Improvement of prediction speed and accuracy with internet enabled networking software

Robert Dzupin,[a] Charles. R. Hurburgh[a] and Sylvie. A. Roussel[b]

[a]*Iowa State University,1563 Food Sciences Building, 50011 Ames, IA, USA. E-mail: slovakia@iastate.edu*

[b]*Agrometrix, Cemagref/Minea,361 rue JF Breton BP 5095, 34033 Montpellier, France. E-mail: sylvie.roussel@montpellier.cemagref.fr*

Introduction

Near-infrared (NIR) instruments are popular for the prediction of chemical composition and biological properties of food and agricultural material. In the agricultural and food industries, NIR instruments are primarily used for the detection of C-H, N-H and O-H bonds, which relate to concentration of oil, protein and moisture. The advantages of using NIR instruments are that near-infrared spectroscopy is an unusually fast technique compare to other analytical techniques (often taking less than 1 minute), it is nondestructive, and minimal sample preparation is required. The standard use of NIR spectroscopic data relies on the development of multivariate calibrations. This has been a serious restriction of NIR spectroscopy applications because of the high cost of calibration development.

NIR spectroscopic data are used to predict analyte values and to construct a calibration model in the form of a regression equation. This equation can then be used to predict unknown samples from NIR measurements. The equation is usually obtained by partial least-squares regression (PLS),[1] a well-established multivariate linear method.

However, this calibration technique cannot model non-linearities. A mayor concern when building a model based on measurements coming from a single master NIR instrument is transferability to the other units. Calibration transfer inherently introduces non-linearities. Non-linear calibration methods could improve the accuracy of prediction models as well as their inter-instrument transferability.

Local modeling avoids the need for expensive calibration.[2] Instead of using a regression equation to summarize the database, the complete database is employed. Alternatively, artificial neural network (ANN) can be used. Both these calibration approaches depend on the accumulation of a very large database, with each item possessing full spectra and analytical data".[3]

Nonlinear and large database models can be implemented over the Internet. Software was designed to provide environment for database analysis calculations in the real time. Beside internet connectivity, the solution assumed that the NIR spectrometer will provide a communication interface to send measured optical data to a personal computer. In the current setting the RS 232 interface (serial port) was used to establish a link between instrument and personal computer. The prediction is done by a remote server. The local PC only provides communication and data management. The centralized calculation of this solution also allows simultaneous prediction of the same constituent by several models. It is likely that individual samples are better predicted by one model over others. If model selection can be developed overall accuracy would be improved by matching samples to models.

Near Infrared Spectroscopy: Proceedings of the 11th International Conference.

The objective of this paper is to evaluate performance of the real-time centralized system for handling of data over Internet developed in Grain Quality Laboratory and explore possibility of improvements of accuracy by merging prediction outputs of several chemometrics models implemented in the system.

Materials and methods

Description of the software

The main concept is to link NIR spectrometers and a commercially available database management system (SQL Server™) with flexible, high capacity numerical software (MATLAB™).

The software has three components:

Client computer – used to retrieve optical data from NIR spectrometer, send them over Internet to central database SQL server using modem, DSL or T1 connection. Client computer requirements: MS Windows9X, ME, NT or 2000 operating system and a PC that can support the selected operating system. In our testing environment IBM PC computers with 66 MHz processor speed running Windows 95 proved to be sufficient.

Computer running model calculations in Matlab™ – Personal computer with fast processor (Pentium III or IV) used to process linear, non-linear or database models and calculate predictions. Matlab server is connected to central database SQL server. If real-time processing is required, connection speed requirements are higher than for client computer. (T1, T3 or LAN) Matlab™ computer requirements: Computer running MATLAB models determines if system can be used in real time, therefore only Pentium III with 800 MHz processor or faster has been used in the software system. Because software is using MS Windows specific API calls, only MS Windows9X, ME, NT or 2000 are supported. Windows NT and 2000 are recommended. During our laboratory testing Windows9X was an unstable platform for running MATLAB™ routines over extended period of time.

SQL server- Database server that stores optical data, sample identification data, and calculated predictions from Matlab™ models SQL server computer requirements: A MS SQL 7.0 or 2000 Database server requirement for small systems (less than 50 concurrent connections) is similar to Matlab™ computer requirements. Database operations are characterized as input output very intensive, therefore SCSI hard drives preferably using RAID arrays[4] are recommended.

Near infrared spectrometers

Three near infrared spectrometers Foss/Tecator Infratec instruments (1225-Infratec serial 0065 and two 1229-Infratec serial 553075 and 243108 were used to collect transmission spectra of whole corn samples. Spectrometers provide 100-wavelength spectra in the 850-1050nm range, with 2-nm resolution.

