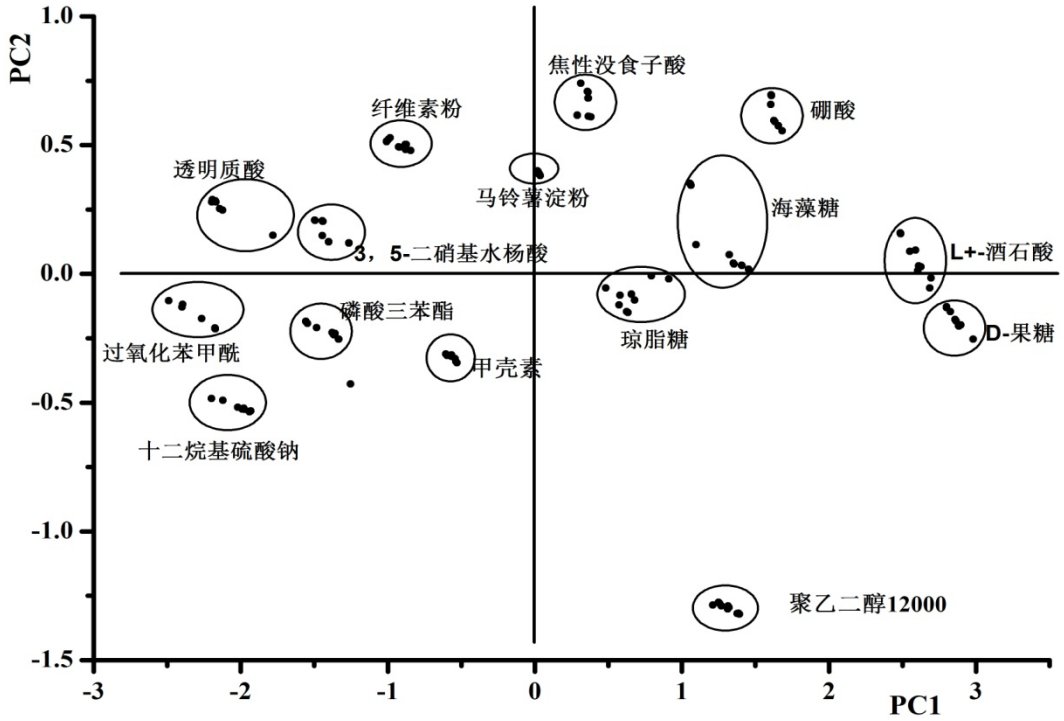


图 2 PCA 得分图（PC1 vs PC2）

4. 2.2 聚类分析

把 150 个样品合并，用聚类分析法分析样品。分类数为 15，迭代次数为 50 次。利用曼哈顿距离分类结果如表 1：

表 1 利用曼哈顿距离聚类分析结果表

样品	正确归类样品		错误归类样品	
	样品数	样品分类号	样品数	样品分类号
聚乙二醇 12000	10	8	0	—
甲壳素	10	4	0	—
硼酸	10	15	0	—
海藻糖	10	2	0	—
透明质酸	10	1	0	—
马铃薯淀粉	10	6	0	—

琼脂糖	10	7	0	—
十二烷基硫酸钠	10	10	0	—
焦性没食子酸	10	13	0	—
L-(+)-酒石酸	10	9	0	—
磷酸三苯酯	10	11	0	—
3，5-二硝基水杨酸	0	5	10	11
D-果糖	0	12	10	9
纤维素粉	6	3	4	12
过氧化苯甲酰	6	14	4	11（1 个）、5（3 个）

从表 1 中可以看出，15 类样品中，有 11 类样品全部正确归类，正确归类率均为 100%；有 4 类样品未实现正确归类，分别为 3，5-二硝基水杨酸（全部错归为磷酸三苯酯），D-果糖（全部错归为 L-(+)-酒石酸），纤维素粉（4 个样品归错为 D-果糖），过氧化苯甲酰（1 个样品错归为磷酸三苯酯，3 个样品错归为 3，5-二硝基水杨酸）。

