Journal of Information Technology Research
Volume 15 • Issue 1

Multi-Class Classification of Agricultural Data Based on Random Forest and Feature Selection

Lei Shi, Henan Agricultural University, China
Yaqian Qin, Zhengzhou University of Science and Technology, China
Juanjuan Zhang, Henan Agricultural University, China
Yan Wang, Zhengzhou University of Science and Technology, China
Hongbo Qiao, Henan Agricultural University, China
Haiping Si, Henan Agricultural University, China*

ABSTRACT

Agricultural production and operation produce a large amount of data, which hides valuable knowledge. Data mining technology can effectively explore the connection between various factors from the massive agricultural data. Classification prediction is one of the most valuable agricultural data mining techniques. This paper presents a new algorithm consisting of machine learning algorithms, feature ranking method, and instance filter, which aims to enhance the capability of the random forest algorithm and better solve the problem of agricultural multi-class classification. The performance of the new algorithm was tested by using four standard agricultural multi-class datasets, and the experimental results showed that the newly proposed method performed well on all datasets. Among them, substantial rise in classification accuracy is observed for Eucalyptus dataset. Applying random forest algorithm on Eucalyptus dataset results in classification accuracy as 53.4%, and after applying the new algorithm (rough set), the classification accuracy significantly increases to 83.7%.

KEYWORDS

Data Mining, Feature Selection, Multi-Class Classification, Random Forest Algorithm, Rough Set

INTRODUCTION

Machine learning (ML) algorithms are essentially processes or sets of procedures that help a model adapt to the data given an objective. Applying machine learning to the process of modern agricultural production can effectively improve the development of modern agriculture, the automation and intelligence of agricultural production. Currently, machine learning algorithms have been successfully and widely used in crop yield prediction (Liu, et al. 2017), crop disease identification (Chaudhary, et al. 2016), agricultural management decision-making (Kassaye, et al. 2020) and other fields. In the prediction problem, the support vector machine (SVM), random forest (RF), artificial neural network (ANN) were utilized for crop yield prediction along with remote sensing, and achieved high accuracy for all cases (Stas, et al. 2016, Heremans, et al. 2015, Liang, et al. 2015). In the classification field, the naive bayes (NB), support vector machine (SVM), random forest (RF) have been successfully applied

DOI: 10.4018/JITR.298618 *Corresponding Author

to provide a solution on these topics, such as crop disease diagnosis (Hill, et al. 2014), agricultural product sorting (Kurtulmus, et al. 2014), and crop identification (Waleed, et al. 2021).

In the actual agricultural production process, the application of computer-related information technology in precision agriculture has become more and more extensive, a large quantity of the attribute data and spatial data closely related to the precision agricultural process have been acquired and accumulated. How to mine hidden relationships from massive agricultural production data, help decision-makers to make accurate agricultural strategies and guide agriculture efficient production is a very important and urgent issue. The classification of interesting agricultural data is often the first step in valuable mining information on agricultural data. Therefore, automatically classifying agricultural data is one of the most significant topics in the field of precision agriculture.

The random forest (RF) algorithm is a new and efficient combination classification method. Its basic idea is to integrate many weak classifiers into one strong classifier. Compared with traditional classifiers, RF has a good tolerance for outliers and noisy data, no over-fitting phenomenon, and good generalization ability (Zhang&Yang, 2020, P&Nair, 2021). Although the RF algorithm has many advantages, the large amount of data and the balance problem greatly affect the performance of the classifier. The large amount of data and imbalance are the challenges of current data classification. When classifying high-dimensional data, the resulting classifier is complex, and the data is prone to overfitting due to the large feature space. Feature selection can reduce the dimensionality of the data, so that the classifier can focus on important features, ignore possible misleading features, reduce computational complexity and improve classification performance. It has been widely used to improve the classification of high-dimensional data (Shi, et al. 2012, Silva, et al. 2013, EI-Bendary, et al. 2015, Rehman, et al. 2018). Instance filtering technology needs to be used in unbalanced data, when the potential value of unbalanced datasets is to be mined (Chaudhary, et al. 2016, Feng, et al. 2018). Rough set is a soft computing method for dealing with fuzzy and uncertain data. Feature selection based on rough set is one core research of the rough set theory. Its basic idea is to select the feature subset with the smallest number of features under the premise that the attribute discrimination ability of the original data is not changed. It eliminates irrelevant and redundant features and improves the performance of the classifier. In the past few decades, rough set has been widely used in classification and feature selection. A single method, such as RFC or rough set theory, is difficult to achieve the goal of accurate data classification, because each method has its own limitations. Therefore, the paper proposed a new algorithm for efficiently catching up with the classification tasks of the agricultural data, which based on random forest and feature selection. The newly method is composed of the computer technology, namely an attribute evaluator method of Gain Ratio, rough set, an instance filter method, random forest algorithm.

The main content of this paper includes: Section 2 introduces the related methods used in this paper. Section 3 describes a newly proposed algorithm for solving the multi-class classification tasks. Section 4 reports the experimental results and analysis based on the four standard agriculture datasets. The last section summarizes this paper and draws the main directions for our next work.

BACKGROUND

Feature Selection

The feature selection phase, also called attribute selection or feature ranking is applied to datasets for choosing a subset or ranking of relevant features. Gain Ratio and rough set are common and more classic attribute selection methods. Hence, an attribute evaluator from Gain Ratio and rough set theory is chosen and used in the design of the proposed approach.

