江西农业大学学报 2012,34(5):1026 - 1031 http://xuebao.jxau.edu.cn
Acta Agriculturae Universitatis Jiangxiensis E - mail: ndxb7775@sina.com

# 近红外光谱结合特征变量筛选方法测定茶汤中的氨基酸含量

吴彦红[1],艾施荣[2],严霖元[1*],杨红飞[1],胡 琪[1]

(1. 江西农业大学 工学院,江西 南昌 330045; 2. 江西农业大学 软件学院,江西 南昌 330045)

**摘要**:采用透射方式获取茶汤的近红外光谱,利用特征变量筛选方法从茶汤的近红外光谱中提取氨基酸光谱信息,建立茶汤中氨基酸含量的快速检测模型。分别利用间隔偏最小二乘法(iPLS)和联合区间偏最小二乘法(siPLS)从茶汤的近红外光谱中提取微弱的氨基酸信息,建立其近红外光谱定量分析模型。结果表明,利用两种方法筛选的特征变量都避开了水的强吸收峰影响,但利用 siPLS 方法建立的模型性能明显好于 iPLS 的。最优的 siPLS 模型对校正集样本的相关系数为 0.912,交互验证均方根误差为 0.185;用预测集中独立样本检验模型性能,其相关系数为 0.887,预测均方根误差为 0.202。研究结果可为液体茶饮料中的成分实时快速检测提供参考。

**关键词**:茶汤;氨基酸;近红外光谱;特征变量筛选
**中图分类号**:O657.33 **文献标志码**:A **文章编号**:1000 - 2286(2012)05 - 1026 - 06

# Determination of Amino Acid Content in Tea Infusion using NIR Spectroscopy Combined with Characteristic Variables Selection Methods

WU Yan-hong[1], AI Shi-rong[2], YAN Lin-yuan[1*], YANG Hong-fei[1], HU Qi[1]

(1. College of Engineering, Jiangxi Agricultural University, Nanchang 330045, China; 2. College of Software, Jiangxi Agricultural University, Nanchang 330045, China)

**Abstract:** The objective of this study was to evaluate the capacity of NIR spectroscopy to rapidly predict the content of amino acid in tea infusion. Transmission mode was used to attain NIR spectroscopy of tea infusion. Interval partial least square (iPLS) and synergy interval partial least square (siPLS) were applied to select the feeble amino acid information from NIR spectroscopy of tea infusion in this study. The optimized characteristic variables were used to develop PLS models. The results show that the selected feature variables based on iPLS and siPLS were not within the range of the strong absorbance for water, but the built model using siPLS had better performance than that of iPLS. The optimal siPLS model was achieved with Rc = 0.912 and RMSECV = 0.185 in the calibration set, Rp = 0.912 and RMSEP = 0.202 in the prediction set. The attained results can provide a reference for the rapid and real - time determination of the components of in liquid tea drinks.

**Key words:** tea infusion; amino acid; NIR spectroscopy; characteristic variables selection

**收稿日期**:2012 - 06 - 28 **修回日期**:2012 - 07 - 26
**基金项目**:江西省科技计划项目(20112BBF60019)和江西省教育厅科学基金项目(GJJ11081)
**作者简介**:吴彦红(1959—),女,教授,主要从事农产品无损检测研究,E - mail: yhwu289@yahoo.com.cn; * **通讯作者**:严霖元,教授,E - mail: yly.jx@163.com。

茶叶内含有茶多酚、氨基酸、咖啡碱等多种对人体有益的成分,既是一种具有营养性和风味性的饮品,又是一种具有降血脂、防辐射、抗癌等药理功效的功能性饮品[1]。随着生活水平的提高,生活节奏的加快,液态茶饮料以其具有方便、消暑解渴、保健疗效、开瓶即饮等特点越来越受到人们的喜爱,目前"即开型"茶饮料已占了茶叶产业的绝大部分市场[2]。而据国家质量监督检验检疫总局对市场上 38 家企业的茶饮料产品质量抽检结果,发现大部分产品中内含成分低于茶饮料标准或根本不含有茶成分[3]。加强茶饮料质量检测力度,已是茶饮料行业迫切需要解决的问题。

