Advance Journal of Food Science and Technology 12(11): 609-616, 2016
DOI:10.19026/ajfst.12.3319
ISSN: 2042-4868; e-ISSN: 2042-4876

Submitted: March 10, 2016 Accepted: May 23, 2016 Published: December 15, 2016

Research Article

Prediction Model of Thermal Properties of Fruits and Vegetables Based on Random Forest and Fusion Model

Xi Xie and Weizhong Jiang
School of Electrical and Information, Jinan University, Zhuhai, 519070, China

Abstract: This study is aimed to model the complicated correlation between physicochemical property and thermal property of fruits and vegetables. And this study predicts thermal property values with higher accuracy based on fusion model and Pearson correlation analysis. Thermal property is important for the storage and transportation of fruits and vegetables. And it's an important factor of fruits and vegetables fresh-keeping. To design a fresh-keeping device, it's necessary to measure the thermal properties of fruits and vegetables. Some people use complicated devices to measure thermal properties directly or establish physical thermal model to analyze thermal properties. But it's difficult to use only physical thermal model to express the complicated correlation between physicochemical properties and thermal properties. Leaning machine is a new way to model the complicated correlation among multiple dimensions attributes. At first, this study uses Pearson correlation analysis to analyze the correlation between physicochemical property and thermal property. This step is to choose predictors which will be used to predict the thermal properties. This study uses BP neural network, random forest and GBDT algorithm to predict thermal properties of fruits and vegetables. And after testing models with 700 sets of original data, the test result shows that all of these three models have good performance. To get higher accuracy of prediction, this study uses BP NN and random forest to establish fusion model which means it uses random forest to fuse the prediction result of random forest, BP neural network and the original predictors. The performance of fusion model is 89.3% (R^2) for prediction of Thermal conductivity and 96.3% (R-square) for prediction of Freezing point. The test result shows that fusion model has better performance and higher accuracy to predict thermal properties of fruits and vegetables. This study was conducted in Electrical and Information School, Jinan University from Apr. 15th 2015 to Mar. 1st 2016.

Keywords: Fruits and vegetables, fusion model, physicochemical properties, Pearson correlation analysis, random forest, thermal properties

INTRODUCTION

The thermal property of foods is directly related to the shelf life of foods. As was the case for fruits and vegetables, thermal property is important for the storage and transportation of fruits and vegetables. The design of Fresh-keeping equipment is based on thermal property values of fruits and vegetables. In China, the loss ratio of fruits and vegetables is between 25-35% when fruits and vegetables are picked, transported or stored (Zhong, 2010). To design a better fresh-keeping device, it's necessary to get the specific values of thermal properties. There are several general methods to get the values of thermal properties. One is using a complex instrument to measure these property values directly or establishing physical thermal model to analyze thermal properties. Another is using learning machine algorithm to model the relationship between physicochemical property and thermal property. Some people use physical thermal model to analyze thermal properties. Cheng *et al.* (2006) compares different equations to measure thermo-physical properties with different equipment and models and gets a certain researching achievement. This study uses learning machine algorithm to predict thermal properties.

Some people have already tried to use learning machine to predict the thermal property values of different kinds of food. Zhong (2010) uses neural network prediction model to predict thermo-physical properties. Zhang (2005) and Zhang *et al.* (2010) studies on Thermal Conductivity Measurement System and Temperature Field Simulation of Postharvest fruits and vegetables and also uses BP neural networks to predict thermal conductivities. Brillante *et al.* (2015) uses random forest and gradient machine to predict skin flavonoid content from berry physical–mechanical characteristics in wine grapes. Sablani and Shafiur Rahman (2003) uses neural networks to predict thermal

Corresponding Author: Weizhong Jiang, School of Electrical and Information, Jinan University, Zhuhai, 519070, China

conductivity of food. BP neural network has the advantage of simple structure and mature technique and it performs well in nonlinear function approximation. But it still doesn't have enough prediction accuracy. Brillante *et al.* (2015) uses random forest to predict flavonoid content of white grape and the model performs well.