分析得知，3，5-二硝基水杨酸、磷酸三苯酯、D-果糖、L-(+)-酒石酸错误率较高，因此将它们的 40 个样品合并，重新用聚类分析法分析，分类数为 4，迭代次数为 50 次。利用曼哈顿距离分类结果如表 2。从表中可以看出，原来相互归错类的样品全部得到了正确归类，说明利用曼哈顿距离聚类分析这四类样品是可行的。

表 2　40 个样品利用曼哈顿距离聚类分析结果表

样品	正确归类样品		错误归类样品	
	样品数	样品分类号	样品数	样品分类号
L-(+)-酒石酸	10	1	0	—
D-果糖	10	2	0	—
3，5-二硝基水杨酸	10	3	0	—
磷酸三苯酯	10	4	0	—

利用曼哈顿距离分类后，我们又利用了欧氏距离进行了聚类分析。15 类样品中，有 10 类样品全部正确归类，正确归类率均为 100%；有 5 类样品未实现正确归类，分别为 D-果糖（全部错归为 L-(+)-酒石酸），硼酸（全部错归为海藻糖），过氧化苯甲酰（1 个样品归错为磷酸三苯酯），透明质酸（有 1 个样品错归为 D-果糖），纤维素粉（4 个样品错归为硼酸）。

分析得知，硼酸、海藻糖、D-果糖、L-(+)-酒石酸错误率较高，因此将它们的 40 个样品合并，重新用聚类分析法分析，分类数为 4，迭代次数为 50 次。利用欧氏距离分类后，原来相互归错类的样品全部得到了正确归类，说明利用欧氏距离聚类分析这四类样品是可行的。

4. 2.3 SIMCA 分类法

把 150 个样品合并，用 SIMCA 建立定性模型。15 类样品总数共 150 个，每类样品随机选取 6 个共 90 个为模型预测集，剩余 60 个样品为校正集建立模型，置信度水平为 5%，对模型进行预测，结果如表 3 和图 3。从图表中可以看出，每一类样品对自身检验集样品的识别率为 100%，并且把其它类样品全部拒绝，拒绝率为 100%，无误判样品，整体预测效果很好。

从图 3 可以看出，每种样品的模型能有效辨别出被检测的样品是否属于同类，不是该种药品的样品不能被该模型识别，不同样品建立的总模型能全部识别所有模型内的样品类型。

表 3　SIMCA 定性模型对本类样品的识别率和拒绝率

序号	样品	识别率（%）	拒绝率（%）
1	聚乙二醇 12000	100	100
2	磷酸三苯酯	100	100
3	甲壳素	100	100
4	硼酸	100	100
5	马铃薯淀粉	100	100
6	琼脂糖	100	100
7	D-果糖，	100	100
8	3，5-二硝基水杨酸	100	100
9	海藻糖	100	100
10	过氧化苯甲酰	100	100
11	焦性没食子酸	100	100
12	十二烷基硫酸钠	100	100
13	L-(+)-酒石酸	100	100
14	纤维素粉	100	100
15	透明质酸	100	100