茶叶中的氨基酸是构成其鲜爽味的主要物质,也是人体所需的主要营养物质,具有预防疾病等保健功效[4]。而茶叶中的氨基酸含量较少,只占干物质含量的 3% 左右,在茶饮料中,氨基酸含量更是甚少。氨基酸常规检测方法主要有茚三酮比色法[5]、高效液相色谱法等[6],这些方法检测精度高,但属于化学方法,检测步骤繁琐、耗时长、费用高,无法满足茶饮料加工和贸易过程中的质量快速检测需要。

近红外光谱技术具有简单、快速、成本低和重现性好等优点,已被广泛应用于液态样品品质的快速检测中[7-9],目前在茶汤的茶多酚成分检测中也得到了广泛应用[10-11],但用于茶汤中氨基酸含量检测还未见相关报道。氨基酸组分复杂,导致了其近红外光谱的复杂性,另外,茶饮料的近红外谱峰中含有水的强吸收峰。因此,如何通过一些数据挖掘方法从复杂的液体光谱信息中提取特征信息,以建立精确、稳定的定量分析模型是必须要解决的关键问题。本研究分别采用间隔偏最小二乘法(interval partial least square,iPLS)和联合区间偏最小二乘法(Synergy interval partial least square, siPLS)从茶汤的近红外光谱中提取氨基酸光谱信息,建立氨基酸的近红外光谱定量分析模型,文中由 iPLS 方法建立的模型简称 iPLS 模型,由 siPLS 方法建立的模型简称 siPLS 模型。

## 1　材料和方法

### 1.1　样本收集与茶汤制备

从茶叶市场收集绿茶、红茶、乌龙茶种类 90 个茶样,茶样原产地为江西、江苏、福建、云南、安微、浙江等国内重要产区,将茶样编号后置于 4 ℃冰柜中保存。试验时,从每个茶样中分别称取 3 g,在室温下放置 12 h 达到室温平衡后,用 150 mL 沸蒸馏水加盖冲泡 5 min,倒出茶汤,用滤纸过滤,将滤液迅速冷却到室温。

### 1.2　茶汤近红外光谱采集

使用 Antaris Ⅱ傅里叶变换近红外光谱仪(Thermo Scientific,USA)采集茶汤光谱,光谱仪带有透射样品池附件。光谱采集条件为: 光谱波数范围 10 000 ~ 4 000 $cm^{-1}$,分辨率为 3.856 $^{-1}$,扫描次数为 32 次。采集光谱时,将备用茶汤注入 5 mm 光程的样品池中,每个样本采集光谱后旋转大约 60°重复采集 3 次,求其平均值作为原始光谱。

### 1.3　氨基酸含量参考测定

为防止因茶汤浓度过高而导致吸光度值过大,用吸管分别从每个样本的茶汤中吸取 10 mL 到 25 mL 容量瓶中,用蒸馏水稀释到刻度。根据酒石酸亚铁比色法(GB/T 8314—2002)测量各稀释后茶汤中的氨基酸含量。从原始样本中随机选出 60 个组成校正集,用来建立校正模型; 余下 30 个组成预测集,用来检验模型性能。

### 1.4　间隔偏最小乘法(iPLS)和联合区间偏最小二乘法(siPLS)

间隔偏最小二乘法(iPLS)是一种有效的波长筛选方法[12],其原理是将整个光谱区等分为若干个子区间,然后在全光谱区和每个子区间内分别建立偏最小二乘回归模型,比较各模型的精度,取精度最高的模型所在的子区间为最终优选的特征波长区间。联合区间偏最小二乘法(siPLS)已广泛应用于光谱数据的特征波长优选上[13-14],其原理是将整个光谱区等分为多个子区间,再联合其中的 2 个、3 个或更多的子区间建立模型,比较各模型的预测误差,误差最小的模型所对应的联合子区间即是被优选的特征光谱区间。

### 1.5　模型性能评价

在校正集中采用交互验证法(leave-one-out cross-validation)优化模型参数,以氨基酸含量的实测值与模型预测值的相关系数 $R_c$、校正集交互验证均方根误差(root mean square error of cross-validation,*RMSECV*)

