Hot Topic: Application of Support Vector Machine Method in Prediction of Alfalfa Protein Fractions by Near Infrared Reflectance Spectroscopy

Z. Nie,* J. Han,*¹ T. Liu,[†] and X. Liu[‡]

*Department of Grassland Science, College of Animal Science and Technology, China Agricultural University, Beijing 10094, China

[†]Beijing Petrochemical Design Institute, Beijing 100101, China

[‡]Department of Chemistry, College of Sciences, Shanghai University, Shanghai 200444, China

ABSTRACT

The object of this study was to explore the potential for support vector machine (SVM) to improve the precision of predicting protein fractions by near infrared reflectance spectroscopy (NIRS). Generally, most protein fractions determined in Cornell Net Carbohydrate and Protein System (CNCPS), especially the neutral detergent insoluble protein (NDFCP) and acid detergent insoluble protein (ADFCP), could not be accurately predicted by the commonly used partial least squares (PLS) method. A recently developed chemometric method, SVM, was applied in NIRS prediction of alfalfa protein fractions in this study. Two hundred thirty alfalfa samples were scanned on a near infrared reflectance spectrophotometer, and analyzed for crude protein (CP), true protein precipitated in tungstic acid (TCP), borate-phosphate buffer-insoluble protein (BICP), NDFCP, and ADFCP. These 5 laboratory proteins and the CNCPS protein fractions A, B1, B2, B3, and C were predicted by NIRS using the PLS and SVM methods. According to PLS-NIRS regression, CP, TCP, BICP, A, and B2 obtained the determination coefficient of prediction (R_p^2) of 0.96, 0.91, 0.94, 0.94, and 0.93, and the ratios of standard deviation of prediction samples:standard error of prediction samples (RPD) values were 5.07, 3.31, 3.98, 3.96, and 3.91. Neutral detergent insoluble protein, ADFCP (fraction C), B1, and B3 were predicted with R_p^2 of 0.75, 0.83, 0.30, and 0.62, and RPD values of 1.98, 2.42, 1.20, and 1.62; Calibrated by the SVM-NIRS method, R_p^2 values of CP, TCP, BICP, NDFCP, ADFCP(C), A, and B2 achieved 0.99, 0.97, 0.97, 0.90, 0.93, 0.97, and 0.97, respectively. The RPD values of those fractions were 8.68, 8.26, 6.11, 3.08, 3.69, 5.97, and 5.81, respectively. The R_p^2 and RPD values of fractions B1 and B3 were 2.67 and 0.87 (B1) and 2.51 and 0.75 (B3) directly predicted by SVM-NIRS

model. In this study, the chemical analysis results of B1 and B3 were also correlated with calculated results from TCP-BICP and NDFCP-ADFCP, which were predicted by SVM-NIRS models. The B1 protein fraction achieved R_p^2 and RPD values of 0.87 and 3.61, whereas values for B2 were 0.75 and 2.00. Data suggested that use of SVM methods in NIRS technology could improve the accuracy of predicting protein fractions. This study showed the potential of increasing the NIRS prediction accuracy to a level of practical use for all protein fractions, except B3.

Key words: support vector machine, protein fraction, alfalfa, near infrared reflectance spectroscopy

INTRODUCTION

Forage protein nutritive value is determined by the content of total CP and protein degradability (Blaxter, 1956). Furthermore, ruminal degradation of feed protein depends upon the rate of degradation of different individual protein fractions. Many studies (Knowlton et al., 1992; Fox et al., 1995) have proposed further subdivision of protein fractions. The Cornell Net Carbohydrate and Protein System (CNCPS) developed new models to separate the feed protein fractions based on the difference in the rate and extent of ruminal protein degradability. In the CNCPS, Sniffen et al. (1992) described 5 protein fractions: NPN is denoted as fraction A; true protein is broken down into fractions B1, B2, and B3, based on different decreasing solubility (rapid, intermediate, and slow, respectively); and fraction C represents the unavailable true protein. The 5 fractions are dependent upon the estimation of insoluble protein, true protein, and the protein residual in NDF and ADF. However, determination of the protein fractions in CNCPS is a tedious process because the feed samples need to be treated with tungstic acid (or trichloroacetic acid), borate-phosphate buffer, neutral detergent or acid detergent solutions, and 5 Kjeldahl procedures (Krishnamoorthy et al., 1982; Licitra et al., 1996). The analysis takes 2 d or more. Compared with these conventional laboratory procedures, near infrared re-

Received January 1, 2008.

Accepted March 10, 2008.

¹Corresponding author: jianguoan2058@126.com

flectance spectroscopy (**NIRS**) offers advantages of high efficiency and low cost. It has been used widely and successfully in evaluation of forage and feedstuff nutrition (Norris et al., 1976; Shenk et al., 1981; Abrams et al., 1987; Givens et al., 1997).

