http://dx.doi.org/10.5012/bkcs.2012.33.12.4267

Random Forest as a Non-parametric Algorithm for Near-infrared (NIR) Spectroscopic Discrimination for Geographical Origin of Agricultural Samples

Sanguk Lee, Hangseok Choi,[†] Kyungjoon Cha,[‡] Mi-Kyeong Kim,[§] Jeong-Soo Kim,[§] Chi Hee Youn,[#] Su-Heon Lee,[¶] and Hoeil Chung[*]

Department of Chemistry and Research Institute for Natural Sciences, Hanyang University, Seoul 133-791, Korea
[*]E-mail: hoeil@hanyang.ac.kr
[†]Department of Long-Term Care Claims Review, National Health Insurance Corporation, Seoul 121-749, Korea
[‡]Department of Mathematics, Hanyang University, Seoul 133-791, Korea
[§]Crop Protection Division, National Academy of Agricultural Science, RDA, Suwon 441-707, Korea
[#]Smarteome. Co., Ltd., Suwon 441-853, Korea
[¶]Department of Applied Bioscience, Kyungpook National University, Daegu 702-701, Korea
Received September 1, 2012, Accepted October 9, 2012

Key Words : Random forest, Discrimination, Model over-fitting, NIR analysis, Geographical origin

Recently, agricultural products from diverse geographical origins are widely available in markets due to increased international trading *via* free trade agreement (FTA). Therefore, there is a demand for analytical methods to correctly identify their geographical origins to ensure public confidence in fair evaluation of product value. The use of chromatographic methods and/or DNA analysis enables the identification of geographical origins; however, these are slow and requiring extensive sample preparation.[1-6] When number of target samples is large and efficient screening of them is necessary, a fast and non-destructive analytical method such as near-infrared (NIR) and Raman spectroscopy is adequate, although the subsequent accuracy of identification would degrade.

Since agricultural products are complex in composition, the corresponding NIR or Raman spectra are highly overlapping without distinct signatures of individual components. In this situation, multivariate discrimination methods are typically employed to establish empirical correlation between spectral features and relevant geographical origins. Although diverse discrimination methods have been developed,[7-12] three methods of principal component analysis-linear discriminant analysis (PCA-LDA),[8,9,13-15] partial least squares-discriminant analysis (PLS-DA),[8,9,16-18] and soft independent modeling of class analogy (SIMCA)[8,10,19-21] have been most frequently adopted for many practical field applications.

Recently, random forest (RF), one of the latest ensemble methods in machine learning, has been recognized as an effective method for discrimination, largely in biomedical fields such as gene analysis, metabolomics and medical image analysis.[22-26] Diverse analytical data such as mass spectra and chromatograms has been directly fed into RF for either discrimination or classification. RF combines many trees to form a forest for analysis. An individual tree represents a model describing the characteristics of an input feature that is present in a subset of the whole dataset. Detail mathematical descriptions of RF can be found in other publications.[27-30]

Here, the procedure for building a RF model is briefly described. To build a tree, initially two thirds (66.7%) of n total spectra are randomly selected. This procedure is called out-of-bag (OOB) sampling.[27] The percentage of the sample selection could be changed, but the selection of two thirds out of the total samples is typical. After the random selection of samples, one third (33.3%) of the wavelengths out of p total wavelengths are randomly selected again and are used to build a tree.

When variables (absorbance values in the case of NIR spectra) are divided into two daughter nodes, a wavelength that can minimize the variances of divided daughter nodes is identified out of the randomly selected wavelengths. The minimal sum of both variances corresponds to the more distinct two daughter nodes. At the next two daughter nodes divided at the previous node are further classified by searching again for the best wavelengths out of the selected wavelengths. The same procedure of building nodes is repeated until the selected samples for a given tree are fully classified. Finally, each classified sample is assigned to the corresponding group (geographical origin). After building the first tree, the samples that remain (one third of n spectra) are passed down the tree and then the resulting discrimination accuracy is calculated. Next, by repeating the building of each tree, k trees are built and finally form a RF model.