To predict thermal properties more accurately, this study uses BPNN, random forest and GBDT to build up a multi-layer fusion model. Random forest is a combined classifier based on statistic studying theory. And it combines bootstrap resample and decision tree algorithm (Cao, 2014). Random forest is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them (Breiman, 2001). So this study also does correlation analysis for the properties of fruits and vegetables.

Correlation rule analysis is one important branchon data mining. And it's widely applied in decision modeling (Jian, 2013). In this study, Pearson Correlation analysis is used to analyze the correlation among physicochemical properties of fruits and vegetables because the correlation among predictors is an important influence factor for model performance. In this study, density, water content and solid content are three based predictors. Thermal conductivity and freezing point are prediction thermal properties. After using person correlation to filter predictors, BP NN, random forest and GBDT will be separate tested by using training dataset and testing dataset. After analyzing the test result of every single model, this study uses random forest to fuse the original predictors and the prediction result of BPNN and random forest in the second layer. After testing fusion model, the results show that the performance of fusion model is better than that of any of those three models. And it can predict thermal properties with higher accuracy.

MATERIALS AND METHODS

Physicochemical properties and thermal properties: This study uses density, water content and soluble solids content to model the relationship between physicochemical properties and thermal properties of fruits and vegetables. Density is an important indicator of the quality of fruits and vegetables. And it reflects the organization porosity of plants. For plants, Organization porosity is directly related to heat transfer process. As density increases, thermal conductivity will increase (Zhang, 2005). Water content is from bound water and free water. And it is directly related to cell activity. As water content increases, thermal conductivity will increase (Zhang, 2005). Soluble solids content contains organic acid, salt, pectin, vitamin and *et al.* The relationship between soluble solids content and thermal conductivity is not so clear because Zhang (2005) and Zhong (2010) have different conclusion about it. And mathematical relationship is not important in this study because learning machine algorithm is responsible for modeling the relationship among input features and outcomes. This study also uses learning machine to model the relationship between physicochemical properties and freezing point. Although thermal property of fruits and vegetables doesn't only contain density, water content and solids content, using these physicochemical properties to predict thermal properties is still effective.

Random forest model and GBDT model: Decision tree is a classical single classifier. And its generation process consists of three phases. In the first phase, tree structure is generated by using Recurrence analysis method to analyze training dataset. In the second phase, a series of rules are generated by analyzing all paths from the root node to the leaf nodes. In the third phase, using these rules and test dataset to do classification or prediction. Generating decision tree uses a greedy local algorithm to generate rules and its classification rules are always complicated. Using pruning for decision tree is an effective way to optimize tree structure. And because ID3 algorithm doesn't backtrack, sometimes decision tree traps/in local optimum. Over-fitting is another problem when using decision tree. And over-fitting is a common problem in learning machine.

Random Forest and GBDT are both combined classifier. They consist of multi decision trees. The biggest difference between the random forest and GBDT is the method to generate training dataset. Random forest use sampling with replacement to generate training dataset which called bagging. Bagging is based on repeatable random sampling. And bagging is an effective way to increase accuracy of learning algorithm. All original datasets have the same probability to be extracted in bagging algorithm. But if some datasets are extracted too many times, they can't be extracted any more. The probability that every dataset can be extracted is $1 - (1\text{-}1/N)^N$. And Nis the amounts of original datasets. GBDT is also called MART or GBRT. It uses weight updating sampling to train the tree model which is called Boosting. Boosting generates training dataset for every decision tree in GBDT model and it also assigns a weight for each decision tree. In every training, weight of each decision tree will be updated according to their contribution to the classification or prediction result. When testing GBDT model, each decision tree will vote to the test