Gain Ratio Gain Ratio (Hall and Smith, 1998) is one of the most popular method to optimize feature selection. For a feature, the amount of information will change when in the amount of information is the amount of information that the feature brings to the system, that is, the information

gain (Eissa, et al. 2016). The Gain Ratio is an extension of information gain, which is the ratio of the information gain to the information entropy of the feature. In information theory, the amount of information is "entropy", and the gain ratio is calculated as the formula:

$$GainRatio\left(\mathrm{C},\mathrm{T_i}\right)=\left[\mathrm{H}\left(\mathrm{C}\right)-{\mathrm{H}\left(\mathrm{C}|\mathrm{T_i}\right)}\Big/{\mathrm{H}\left(\mathrm{T_i}\right)}\right]$$

where C represents the category, $\mathrm{T_i}$ represents the feature, H(C) is the entropy of the C class, H($\mathrm{T_i}$) is the entropy of the feature $\mathrm{T_i}$, and H(C| $\mathrm{T_i}$) is the entropy of the C class given the $\mathrm{T_i}$ feature condition.

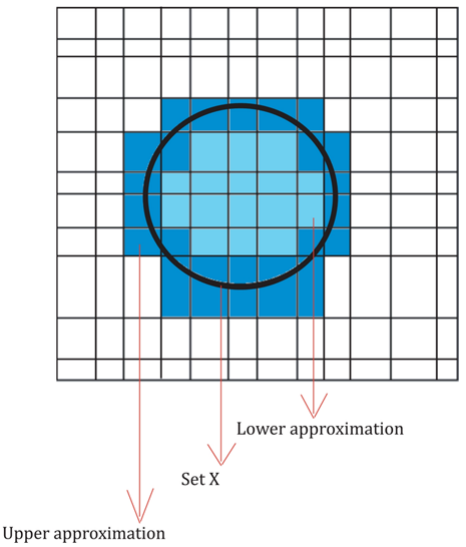
Rough Set Rough set theory (Pawlak, 1982) is a mathematical tool that can quantitatively analyze and deal with fuzzy, uncertain and incomplete information. It has been widely used in the field of machine learning, such as decision analysis, data mining and knowledge discovery (Shakiba & Hooshmandasl, 2016, Huang, et al. 2017, Chen, et al. 2017). The principle of rough set attribute reduction algorithm is to delete redundant knowledge, that is, redundant attributes, while keeping the classification results unchanged. This section mainly introduces the rough set theory related to the current work.

Definition 1. Information system. The information system is formulated as a 4-tuple as follows:

$$IS=\left(U,A,V,f\right):$$

where U is a set of finite objects; A is a nonempty finite set of attributes characterizing objects; Q is a set of finite properties, divided into a set of conditional attributes C and a set of decision attributes D, $\mathrm{Q}=\mathrm{C}\bigcup\mathrm{D}$, $\mathrm{C}\bigcap\mathrm{D}=\Phi$; $\mathrm{V}=\mathrm{U}_{\mathrm{a}\in\mathrm{A}}\mathrm{V}_{\mathrm{a}}$, which is a collection of attribute values, V_{a} represents the range of attributes $\mathrm{a}\in\mathrm{Q}$; f: $\mathrm{U}\times\mathrm{A}\rightarrow\mathrm{V}$ is called an information function.

Figure 1. Lower and upper approximation



Definition 2. Indiscernibility relation. In the information system IS, for each attribute subset, the indistinguishable relationship of the B construct is defined as follows:

$$IND\left(\mathrm{B}\right)=\left\{\left(\mathrm{x},\mathrm{y}\right)\epsilon\mathrm{U}^{2}\mid\forall\epsilon\mathrm{B},\mathrm{B}\left(\mathrm{x}\right)=\mathrm{b}\left(\mathrm{y}\right)\right\}$$

Definition 3. Lower approximation and Upper approximation. Figure 1provides a schematic diagram of a rough set X within the upper and lower approximations. In the information system IS, let attribute set $x\in U$ and R be an equivalence relationship. The R lower approximation of x and the R upper approximation of x are denoted by:

$$R_{x^{-}}=\left\{Y\in U\Big/_{R}\mid Y\subseteq x^{-}\right\}$$
$$R_{x^{+}}=\left\{Y\in U\Big/_{R}\mid Y\cap x^{+}\neq 0\right\}$$

Definition 4. Equivalence class. Let $\mathrm{X}\subseteq\mathrm{U}$ and attribute subset $\mathrm{B}\subseteq\mathrm{A}$, X be approximated by B-lower approximation and B-upper approximation. The equivalence classes of the indiscernibility relation $\mathrm{IND}\left(\mathrm{B}\right)$ defined as follows:

$$\left[x\right]_{b}=\left\{y\middle|\left(x,y\right)\in IND\left(B\right),y\in U\right\}$$

Instance Filter - Simple Random Sample

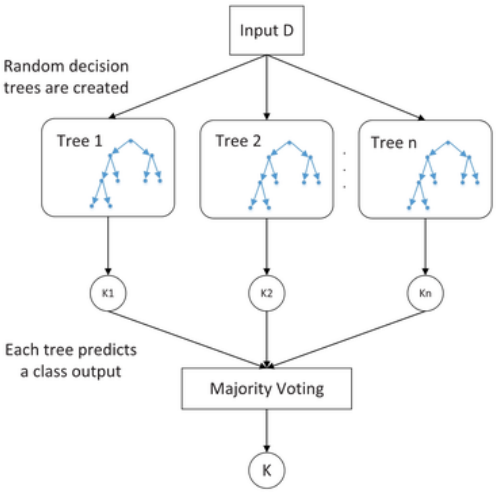
The data collected from practical applications is often unbalanced. In other words, the sample distribution of the classes in the data is uneven, which greatly affects the classification results. When the data is unbalanced, the traditional classification algorithm that takes the overall classification accuracy as the learning objective will pay much attention to the majority class and often ignore the minority class, which is likely to cause judgment errors and falsely high results accuracy. Therefore, dealing with unbalanced data is an important and arduous task in data mining. Resampling is one of the most important filtering techniques (Ismail, et al. 2016). Since the phenomenon of data imbalance is common in many current problems, rebalancing the sample space is a common initial step before executing data mining algorithms. There are currently two resampling methods that can rebalance the data: with replacement and without replacement. The difference between the two is the number of sample selections. The research results show that resampling with replacement can improve the classification accuracy of machine learning algorithms (Chaudhary, et al. 2016, Khaldy, et al. 2020). Therefore, in this paper, the resampling method with replacement is used for instance filtering.