The calibration database contains measurements provided by 2 *Master* instruments (1225-Infratec #0065 with a cuvette sample presentation and 1229-Infratec #553075 with a flow presentation). Five constituents (moisture, protein, oil, starch, and density) were reported.

Processing algorithms

Three processing algorithms were implemented in this test:

- a linear regression model (Partial Least-Squares Regression: PLS),
- a local regression model (Locally Weighted Regression: LWR) and

- a non-linear model (Artificial Neural Networks: ANN). There is no restriction on the number of models or processing algorithms that could be used

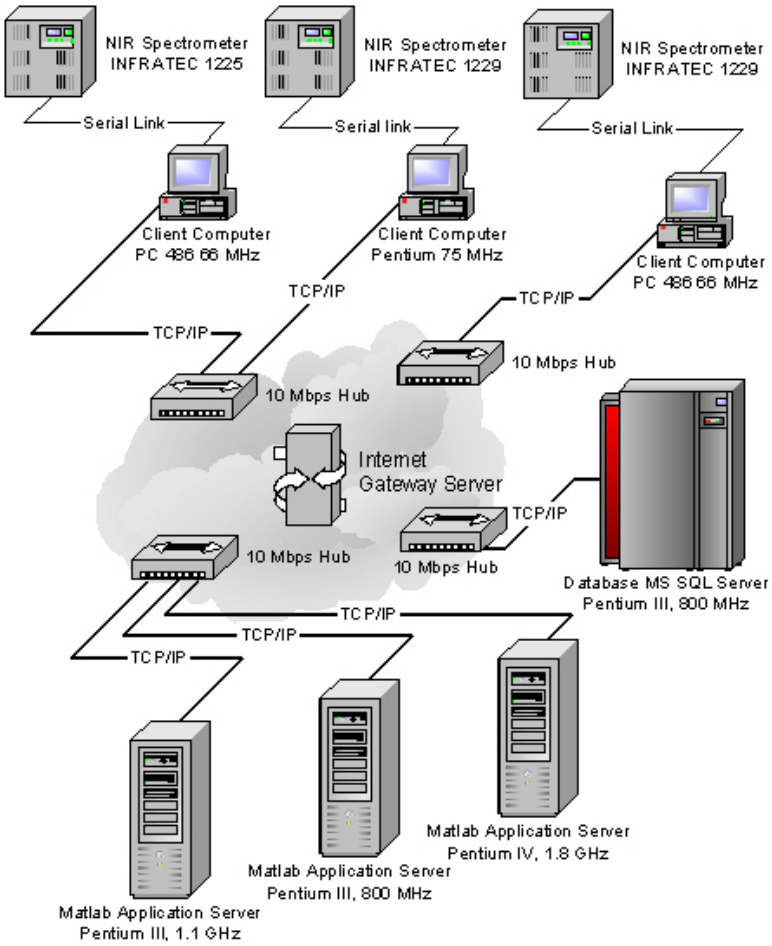

Figure 1. GrainNet configuration

Processing algorithms

Three processing algorithms were implemented in this test:

- a linear regression model (Partial Least-Squares Regression: PLS),
- a local regression model (Locally Weighted Regression: LWR) and
- a non-linear model (Artificial Neural Networks: ANN). There is no restriction on the number of models or processing algorithms that could be used

Partial least squares

Partial Least Squares Regression (PLS) is a well documented multivariate linear model that is well documented and commonly applied in the NIR area.[5] In this study, PLS is the reference model for comparison. All the data were mean-centered and the number of latent variables was tuned (lvs ≤ 15). In our model, 13 latent variables were used.

To reduce the number of wavelengths and increase the robustness and transferability, the Standard Normal Variate (SNV)[6] pre-processing technique was applied.

Locally weighted regression

The Locally Weighted Regression (LWR)[7] builds local linear regressions that enable the model to fit non-linearities. For each sample, its neighborhood is determined by the Mahalanobis distance computed on the first principal components issued from x-values (spectra) and the Euclidean y-distances. Since the y-values of the samples to be predicted are unknown, the distance and the neighborhood are computed iteratively. The neighborhood size as well as the weighting given to the distance in y (alpha) must also be tuned carefully.[8]

In the GrainNet software implementation, the lwrxy function from PLS toolbox was used as the LWR model. Input parameters usedare shown in Table 1.