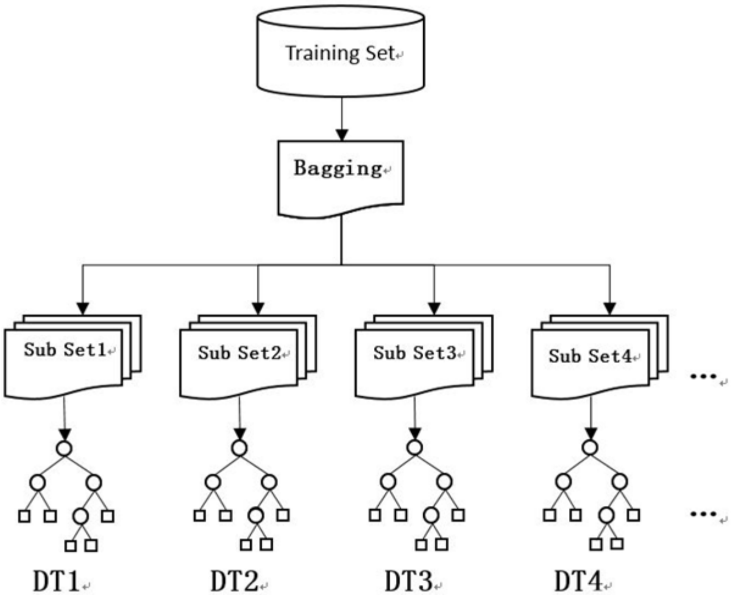

Fig. 1: Training random forest

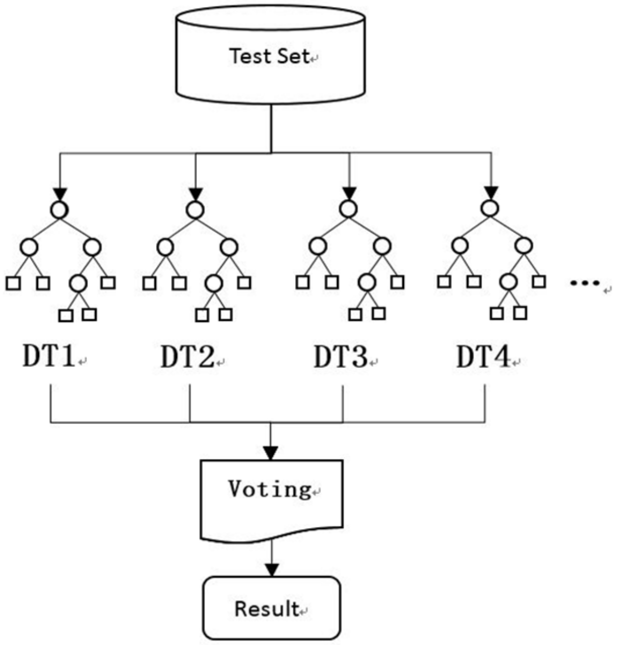

Fig. 2: Test random forest

result with their weight. The training process of random forest is showed in Fig. 1. And testing process is showed in Fig. 2.

Fusion model and tune models: In general, feature construction is the most important part of learning machine, especially when original dataset lacks of input feature. In this study, density, water content and solid content are three based input features. Thermal conductivity and freezing point are prediction properties. Tree model is sensitive to input features. Using right features will increase the performance of tree model. To increase accuracy of prediction, it's necessary to analyze the relationship among these features, especially the correlation between input features and outcomes. This study chooses Pearson correlation coefficient to analyze the features from original dataset. The result is showed in Table 2. Another important way to increase the capability of nonlinear approximation and this study tries to build up a fusion model to increase the capability. To build up a fusion model, tuning based