Random Forest Algorithm

The random forest (RF) algorithm takes decision trees as the basic unit, and integrates multiple decision trees through the idea of ensemble learning. In essence, it is an ensemble learning algorithm based on machine learning, which overcomes many shortcomings of a single classifier and has good accuracy. The idea of the random forest algorithm is to use the Bootstrap re-sampling method to extract multiple samples from the original samples, and to construct a decision tree on the extracted samples respectively, and then combine the predictions of multiple decision trees, and finally obtain the final prediction result through the voting method. The specific working principle of the random forest classifier is shown in Figure 2. Among them, D is the input training sample, Tree1, Tree2,

..., Treen are randomly generated decision trees, {K1, K2, ..., Kn} are the output categories of each decision tree, and K is the last output of the sample determined by majority voting.

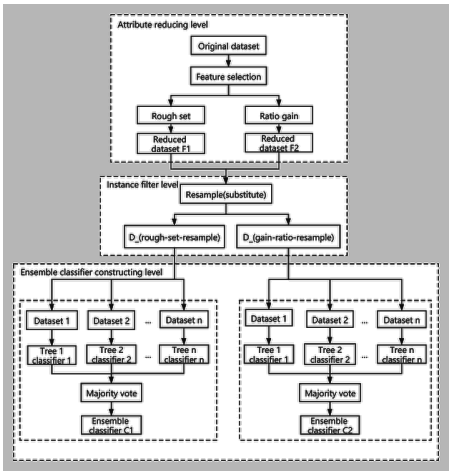
Figure 2. Architectural design of the random forest Classification algorithm



THE PROPOSED APPROACH

In this section, the proposed approach is introduced in detail. It consists of three levels, i.e., attribute reducing, instance filter and ensemble constructing. Attribute reducing includes rough set and ratio gain. The architectural design of the proposed approach is shown in Figure 3. Among them, the important algorithm is specifically described as follows.

Figure 3. Architectural design of the proposed approach



In the first attribute reducing level, the original dataset is used to construct a decision table, and then two algorithms, rough set and ratio gain, are used to select the features of the decision table respectively. The purpose is to remove redundant attributes in the dataset. The algorithm is described as follows:

Algorithm 1: Attribute reducing based on rough set

Input: Original agricultural dataset D
Output: Reduction dataset $F1$
1) Read the Original agricultural dataset D in decision information table IS
2)Build indiscernibility matrix $M\left(IS\right)$
3)Reduce M using absorbtion laws
4)Obtain d no-empty fields of reduced M
5)Build families of sets
6)Remove dispensable attributes from each element of family sets
7)Remove redundant elements from families of sets
8)Obtain the reduction dataset $F1$

Algorithm 2: Attribute reducing based on ratio gain

Input: D , the training dataset, which contains a set of training examples and their related class labels, the total number of features is n
Output: Reduction dataset $F2$
1)Compute information entropy of D $Entropy\left(D\right)$
2)for i =1 to D do:
Compute information entropy of D $Entropy\left(D,D_i\right)$
Compute information entropy of D_i $IG\left(D_i\right)=Entropy\left(D\right)-Entropy\left(D,\ D_i\right)$
Compute the total amount of information for D_i $I\left(D_i\right)$
Define the information gain rate of feature D_i as $IGR\left(D_i\right)$
If $I\left(D_i\right)=0$ then
Return

Algorithm 2 continued on next page

Algorithm 2 continued

else
compute $IGR\left(D_i\right)$
3) Count the information gain rate values of all features D_i and store them in the dictionary, assuming $Key = D_i$, $value = IGR\left(D_i\right)$, $dict\left[Key\right] = value$
4) Sort the array in descending order, filter the remaining features
5) Obtain the reduction dataset $F2$
In the two level, the resampling method with replacement is used to solve the problem of data imbalance, the results are D_(gain-ratio-resample) and D_(rough-set-resample). The basic idea is to re-sampling while keeping the category distribution of the sub-sample unchanged.
In the three level,the random forest algorithm is used for classification. First, the random forest classifier is constructed on the reduction training dataset obtained in the two level. The final classification result is determined by combining the prediction results of the decision tree. The specific description of the algorithm is as Algorithm 3.

Algorithm 3: Random forest constructing algorithm

Input: D_(gain-ratio-resample) and D_ (rough-set- resample)
Output: Ensemble classifier C1 C2
Generate c bootstrap samples in the following way:
for $i = i$ to c do:
Randomly sample the reduced training dataset with replacement to produce D_i
Create a root node, N_i containing D_i
Call BuildTree(N_i)
end for
BuildTree(N):
If N contains instances of only one class then
return
else
Randomly select x% of the possible splitting features in N
Select the feature F with the highest information gain to split on
Create f child nodes of N, N_1, N_2,..., N_f, where F has f possible values(F_1, F_2,..., F_f)

Algorithm 3 continued on next page

Algorithm 3 continued

for $i = 1$ to f do
Set the contents of N_i to D_i, where D_i is all instances in N that match F_i
Call BuildTree(N_i)
end for
end if

EXPERIMENT RESULT AND DISCUSSION

This article uses two data analysis tools, Weka and ROSE. Weka is an open work platform of data mining that integrates a large number of machine learning algorithms that can undertake data mining tasks. It can not only perform data preprocessing, classification, regression and other processing operations, but also evaluate algorithm performance and result visualization functions. ROSE is a microcomputer software designed to analyze data by means of the rough set theory. In the course of this experiment, the parameter settings of two tools are using default values. In order to verify the performance of the newly proposed method in the classification of agricultural data, four standard agriculture datasets are selected for verification, namely Eucalyptus, Pasture, White-clover and Squash-stored, which came from agricultural researchers in New Zealand (http://www.cs.waikato.ac.nz/ml/weka/index.html) and UCI machine learning repository (http://archive.ics.uci.edu). A detailed description of these standard agriculture datasets is shown in Table 1.

