Available at www.sciencedirect.com

INFORMATION PROCESSING IN AGRICULTURE 3 (2016) 215–222

journal homepage: www.elsevier.com/locate/inpa

An improved random forest classifier for multi-class classification

Archana Chaudhary [a,*], *Savita Kolhe* [b], *Raj Kamal* [c]

[a] *School of Computer Science and IT, Devi Ahilya University, Khandwa Road, Indore, Madhya Pradesh, India.*
[b] *ICAR-Directorate of Soybean Research, Khandwa Road, Indore, Madhya Pradesh, India*
[c] *Medicaps Institute of Science and Technology, Indore, Madhya Pradesh, India*

ARTICLE INFO

Article history:
Received 3 February 2016
Accepted 23 August 2016
Available online 1 September 2016

Keywords:
Groundnut disease
Improved-RFC
Machine learning
Multi-class classification

ABSTRACT

The paper presents an improved-RFC (Random Forest Classifier) approach for multi-class disease classification problem. It consists of a combination of Random Forest machine learning algorithm, an attribute evaluator method and an instance filter method. It intends to improve the performance of Random Forest algorithm. The performance results confirm that the proposed improved-RFC approach performs better than Random Forest algorithm with increase in disease classification accuracy up to 97.80% for multi-class groundnut disease dataset. The performance of improved-RFC approach is tested for its efficiency on five benchmark datasets. It shows superior performance on all these datasets.

1. Introduction

Diseases, pests and uneven rainfalls are vital reasons for yield losses in crops. Significant crop losses by pests and diseases have been accounted from many countries [1,24,38]. Yield losses due to some diseases are to an extent of 70% [14]. The degree of economic losses due to diseases is much more than the reported global yield losses of 600 million US$ [28]. Groundnut (*Arachis hypogaea* L.) is an important oilseed crop and a vital source of protein. More than fifty-five pathogens along with viruses have been reported to affect groundnut crop. Some diseases are extensively distributed and cause more financial losses while others are confined in distribution and are not considered to be reasonably significant at the present time. Proper diagnosis of disease(s) is the first step in planning a Disease Intelligent System [27]. Symptoms on plant parts along with the congenial climatic conditions can be used to identify most of the diseases [43–45].

Classification is a basic task in the field of machine learning. It is the recognition of the category labels of instances that are normally described by a set of attributes (features) in a dataset. The aim of classification is to accurately predict class labels of instances whose values of attributes are known, but labels of classes are unknown [12]. Classification task in the field of machine learning is binary, multi-class, multi-labeled and hierarchical. Disease diagnosis is a multi-class classification problem which deals with high dimensional datasets. The classification task with disease diagnosis problem is to assign a disease label to a particular instance. High dimensional datasets have the problem of presence of irrelevant or redundant features which often lowers the performance of machine learning algorithms. Hence, the use of suitable feature selection methods becomes essential for classification tasks that deal with high dimensional data [11,21].

Several machine learning algorithms are successfully used for the problems of classification and prediction [2,30]. Machine learning algorithms are applied to identify Mastitis

* Corresponding author.
E-mail addresses: archana_scs@yahoo.in, archana227@gmail.com (A. Chaudhary).
Peer review under responsibility of China Agricultural University.
http://dx.doi.org/10.1016/j.inpa.2016.08.002

in the field of dairy farming [26,35] and Estrus [31], forecast production of milk [29] and to find out reasons for culling [30]. Machine learning algorithms are applied for accurate identification of crop diseases Leaf brown spot, Rice blast, Sheath rot, Bacterial blight, Cercospora leaf spot, Leaf rust, Potato late blight and Powdery mildew [4,34]. Genetic Algorithms and Multilayer Neural Networks are applied for identification of Tobacco rattle and Cucumber green mottle mosaic plant viruses to solve productivity problems [17]. Random Forest algorithm is successfully used for accurate identification of diseases in disease diagnosis problems [3,33,36].

The performance of Random Forest algorithm is improved by using a combination of an attribute evaluator method and an instance filter method in the present work. The paper is arranged as follows: Section 2 portrays materials and methods. Section 3 describes the improved-RFC approach. Section 4 presents results and discussions. Section 5 gives the conclusions drawn.

2. Materials and methods

We have used the WEKA [19] open source software with default parameter settings to conduct the present work. WEKA includes multiple supervised and unsupervised machine learning algorithms. Additionally, it also has a wide set of techniques for data preprocessing and modeling, with a user friendly interface for training and testing machine learning models.