For an unknown sample spectrum, the wavelengths employed in a given tree are again selected, and the corresponding absorbance values are passed through the tree to decide the group of a sample. Therefore, k predicted groups individually determined by each tree are available for each sample. Ultimately, the final predicted group of an unknown sample is obtained by averaging these k predicted results. As described above, each tree in a forest is developed by employing randomly selected samples in a given dataset, and variables (wavelengths or wavenumbers) used to build a tree are also randomly selected, so each tree can be regarded as indepen-

dent. When many independent trees are combined for analysis, the risks of biased decisions or overfitting would greatly decrease.

Although applications of RF have been rapidly increased in diverse areas, the potential of RF for vibrational spectroscopic discrimination for geographical origins of agricultural products has been rarely investigated. In this study, we have explored RF as a potential multivariate method for discrimination using three different NIR spectral datasets collected from sesame, *Angelica gigas* and rice samples with different geographical origins.

To initially examine the spectral features of samples between different geographical origins, all raw spectra were converted to second derivative (2D) spectra to enhance the spectral variations. Figure 1(a) shows the average 2D spectra of the domestic (black) and imported (red) sesame samples. As shown, the spectral features of the two groups are quite similar to each other; while, minor spectral differences are observed especially around 6500 and 5000 cm^{-1}. Figure 1(b) displays the average 2D spectra of domestic (black) and imported (red) *Angelica gigas* samples. The spectral difference between the *Angelica gigas* samples of two different origins is relatively more distinct compared with that of the sesame samples. Figure 2(c) presents the average 2D spectra of the domestic (black) and imported (red) rice samples. The spectral difference is very small and only minute differences are observed around 8300 and 7010 cm^{-1}. In overall, the spectral differences between two geographical origins for these three samples are not significant, so a method able to effectively reflect minor spectral differences in discrimination analysis becomes more beneficial.

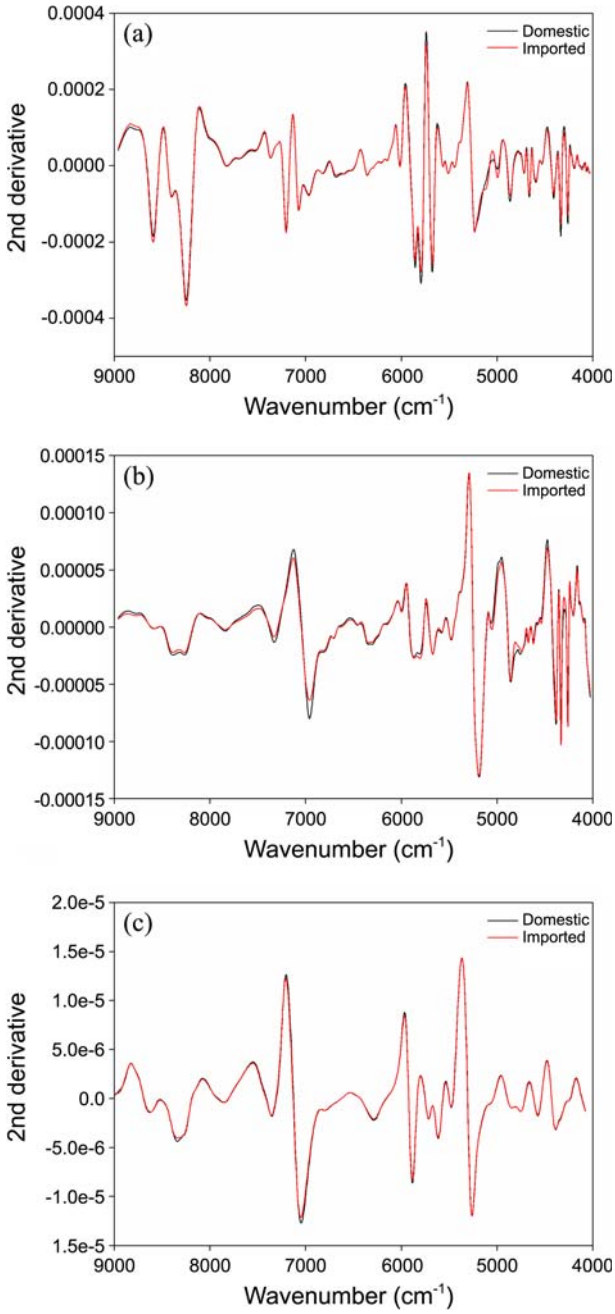

Figure 1. Average 2D spectra of domestic (blue) and imported (red) sesame (a), *Angelica gigas* (b) and rice (c) samples.