Table 1. Description of standard agriculture datasets

Datasets	Classes	Attribute Type	Instance	No.of Attribute
Eucalyptus	5	Numeric and Nominal	736	20
Pasture	3	Nominal	36	23
White-clover	4	Numeric and Nominal	63	32
Squash-stored	3	Numeric and Nominal	52	25

Performance Evaluation Indices

The selection of the evaluation indices of the classification model is an important part of the classification problem research. Choosing the appropriate evaluation indices can objectively and accurately evaluate the performance of the classification model. In the experiment, four evaluation indices of accuracy, precision, F-measure and AUC are selected to evaluate the performance of the newly proposed method in the application of agricultural data classification. The evaluation index is calculated by the confusion matrix and defined as follows. In the two classification problems, the samples can be divided into four situations: true positive (TP), false positive (FP), true negative (TN) and false negative (FN) based on the combination of their true category and the learner's predicted

category. Among them, TP, FP, FN, and TN respectively represent the number of relevant samples, which the total number of samples is N. The "confusion matrix" of the classification results is shown in Table 2.

Table 2. Confusion matrix

		Predicted label	
		Belong	**Not Belong**
Real label	Belong	TP	FN
	Not belong	FP	TN

True Positive (TP): the number of positive examples that are correctly classified, that is, the number of instances that are actually positive and are classified as positive by the classifier.

True Negative (TN): the number of false positives, that is, the number of instances that are actually negative but are classified as positive by the classifier.

False Positive (FP): the number of false negatives, that is, the number of instances that are actually positive but are classified as negative by the classifier.

False Negative (FN): the number of negative examples that are correctly classified, that is, the number of cases that are actually negative and are classified as negative by the classifier.

$$TP + FP + TN + FN = N$$

The four evaluation indices are specifically defined as follows: The accuracy of the classifier is the most common evaluation index, that is, the number of samples to be matched divided by the number of all samples. Generally speaking, the higher the correct rate, the better the classifier.

$$Acc = \frac{TP + TN}{TP + FP + TN + RN}$$

Precision is a measure of precise, which represents the rate of actual True Positive among the positive instance.

$$Pre = \frac{TP}{TP + FP}$$

Recall is a measure of coverage, it measures the number of Ture Positive that are correctly classified as positive cases.

$$Rec = \frac{TP}{TP + FN}$$

F-measure is the weighted harmonic average of Pre and Rec, which combines the results of precision and recall to make the evaluation more comprehensive and accurate. When the value of F-1 is high, the experimental method is ideal

$$F-1 = 2 \times \frac{Pre \times Rec}{Pre + Rec}$$

The receiver operating characteristic (ROC) curve can directly reflect the diagnostic ability of the classifier without considering other issues such as class distribution or error cost. It is defined as a plot of the true positive rate (TPR) against the false positive rate (FPR) (Charles, 1978). In ROC space, the TP rate and FP rate (formulas such as ERP and TPR) are used as the horizontal and vertical coordinates. The closer the curve is to the upper left corner, the better the performance of the classifier is. Hence, ROC curve is a true representative of classifier performance (Wang, et al. 2014).

$$ERP = \frac{FP}{TN + FP}$$

$$TPR = \frac{TP}{TN + FP}$$

The area under the curve (AUC) represents the degree or measure of separability and tells how much the model is capable of distinguishing between classes. The score of AUC is between 0 and 1 for evaluating the performance of the model. AUC has been found to be more sensitive in the analysis of variance tests and independent to the decision threshold. The larger the value of AUC is, the better the model can distinguish between positive and negative instance.

RESULT AND DISCUSSION

In order to evaluate the capabilities and effectiveness of the proposed method, two relatively popular machine learning algorithms C4.5 and naive bayes (NB) are selected for comparative research, and the cross-validation method is used to evaluate the model to test the performance of the algorithm. The specific process is to first divide the experimental datasets into 10 equally, and take turns using 9 of them as training data and 1 as test data. Repeatedly obtain ten training results under the same parameter settings and operation steps, and use the average value as the experimental result. Previous studies have proved that ten-fold cross validation can be better for evaluating the performance of the machine learning algorithms (Sun, et al. 2018, Roimi, et al. 2020).

Table 3 shows the number and proportion of attributes in the dataset after feature selection. The retained attributes are only those that have a positive impact on classification. Among them, the Pasture dataset reduces one unrelated attribute through the rough set theory, indicating that the Pasture dataset attribute is highly correlated with the class label compared with the other three datasets. It also provides proof that the random forest classifier performed best for the Pasture dataset relative to the remaining datasets. The results in Table 3-5 apparently indicated that the proposed methodology not only reduces the number of attributes effectively but also can improve classification performance considerably.