和预测集均方根误差(root mean squared error of prediction,*RMSEP*)作为模型性能的评价指标,所有数据分析基于 Matlab V7.8.0 平台完成。

## 2 结果与讨论

### 2.1 茶汤近红外原始光谱分析及光谱预处理

图1为茶汤近红外原始光谱图,从图1可看出,各样本茶汤的光谱信息几乎没有差异,且在 6 900 ~ 7 140 $cm^{-1}$ 和 5 155 $cm^{-1}$ 附近存在强吸收峰,这是茶汤中水的O-H伸缩振动的合频和一级倍频吸收[15],因此必须从整个光谱信息中提取与氨基酸相关的特征信息,消除水峰及其他无关信息对模型性能的影响。另外,因原始光谱中夹带有噪声信息及其他无关信息,文献研究表明标准正态变量变换(Standard Normal Variate Transformation,SNV)能有效去除光谱中的噪音信息[16],本研究采用SNV方法预处理原始光谱图。

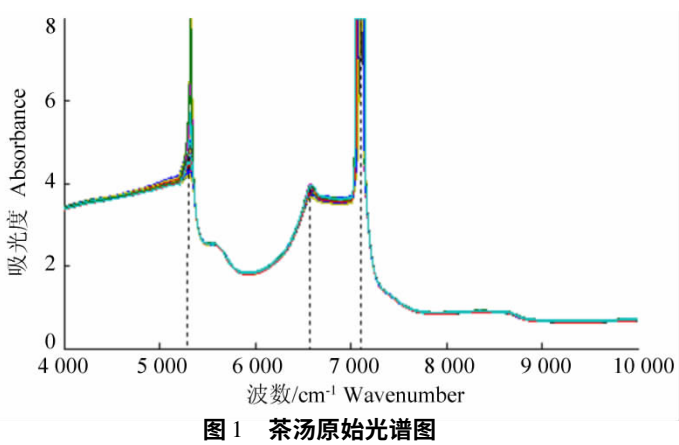



图1 茶汤原始光谱图

Fig. 1 NIR original spectra of tea infusion

### 2.2 间隔偏最小二乘法(iPLS)筛选特征光谱区间及iPLS模型建立

为考察子区间数划分对优选结果的影响,试验将整个光谱区分别等分为11~25个子区间,在校正集中用交互验证法优化模型参数,模型性能衡量标准由交互验证均方根误差值(*RMSECV*)决定,优选结果见图2。从图中可看出,当整个光谱区等分为15子区间时,所建模型的*RMSECV*值最小,因此研究将整个光谱区等分为15个子区间来优选特征光谱区间。图3是在15个子区间上优选的区间结果图,当选择第12个子区间,使用前5个主成分建立模型时,其*RMSECV*最小,对应的波数范围为8 412.0~8 809.2 $cm^{-1}$。模型对校正集的相关系数($R_c$)和交互验证均方根误差(*RMSECV*)分别为0.788和0.278;预测集的中相关系数($R_p$)和预测均方根误差(*RMSEP*)分别为0.724和0.299。图中表明,用iPLS优选出来的特征变量不在水的强吸收峰内,从而避免了水峰的影响。