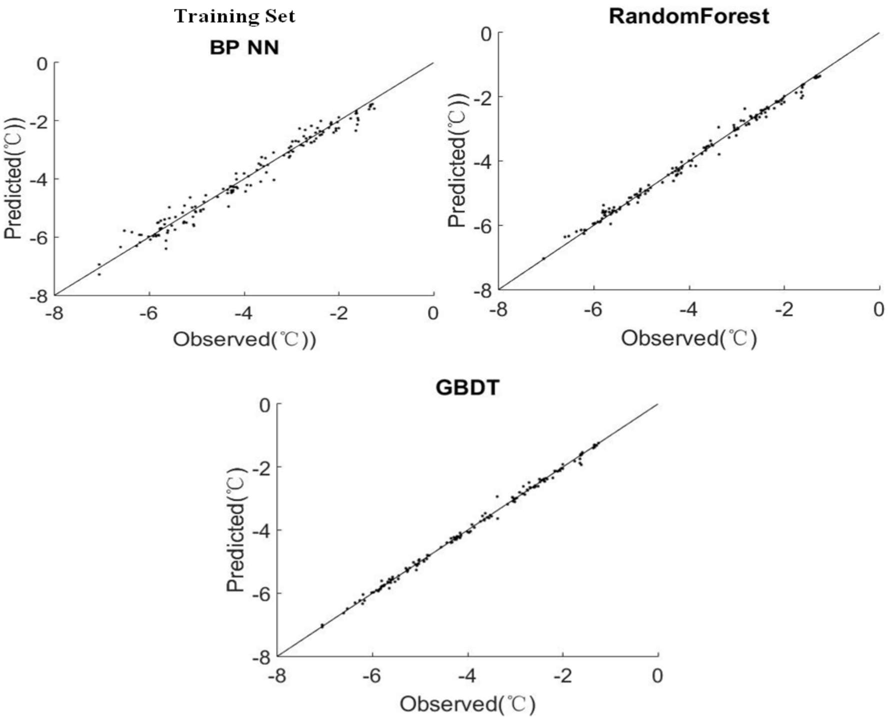

Fig. 3: Training result of three models

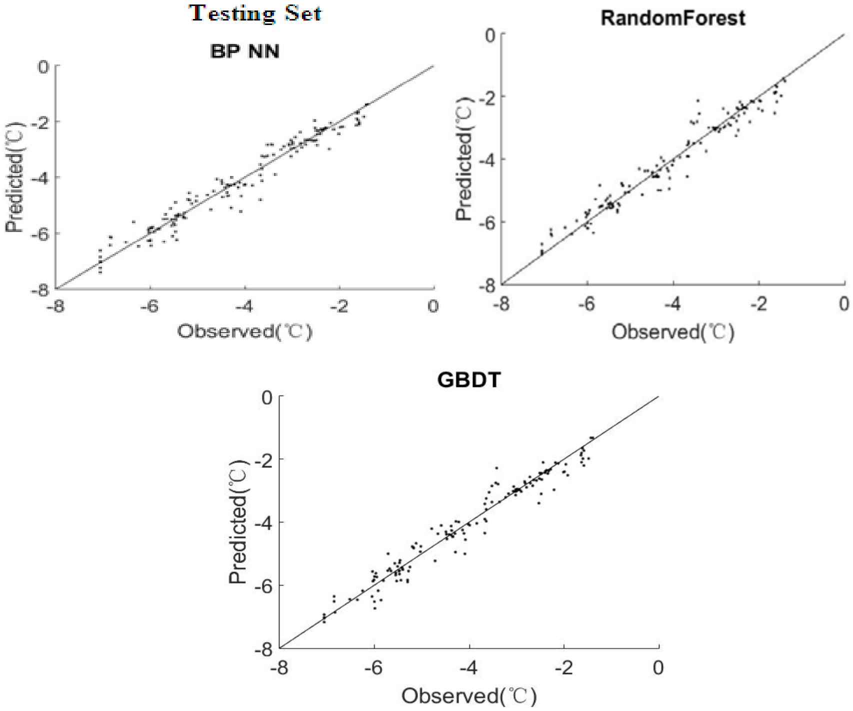

Fig. 4: Testing result of three models

model is the first thing to do. This study chooses BPNN, random forest and GBDT as based model. Model tuning suggests those 3000 trees and 0.005 of shrinkage for GBDT model. And model parameters for BPNN are 30 hidden nodes and 100 iterations. After model parameters are determined, testing model will be next to do. To accurately analyze these three based models, 5-fold cross validation is used to test these models. And freezing point is the prediction object.