This paper uses four indices of accuracy, precision, F-measure and AUC to evaluate the performance of the newly proposed method, and compares it with two popular machine learning algorithms, C4.5 and naive bayes (NB). The test results are shown in Table 4-5. The results show that the four indices obtained by the newly proposed method in the application of the four standard

agricultural datasets are all higher than the C4.5 and naive bayes (NB). Among them, substantial rise in classification accuracy is observed for White-clover dataset. Applying naive bayes (NB) on White-clover dataset results in classification accuracy as 57.1% and after applying the new algorithm the classification accuracy significantly increases to 92.1%.

Table 4 shows the performance index values for random forest algorithm and the proposed approach using four standard agriculture datasets. Figure 4-7 shows the results of the newly proposed method using different feature selection methods to classify and the comparison method random forest algorithm classification results. Obviously, the proposed method has a good classification result on the four standard agriculture datasets. The feature selection method using rough set theory is higher, but the difference between the two is not obvious, and both are higher than the random forest algorithm. Among them, substantial rise in classification accuracy is observed for Eucalyptus dataset. In the Eucalyptus dataset, the accuracy of the proposed method is 82.3% (gain ratio) and 83.7% (rough set), which is at least 28.9% higher than random forest (RF) algorithm. Moreover, the gap between the two is not obvious, only 1.4%, and the feature selection method using rough set theory is higher. Slight rise in classification accuracy is observed for Pasture dataset. In the Eucalyptus dataset, the accuracy of the proposed method is 97.2% (gain ratio) and 100% (rough set), which is at least 13.9% higher than random forest (RF) algorithm. The paper has shown that the classification performance of the integrated algorithm compared with a single random forest algorithm is significantly improved, and it has good applicability in the application of multi-class classification of agricultural data.

Table 3. Number of retained attributes for the proposed algorithm.

Datasets	Number of original attributes	Number of retained attributes	Percentage (%)
Eucalyptus	20	15	75.0
Pasture	23	22	95.7
White-clover	32	28	87.5
Squash-stored	25	23	92.0

Table 4. The performance index values for random forest algorithm and the proposed approach using benchmark datasets

Benchmark datasets	Performance indices	Random forest algorithm	Proposed approach	
			Gain Ratio	Rough set
Eucalyptus	Accuracy	0.534	0.823	0.837
	Precision	0.529	0.825	0.838
	F-measure	0.531	0.824	0.837
	AUC	0.829	0.959	0.965
Pasture	Accuracy	0.833	0.972	1.000
	Precision	0.825	0.974	1.000
	F-measure	0.833	0.972	1.000
	AUC	0.911	0.998	1.000

Table 4 continued on next page

Table 4 continued

Benchmark datasets	Performance indices	Random forest algorithm	Proposed approach	
			Gain Ratio	Rough set
White-clover	Accuracy	0.667	0.905	0.921
	Precision	0.669	0.911	0.923
	F-measure	0.652	0.904	0.918
	AUC	0.708	0.989	0.971
Squash-stored	Accuracy	0.577	0.769	0.808
	Precision	0.628	0.788	0.823
	F-measure	0.566	0.758	0.799
	AUC	0.717	0.906	0.902

Table 5. Comparison of the performance index values of prediction on datasets

datasets	Performance indices	A proposed approach (rough set)	C4.5	NaiveBayes
Eucalyptus	Accuracy	0.837	0.637	0.556
	Precision	0.838	0.651	0.628
	F-measure	0.837	0.634	0.559
	AUC	0.965	0.842	0.863
Pasture	Accuracy	1.000	0.778	0.722
	Precision	1.000	0.772	0.724
	F-measure	1.000	0.772	0.722
	AUC	1.000	0.841	0.837
White-clover	Accuracy	0.921	0.635	0.571
	Precision	0.923	0.649	0.635
	F-measure	0.918	0.638	0.574
	AUC	0.971	0.686	0.706
Squash-stored	Accuracy	0.808	0.654	0.615
	Precision	0.823	0.660	0.630
	F-measure	0.799	0.646	0.614
	AUC	0.902	0.711	0.803

Figure 4. Performance graph of random forest algorithm, a new algorithm (Gain Ratio) and a new algorithm (rough set theory) for identification of Eucalyptus.

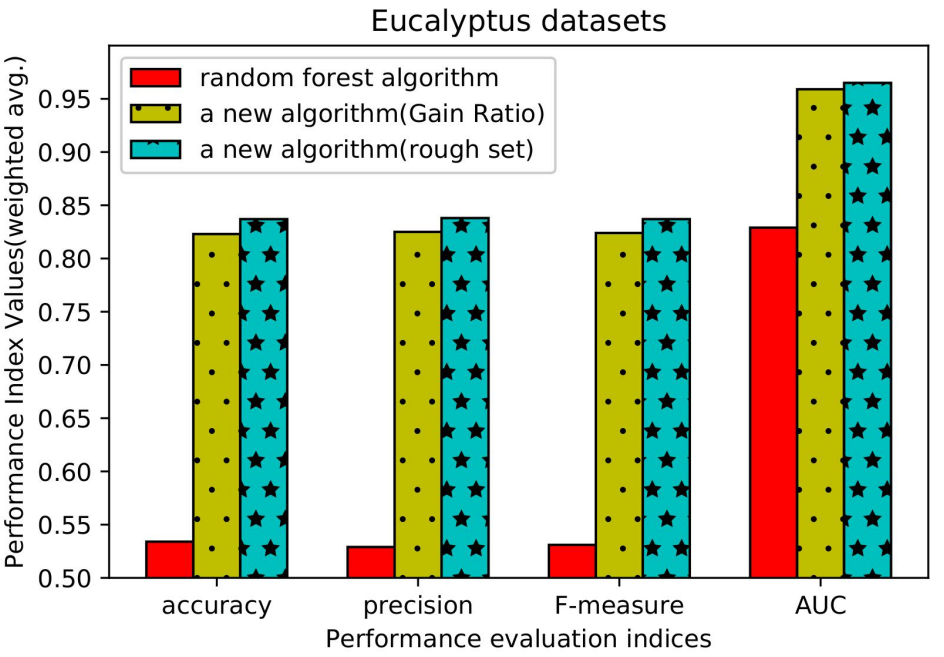


Figure 5. Performance graph of random forest algorithm, a new algorithm (Gain Ratio) and a new algorithm (rough set theory) for identification of Pasture.