2.1. Datasets

The improved-RFC approach is applied to groundnut disease multi-class dataset. The performance of improved-RFC is also examined on five benchmark datasets.

2.1.1. Real-life groundnut disease dataset

The real-life groundnut disease dataset is developed using different sources [8,9,14,25,38,41] by taking into account the symptoms of disease(s), climatic conditions favoring the disease and crop part(s) affected. The dataset consists of 1080 instances with no missing values. It is a multi-class dataset with 13 disease classes. It has 26 attributes and one disease target class as shown in Table 1. All the attributes are nominal.

2.1.2. Real benchmark datasets

Five real benchmark datasets from UCI machine learning repository [13] are used for the purpose of testing the performance of improved-RFC approach. The structure of these benchmark datasets is shown in Table 2.

2.2. Feature selection methods

Feature selection or attribute evaluator or filter method consists of identifying the relevant features and ignoring the irrelevant ones from a dataset [6]. The use of attribute evaluator methods enhances the performance of machine learning algorithms. These methods offer better understanding of data and permit capability of data reduction. The attribute evaluator methods used in the design of improved-RFC approach are.

2.2.1. Correlation-based feature selection (CFS)

It is a simple attribute evaluator method that grades feature subsets on the basis of a correlation based heuristic estimation function [6]. The bias associated with the function is towards the subsets containing features that have high correlation with the class but are not correlated with each other [20]. It is used in the design of improved-RFC approach as it ignores irrelevant features as they have low correlation with the class. Redundant features are not considered in the resultant feature subset as they are highly correlated with one or other features.

2.2.2. Symmetrical uncertainty (SU)

SU [6] is an attribute evaluator method. It is used in the design of improved-RFC approach as it provides a symmetrical measurement for correlation between features and also balances the bias of mutual information. SU is defined as the fraction between the Information Gain (IG) and the Entropy (H) of two features, x and z such that

$$SU(x,z) = 2 \times IG(x|z)/[H(x) + H(z)] \quad (1)$$

where the IG is described as

$$IG(x|z) = H(z) + H(x) - H(x,z) \quad (2)$$

where $H(x)$ and $H(x, z)$ represent the entropy and joint entropy respectively.

2.2.3. Gain ratio

Gain Ratio attribute evaluator is used in the design of improved-RFC approach as it is an improvement to Information Gain which resolves the matter of bias towards attributes with a larger set of values [23]. It measures gain in information for the purpose of classification with respect to the entropy of feature Fe_i:

$$\text{Gain Ratio}(C,Fe_i) = [H(C) - H(C|Fe_i)]/H(Fe_i) \quad (3)$$

where $H(C)$ represents entropy of class C, H ($C|Fe_i$) represents the entropy of class C given feature Fe_i and H (Fe_i) is the entropy measure of feature Fe_i.

2.3. Simple random sampling – instance filter

Real-world datasets, as groundnut disease dataset have non-uniform class distributions. This non-uniformity of class distributions considerably influences the performance of a classification algorithm in the training phase. The two strategies in training machine learning algorithms are as follows – (i) training the algorithm by considering original class distribution and (ii) training the algorithm through minority class representations balanced through a sampling strategy [33,42]. Simple random sampling is one of the fundamental sampling techniques of statistics that gives a fair sample from the original data. The two ways of selecting samples are – (i) with replacement – a sample can be picked more than once (ii) without replacement – a sample can be chosen only once [32]. The unbalanced nature of distribution of groundnut disease classes makes the dataset suitable to test the result of

Table 1 – Groundnut disease dataset description.