For the purpose of comparison, PCA-LDA and PLS-DA were also performed using the same datasets and the resulting discrimination accuracies were compared with those acquired using RF. PCA was initially performed using each spectral dataset and the resulting scores were used for the LDA. A combination of two scores was employed since it was easy to visualize the discrimination performance in the two-dimensional domain. For the PLS-DA, discrimination models were developed by assigning one group as 1 and the other group as 2. The prediction accuracy was obtained by evaluating the predicted values of the samples, whether these were below or above 1.5. RF was performed as described earlier.

The discrimination errors (percent errors) obtained by predicting the samples in each validation set are summarized in Table 1. The numbers in parentheses in the columns for the PCA-LDA, PLS-DA, and RF correspond to the selected optimal two-score combination, number of factors, and number of trees, respectively, determined by leave-one-out cross-validation using the corresponding calibration dataset. As shown in the table, the discrimination accuracies improved in all three cases when RF was used.

In the case of factor-based analysis such as PCA-LCA and PLS-DA, determination of an optimal number of factors (eigenvectors) is inevitably required, since a model can be easily overfit when excess factors are used. While, a RF model would not be overfit even with the inclusion of large number of trees. This non-overfitting nature of RF has been

Table 1. The discrimination errors (percent errors) acquired using PCA-LDA, PLS-DA and RF for sesame, *Angelica gigas* and rice samples. The numbers in parentheses in the columns for the PCA-LDA, PLS-DA and RF correspond to the selected optimal two-score combination, the number of factors and the number of tress, respectively

	PCA-LDA	PLS-DA	RF
Sesame	9.8 ($1^{st}/4^{th}$)	9.4 (3)	6.7 (10000)
Angelica gigas	10.9 ($1^{st}/2^{nd}$)	8.6 (3)	5.5 (10000)
Rice	22.2 ($2^{nd}/3^{rd}$)	5.6 (7)	0.0 (10000)

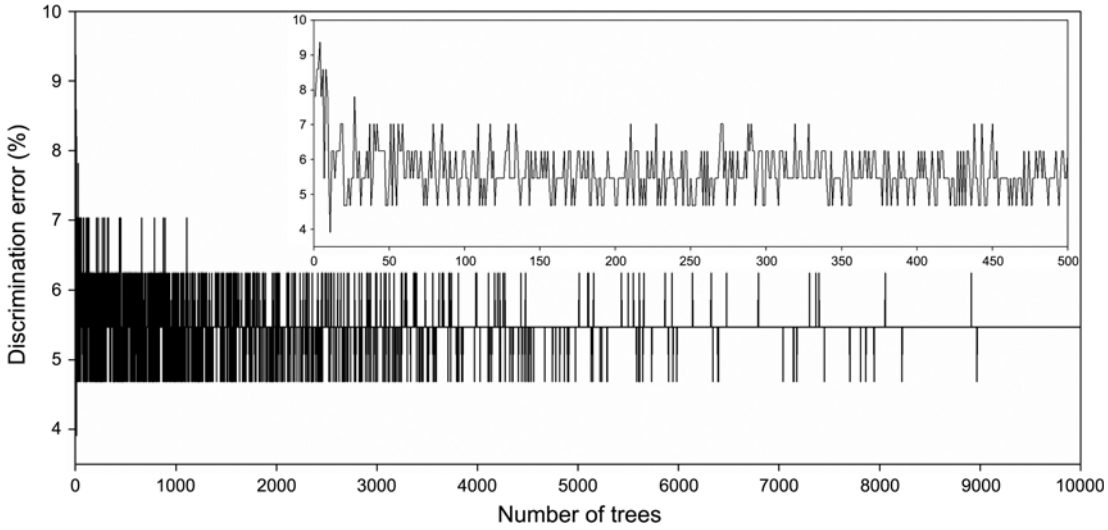

Figure 2. Variation in discrimination error as a function of the number of trees used to discriminate the imported and domestic *Angelica gigas* samples. For the examination of detail, the range up to 500 trees is enlarged.