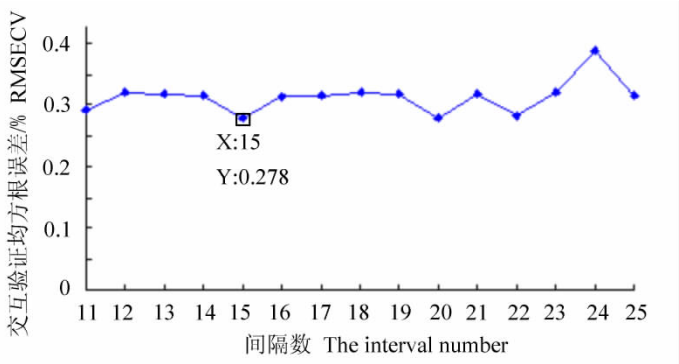



图2 各划分子区间上所建最优模型的交互验证均方根误差

Fig. 2 RMSECV of the optimal models based on all divided interval numbers

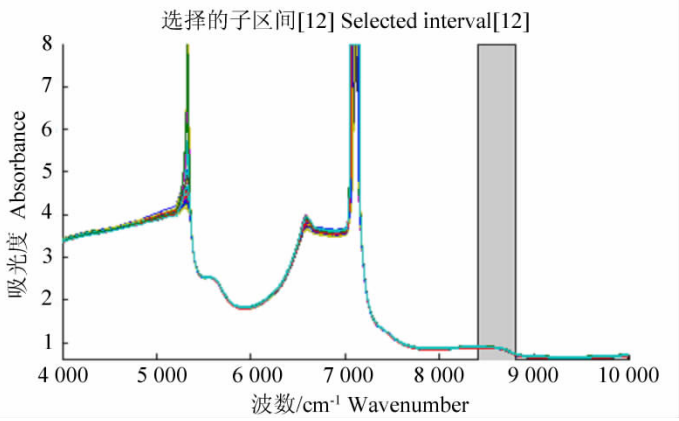



图3 iPLS优选的间隔数

Fig. 3 The selected interval by iPLS

### 2.3 联合区间偏最小二乘法(siPLS)筛选特征光谱区间及siPLS模型建立

由上述结果可知,建立的iPLS模型性能较差,研究采用另一种特征变量筛选方法(联合区间偏最小二乘法,siPLS)优选特征变量,试验将整个光谱区分别等分为11~25个子区间,在各等分的子区间内,又分别联合其中的2、3和4个子区间建立模型。在校正集中交互验证法优化模型参数,由最小的交互验证均方根误差值(*RMSECV*)作为衡量标准,优选结果见表1。从表中结果可知,当整个光谱区被划分

表 1　由 siPLS 方法在不同子区间划分的特征光谱区间优选结果

Tab. 1　Results of feature spectral regions selected by siPLS

| 区间数 Interval numbers | 主成分因子数 PLS factors | 被选子区间 Selected interval numbers | *RMSECV*/% |
|---|---|---|---|
| 11 | 10 | [4 9 10 11] | 0.220 |
| 12 | 7 | [4 8 9] | 0.208 |
| 13 | 5 | [5 9 12] | 0.228 |
| 14 | 8 | [5 12] | 0.196 |
| 15 | 9 | [5 9 12] | 0.205 |
| 16 | 7 | [5 7 11 13] | 0.219 |
| 17 | 7 | [6 10 14] | 0.218 |
| 18 | 8 | [6 11 13] | 0.191 |
| 19 | 7 | [6 12 13 15] | 0.208 |
| 20 | 7 | [6 8 14 17] | 0.209 |
| 21 | 7 | [7 13 17] | 0.185 |
| 22 | 8 | [7 8 13 15] | 0.205 |
| 23 | 7 | [7 9 15 20] | 0.189 |
| 24 | 7 | [8 15 19] | 0.199 |
| 25 | 8 | [8 16 21 24] | 0.197 |

成 21 个子区间,联合其中的第 7、13 和第 17 个子区间,使用前 7 个主成分数时建立模型的 *RMSECV* 值最小。图 4 是优选结果图,图中的灰色区是筛选出的 3 个特征子区间,各子区间对应的波数范围分别为 5 723.7 ~ 6 005.2 $cm^{-1}$、7 436.2 ~ 7 717.7 $cm^{-1}$ 和 8 577.8 ~ 8 859.4 $cm^{-1}$,共 222 个变量。由图也可看出,优选的特征变量不在水的强吸收峰内。