The analysis of protein fractions with NIRS relies on both the ability of laboratory procedures to precisely measure protein fractions of interest and the significant correlation between particular amide bonds of protein fractions and sample spectra. Some protein fractions, however, cannot be accurately predicted by NIRS technology because their detection is affected by interactions between N and cell wall constituents or carbohydrates. Therefore, it is necessary to apply chemometric methods to analyze and extract the particular information. Classical chemometric methods such as partial least squares (**PLS**) are widely used in classification and regression analysis. There has been an increasing awareness of the potential of using a more recent technique, support vector machine (**SVM**), in NIRS analysis. Support vector machine, proposed by Vapnik (1998), is based on the statistical learning theory of structural risk minimization in Vapnik-Chervonenkis dimension. Support vector machines map input vectors into a higher dimensional space where maximal separating hyperplanes are constructed; the generalisation error decreases as the margin or distance between separated hyperplanes increases. It offers advantages over conventional statistical learning algorithms: 1) high generalization performance even with high dimension feature vectors; 2) the ability to manage kernel functions that map input data to higher dimensional space without increasing computational complexity; 3) better performance in dealing with nonlinear data. Compared with artificial neural network, which uses the empirical risk minimization principle, SVM is superior in avoiding overfitting and multidimensional problems, especially in a small sample set. According to the literature, SVM has been successfully applied to many applications such as drug design and combinatorial chemistry, but there are few reports in the animal and feed science area. Fernández Pierna et al. (2004) studied the SVM-NIRS measurement of meat and bone in compound feeds, and Wu et al. (2007) reported on the successful prediction of the content in milk powder by infrared spectroscopy.

At present, estimation of CNCPS protein fractions of forages and TMR by NIRS has been studied (Hoffman et al., 1999; Mentink et al., 2006; Valdés et al., 2006). However, most of these studies could only predict part of these fractions with an acceptable accuracy. As far as we know, there is no research showing all of the protein fractions predicted accurately in the same study. On the other hand, no research has described

the accuracy of NIRS prediction of protein fractions by SVM method.

To explore the potential for SVM regression (**SVR**) method to improve the precision of predicting protein fractions by NIRS technology, SVM was applied in NIRS prediction of protein fractions in the study. Moreover, PLS models were calibrated to compare with SVM-NIRS models.

MATERIALS AND METHODS

Sample Preparation and Laboratory Analyses

Alfalfa samples ($n = 230$) were collected from 5 sites of experimental and commercial farms in Gansu Province (northwest China, 39.81°N, 97.78°E to 39.14°N, 99.84°E) during 2005 and 2006. The detailed information of sample origins is shown in Table 1. Samples included 4 varieties, multiple maturity stages, and 3 harvests. Approximately 2 kg of fresh alfalfa sample was collected at 50 mm above the soil level and then oven-dried at 65°C for 48 h. Dry samples were ground in a cyclone mill (Cyclotec 1093, Foss, Hillerød, Denmark) fitted with a 1-mm screen for chemical analyses and NIRS scan.

According to Licitra et al. (1996), 4 steps are used to separate CP. First, the sample is treated with tungstic acid to precipitate out true protein (**TCP**). Fraction A is $\text{NPN} \times 6.25$ and is calculated by subtracting TCP from total CP. Second, samples are treated with borate-phosphate buffer solution. Fraction B1 is buffer-soluble true protein and calculated by TCP minus buffer-insoluble protein (**BICP**). Third, the sample is refluxed in neutral detergent solution to separate sample protein; B2 is the part dissolved in neutral detergent solution and calculated by BICP minus neutral detergent-insoluble protein (**NDFCP**). In the last step, residual NDFCP of sample is put in acid detergent solution and fraction B3 is acid detergent-soluble protein, calculated by NDFCP minus acid detergent-insoluble protein (**ADFCP**); fraction C is defined as ADFCP.

Samples were analyzed for DM at 105°C for 3 h. Total CP and the protein of the other 4 precipitated parts (TCP, BICP, NDFCP, and ADFCP) were determined by the Kjeldahl method described by AOAC (2000). The NDF and ADF concentrations were determined according to Van Soest et al. (1991). All of the wet chemistry determinations were made in 3 replications.

NIRS Analysis. Alfalfa samples were scanned on a Nicolet Antaris FT-Near Infrared Analyzer (Thermo Electron Corp., Madison WI) equipped with an integrating sphere. Samples were put in a quartz cup (48-mm inner diameter) and subsequently scanned 64 times with 3 replications over 1,000 to 2,500 nm at the resolution of 8 cm^{-1} . All spectra data were recorded as

Table 1. Description of the origin of alfalfa samples used in the present study

Origin	n ¹	Year	Harvest	Variety	Location
Experimental farm Jiuquan	60	2005	1	Algonquin	39.71°N 98.50°E
Commercial farm Yumen	6	2005	1	Zhongmu No. 1	39.81°N 97.58°E
Commercial farm Jinta	6	2006	2	Gannong No. 3	39.97°N 98.92°E
Commercial farm Gaotai	6	2006	3	Derby	39.14°N 99.84°E
Experimental farm Jiuquan	74	2006	1-3	Algonquin	39.71°N 98.50°E
Experimental farm Yumen	74	2006	1-3	Zhongmu No. 1	39.81°N 97.58°E
Commercial farm Jiuquan	4	2006	2	Algonquin	39.71°N 98.50°E

¹n = number of samples.

log (1/R) and a mean spectrum was calculated for each sample (Nie et al., 2007). Mathematical processing was performed using TQ Analyst software (Thermo Nicolet Corp., Madison, WI). All 230 samples were randomly divided into 2 groups for calibration (n = 150) and prediction (n = 80).