Table 1. Parameters for locally weighted regression

lvs	the number principal components used to model the independent variables
npts	the number of points defined as local
alpha	the weighting given to the distance in y
iter	the number of iterations to use

Artificial neural network

Artificial Neural Networks (ANN) are able to fit non-linear relationships between multivariate x and y-values. In this study, supervised 3-layer feed-forward neural networks are trained with dynamic learning using error-gradient back-propagation algorithms.[9] The inputs (and the outputs) are scaled between –1 and +1 to fit to the range of the hyperbolic tangent activation functions. The *master* database is used to train the ANN, with no early stopping method (they stop the training too early because the error descent is not monotonous). Thus, the number of epochs has to be tuned.[10]

In our model, the neural network contained 30 inputs, 10 hidden layers, and 2500 epochs were used.

Calibration databases

Original databases:

6442 corn samples : 2762 from unit serial 0065, 2823 from unit serial 553075, and 857 from unit serial 0350.

Database cleaning (outlier removal):

- with PCA (spectral outliers) and prediction residuals (chemistry value outliers) for every constituent.

Table 2. Corn calibration database and models

CORN	Moisture	Protein	Oil	Starch	Density
Initial database	5782	2138	2137	2127	1925
PCA outliers	3	1	1	1	1
Residual outliers	9	21	12	12	64
Final database	5782	2116	2124	2062	1857
Calibration set	4625	1693	1699	1649	1485
Test set	1157	423	425	413	372
SEP for LWR model	0.32%	0.33%	0.30%	0.70%	1.63%
SEP for NN model	0.31%	0.28%	N/A	N/A	N/A
SEP for PLS Model	0.41%	0.34%	N/A	N/A	N/A

*Model was not developed

Model comparison

The corrected standard error of prediction (SEP corrected) was calculated from a verification set of 30 samples with wet chemistry references provided by Woodson–Tenent Laboratories, Inc. (Des Moines, IA). These samples had replicated chemistry values and were laboratory transfer standards. Possibilities for improvement in precision of the models were explored.

Bias corrected standard error of prediction was calculated by the equation:

$$SEP(corrected) = \sqrt{\frac{\Sigma(y-x)^2 - (\Sigma(y-x))^2 / n}{n-1}} \quad (1)$$

where: y is the result from the chemical analysis
x is the result predicted from NIR measurements
n is the number of samples in the validation set

SEP(corrected) was calculated for PLS, ANN and LWR models. SEP was also calculated for the average of prediction differences of all three models.

An optimal SEP was manually calculated. From the model that was closest to the chemical analysis result for each sample individually.

System performance

To estimate the number of instruments that could supported by GrainNet software, the throughput of the system (number of processed samples per minute) was calculated:

$$Throughput = \frac{60}{t} t_s \quad (2)$$

where:

$t = t_1+t_2+t_3+t_4+t_2$

t_1 – time (in seconds), necessary to retrieve data from SQL Server™ database to computer running Matlab™

t_2 – network delay (in seconds) between SQL Server™ database and Matlab™ computer

t_3 – time (in seconds), needed to processing data in to Matlab™ environment

t_4 – time (in seconds), necessary to update SQL Server™ database with output from Matlab™

t_s – time (in seconds) to measure one sample on the NIR spectrometer (load, measure and unload sample from spectrometer)

Results

Model performance evaluation

To compare the performance of the PLS, ANN and LWR models, SEP(Corrected) was calculated (Table 4). As expected, the ANN and LWR models were more accurate than the PLS model. SEP was also calculated for the average prediction of all models. LWR was the model with lowest SEP when processing optical data collected from 1225-Infratec #0065 spectrometer. ANN method had lowest SEP for 1229-Infratec 553075 and 243108 spectrometers, but in the same time the SEP for 1225-Infratec 0065 using ANN was the highest of all three models even though this instrument was in the training database. The SEP for averaged prediction differences was more consistent across units. To represent what might be ideally achieved with model selection, the optimal model concept was introduced. In the optimal model, the prediction closest to the reference value is manually selected from the pool of models.

Table 4. Corrected Standard error of predictions for corn protein

SEP	PLS model	LWR model	ANN model	Model with Averages	Optimal model
Spectrometer 0065	0.31	**0.28**	0.31	0.29	0.24
Spectrometer 553075	0.28	0.26	**0.24**	0.25	0.22
Spectrometer 243108*	0.30	0.29	**0.27**	0.28	0.22
Average	0.30	0.28	0.27	0.27	0.23

Number of samples: 30

*Not in the calibration pool

System performance evaluation

Throughput of the models is reported in Table 3. The first line shows throughput when all 3 processing algorithms were used. The second line of the table shows throughput of the system, using only database processing algorithm (LWR) to calculate five constituents (moisture, protein, oil, starch, and density). The database throughput was also measured. Database throughput is the number of samples that can be processed by computer used in the system if no model calculation is performed. Database throughput accounts for network delays between the database and servers with Matlab™ routines. Because the computers are using same network connection to the database server, database throughput is same for all three computers. Table 3 can predict the number of

computers for processing selected calculations in Matlab™, in for real time support of the NIR spectrometers. The assumption is that the new optical data are send from the NIR spectrometer once per minute. If this time is different, the estimated number of processed samples needs to be multiplied by the appropriate ratio. For example, if the processing time for one sample is three minutes, number of users that computer can handle would be three times higher.