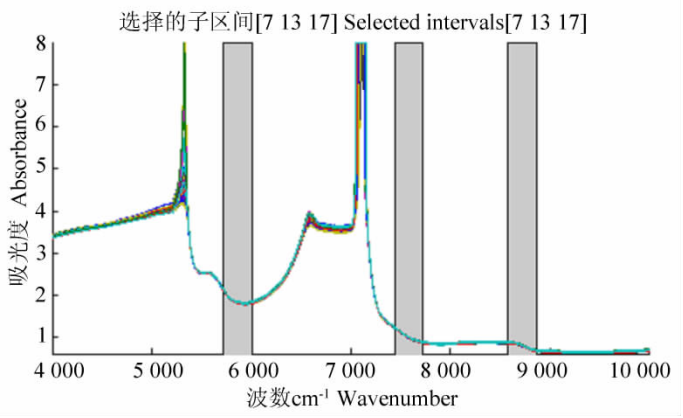



图 4　由 siPLS 筛选的特征光谱区间 [7 1317]

Fig. 4　Optimal spectral regions [7 1317] selected by siPLS

由优选的 222 个变量建立 PLS 模型,模型对校正集样本的相关系数($R_c$)和交互验证均方根误差(*RMSECV*)分别为 0.912 和 0.185;预测集的中相关系数($R_p$)和预测均方根误差(*RMSEP*)分别为 0.887 和 0.202。图 5 校正集和预测集中样本的实测值与模型预测值之间的散点图。

## 2.4　模型性能比较与讨论

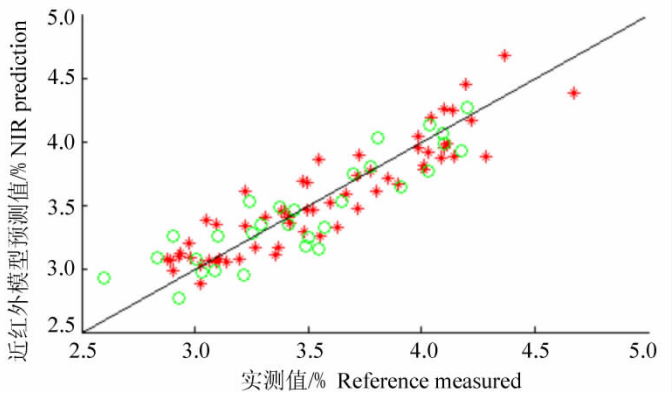
本研究比较经典偏最小二乘(PLS)模型、iPLS 模型和 siPLS 模型性能,比较结果见表 2。从表 2 可看出,siPLS 模型性能最好,经典 PLS 模型较差。这是因为经典 PLS 模型是利用茶汤的近红外全光谱变

表 2　不同模型结果比较

Tab. 2　Results and comparison from different models

| 模型 Model | 主成分因子数 PLS factors | 变量数 The number of variables | 校正集 Calibration set | | 预测集 Prediction set | |
|---|---|---|---|---|---|---|
| | | | $R_c$ | *RMSECV* | $R_p$ | *RMSEP* |
| PLS | 6 | 1 557 | 0.400 | 0.459 | 0.574 | 0.356 |
| iPLS | 5 | 103 | 0.788 | 0.278 | 0.724 | 0.299 |
| siPLS | 7 | 222 | 0.912 | 0.185 | 0.887 | 0.202 |

量建模，全光谱区内含有大量与氨基酸成分无关的信息，尤其是水的吸收峰，这些信息参与模型建立必然会严重影响模型的精度和稳定性。而利用 iPLS 建模时，优选出与氨基酸成分最相关的特征变量建立模型，且优选得到的特征变量不在水的强吸收峰内，因此其模型性能要好于全光谱区的 PLS 模型性能；但仅优选其中的一个子区间来建立模型，其他子区间中也含有一些与氨基酸成分最相关的信息。SiPLS 模型是通过优选出几个与氨基酸成分最相关的光谱子区间建立的模型，其性能明显要好于 iPLS 模型。




“* ”表示校正集样本，“o”表示预测集样本。

“* ” is calibration set，“o” is prediction set.