PLS Regression. The PLS regression was performed using TQ Analyst software. According to the Diagnostics module of the TQ Analyst software, parameters in mathematical processing were sought through trials to optimize the models. Main parameters in this software involved data format with the options of original spectrum, first derivative, and second derivative spectrum. The subsequent option of derivative was the filter to smooth data (Nie et al., 2007). Norris derivative (ND) filter was applied to smooth the samples spectrum in this study and it was followed with options of segment length and gap between segments, which were the main selected parameters in this study.

All parameters for a new model were tested using leave-one-out cross-validation and best model should be selected with the minimum root mean square error of cross-validation (**RMSECV**). In addition, the ratio of standard error of performance to the standard deviation of the reference data (**RPD**), which is calculated as standard deviation of prediction sets (**SD**)/standard error of prediction (**SEP**), was considered as a crucial criterion of the NIRS equations (Williams and Norris, 2001). Root mean SEP was calculated from (Naes et al., 2002):

$$\text{RMSEP} = \sqrt{1/N \sum (x_i - y_i)^2}.$$

Standard error of prediction was calculated by the following formula (Naes et al., 2002):

$$\text{SEP} = \sqrt{1/N - 1 \sum (x_i - y_i - \text{Bias})^2},$$

where $x_i - y_i$ = difference between results obtained by routine method (x_i) and reference method (y_i) on sample I , and bias is given by

$$\text{Bias} = 1/N \sum (x_i - y_i)$$

where N = total number of samples in the test. The root mean square error of calibration (**RMSEC**) and **RMSECV** were obtained from calibration and cross-validation data sample sets, respectively, and calculated in the same equation as **RMSEP** (Naes et al., 2002). The repeatability standard deviation (**Sr**) of the reference methods (laboratory procedures) was calculated to provide the laboratory determination error in reference methods. A lower **Sr** value indicates a more precise laboratory measurement. The range (maximum minus minimum) to **Sr** ratio (**RSrR**) value was another useful indicator of the probability of successful NIRS calibration and performance, according to Williams (1987).

SVM Regression. According to Vapnik (1998), the basic idea of SVM method is to map the data into a higher dimensional feature space via nonlinear mapping (Φ) and then to do regression in this space. Therefore, regression approximation addresses the problem of estimating a function based on a given data set $G = \{(\mathbf{x}_i, d_i)\}_{i=1}^l$, where $\mathbf{x}_i$ is the input vector and d_i is the desired value; SVM approximates the function in the following form:

$$y = \sum_{i=1}^l w_i \Phi_i(\mathbf{x}) + b \quad [1]$$

where $\{\Phi_i(\mathbf{x})\}_{i=1}^l$ is the set of mappings of input features, and $\{w_i\}_{i=1}^l$ and b are coefficients. They are estimated by minimizing the regularized risk function $R(C)$:

$$R(C) = C \frac{1}{N} \sum_{i=1}^N L_\varepsilon(d_i, y_i) + \frac{1}{2} \|\mathbf{w}\|^2 \quad [2]$$

where

$$L_\varepsilon(d, y) = \begin{cases} |d - y| - \varepsilon & \text{for } |d - y| \geq \varepsilon \\ 0 & \text{otherwise} \end{cases} \quad [3]$$

and ε is a prescribed parameter. In Equation [2], $C \frac{1}{N} \sum_{i=1}^N L_\varepsilon(d_i, y_i)$ is the so-called empirical error (risk), which

is measured by ε -insensitive loss function $L_\varepsilon(d, y)$. This equation indicates that C does not penalize errors below ε . The second term, $\frac{1}{2}\|\mathbf{w}\|^2$, is used as a measurement of function flatness. C is a regularized constant determining the tradeoff between the calibration error and the model flatness. Introduction of slack variables “ ξ ” leads Equation [2] to the following constrained function:

$$\text{Max } R(\mathbf{w}, \xi^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C^* \sum_{i=1}^n (\xi_i + \xi_i^*) \quad [4]$$

$$\text{subject to } w\Phi(x_i) + b - d_i \leq \varepsilon + \xi_i$$

$$d_i - w\Phi(x_i) - b_i \leq \varepsilon + \xi_i \quad [5]$$

$$\xi_i, \xi_i^* \geq 0.$$

Thus, decision function Equation [1] becomes the following form:

$$f(\mathbf{x}, \alpha_i, \alpha_i^*) = \sum_{i=1}^l (\alpha_i - \alpha_i^*)K(\mathbf{x}, \mathbf{x}_i) + b. \quad [6]$$

In Equation [6], α_i, α_i^* are the introduced Lagrange multipliers. They satisfy the equality $\alpha_i \cdot \alpha_i^* = 0$, $\alpha_i \geq 0$, $\alpha_i^* \geq 0$; $i = 1, \dots, l$, and are obtained by maximizing the dual form of Equation [4] which has the following form:

$$\begin{aligned} \Phi(\alpha_i, \alpha_i^*) &= \sum_{i=1}^l d_i(\alpha_i - \alpha_i^*) - \varepsilon \sum_{i=1}^l (\alpha_i + \alpha_i^*) \\ &\quad - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\mathbf{x}_i, \mathbf{x}_j) \end{aligned} \quad [7]$$

with the following constraints:

$$\begin{aligned} 0 &\leq \alpha_i \leq C \quad i = 1, \dots, l \\ 0 &\leq \alpha_i^* \leq C \quad i = 1, \dots, l \\ \sum_{i=1}^l (\alpha_i - \alpha_i^*) &= 0. \end{aligned} \quad [8]$$

Based on the Karush-Kuhn-Tucker conditions, only several coefficients $(\alpha_i - \alpha_i^*)$ will assume nonzero values, and the data points associated with them could be referred to as support vectors. In Equation [6], $K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function. The value is equal to the inner product of 2 vectors $\Phi(\mathbf{x}_i)$ and $\Phi(\mathbf{x}_j)$ in the feature space $\Phi(\mathbf{x})$. That is, $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$. The elegance of using kernel function lies in the fact that one can deal with feature spaces of arbitrary dimensionality without

having to compute the map $\Phi(\mathbf{x})$. Any function that satisfies Mercer's condition can be used as kernel function.