The test results are showed in Fig. 3 and 4. In training dataset, the performance of BPNN is the worst among these three models and the performance of

Table 1: Training and testing result for three models

Prediction method	CV train RMSE	CV train R2	CV test RMSE	CV test R2
BP NN	0.005	0.970	0.015	0.948
GBDT	1.30×10^{-4}	0.995	0.010	0.957
Random forest	6.53×10^{-4}	0.989	0.011	0.955

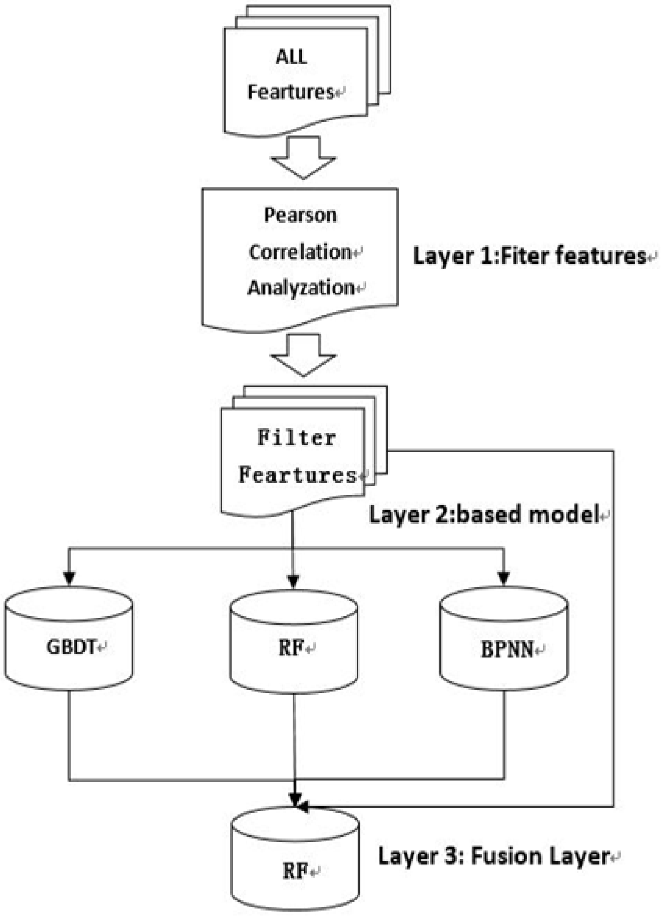

Fig. 5: Structure of fusion model

GBDT is almost the same as that of random forest. In testing dataset, the performance of these three models decreases. But performance of tree model is still better than that of BPNN. To accurately compare performance of these models, the average r-square of 5-fold cross validation is showed in Table 1.

In Table 1, all of three models have good performance when they model the relationship between physicochemical properties and freezing point. By comparing the R-square value between train set and test set, Over-fitting problem is found in tree model and the problem is worse in GBDT. All three based models are analyzed above. To increase accuracy of prediction, this study builds up fusion model.

Figure 5 shows the structure of fusion model. It uses Pearson Correlation Analysis to choose features as predictors at layer 1 because the correlation between physicochemical properties and thermal properties is an important influence factor for model performance. At layer 2, it uses three based models mentioned above and filter features to predict thermal properties. And the output of these three models will be added to input features of layer 3. At layer 3, it uses random forest to fuse all the features include filter features and the output of based models. Multi layers have better non linearity and it will dig up more information for prediction. But in this study, GBDT is removed from layer 2 after experiments because its over-fitting problem reduces the performance of fusion model rather than increase the accuracy of prediction.

RESULTS AND DISCUSSION

Table 2. Pearson Correlation for global dataset of physicochemical properties and thermal properties of fruits and vegetables from 700 sets of original data. Numbers stay for the Pearson correlation coefficient r, while stars are for p-value.

Table 2: Pearson correlation in the global dataset

Properties	Density	Water content	Solids content	Thermalconductivity	Freezing point
Density	1	-0.426***	0.266***	0.696***	-0.188***
Water content	-0.426***	1	-0.707***	0.045	0.679***
Solids content	0.266***	-0.707***	1	-0.036	-0.867***
Thermal conductivity	0.696***	0.045	-0.036	1	0.191***
Freezing point	-0.18***	0.679***	-0.867***	0.191***	1

The symbolic meaning of asterisk: *** p-value<0.001. ** p-value<0.01. * p-value<0.05. p-value<0.1

Table 3: 5-fold crossed validation for fusion model to predict thermal conductivity