Table 3. Throughput of implemented models

Model	**Pentium III 0.8 GHz (Number of processed samples per minute)**	**Pentium III 1.1 GHz (Number of processed samples per minute)**	**Pentium IV 1.8 GHz (Number of processed samples per minute)**	**Database throughput (Number of processed samples per minute)**
PLS ANN LWR (2 constituents)	11.6	13.8	16.2	35.0
LWR corn (5 constituents)	16.1	16.4	22.1	35.0

Discussion

An "ultimate" system where calibration is based on samples supplied by diverse clients to a host laboratory, and is used to predict results upon receipt of spectra by e-mail, using the local or ANN models, was proposed by Phil Williams[3]. GrainNet software is extending the idea of the "ultimate" system to real-time and the possibility of improving accuracy of prediction by center averaging the results of several models or choosing models based on sample properties.

Because the NIR instruments collect raw optical data, GrainNet software is not limited to any particular NIR instrument manufacturer. The only implementation requirement of the instrument is the capability of the spectrometer to send raw optical data to a standard communication port. (RS 232, USB, ect.)

The software requires a fast network connection between the database server and computers that process the models in Matlab™. A fast network connection is especially necessary if several computers are used to calculate prediction. For example, to predict 5 constituents using LWR, we can process 54 samples per minute with three computers instead of 16 or 22 samples per minute if only one computer is used.

Data in Table 4 suggests that real-time access to rapid computing can improve accuracy by merging or selecting among prediction outputs of several chemometrics models. Using the optimal model to estimate the potential improvement beyond the PLS, LWR, or ANN models, the accuracy of all three models can be improved. The accuracy of the PLS model was improved by 23 percent. The accuracy of the LWR model was improved by 18 percent and the accuracy of the ANN model was improved by 15 percent.

Optical data retrieved from NIR instruments are accompanied by Instrument ID, Time, Computer ID, User Name, and by data manually entered by the operator (Sample ID, Variety, etc.). Therefore, each set of optical data in the SQL Server© database can be uniquely identified, as required for instrument network management.

References

1. Sylvie A. Roussel, Glen R. Rippke, Charles R. Hurburgh, Jr., Accuracy and robustness comparison of different processing algorithms for grain quality assessment, Proceedings Pittcon 2000., www.pittcon.org (2000).
2. Tony Davies, Some variations on a 'Local' theme, Spectroscopy Europe, July/August 1999, Norwich Near Infrared Consultancy, 75 Intwood Road, Cringleford, Norwich NR4 6AA, UK. E-mail: td1@nir-pub.demon.co.uk (1999).
3. Phil Williams and Karl Norris, Near-Infrared Technology in the Agricultural and Food Industries, American Association of Cereal Chemists, Inc. St. Paul, Minnesota, USA, p. 145 (2001).
4. Brad M. McGehee, How to Performance Tune Microsoft SQL Server During Setup, http://www.sql-server-performance.com/ sql_server_setup.asp, © 2000 - 2003 SQL-Server-Performance.Com (2003).
5. Geladi P, Kowalski B.,Partial Least Squares regression: a tutorial, Anal. Chim. Acta, 185, pp1-17 (1986).
6. Barnes RJ, Dhanoa MS, Lister SJ,Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra, Appl. Spectrosc., 43(5), pp772-777 (1989).
7. Wang Z, Isaksson T, Kowalski BR., New Approach for Distance Measurement in Locally Weighted Regression, Anal. Chem., 66, pp294-260 (1994).
8. Sylvie A. Roussel, Glen R. Rippke, Charles R. Hurburgh, Jr., Accuracy and transferability comparison of various multivariate processing algorithms for corn quality assessment, The 10th International Diffuse Reflectance Conference (2000).
9. Rumelhart D E, Learning and generalization, In Proceedings of IEEE International Conference on Neural Networks, San Diego, CA, USA (1988).
10. Sylvie A. Roussel, Glen R. Rippke, Charles R. Hurburgh, Comparison of Global Linear, Local Linear and Non-Linear Models for Grain Quality Assessment, Proceedings Pittcon 2003., www.pittcon.org (2003).