In the work, the SVM regression was carried out on the MASTER software package (Beijing Petrochemical Design Institute, Beijing, China). Best models and the relative parameters were also selected by minimum RMSECV and RPD. In the MASTER software, parameters C and ε were optimized in 4 types of kernel functions, which include linear, polynomial, radial basis function (RBF), and S-type kernel function. For parameter selection, C values ranged from 1 to 50 with incremental steps of 1, and ε values ranged from 0.01 to 0.1 with incremental steps of 0.01 (Gu et al., 2006; Liu et al., 2006).

RESULTS AND DISCUSSION

Alfalfa Protein Fractions

Results of analysis of laboratory and CNCPS protein fractions of alfalfa samples in the calibration and validation sets are presented in Table 2. Calibration and prediction sets were similar in range, average, and SD for all variables. Crude protein, TCP, BICP, and CNCPS fractions A and B2 had wider range of contents, higher average value (>6%) and SD (>10%) than other protein fractions. From the error statistics of laboratory protein fractions, Table 2 shows the lower Sr and higher RSrR values compared with the report of TMR in Mentink et al. (2006). Precise wet chemistry measurements increase the probability of successful NIRS calibration and prediction (Williams, 1987). In the current study, the NDFCP and ADFCP had lower Sr than the other 3 laboratory protein fractions, but the RSrR values were similar to those of other protein fractions, which ranged from 126.11 to 178.75.

PLS-NIRS Models of Protein Fractions. According to Table 3, the PLS-NIRS equations indicated the most outstanding ability of prediction of CP content of alfalfa hays ($R_p^2 = 0.96$, RPD = 5.07). The laboratory protein fraction TCP achieved R_p^2 and RPD values of 0.91 and 3.31, which showed the practicability of NIRS model. Borate-phosphate buffer-insoluble protein was estimated accurately by PLS regression, with the highest R_p^2 and RPD values (0.94, 3.98) other than those for CP in this study. Prediction of NDFCP for alfalfa hays by NIRS was considerably less accurate ($R_p^2 = 0.75$, RPD = 1.98). The last laboratory protein fraction ADFCP, which is also the CNCPS fraction C, was estimated with medium accuracy by NIRS ($R_p^2 = 0.83$, RPD = 2.42) in this study. According to Williams and Norris (2001), NIRS equations with RPD <3 are not

Table 2. Statistics¹ of chemical analysis for laboratory and Cornell Net Carbohydrate and Protein System protein fractions in calibration (Cal.) and prediction (Pred.) samples sets

Fraction, ² % of DM	Set	n ³	Average	Minimum	Maximum	SD	Sr	RSrR
CP	Cal.	150	17.79	11.46	25.47	3.89	0.08	178.75
	Pred.	80	17.69	11.17	24.60	3.93		
TCP	Cal.	150	10.85	6.62	17.97	2.29	0.09	126.11
	Pred.	80	10.80	6.99	16.61	2.06		
BICP	Cal.	150	8.15	4.78	13.29	1.87	0.05	170.20
	Pred.	80	8.12	5.21	12.01	1.74		
NDFCP	Cal.	150	1.31	0.66	2.18	0.30	0.01	170.33
	Pred.	80	1.28	0.63	1.87	0.31		
ADFCP (C)	Cal.	150	1.08	0.66	1.55	0.21	0.01	139.44
	Pred.	80	1.05	0.56	1.51	0.23		
A	Cal.	150	6.94	2.02	12.11	2.56		
	Pred.	80	6.90	2.35	12.83	2.61		
B1	Cal.	150	2.70	0.38	5.05	0.88		
	Pred.	80	2.68	1.08	4.60	0.86		
B2	Cal.	150	6.84	3.09	11.83	1.91		
	Pred.	80	6.84	3.58	10.28	1.79		
B3	Cal.	150	0.23	0.00	0.97	0.20		
	Pred.	80	0.23	0.01	0.92	0.18		

¹SD = standard deviation of samples set; Sr = repeatability standard deviation of the reference method of laboratory protein fractions; RSrR = (maximum – minimum)/Sr.

²TCP = true CP; BICP = buffer-insoluble CP; NDFCP = neutral detergent insoluble CP; ADFCP (C) = acid detergent insoluble CP or protein fraction C.

³n = number of samples.

recommended in application. As a result, the ADFCP (C) NIRS equation still could not be taken to practical use. The PLS-NIRS equations of CNCPS fraction A and B2 were developed accurately with better R_p^2 (0.94, 0.93) and RPD values (3.96, 3.91). Fractions B1 and B3 were estimated less accurately by PLS-NIRS in this study ($R_p^2 < 0.62$, RPD < 1.62).