Prediction performance	CV1	CV2	CV3	CV4	CV5	Average
Train set R2	0.9909	0.9908	0.9904	0.9897	0.9892	0.9902
Test set R2	0.8726	0.8643	0.8995	0.9149	0.9182	0.8939
Train set RMSE	8.24×10^{-10}	7.91×10^{-10}	8.71×10^{-10}	1.09×10^{-9}	1.03×10^{-9}	9.22×10^{-10}
Test set RMSE	1.37×10^{-7}	2.00×10^{-7}	9.98×10^{-8}	4.90×10^{-8}	8.54×10^{-8}	1.14×10^{-7}

Table 4: 5-fold crossed validation for fusion model to predict freezing point

	CV1	CV2	CV3	CV4	CV5	Average
Train set R2	0.9969	0.9971	0.9966	0.9969	0.9969	0.9969
Test set R2	0.9697	0.9488	0.9691	0.9681	0.9619	0.9635
Train set RMSE	5.59×10^{-5}	5.15×10^{-5}	6.30×10^{-5}	5.25×10^{-5}	5.45×10^{-5}	5.55×10^{-5}
Test set RMSE	0.005	0.011	0.006	0.006	0.009	0.007

Table 2 shows Pearson Correlation in the global dataset. By analyzing the relationship between these properties, all prediction effect can be explained more clearly. Sensibility to correlated predictors depends on the used statistical learning technique, but it is generally not welcomed because redundant and non-informative inputs reduce model performance (Brillante *et al.*, 2015). When inference is the objective, the negative effect of correlated variables is even worse than for predictions alone. It's important to choose the right predictors because any change of the model input will change the structure of the random forest and GBDT. Water content and solid content have the high correlations with density (r values of _-0.42 and _0.26, respectively, p-value<0.001). And Thermal conductivity has the high correlations with density (r values of_0.696, p-value<0.001). High correlation between input models among predictors are not welcomed because Decision Tree and Forest models completely fail when correlated predictors are present (Maloney *et al.*, 2012; Strobl *et al.*, 2007). So at first, water content and solid content properties are removed from the predictors because they are highly correlated to density and their contribution to predict Thermal conductivity seems to be zero according to the Pearson correlation (p-value>0.05). But after experiment, the model with water content and solid content performs a little better than that without these two predictors. This may be because the weak correlation between Thermal conductivity and those two properties still has contribution to prediction model. But most of the contribution is still from density predictor. Therefore all of three predictors should be all retained to predict Thermal conductivity. Freezing point also shows high correlations with all three predictors (r values of _-0.18, _0.67 and _-0.86, respectively, p-value<0.001). There are three predictors with high correlation with freezing point while there is only one predictor with high correlation with Thermal conductivity. This will make the performance of freezing point prediction a lot better than that of thermal conductivity prediction.

Table 3 and 4 show the test result of fusion model by using 5-fold crossed validation. All 700 sets of data are randomly divided into 5 folds. Every crossed validation uses 4 folds to train the fusion model and 1 fold to test the fusion model. Because the major structure of fusion model is random forest. That means different data sets used to train will generate a totally different tree structure and prediction result may be changed a lot. In this study, R-square is used to evaluate performance of prediction model. All training R-square is up to 99% while testing R-square values for thermal conductivity and freezing point respectively are 89% and 96%. That means the fusion model tended to over-fit the training data set. The over-fitting problem in prediction of thermal conductivity is worse than that in prediction of freezing point. But Table 5 and 6 shows that over-fitting problem exists in every prediction model in this study. So the problem is from the data set rather than the fusion model itself. It's probably that data set is not large enough for prediction model when there is high correlation between predictors. And the over-fitting problem in prediction of thermal conductivity is worse because there is only one predictor with high correlation with outcome which means predictors are not rich enough. More physicochemical properties predictors of fruits and vegetables and more sample data sets may can fix this over-fitting problem. The lowest R-square is 86% and the highest R-square is 91% in test of Thermal conductivity prediction. That means different data set have influence on performance of fusion model and cross validation is necessary to estimate the model performance.