图 5 校正集和预测集中实测值与模型预测值之间的散点图

Fig. 5 Reference measurement versus NIR prediction in the calibration and prediction sets

## 3 结 论

本文研究利用近红外光谱技术结合特征变量筛选方法快速测定茶汤中的氨基酸含量。研究结果发现，利用 iPLS 方法筛选的特征变量建立模型避开了水的强吸收峰影响，但模型精度较差，而利用 siPLS 方法筛选的特征变量建立模型同样避开了水的强吸收峰影响，但模型性能明显好于 iPLS 的。研究表明利用近红外光谱技术结合 siPLS 方法测定茶汤中的氨基酸含量是可行的。

茶汤中含有大量水分，由于水分子的强吸收使茶汤近红外原始光谱在 6 900 ~ 7 140 $cm^{-1}$ 波段产生吸收峰饱和现象；另外，氨基酸在茶汤中的含量很低，相对于水的强吸收，氨基酸的近红外光谱非常微弱，其吸收信息容易被水的强吸收信号掩盖和干扰。姜礼义等[10]探讨利用合适的光程来减少水溶液干扰所产生的误差，得出使用 1 mm 光程的透射附件所建立模型是可行的。但透射附件的光程太小，在清洗、制作工艺等方面会存在一些相应的负面问题。本文探讨使用化学计量学方法从茶汤近红外光谱中提取氨基酸的微弱光谱信息，茶汤原始光谱使用 5 mm 光程石英比色皿采集，其中有部分光谱区产生饱和现象，严重影响模型精度。而 siPLS 方法通过将整个光谱区等分成多个子区间，再分别联合其中的几个精度较高的子区间建立模型，这些精度较高子区间内的变量跟待测成分最相关，直接剔除了一些无关信息且避开了水峰影响。与利用小光程的近红外光谱分析方法相比，该方法只使用了全光谱变量（1 557 个变量）中的 222 个变量建立模型，模型更简洁，且避免利用小光程来减少水的强吸收影响所带来的缺陷。

**参考文献：**

[1] Bettuzzi S，Brausi M，Rizzi F. Chemoprevention of human prostate cancer by oral administration of green tea catechins in volunteers with high – grade prostate intraepithelial neoplasia：a preliminary report from a one – year proof – of – principle study [J]. Cancer Research，2006，66(2)：1234 – 1240.

[2] 艾施荣，吴瑞梅，吴燕. 基于神经网络近红外光谱鉴别茶饮料的研究 [J]. 安徽农业科学，2010，38(14)：7658 – 7662.

[3] 谷晓君. 来自茶饮料的困惑 [J]. 上海标准化，2004，5：36 – 38.

[4] 郭志明. 近红外光谱法测定茶叶中游离氨基酸的研究 [J]. 光谱仪器与分析，2011(1)：105 – 109.

[5] 张正竹. 茶叶生物化学实验教程 [M]. 北京：中国农业出版社，2009：42 – 43.

[6] 郭升平. 高效液相色谱法测定茶叶中氨基酸的研究 [J]. 色谱，1996，14(6)：464 – 466.

[7] Yu H Y，Ying Y B，Fu X P，et al. Quality determination of Chinese rice wine based on Fourier transform near infrared spectroscopy [J]. Journal of Near Infrared Spectroscopy，2006，14(1)：37 – 44.

[8] Cozzolino D，Kwiatlowski M J，Waters E J，et al. A feasibility study on the use of visible and short wavelengths in the near – infrared region for the nondestructive measurement of wine composition [J]. Analytical and Bioanalytical Chemistry，2007，

387(6):2289 – 2295.

[9] Galtier N, Dupuy Y, Le D, et al. Geographic origins and compositions of virgin olive oils determinated by chemometric analysis of NIR spectra [J]. Analytica Chimica Acta,2007,595(1):136 – 144.

[10] 姜礼义,刘福莉,陈华才,等. 绿茶汤中茶多酚近红外定量分析的光程选择 [J]. 中国计量学院学报,2009,20(2):135 – 138.

[11] 吴瑞梅,赵杰文,陈全胜,等. 特征变量筛选在绿茶汤中茶多酚近红外光谱定量分析中的应用 [J]. 农业机械学报,2011,42(12):173 – 176.

[12] Norgaard L, Saudland A, Wagner J, et al. Interval partial least – squares regression (iPLS): a comparative chemometric study with an example from near – infrared spectroscopy [J]. Applied Spectroscopy, 2000,54(3):413 – 419.