On the whole, CP and laboratory protein fractions TCP, BICP, and CNCPS fractions A and B1 could be

predicted with an acceptable degree of accuracy (RPD > 3), whereas NDFCP, ADFCP, B1, and B3 could not be estimated accurately by PLS-NIRS in this study. Based on RPD values, we deduced the ranking of accuracy of protein fractions predicted by PLS-NIRS equations to be CP > BICP > A > B2 > TCP > ADFCP (C) > NDFCP > B3 > B1.

SVM-NIRS Models of Protein Fractions. It is important for SVM regression to choose the correct num-

Table 3. Statistics¹ for partial least squares-near infrared reflectance spectroscopy model analysis of laboratory and Cornell Net Carbohydrate and Protein System protein fractions

Fraction, ² % of DM	Parameters ³	R_{CAL}^2	RMSEC	R_{CV}^2	RMSECV	R_p^2	RMSEP	RPD
CP	2, 13, 1	0.99	0.46	0.96	0.75	0.96	0.78	5.07
TCP	2, 15, 5	0.95	0.49	0.92	0.64	0.91	0.63	3.31
BICP	2, 15, 5	0.95	0.41	0.92	0.52	0.94	0.44	3.98
NDFCP	2, 19, 1	0.79	0.14	0.68	0.17	0.75	0.16	1.98
ADFCP (C)	2, 19, 1	0.85	0.08	0.81	0.10	0.83	0.09	2.42
A	2, 17, 2	0.95	0.54	0.94	0.63	0.94	0.66	3.96
B1	2, 19, 1	0.45	0.65	0.41	0.69	0.30	0.72	1.20
B2	2, 15, 1	0.95	0.43	0.94	0.48	0.93	0.46	3.91
B3	2, 15, 1	0.64	0.12	0.32	0.20	0.62	0.12	1.62

¹ R_{CAL}^2 = coefficient of determination of calibration; RMSEC = root mean square error of calibration; R_{CV}^2 = coefficient of determination of cross-validation; RMSECV = root mean square error of cross-validation; R_p^2 = coefficient of determination of prediction; RMSEP = root mean square error of prediction; RPD = standard deviation of prediction sets/standard error of prediction.

²TCP = true CP; BICP = buffer-insoluble CP; NDFCP = neutral detergent insoluble CP; ADFCP (C) = acid detergent insoluble CP or protein fraction C.

³Parameters: parameters in mathematical processing, where the first number is the data format (order of derivative), and the second and third numbers are segment length and gap between segment in spectrum data smooth, respectively.

Figure 1. Root mean square error of cross-validation (RMSECV) of different numbers of latent variables in support vector machine-near infrared reflectance spectroscopy model of fraction A.

ber of latent variables. In this research, optimal latent variables of these protein fractions were selected according to the minimum RMSECV. Taking into account all of the protein fractions, the number of latent variables of SVM-NIRS models in this study was 15, which meant that latent variables 1 to 15 were used in SVM regression. For instance, Figure 1 showed the values of the RMSECV using different numbers (1 to 15) of latent variables of fraction A.

After choosing the correct number of latent variables, leave-one-out cross-validation was carried out to select the proper kernel functions by lowest RMSECV between various parameter combinations and 4 types of kernel functions. The best modeling modulus used in SVR-NIRS models of all these protein fractions are shown in Table 4. All protein fractions obtained lowest RMSECV from linear kernel function. All of the optimized C values were less than 10 and the ϵ values were below 0.05. Compared with the PLS models in Table 3, the SVM-NIRS models in this study strongly decreased

the RMSECV value, which is an obvious signal of accuracy improvement of SVM-NIRS models.

The statistical result of SVM-NIRS model calibration and prediction are shown in Table 5. The R_p^2 and RPD values were 0.90 and 3.08 for NDFCP and 0.93 and 3.69 for ADFCP (C), respectively. The CP, TCP, BICP, A, and B2 fractions achieved greater RPD values ranging from 5.81 to 8.68. According to Williams and Norris (2001), NIRS equations with RPD values greater than 3 indicate efficient NIRS predictions. The model can be used in quality control and even process control when the RPD value is greater than 5. Therefore, the SVM-NIRS models of CP and all protein fractions except B1 and B3 achieved sufficient accuracy for relative research and practical applications. The RPD values of B1 and B3 were lower than 3 and they were not recommended for practical use.

The calibration and prediction of SVM-NIRS models were improved to varying extents compared with PLS models. The accuracy of the SVM model of CP increased with RMSEP = 0.60% and RPD = 8.68 (Table 5), compared with 0.78% and 5.07 with the PLS model (Table 3). This result was similar to the reports in many other studies (Hoffman et al., 1999; Andres et al., 2005; Mentink et al., 2006; Valdés et al., 2006). The reason for the accurate prediction of CP content by NIRS technology is the highly significant correlation between the absorption of amide bonds, involving protein content analyzed by the Kjeldahl method (Shenk and Westerhaus, 1994).