Table 5: Test result of prediction of thermal conductivity for all models (using 5-fold crossed validation)

Prediction method	CV train RMSE	CV train R2	CV test RMSE	CV test R2
BP NN	7.88×10^{-8}	0.910	1.95×10^{-7}	0.861
GBDT	2.16×10^{-9}	0.985	1.43×10^{-7}	0.880
Random forest	7.08×10^{-9}	0.972	1.27×10^{-7}	0.888
Fusion model	9.22×10^{-10}	0.990	1.14×10^{-7}	0.893

Table 6: Test result of prediction of freezing point for all models (using 5-fold crossed validation)

Prediction method	CV train RMSE	CV train R2	CV test RMSE	CV test R2
BP NN	0.005	0.970	0.015	0.948
GBDT	1.30×10^{-4}	0.995	0.010	0.957
Random forest	6.53×10^{-4}	0.989	0.011	0.955
Fusion model	5.55×10^{-5}	0.996	0.007	0.963

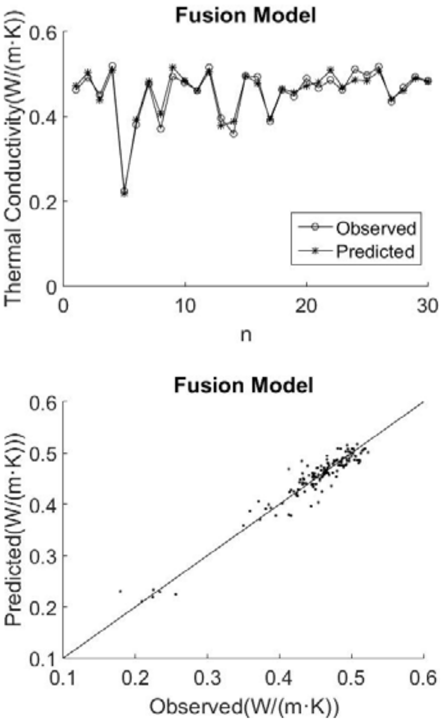

Fig. 6: Fusion model predicts thermal conductivity

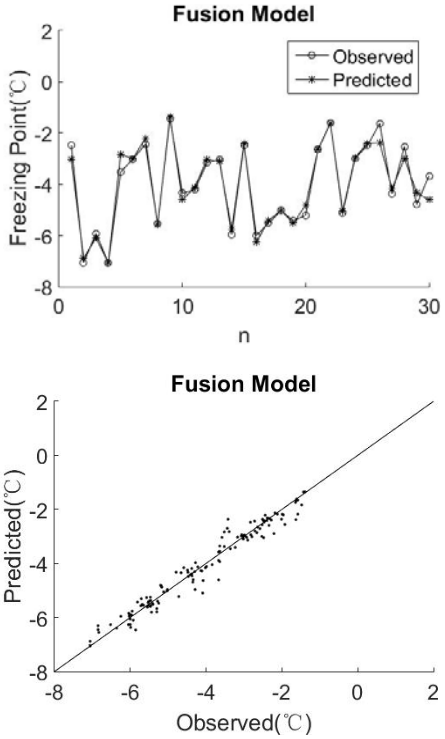

Fig. 7: Fusion model predicts freezing point

Table 5 and 6 show the results of the all models. Although over-fitting problem exists in these models, they still can accurately predict the test data set. The average R-square values of test data sets are considered as the proof of model performance. The performance of Fusion Model is the best, 89.3% (R-square) for prediction of Thermal conductivity and 96.3% (R-square) for prediction of Freezing point. All three predictors have high correlation with freezing point and that makes the prediction of freezing point have better performance. Performance of GBDT is almost the same as that of random forest. BP neural network has lowest R-square value. Random forest is simpler to perform and accurately tune than GBDT. And its training speed is faster than GBDT's. Furthermore, its over-fitting problem is slightly smaller than GBDT's. In this study, prediction result of GBDT model is considered as predictor for fusion model at first. But the fusion model with GBDT doesn't perform well because GBDT exacerbates the over-fitting problem of fusion mode. And also the R-square value of fusion model is reduced. That's why GBDT prediction result is removed from the predictors of fusion model.