Attribute number	Attribute description	Possible values of attributes	Assigned values
1.	Temperature	Normal, lower-than-normal, greater-than-normal	1–3
2.	Soil-moisture	High, normal, low	1–3
3.	Relative-humidity	High, normal, low	1–3
4.	Severity	Minor, severe	1–2
5.	Leaf	Normal, abnormal	1–2
6.	Leaf-lesions	Black, brown, chlorotic, circular, dark-brown, dark-brown-to-black, grayish-green, irregular, light-brown-centre-and-yellow-halo, marginal-irregular, necrotic, orange-colored-pustules, powdery-white, small, sub-circular, water-soaked, wilting, zonate-appearance, does-not-apply	1–19
7.	Seed	Normal, abnormal	1–2
8.	Seed-lesions	Rotten, shriveled, yellow-and-wilted, does-not-apply	1–4
9.	Hypocotyl	Normal, abnormal	1–2
10.	Hypocotyl-lesions	Brown-to-dark-brown, damping-off, light-brown, rotten, sunken, water-soaked, does-not-apply	1–7
11.	Pod	Normal, abnormal	1–2
12.	Pod-lesions	Discrete, rotten, does-not-apply	1–3
13.	Stem	Normal, abnormal	1–2
14.	Stem-lesions	Black-and-sooty, chlorotic, internal-vascular-browning-and-discoloration, necrotic, oval-to-elongate, shredded, water-soaked, wilting, does-not-apply	1–9
15.	Root	Normal, abnormal	1–2
16.	Root-lesions	Black, rotten, shredded, internal-vascular-browning-and-discoloration, does-not-apply	1–5
17.	Collar	Normal, abnormal	1–2
18.	Collar-lesions	Shredded-and-dark-brown, does-not-apply	1–2
19.	Peg	Normal, abnormal	1–2
20.	Peg-lesions	Discrete, oval-to-elongate, rotten, does-not-apply	1–4
21.	Leaf-surface	Upper, lower	1–2
22.	Mycelia	Sporulating, white, does-not-apply	1–3
23.	Sclerotia	Mustard-sized-and-color, does-not-apply	1–2
24.	Fruiting-bodies	Black, concentric-rings, reddish-orange, does-not-apply	1–4
25.	Plant-effect	Chlorotic, death, drying, normal, stunted-growth	1–5
26	Leaf-wetness	Present, absent	1–2
27.	Target class	Alternaria leaf spot, Charcoal rot, Collar rot, Cylindrocladium black rot, Early leaf spot, Fusarium rot, Late leaf spot, Myrothecium leaf blight, Powdery Mildew, Rust, Stem rot, Yellow mold, Zonate leaf spot	1–13

Table 2 – Description of benchmark datasets.

Datasets	Classes	Attribute types	Instances	No. of attributes
Audiology	24-Class	Nominal	226	69
Breast Cancer	2-Class	Nominal	286	9
Diabetes	2-Class	Real	768	8
Soybean	19-Class	Nominal	683	35
Vote	2-Class	Nominal	435	16

simple random sampling strategy with the help of an instance filter. The instance filter-Resample scales up the classification accuracy obtained by Random Forest algorithm [33]. Hence we have used instance filter-Resample in the present work.

2.4. *Selection of classification algorithm*

A comparison of machine learning algorithms such as Neural Network (NN), Logistic Regression (LR) and Support Vector Machine (SVM) is conducted for choosing suitable classification algorithm in the present work. It is observed that NN, LR and SVM when applied to groundnut disease dataset showed comparable performance as that of Random Forest. But Random Forest shows a greater increase in disease classification accuracy as 97.80% as compared to NN, LR and SVM as 92.20%, 94.80% and 95.70% respectively. Random Forest algorithm has already shown outstanding performance for many disease diagnosis problems [3,33,36,37,39,40]. Hence for the above mentioned reasons we have selected Random Forest algorithm for groundnut disease classification.

2.5. *Random Forest classification algorithm*

Random Forest is a popular machine learning algorithm used for several types of classification tasks [10,15,16,18,22,33,37,39]. A Random Forest is an ensemble of tree-structured classifiers [7]. Every tree of the forest gives a unit vote, assigning each input to the most probable class label. It is a fast method, robust to noise and it is a successful ensemble which can identify non-linear patterns in the data. It can easily handle both numerical and categorical data [39]. One of the major advantages of Random Forest is that it does not suffer from over fitting, even if more trees are appended to the forest.

3. Improved-RFC approach

Improved-RFC approach uses Random Forest algorithm, an attribute evaluator method and an instance filter method-Resample. The aim of the approach is to improve classification accuracy of Random Forest algorithm for multi-class classification problems.