All of the SVM-NIRS models of protein fractions were improved, because the coefficient of determination, root mean square error, and the RPD values of their SVM models were all better than PLS models. Between those fractions, TCP, BICP, A, and B2 performed much better with R_p^2 of 0.97, 0.97, 0.97, and 0.97, and RPD of 8.26, 6.11, 5.97, and 5.81, respectively. In contrast, the PLS-NIRS model R_p^2 values of those fractions in this study

Table 4. Parameters¹ used in support vector machine-near infrared reflectance spectroscopy models of protein fractions

Fraction, ² % of DM	Kernel function	C	ϵ	RMSECV ³
CP	Linear kernel function	7	0.02	0.64
TCP	Linear kernel function	1	0.03	0.39
BICP	Linear kernel function	1	0.01	0.31
NDFCP	Linear kernel function	10	0.03	0.13
ADFCP (C)	Linear kernel function	1	0.05	0.05
A	Linear kernel function	1	0.01	0.37
B1	Linear kernel function	2	0.04	0.35
B2	Linear kernel function	1	0.01	0.33
B3	Linear kernel function	5	0.04	0.12

¹C = penalty factor; ϵ = insensitive factor.

²TCP = true CP; BICP = buffer-insoluble CP; NDFCP = neutral detergent insoluble CP; ADFCP (C) = acid detergent insoluble CP or protein fraction C.

³RMSECV = root mean square error of cross validation.

Table 5. Statistics¹ for calibration and prediction of support vector machine-near infrared reflectance spectroscopy models of laboratory and the Cornell Net Carbohydrate and Protein System protein fractions

Fraction, ² % of DM	n ³	R _{CAL} ²	RMSEC	R _p ²	RMSEP	RPD
CP	230	0.99	0.58	0.99	0.60	8.68
TCP	230	0.99	0.27	0.97	0.35	8.26
BICP	230	0.98	0.29	0.97	0.31	6.11
NDFCP	230	0.90	0.11	0.90	0.10	3.08
ADFCP (C)	230	0.93	0.06	0.93	0.06	3.69
A	230	0.98	0.31	0.97	0.33	5.97
B1	230	0.86	0.31	0.83	0.36	2.67
B2	230	0.98	0.29	0.97	0.35	5.81
B3	230	0.84	0.10	0.73	0.08	2.51

¹R_{CAL}² = coefficient of determination of calibration; RMSEC = root mean square error of calibration; R_p² = coefficient of determination of prediction; RMSEP = root mean square error of prediction; RPD = standard deviation of prediction sets/standard error of prediction.

²TCP = true CP; BICP = buffer-insoluble CP; NDFCP = neutral detergent insoluble CP; ADFCP (C) = acid detergent insoluble CP (protein fraction C).

³n = number of samples.

were between 0.91 and 0.94, and the RPD values ranged from 3.31 to 3.98. To our knowledge, the only related report in Valdés et al. (2006) showed that fractions BICP, A, and B2 could not be accurately predicted in botanically heterogeneous permanent meadows by PLS-NIRS, because of the low R_p² values of 0.66, 0.42, and 0.00, and RPD values of 1.23, 1.79, and 0.96, respectively.

It is important to point out that the SVM-NIRS model accuracy of NDFCP and ADFCP (C) were increased to a practical extent, with R_p² and RPD values of 0.90 and 3.08 and 0.93 and 3.69 for NDFCP and ADFCP (C), respectively (Table 5). In PLS models in this study, the R_p² and RPD values were 0.75 and 1.98 for NDFCP and 0.83 and 2.42 for ADFCP (Table 3). In other studies, predictions of NDFCP and ADFCP (C) by NIRS were also less accurate. Hoffman et al. (1999) observed an unsuccessful NDFCP and ADFCP PLS-NIRS model for legume and grass silage with low R_{CAL}² (0.72, 0.77) and R_p² (0.84, 0.42) values. Valdés et al. (2006) obtained an accurate NDFCP PLS-NIRS equation with R_p² and RPD values of 0.91 and 3.33, but the PLS-NIRS model had low accuracy for ADFCP, with R_p² and RPD values of 0.72 and 2.14.

According to Tables 3 and 5, the SVM-NIRS model accuracy of fraction B1 and B3 were also improved, in contrast with PLS models in this study and in Valdés et al. (2006). However, with R_p² values of 0.83 and 0.73 and RPD values of 2.67 and 2.51, 2 SVR-NIRS equations were still not practicable.

Because protein fractions B1 and B3 were calculated from the laboratory fractions TCP, BICP, NDFCP, and ADFCP, which were accurately predicted by SVR-NIRS equations in this study, they were recalculated by 4

laboratory fractions from the SVR-NIRS predicted data and correlated with original calculation results from laboratory-analyzed data. Table 6 shows the statistics of 2 types of calculation results. It was obvious that B1 achieved better performance from the calculation based on SVM-NIRS-predicted data, compared with the directly predicted results by SVR-NIRS equations. The R² of prediction increased from 0.83 to 0.87, and the RPD value reached 3.61, making it adequate for practical use. Meanwhile, the new calculation result of B3 did not improve and actually decreased, with R² = 0.75 and RPD = 2.00.