mathematically proven.[27] Figure 2 shows the variation in discrimination error as a function of the number of trees used to discriminate the imported and domestic *Angelica gigas* samples. For the examination of detail, the range up to 500 trees is enlarged in the figure. When the number of trees used in an RF model is small, *i.e*., below 500, the resulting discrimination errors are largely fluctuating with the continual addition of trees, indicating insufficient robustness of the built RF model. The fluctuation of error continues up to 4800 trees; while, the frequency of error fluctuation largely decrease. After 4800 trees, the frequency of error fluctuation substantially decreases further and no change in the error is observed after the inclusion of 9000 trees. This insensitivity of error to the number of trees implies that the prediction performance of the model is now robust. In this case, 10000 trees are used in the RF model, since the discrimination errors have sufficiently converged to a constant level. The similar trends were also observed when the number of trees varied in the RF models for geographical discrimination of sesame as well as rice samples. Therefore, also 10000 trees were used in both RF models as described in Table 1.

In conclusion, the potential of RF has been demonstrated for NIR spectroscopic discrimination of agricultural samples according to their geographical origins. Free of over-fitting is the most important advantage of RF since many analysts do not want to involve in argument for possible over-fitting of a model when factor-based multivariate method is used. The decreased discrimination error by the use of RF in this study would not be generalized since only three different samples were employed; however, it is clear that RF could be an alternative potential discrimination method worthwhile to consider along with existing factor-based multivariate methods.

Experimental

In the case of sesame and *Angelica gigas* samples, diffuse reflectance NIR spectral datasets were kindly supplied by the National Agricultural Products Quality Management Service (NAQS), Seoul, Korea. Also, NAQS also provided the rice samples. All of the NIR spectra were collected with a Foss NIRSystems Model 6500 spectrometer equipped with a quartz halogen lamp and PbS detector. All of the samples were ground into powders (20-mesh) for the collection of diffuse reflectance spectra.

Table 2. Description for the division of dataset into calibration and validation set for sesame, Angelica gigas, and rice samples. The numbers in the table indicate the number of spectra assigned to the corresponding calibration and validation set

	Total	Calibration (Domestic/Imported)	Validation (Domestic/Imported)
Sesame	1122	898 (552/346)	224 (138/86)
Angelica gigas	647	519 (293/226)	128 (72/56)
Rice	60	48 (24/24)	12 (6/6)

Table 2 describes the division of dataset into calibration and validation set for three different agricultural samples. The numbers in the table indicate the number of spectra assigned to the corresponding calibration and validation set. In all three cases, the calibration samples were randomly selected. All of the spectral pre-treatments as well as the discrimination analyses, including second derivative, PCA-LDA, PLS-DA and RF, were performed using MATLAB Version 7.0 (Math Works Inc., MA, USA).

Acknowledgments. This work was carried out with the support of "Cooperative Research Program for Agriculture Science & Technology Development (Project No. PJ007511)" Rural Development Administration, Republic of Korea.