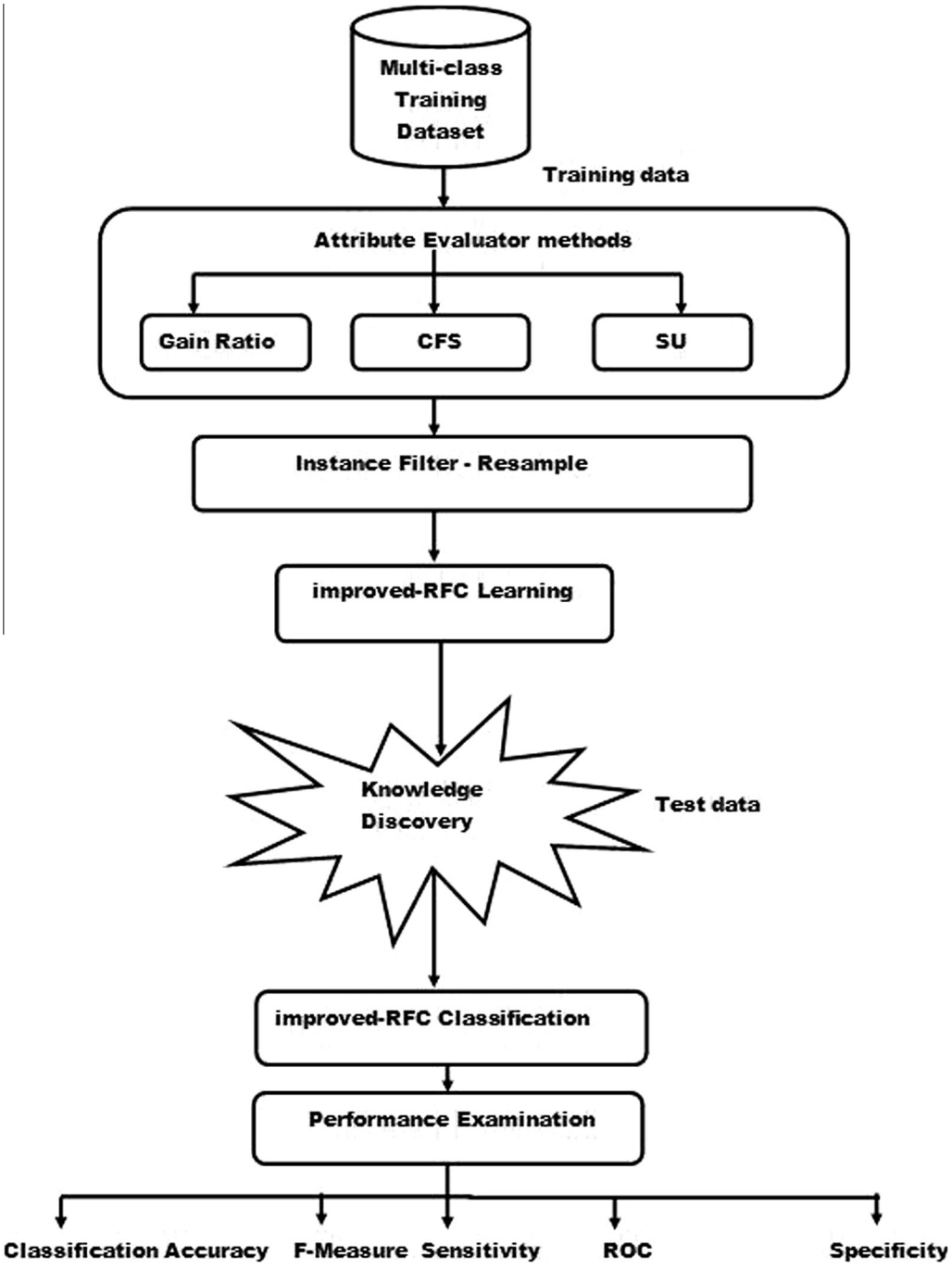


Fig. 1 – Architectural design of improved-RFC approach.

Table 3 – Class distributions of groundnut disease dataset before and after sampling.

Class	Class labels of groundnut diseases	Before sampling	After sampling
01.	Alternaria leaf spot	105	86
02.	Charcoal rot	102	84
03.	Collar rot	82	84
04.	Cylindrocladium black rot	70	81
05.	Early leaf spot	72	84
06.	Fusarium rot	78	80
07.	Late leaf spot	82	83
08.	Myrothecium leaf blight	79	80
09.	Powdery mildew	83	82
10.	Rust	107	87
11.	Stem rot	79	82
12.	Yellow mold	71	86
13.	Zonate leaf spot	70	81

3.1. *Algorithm of improved-RFC approach*

The pseudo-code of improved-RFC approach is given below.

Algorithm 1. Improved-RFC
Input: $D_{Train} = \{x_1, x_2 \ldots x_n\}$ // Training dataset which includes a set of training examples and their related class labels.
Output: classification-accuracy A.
Method:
step 1: Select an attribute evaluator method and apply it on training dataset-D_{train} to obtain a subset of attributes A_m.
step 2: Apply instance filter-Resample for Am of Dtrain and obtain Dtrain-resample.
step 3: Select Random Forest classification algorithm on $D_{train\text{-}resample}$ and obtain classification-accuracy A
step 4: Output classification-accuracy A.