Figure 6 and 7 show part of prediction results of fusion model by using test data set. They intuitively show the high accuracy of prediction result of fusion model. If most of the ratio between predicted values and observed values is nearly 1, the fusion model has good performance. The prediction results of thermal conductivity and freezing point both have high accuracy by using fusion model. And fusion model performs better than any one of those three models. But it's also true that the accuracy is not high enough for real application. To further increase the accuracy of prediction and fix the over-fitting problem, the model acquires more data sets and more physicochemical properties of fruits and vegetables.

CONCLUSION

In this study, this study collects a varied dataset of characteristics of fruits and vegetables. It uses three single models and a fusion model to model the relationship between physicochemical properties and thermo physical properties of fruits and vegetables. Random forest and GBDT models are used to predict thermo physical properties at first. And the performance is better than BP neural network. To achieve higher accuracy of prediction, this study builds up a fusion model by using random forest to fuse the prediction result which is generated by random forest and BP neural network. Fusion model has multilayer structure which has better non-linear performance. After using 5-fold cross validation to test all models, the result shows that fusion model has the best performance among 4 models (BPNN, Random Forest, GBDT and Fusion Model). And the prediction result of fusion model has high accuracy. By comparing the performance among 4 models, over-fitting problem is found in those models with tree structure. Tree structure model tends to over-fit training data while there is less of over-fitting problem with neural network. To fix this problem, a larger dataset and more variety of characteristics of fruits and vegetables are needed to optimize the tree structure model. Despite this, fusion model is an effective model to model the relationship between physicochemical properties and thermo physical properties. And it's an effective way to predict thermo physical properties of fruits and vegetables.

ACKNOWLEDGMENT

Thanks R development team provide R open source software and tools. Thanks Jinan University to provide the experimental environment.

Conflict of interest: We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

REFERENCES

Breiman, L., 2001. Random forests. Mach. Learn., 45(1): 5-32.

Brillante, L., F. Gaiotti, L. Lovat, S. Vincenzi, S. Giacosa, F. Torchio, S.R. Segade, L. Rolle and D. Tomasi, 2015. Investigating the use of gradient boosting machine, random forest and their ensemble to predict skin flavonoid content from berry physical–mechanical characteristics in wine grapes. Comput. Electron. Agr., 117: 186-193.

Cao, Z.F., 2014. Study on optimization of random forests algorithm. Ph.D. Thesis, Capital University of Economics and Business, Henan.

Cheng, Y.J., B. Liu, Q. Wang, J. Liu and R.X. Chai, 2006. Investigation and comparation of fruits and vegetables thermophysical property. Storage Process, 6(3): 26-28.

Jian, Y.J., 2013. Research on correlation rules mining algorithm based on matrix. M.S. Thesis, Lanzhou University, Gansu.

Maloney, K.O., M. Schmid and D.E. Weller, 2012. Applying additive modelling and gradient boosting to assess the effects of watershed and reach characteristics on riverine assemblages. Method. Ecol. Evol., 3(1): 116-128.

Sablani, S.S. and M. Shafiur Rahman, 2003. Using neural networks to predict thermal conductivity of food as a function of moisture content, temperature and apparent porosity. Food Res. Int., 36(6): 617-623.

Strobl, C., A.L. Boulesteix, A. Zeileis and T. Hothorn, 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8(1): 1-25.

Zhang, M., 2005. Study of thermal conductivity measurement system and temperature field simulation of postharvest fruits and vegetables. Ph.D. Thesis, Henan Agricultural University, Henan.

Zhang, M., Z.Y. Zhong, L. Yang, H.Z. Zhao, J.H. Chen and Z.H. Che, 2010. Prediction model of thermal conductivities of fruits and vegetables based on BP neural networks. T. CSAE, 41: 117-121.

Zhong, Z.Y., 2010. Relationships between thermo-physical properties and physiological and biochemical parameters of fruits and vegetables and neural network prediction model on thermo-physical properties. M.Sc. Thesis, Shanghai Ocean University, Shanghai.