Generally, high accuracy of NIRS prediction has 3 basic requirements: 1) precise measurement of the laboratory procedures; 2) exact quality and quantity analyses for the characteristic chemical bonds (such as nitrogen). Because detection of character information by NIRS could be strongly interfered with difference between similar chemical bonds or other intense absorbance bonds; and 3) a certain content and range of nutrient values over the NIRS limit of determination (0.1%). According to the error statistics in Table 2, the Sr values of all CP and laboratory protein fractions

Table 6. Correlation statistics¹ between the directly predicted results by support vector machine-near infrared reflectance spectroscopy (SVM-NIRS) and indirectly calculated results from other laboratory fraction constituents based on SVM-NIRS prediction in the prediction sample set

Fraction	R ²	RMSE	RPD
B1	0.87	0.25	3.61
B3	0.75	0.10	2.00

¹R² = coefficient of determination; RMSE = root mean square error of prediction; RPD = standard deviation of prediction sets/standard error of prediction.

ranged from 0.01 to 0.09%, it showed the qualified wet chemistry determinations of those fractions in this study, which showed that the wet chemistry determinations of those fractions were done accurately in the present study. Therefore, errors of wet chemistry determinations were not the main restricted reason for different accuracy of NIRS equations. However, because of the low range expressed by the low SD value (<1%) and low content of NDFCP and ADFCP (C), those 2 fractions did not derive precise NIRS equations by PLS regression, although the RSrR values of NDFCP (170.33) and ADFCP (C) (139.44) were similar to those of other protein fractions. Results in Valdés et al. (2006) could be helpful for explaining our results, because they obtained an accurate NDFCP NIRS equation with mean and SD values of 12.3 and 2.26%, while gaining an inaccurate equation for ADFCP with lower mean and SD values of 4.1 and 1.44%. The CNCPS fractions were calculated from laboratory protein fractions, so the errors, ranges, and contents of CNCPS protein fractions were affected by the relative laboratory protein fractions. As a result, the performance of NIRS prediction of CNCPS protein fractions was directly related to the accuracy of NIRS equations of the certain laboratory protein fractions. Therefore, protein fractions A and B2, which were calculated from CP – TCP and BICP – NDIP, obtained more precise NIRS equations, whereas PLS-NIRS prediction of fractions B1 and B3 were not accurate with lower range, SD, and average values.

Support vector machine regression improved the accuracy of NIRS prediction of all protein fractions based on the structural risk minimization statistical learning theory. It is more significant that performance of protein fractions NDFCP, ADFCP (C), and B1 were increased to an accurate level with $R^2 > 0.87$ and $RPD > 3.0$ by direct and indirect NIRS prediction. This mainly resulted from the better performance of SVM regression when treating nonlinear data such as those 3 fractions. Because the range (from 0.89 to 4.67%) and SD (0.30 to 1.91%) were low, nonlinear relationships were produced between spectral absorbance data and the content of those nutrients. These fractions were too concentrated to be predicted precisely by PLS-NIRS, whereas a better result was achieved by application of the SVM method. However, both the PLS and SVM methods failed to perform an accurate NIRS prediction for protein fraction B3, even though it was improved by SVM regression compared with PLS-NIRS prediction. This was a comprehensive result of lowest content of fraction B3 (average 0.23%), which was already close to the limit of determination of NIRS technology, the smallest variability (SD = 0.20%) in all the fractions, and potential errors transferred from laboratory procedures (NDFCP and ADFCP).

CONCLUSIONS

In general, the SVM method was preferred to PLS in calibrating NIRS models of protein fractions. In many related studies, the PLS-NIRS solution suffered difficulties with producing models and generalization ability when dealing with nonlinear data sets. For instance, all the protein fractions (NDFCP, ADFCP, B1, B3, and C) with smaller SD values (<1%) in this study calibrated unavailable PLS models with low accuracy ($R_p^2 < 0.85$, $RPD < 3$). Protein fractions could probably be predicted more accurately by NIRS in forages and feedstuffs that have greater N concentrations. However, the application of the SVM regression method in NIRS has made considerable progress in improving the accuracy of predicting laboratory and CNCPS protein fraction content. Compared with the known research, this study showed the potential of accurate and rapid prediction of all the laboratory protein fractions, TCP, BICP, NDFCP, ADFCP, and primary CNCPS fractions A, B2, and C by SVM-NIRS technology. In addition, fraction B1 was accurately calculated indirectly based on the SVR-NIRS prediction results. Those better results are explained by using SVM's specific optimization procedure that avoids overfitting. Because of this advantage, it is possible to perform with greater data generalization ability, especially in these nonlinear data sets such as NDFCP and ADFCP content in this study. This result will be useful for applications of both the CNCPS and NIRS technology in dairy nutrition research. Further work should be carried on broadening the sample variance (such as species and cultivars) to increase the nitrogen content and range, particularly for NDFCP, ADFCP, and B3.

ACKNOWLEDGMENTS

This work was supported by the National Agricultural Science and Technology Program (No.nyhyzx07–022) and 948 Program of China Agriculture Ministry (No. 2006-G38). The authors thank the Institute of Forage Production of Gansu Branch of Chengdu Daye International Investment Co. Ltd (Jiuquan, Gansu, China) for the chemical analysis work.

REFERENCES

- Abrams, S. M., J. S. Shenk, M. O. Westerhaus, and F. E. Barton. 1987. Determination of forage quality by near infrared reflectance spectroscopy: Efficacy of broad-based calibration equations. *J. Dairy Sci.* 70:806–813.
- Andrés, S., A. Calleja, S. López, Á. R. Mantecón, and F. J. Giráldez. 2005. Nutritive evaluation of herbage from permanent meadows by near-infrared reflectance spectroscopy: 2. Prediction of crude protein and dry matter degradability. *J. Sci. Food Agric.* 85:1572–1579.
- AOAC. 2000. Official Methods of Analysis. 17th ed. Association of Official Analytical Chemists, Washington, DC.