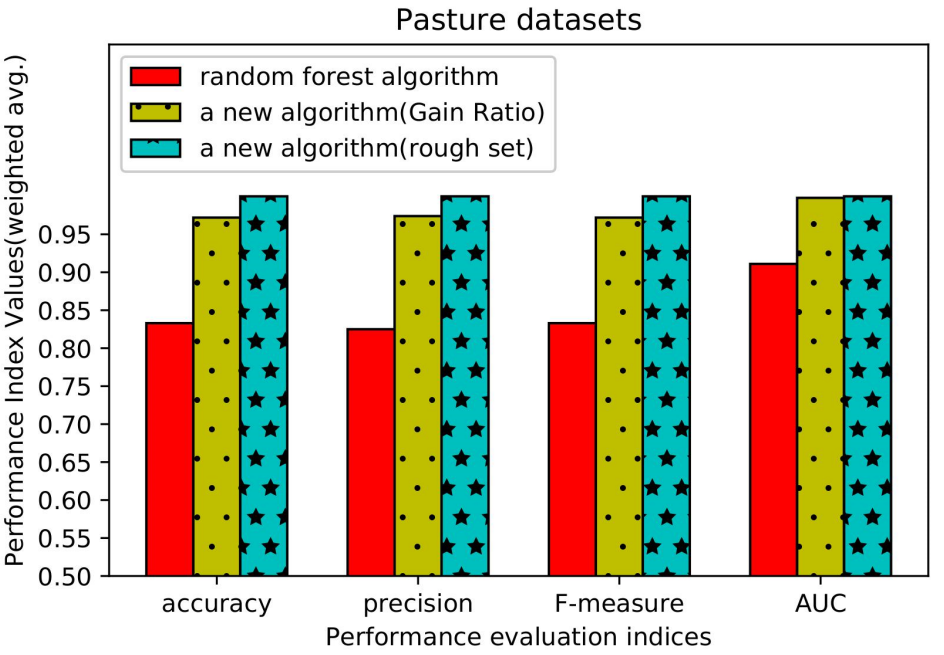


Figure 6. Performance graph of random forest algorithm, a new algorithm (Gain Ratio) and a new algorithm (rough Set theory) for identification of White clover.

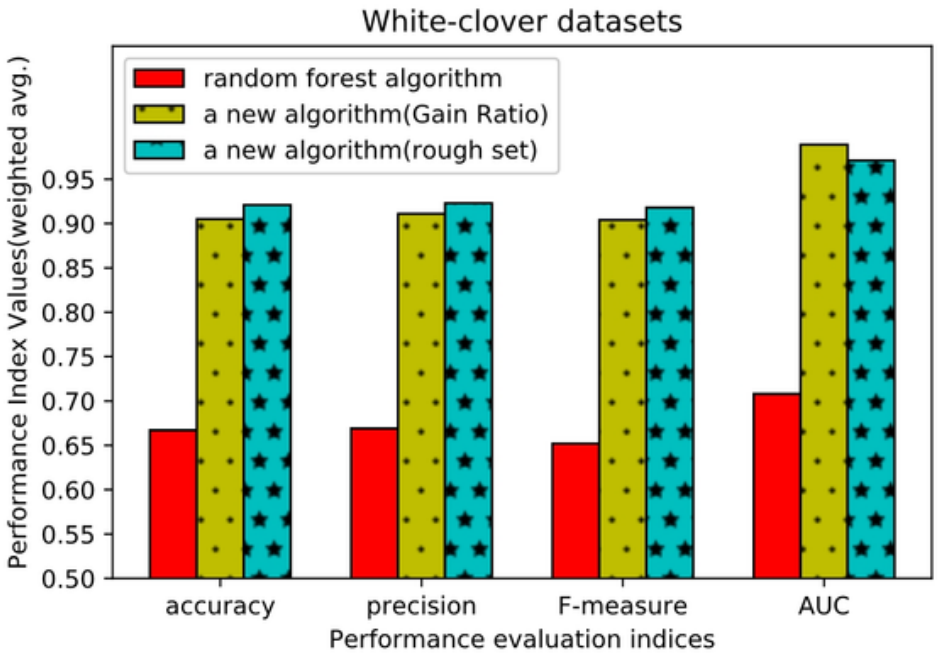
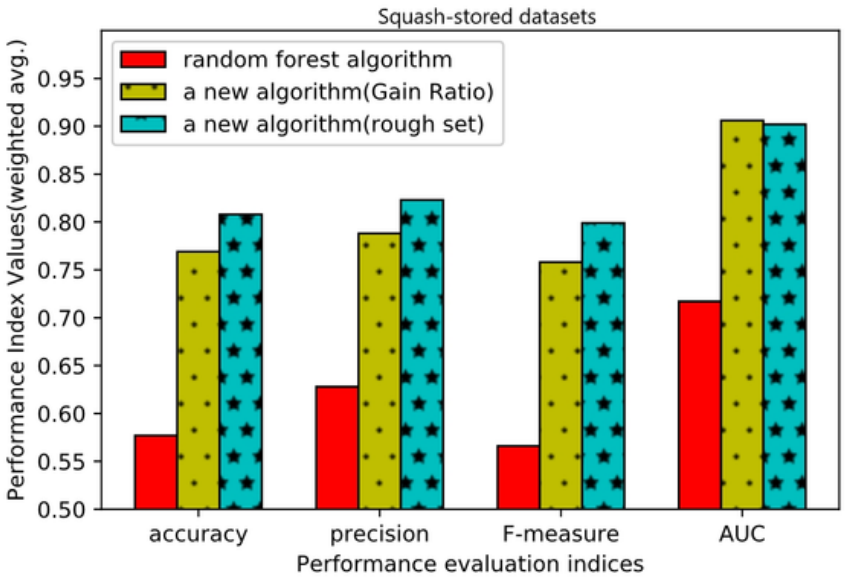