Classification Table

Sample	RESULT1	RESULT10	RESULT11	RESULT12	RESULT13	RESULT14	RESULT15	RESULT2	RESULT3	RESULT4	RESULT5	RESULT6	RESULT7	RESULT8	RESULT9
3，5-二硝基水杨酸 (2) - 0	*														
3，5-二硝基水杨酸 - 0	*														
3，5-二硝基水杨酸-12 - 0	*														
3，5-二硝基水杨酸-2 (2) - 0	*														
D-果糖-1 - 0								*							
D-果糖-10 - 0								*							
D-果糖-12 - 0								*							
D-果糖-14 - 0								*							
L+-酒石酸-1 - 0									*						
L+-酒石酸-12 - 0									*						
L+-酒石酸-14 - 0									*						
L+-酒石酸-16 - 0									*						
过氧化苯甲酰-1 - 0										*					
过氧化苯甲酰-10 - 0										*					
过氧化苯甲酰-12 - 0										*					
过氧化苯甲酰-14 - 0										*					
海藻糖-1 - 0											*				
海藻糖-11 - 0											*				
海藻糖-13 - 0											*				
海藻糖-14 - 0											*				
甲壳素-1 - 0												*			
甲壳素-11 - 0												*			
甲壳素-12 - 0												*			
甲壳素-14 - 0												*			
焦性没食子酸-1 - 0													*		
焦性没食子酸-10 - 0													*		
焦性没食子酸-12 - 0													*		
焦性没食子酸-14 - 0													*		
聚乙二醇-1 - 0														*	
聚乙二醇-11 - 0														*	
聚乙二醇-14 - 0														*	
聚乙二醇-16 - 0														*	
磷酸三苯酯-1 - 0															*
磷酸三苯酯-11 - 0															*
磷酸三苯酯-13 - 0															*
磷酸三苯酯-15 - 0															*
马铃薯淀粉-1 - 0		*													
马铃薯淀粉-10 - 0		*													

RESULT4, Significance = 5.0%

马铃薯淀粉-2 - 0		*													
马铃薯淀粉-3 - 0		*													
硼酸-10 - 0			*												
硼酸-11 - 0			*												
硼酸-12 - 0			*												
硼酸-14 - 0			*												
琼脂糖-1 - 0				*											
琼脂糖-11 - 0				*											
琼脂糖-13 - 0				*											
琼脂糖-15 - 0				*											
十二烷基硫酸钠-1 - 0					*										
十二烷基硫酸钠-11 - 0					*										
十二烷基硫酸钠-13 - 0					*										
十二烷基硫酸钠-15 - 0					*										
透明质酸-1 - 0						*									
透明质酸-11 - 0						*									
透明质酸-12 - 0						*									
透明质酸-14 - 0						*									
纤维素粉-1 - 0							*								
纤维素粉-10 - 0							*								
纤维素粉-2 - 0							*								
纤维素粉-3 - 0							*								

RESULT3, Significance = 5.0%

图 3 SIMCA 模型结果

4.2.4 小结

JDSU MicroNIR™ 便携式近红外光谱分析法能有效判别本研究中的 15 类化学药品。在 PCA 分析中，15 类样品分别聚类并获得明显分离；聚类分析中，不论是利用曼哈顿距离还是欧氏距离聚类，均有 10 类以上样品能获得 100%的正确归类，即使归类错误较多的样品集，在将它们的光谱合并并重新聚类后，其正确归类率亦可达 100%；SIMCA 判别分析中，不同类样品的模型能有效辨别出被检测的样品是否属于同类，不是该类的样品不能被该模型识别，所有模型的正确识别率均为 100%。因此，以实际生产中标准的各类化学样品建立定性分析模型，运用近红外光谱法定性分析化学药品是否合格，可对产品质量作实时监测。

5 结论与讨论

5.1 样品的代表性研究

本研究选取的样品代表性较好，充分保证了模型的稳定性与适应性。

5.2 JDSU MicroNIR™ 近红外光谱分析技术的可用性分析

JDSU MicroNIR™ 便携式近红外光谱分析法能有效判别本研究中的 15 类化学药品。本研究中采用的模式识别方法均给出了较好的结果。在现有近红外光谱分析法中，已有模型需要不断地维护与优化，使建模样品集符合“多而精”的原则。通过提高总体样品的代表性，可提高模型的稳定性，以此实现“精”；而校正样品集在具有代表性的基础上应尽量扩充样品数目，以增加模型样品的代表性，从而提高模型的预测准确度，以此实现“多”。不同时期对模型适应性的检验与校正可通过模式识别分析软件中模型检验与模型转移功能来实现。通过建立标准品的近红外光谱定性分析模型，可以实现对化学药品是否合格的定性分析，并对产品质量作实时监测。