- Blaxter, K. L. 1956. The nutritive values of foods as sources of energy: A review. *J. Dairy Sci.* 39:1396–1424.
- Fernández Pierna, J. A., V. Baeten, A. Michotte Renier, R. P. Cogdill, and P. Dardenne. 2004. Combination of support vector machines (SVM) and near-infrared (NIR) imaging spectroscopy for the detection of meat and bone meal (MBM) in compound feeds. *J. Chemometr.* 18:341–349.
- Fox, D. G., M. C. Barry, R. E. Pitt, D. K. Roseler, and W. C. Stone. 1995. Application of the Cornell Net Carbohydrate and Protein Model for cattle consuming forages. *J. Anim. Sci.* 73:267–277.
- Givens, D. I., J. L. De Boever, and E. R. Deaville. 1997. The principles, practices and some future applications of near infrared spectroscopy for predicting the nutritive value of foods for animals and humans. *Nutr. Res. Rev.* 10:83–114.
- Gu, T., W. Lu, X. Bao, and N. Chen. 2006. Using support vector regression for the prediction of the band gap and melting point of binary and ternary compound semiconductors. *Solid State Sci.* 8:129–136.
- Hoffman, P. C., N. M. Brehm, L. M. Bauman, J. B. Peters, and D. J. Undersander. 1999. Prediction of laboratory and in situ protein fractions in legume and grass silages using near-infrared reflectance spectroscopy. *J. Dairy Sci.* 82:764–770.
- Knowlton, K. F., R. E. Pitt, and D. G. Fox. 1992. Dynamic model prediction of the value of reduced solubility of alfalfa silage protein for lactating dairy cows. *J. Dairy Sci.* 65:1507–1516.
- Krishnamoorthy, U., T. V. Muscato, C. J. Sniffen, and P. J. Van Soest. 1982. Nitrogen fractions in selected feedstuffs. *J. Dairy Sci.* 65:217–225.
- Licitra, G., T. M. Hernandez, and P. J. Van Soest. 1996. Standardization of procedures for nitrogen fractionation of ruminant feeds. *Anim. Feed Sci. Technol.* 57:347–358.
- Liu, X., W. C. Lu, S. L. Jin, Y. W. Li, and N. Y. Chen. 2006. Support vector regression applied to materials optimization of sialon ceramics. *Chem. Intell.* 82:8–14.
- Mentink, R. L., P. C. Hoffman, and L. M. Bauman. 2006. Utility of near-infrared reflectance spectroscopy to predict nutrient composition and in vitro digestibility of total mixed rations. *J. Dairy Sci.* 89:2320–2326.
- Naes, T., T. Isakson, T. Fearn, and T. Davies. 2002. A Userfriendly Guide to Multivariate Calibration and Classification. NIR Publications, Chichester, UK.
- Nie, Z. D., J. G. Han, Z. Yu, L. D. Zhang, J. H. Li, Y. Zhong, and F. Y. Liu. 2007. Quality prediction of alfalfa hay using Fourier transform near infrared reflectance spectroscopy. *Spectrosc. Spect. Anal.* 27:1308–1311.
- Norris, K. H., R. F. Barnes, J. E. Moore, and J. S. Shenk. 1976. Predicting forage quality by infrared reflectance spectroscopy. *J. Anim. Sci.* 43:889–897.
- Shenk, J. S., I. Landa, M. R. Hoover, and M. O. Westerhaus. 1981. Description and evaluation of a near infrared reflectance spectrocomputer for forage and grain analysis. *Crop Sci.* 21:355–358.
- Shenk, J. S., and M. O. Westerhaus. 1994. Forage Quality, Evaluation and Utilization. American Society of Agronomy, Crop Science Society of America and Soil Science Society of America, Madison, WI.
- Sniffen, C. J., J. D. O'Connor, P. J. Van Soest, D. G. Fox, and J. B. Russell. 1992. A net carbohydrate and protein system for evaluating cattle diets. II. Carbohydrate and protein availability. *J. Anim. Sci.* 70:3562–3577.
- Valdés, C., S. Andres, F. J. Giraldez, R. Garcia, and A. Calleja. 2006. Potential use of visible and near infrared reflectance spectroscopy for the estimation of nitrogen fractions in forages harvested from permanent meadows. *J. Sci. Food Agric.* 86:308–314.
- Van Soest, P. J., J. B. Robertson, and B. A. Lewis. 1991. Methods for dietary fiber, neutral detergent fiber and nonstarch polysaccharides in relation to animal nutrition. *J. Dairy Sci.* 75:3583–3597.
- Vapnik, V. 1998. Statistical Learning Theory. John Wiley and Sons Inc., New York, NY.
- Williams, P., and K. Norris. 2001. Near-infrared technology in the agricultural and food industries. *Am. Assoc. Cereal Chem.*, St. Paul, MN.
- Williams, P. C. 1987. Variables affecting near-infrared reflectance spectroscopic analysis. Pages 143–168 in *Near-infrared Technology in the Agricultural and Food Industries*. Am. Assoc. Cereal Chem., St. Paul, MN.
- Wu, D., S. Feng, and Y. He. 2007. Infrared spectroscopy technique for the nondestructive measurement of fat content in milk powder. *J. Dairy Sci.* 90:3613–3619.