Figure 7. Performance graph of random forest algorithm, a new algorithm (Gain Ratio) and a new algorithm (rough set theory) for identification of Squash-stored.



CONCLUSION

multi-class classification problem was a hot topic and had great potential applications in precision agriculture. This paper proposed multi-class classification algorithm based on random forest, Gain Ratio and rough set theory. In the experiment, four standard agricultural multi-class datasets were used to test the performance of the proposed algorithm. The results show that the proposed algorithm achieves significant performance improvement. Ensemble learning and reinforcement learning are powerful techniques and may have good results in the algorithm. Future work will continue to improve the performance of the proposed algorithm with ensemble learning and reinforcement learning.

CONFLICTS OF INTEREST

The authors declare there is no conflicts of interest regarding the publication of this paper.

FUNDING AGENCY

Publisher has waived the Open Access publishing fee

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (Grant No.31501225), the Modern Agriculture Industry Technology System in Henan Province (Grant No.S2010-01-G04), the Innovative Training Program for College Students in Henan Province (Grant No.201910466018). We also thank to the editor's hard work and anonymous referees' comments and suggestions on this paper.

REFERENCES

Chaudhary, A., Kolhe, S., & Kamal, R. (2016). A hybrid ensemble for classification in multi-class datasets: An application to oilseed disease dataset. *Computers and Electronics in Agriculture*, *124*, 65–72. doi:10.1016/j.compag.2016.03.026

Chen, H. M., Li, T. R., Cai, Y., Luo, C., & Fujita, H. (2016). Parallel attribute reduction in dominance-based neighborhood rough set. *Information Sciences*, *373*, 351–368. doi:10.1016/j.ins.2016.09.012

Eissa, M. M., Elmogy, M., & Hashem, M. (2016). Rough-granular computing knowledge discovery models for medical classification. *Egyptian Informatics Journal*, *17*(3), 265–272. doi:10.1016/j.eij.2016.01.001

El-Bendary, N., Hariri, E. E., Hassanien, A. E., & Badr, A. (2015). Using machine learning techniques for evaluating tomato ripeness. *Expert Systems with Applications*, *42*(4), 1892–1905. doi:10.1016/j.eswa.2014.09.057

Feng, W., Huang, W. J., Ye, H. C., & Zhao, L. (2018). Synthetic minority over-Sampling technique based rotation forest for the classification of unbalanced hyperspectral data. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*.

Hall, M. A., & Smith, L. A. (1998). Practical feature subset selection for machine learning. *Proceedings of 21st Australasian Computer Science Conference (ACSC'98)*, 181-191.

Heremans, S., Dong, Q., Zhang, B., Bydekerke, L., & Orshoven, J. V. (2015). Potential of ensemble tree methods for early-season prediction of winter wheat yield from short time series of remotely sensed normalized difference vegetation index and in situ meteorological data. *Journal of Applied Remote Sensing*, *9*(1), 097095. doi:10.1117/1.JRS.9.097095

Hill, M. G., Connolly, P. G., Reutemann, P., & Fletcher, D. (2014). The use of data mining to assist crop protection decisions on kiwifruit in New Zealand. *Computers and Electronics in Agriculture*, *108*, 250–257. doi:10.1016/j.compag.2014.08.011

Huang, Y., Li, T., Luo, C., Fujita, H., & Horng, S. J. (2017). Dynamic variable precision rough set approach for probabilistic set-valued information systems. *Knowledge-Based Systems*, *122*, 131–147. doi:10.1016/j.knosys.2017.02.002

Ismail, R., Abuelenin, S., & Aboelfetouh, A. (2016). Improving classification performance by using feature selection with resampling. *International Journal of Soft Computing*, *11*(4), 255–269.

Kassaye, K. T., Boulange, J., Lam, V. T., Saito, H., & Watanabe, H. (2020). Monitoring soil water content for decision supporting in agricultural water management based on critical threshold values adopted for Andosol in the temperate monsoon climate. *Agricultural Water Management*, *229*, 229. doi:10.1016/j.agwat.2019.105930

Khaldy, M. A., Alauthman, M., Alsanea, M., & Samara, G. (2020). Improve class prediction by balancing class distribution for diabetes dataset. *International Journal of Scientific & Technology Research*, *9*(4), 823–827.

Kurtulmus, F., Lee, W. S., & Vardar, A. (2014). Immature peach detection in colour images acquired in natural illumination conditions using statistical classifiers and neural network. *Precision Agriculture*, *15*(1), 57–79. doi:10.1007/s11119-013-9323-8

Liang, L., Di, L. P., Zhang, L. P., Deng, M. X., Qin, Z. H., Zhao, S. H., & Lin, H. (2015). Estimation of crop LAI using hyperspectral vegetation indices and a hybrid inversion method. *Remote Sensing of Environment*, *165*, 123–134. doi:10.1016/j.rse.2015.04.032

Liu, S., Cossell, S., Tang, J., Dunn, G., & Whitty, M. (2017). A computer vision system for early stage grape yield estimation based on shoot detection. *Computers and Electronics in Agriculture*, *137*, 88–101. doi:10.1016/j.compag.2017.03.013

Metz, C. E. (1978). Basic Principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*(4), 283–298. doi:10.1016/S0001-2998(78)80014-2 PMID:112681

P, P. K., V, M. A. B., & Nair, G. G. (2021) An efficient classification framework for breast cancer using hyper parameter tuned Random Decision Forest Classifier and Bayesian Optimization. *Biomedical Signal Processing and Control, 68*(3).