References

1. Kim, E. J.; Kwon, J.; Park, S. H.; Park, C.; Seo, Y.-B.; Shin, H.-

K.; Kim, H. K.; Lee, K.-S.; Choi, S.-Y.; Ryu, D. H.; Hwang, G.-S. *J. Agric. Food Chem.* **2011**, *59*, 8806.
2. Ye, N.; Zhang, L.; Gu, X. *Analytical sciences* **2011**, *27*, 765.
3. Pizarro, C.; Rodríguez-Tecedor, S.; Pérez-del-Notario, N.; González-Sáiz, J. M. *J. Chromatography A* **2011**, *1218*, 518.
4. Jaitz, L.; Siegl, K.; Eder, R.; Rak, G.; Abranko, L.; Koellensperger, G.; Hann, S. *Food Chemistry* **2010**, *122*, 366.
5. Ding, X.; Hou, Y.-L. *Biochem. Syst. Ecol.* **2012**, *44*, 233.
6. Barra, A.; Garau, V. L. Dessi, S.; Sarais, G.; Cereti, E.; Arlorio, M.; Daniel, J.; Cabras, P. *J. Agric. Food Chem.* **2008**, *56*, 10847.
7. Lavine, B.; Workman, *J. Anal. Chem*. **2010**, *82*, 4699.
8. Tominaga, Y. *Chemometr. Intell. Lab.* **1999**, *49*, 105.
9. Ciosek, P.; Brzózka, Z.; Wróblewski, W.; Martinelli, E.; Di Natale, C.; D'Amico, A. *Talanta* **2005**, *67*, 590.
10. Kizil, R.; Irudayaraj, J. *J. Agr. Food Chem.* **2006**, *54*, 13.
11. Donald, D.; Coomans, D.; Everingham, Y.; Cozzolino, D.; Gishen, M.; Hancock, T. *Chemometr. Intell. Lab.* **2006**, *82*, 122.
12. Menze, B. H.; Petrich, W.; Hamprecht, F. A. *Anal. Bioanal. Chem.* **2007**, *387*, 1801.
13. Notingher, I.; Green, C.; Dyer, C.; Perkins, E.; Hopkins, N.; Lindsay, C.; Hench, L. L. *J. R. Soc. Interface* **2004**, *1*, 79.
14. Ami, D.; Natalello, A.; Mereghetti, P.; Neri, T.; Zanoni, M.; Monti, M.; Doglia, S. M.; Redi, C. A. *Spectroscopy* **2010**, *24*, 89.
15. Lee, S.; Chung, H.; Choi, H.; Cha, K. *Microchem. J.* **2010**, *95*, 96.
16. Xie, L.; Ying, Y.; Ying, T.; Yu, H.; Fu, X. *Anal. Chim. Acta* **2007**, *584*, 379.
17. Lee, J. H.; Choung, M.-G. *Food Chem.* **2011**, *126*, 368.
18. Cozzolino, D.; Smyth, H. E.; Gishen, M. *J. Agric. Food Chem.* **2003**, *51*, 7703.
19. Woo, Y.; Kim, H.-J.; Zeb, K.; Chung, H. *J. Pharmaceut. Biomed.* **2005**, *36*, 955.
20. Meza-Márquez, O. G.; Gallardo-Velázquez, T.; Osorio-Revilla, G. *Meat Science* **2010**, *86*, 511.
21. Checa-Moreno, R.; Manzano, E.; Mirón, G.; Capitan-Vallvey, L. F. *Talanta* **2008**, *75*, 697.
22. Díaz-Uriarte, R.; Alvarez de Andrés, S. *BMC Bioinformatics* **2006**, *7*, Article 3.
23. Zucknick, M.; Richardson, S.; Stronach, E. A. *Stat. Appl. Genet. Mol.* **2008**, *7*, 7.
24. Beckmann, M.; Enot, D. P.; Overy, D. P.; Scott, I. M.; Jones, P. G.; Allaway, D.; Draper, J. *Brit. J. Nutr.* **2010**, *103*, 1127.
25. Romero, R.; Mazaki-Tovi, S.; Vaisbuch, E.; Kusanovic, J. P.; Chaiworapongsa, T.; Gomez R.; Nien, J. K.; Yoon, B. H.; Mazor, M.; Luo, J.; Banks, D.; Ryals, J.; Beecher, C. *J. Matern-fetal Neo. M.* **2010**, *23*, 1344.
26. Menze, B. H.; Kelm, B. M.; Weber, M.; Bachert, P.; Hamprecht, F. A. *Magn. Reson. Med.* **2008**, *59*, 1457.
27. Breiman, L. *Mach. Learn.* **2001**, *45*, 5.
28. Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Champ and Hall Inc.: 1984.
29. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*; Springer: 2009.
30. Bauer, E.; Kohavi, R. *Mach. Learn.* **1999**, *36*, 105.