[13] Chen Q S, Zhao J W, Liu M H, et al. Determination of total polyphenols content in green tea using FT – NIR spectroscopy and different PLS algorithms [J]. Journal of Pharmaceutical and Biomedical Analysis,2008,46(3):568 – 573.

[14] 朱向荣,李娜,史新元,等. 近红外光谱与组合的间隔偏最小二乘法测定清开灵四混液中总氮和栀子苷的含量 [J]. 高等学校化学学报,2008,29(5):906 – 911.

[15] Chen Q S, Zhao J W, Huang X Y, et al. Simultaneous determination of total polyphenols and caffeine contents of green tea by near – infrared reflectance spectroscopy [J]. Microchemical Journal,2006,83(1):42 – 47.

[16] Liu F, He Y, Wang L, et al. Detection of organic acids and pH of fruit vinegars using near – infrared spectroscopy and multivariate calibration [J]. Food and Bioprocess Technology, 2011,4(8):1331 – 1340.

(上接第 1007 页)

[10] Smith L W, Goering H K, Gordon C H. Relationships of forage compositions with rates of cell wall digestion and indigestibility of cell walls [J]. Journal of Dairy Science, 1972,55(2):1140 – 1147.

[11] Pichard D G, Van Soest P J. Protein solubility of ruminant feeds [M]. Ithaca N Y: Proceedings of Cornell Nutrition Conference,1977:91.

[12] 张吉鹃,卢德勋,李龙瑞,等. 浅析粗饲料品质评定指数及其模型 [J]. 畜牧与兽医,2004,36(4):23 – 25.

[13] 张吉鹃,黄光明,邹庆华,等. 几种奶牛用粗饲料品质的综合评定研究 [J]. 饲料工业,2008,29(21): 34 – 37.

[14] 张吉鹃,邹庆华,王金芬,等. 稻草添补百脉根瘤胃体外发酵及微生物蛋白合成的组合效应研究 [J]. 江西农业大学学报,2011,33(5):942 – 948.

[15] 张吉鹃,李龙瑞,邹庆华. 稻草与不同饲料混合在体外消化率上的组合效应研究 [J]. 草业科学,2010,27(11):137 – 144.

[16] Zhang Ji-kun, Chen Kai-wen, Xie Jin-fang, et al. A comparison of grading index and relative feed value in forage quality evaluation [J]. China Feed Industry, 2008,1(2):30 – 32.

[17] 韦升,杨纯,邹彩霞,等. 应用体外产气法评定广西区内豆腐渣、木薯渣、啤酒糟的营养价值 [J]. 饲料工业,2011,32(7):46 – 48.

[18] 曲永利,吴健豪,李铁. 应用康奈尔净碳水化合物 – 蛋白质体系评定东北农区奶牛饲料营养价值 [J]. 动物营养学报2010,22(1):201 – 206.

[19] 赵广永. 用净碳水化合物—蛋白质体系评定反自动物饲料营养价值 [J]. 中国农业大学学报, 1994,4(增刊):71 – 76.

[20] 郭冬生,彭小兰. 用 CNCPS 方法评定反刍动物常用饲料的营养价值 [J]. 粮食与饲料工业,2010(10):41 – 42.

[21] 张吉鹃,谢金防,肖海红,等. 分级指数与相对值在奶牛用粗饲料品质评定上的比较研究 [J]. 中国奶牛,2008(8):15 – 19.

[22] 张吉鹃,卢德勋,刘建新,等. 粗饲料品质评定指数的研究现状及其进展 [J]. 草业科学,2004,21(9):55 – 61.

[23] 李威,高民,卢德勋,等. CNCPS 与 NRC 在反刍动物方面的分析比较及其研究进展 [J]. 饲料工业,2008,29(13):45 – 48.

[24] 张吉鹃. 粗饲料品质评定指数——产奶二千 [J]. 江西饲料,2005(6):20 – 25.

[25] 李威. 利用 CNCPS 和 GI 进行乳牛日粮优化设计及其应用效果研究 [D]. 呼和浩特: 内蒙古农业大学,2009.

[26] 张吉鹃. 饲料间的组合效应及其在配方设计中的应用 [J]. 草业科学,2009,26(12):113 – 117.