Pawlak, Z. (1982). Rough sets. *International Journal of Computer and Information Sciences*, *11*(5), 341–356. doi:10.1007/BF01001956

Rehman, T. U., Mahmud, M. S., Chang, Y. K., Jin, J., & Shin, J. (2018). Current and future applications of statistical machine learning algorithms for agricultural machine vision systems. *Computers and Electronics in Agriculture*, *156*, 585–605. doi:10.1016/j.compag.2018.12.006

Roimi, M., Neuberger, A., Shrot, A., Paul, M., & Bar-Lavie, Y. (2020). Early diagnosis of bloodstream infections in the intensive care unit using machine-learning algorithms. *Intensive Care Medicine*, *46*(3), 454–462. doi:10.1007/s00134-019-05876-8 PMID:31912208

Saez, J., Wozniak, M., & Krawczyk, B. (2016). Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets, Pattern Recognition. *The Journal of the Pattern Recognition Society*, *57*, 164–178. doi:10.1016/j.patcog.2016.03.012

Shakiba, A., & Hooshmandasl, M. R. (2016). Neighborhood system s-approximation spaces and applications. *Knowledge and Information Systems*, *49*(2), 749–794. doi:10.1007/s10115-015-0913-9

Shi, L., Ma, X. M., Duan, Q. G., Weng, M., & Qiao, H. (2012). Agricultural Data Classification Based on Rough Set and Decision Tree Ensemble. *Sensor Letters*, *10*(1), 271–278. doi:10.1166/sl.2012.1857

Silva, L., Koga, M. L., Cugnasca, C. E., & Costa, A. (2013). Comparative assessment of feature selection and classification techniques for visual inspection of pot plant seedlings. *Computers and Electronics in Agriculture*, *97*, 47–55. doi:10.1016/j.compag.2013.07.001

Stas, M., Orshoven, J. V., Dong, Q., Heremans, S., & Zhang, B. (2016). A comparison of machine learning algorithms for regional wheat yield prediction using NDVI time series of SPOT-VGT. *2016 5th International Conference on Agro-geoinformatics (Agro-geoinformatics)*.

Sun, B. Z., Lam, D., Yang, D. S., Grantham, K., Zhang, T. Z., Mutic, S., & Zhao, T. Y. (2018). A machine learning approach to the accurate prediction of monitor units for a compact proton machine. *Medical Physics*, *45*(5), 2243–2251. doi:10.1002/mp.12842 PMID:29500818

Waleed, M., Um, T. W., Kamal, T., & Usman, S. M. (2021). Classification of agriculture farm machinery using machine learning and internet of things. *Symmetry*, *13*(3), 403. doi:10.3390/sym13030403

Wang, P., Emmerich, M., Li, R., Tang, K., Baeck, T., & Yao, X. (2013). Convex hull-based multi-objective genetic programming for maximizing roc performance. *Neurocomputing*, *125*(3), 102–118.

Zhang, F., & Yang, X. J. (2020). Improving land cover classification in an urbanized coastal area by random forests: The role of variable selection. *Remote Sensing of Environment, 251*(10).

Lei Shi was born in Zhumadian Henan. She Graduated from Henan Agricultural University in 2013 and got Doctor's degree. She worked as a teacher at Henan Agricultural University. And she is an associate professor in the department of computer science, colleges of information and management science. She has published 28 papers, holds one patent and one invention. Her research interests include machine learning, data mining, and computer applications.

Yaqian Qin was born in 1997 and got the M.S.degree from Henan Agriculture University in 2021. She is currently an teacher in the College of Information Engineering of Zhengzhou University of Science and Technology in China.

Juanjuan Zhang was born in 1979 and got the Ph.D. degree from Nanjing Agricultural University in 2009. She is currently an Associate Professor in the College of Information and Management Science of Henan Agricultural University in China. She has published more than 20 papers and won the second prize of National Science and Technology Progress Award and the second prize of Henan Science and Technology Progress Award. Her research direction is agricultural remote sensing monitoring.

Yan Wang was born in 1995 and got the M.S.degree from Henan Agriculture University in 2021. He is currently an teacher in the College of Information Engineering of Zhengzhou University of Science and Technology in China.

Hongbo Qiao was born in Nanyang Henan, China, in 1978. He graduated from Chinese Academy of Agricultural Sciences with a Doctor's degree in 2007. He is an associate professor and the director at the Department of Computer Science, Colleges of information and Management Science in Henan Agricultural University. He is the Director of computer science. He has published 38 papers, holds 2patents, and his research direction is Remote Sensing, Big data.

Haiping Si was born in Henan, China, in 1978. After graduating from Chinese Academy of Agricultural Sciences in 2011 with a Doctor's Degree in July 2011, he worked as a teacher in Henan Agricultural University from Sept. 2011 until now. Since 2015, he is an associate professor at the Department of Computer Science, Colleges of information and Management Science in Henan Agricultural University. And he was a visiting scholar at University of Wisconsin-Milwaukee, USA, from Aug 2017 to Aug 2018. He has published 36 papers and holds 2 patents and 2 inventions. His research interests include Computer Science, Computer Application and Data Processing.