3.2. *Architecture of improved-RFC approach*

The architectural design of improved-RFC is shown in Fig. 1. The improved-RFC approach begins with selecting the multi-class training dataset for classification. An attribute evaluator method from CFS, SU and Gain Ratio is chosen and applied on the training dataset to obtain the relevant attributes for classification (step1 of algorithm 1). After applying an attribute evaluator, instance filter-Resample is applied successfully for balancing the class distributions of the multi-class dataset (step 2 of algorithm 1). The use of instance filter-Resample is voluntary in the improved-RFC approach. If the class distributions of the dataset are already uniform then step 2 can be skipped.

Subsequently Random Forest classification algorithm (step 3 of algorithm 1) is applied on the result obtained from step 2 of algorithm 1. The resultant classification accuracy is obtained from step 4 of algorithm 1. Finally the performance of improved-RFC approach is examined with respect to each attribute evaluator method – CFS, SU and Gain Ratio. Performance metrics – classification accuracy, F-measure, ROC, sensitivity and specificity are noted.

4. Results and discussions

Ten-fold cross validation is appropriate strategy for evaluating the performance of a machine learning algorithm as it offers consistent approximates for classification accuracy for every classification task [3,5]. Therefore, each experiment is conducted with 10-fold cross validation in the present work.

4.1. *Performance evaluation metrics*

The performance of the improved-RFC approach is evaluated using performance metrics-classification accuracy, specificity, sensitivity, Receiver Operating Characteristics (ROC) and

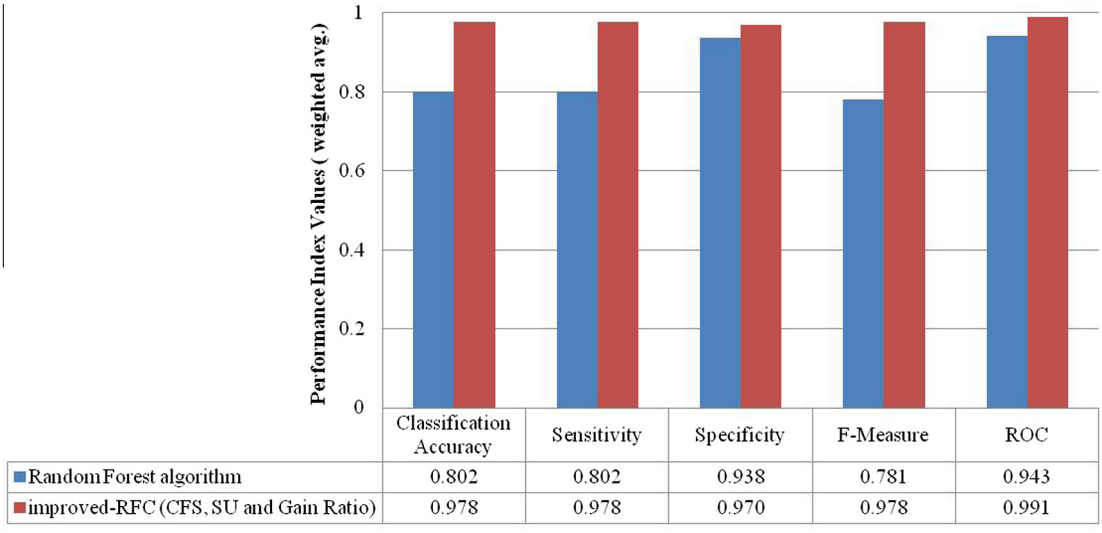


Fig. 2 – Performance comparison of Random Forest algorithm and improved-RFC approach for groundnut disease diagnosis.

F-measure [3,12,33]. ROC performance metric is also useful in assessing the performance of a disease diagnosis test [3]. It has adequate information for clarity and improving the performance of any machine learning algorithm. It provides a trade-off between sensitivity and specificity. It is also observed when the improved-RFC approach is applied on multi-class groundnut disease dataset.

4.2. *Application of improved-RFC approach on multi-class groundnut disease dataset*

Improved-RFC approach is applied on multi-class groundnut disease dataset for exact classification of groundnut disease (s). An attribute evaluator is selected in (step 1 of Algorithm 1). The function of attribute evaluators in the present work is to reduce the high dimensional groundnut disease dataset. CFS finds a subset of attributes considering an attribute is good if it is related to the disease class but is not essential to any of the other relevant attributes where as SU and Gain Ratio work by finding an appropriate ranking of the attribute subsets of groundnut disease dataset.

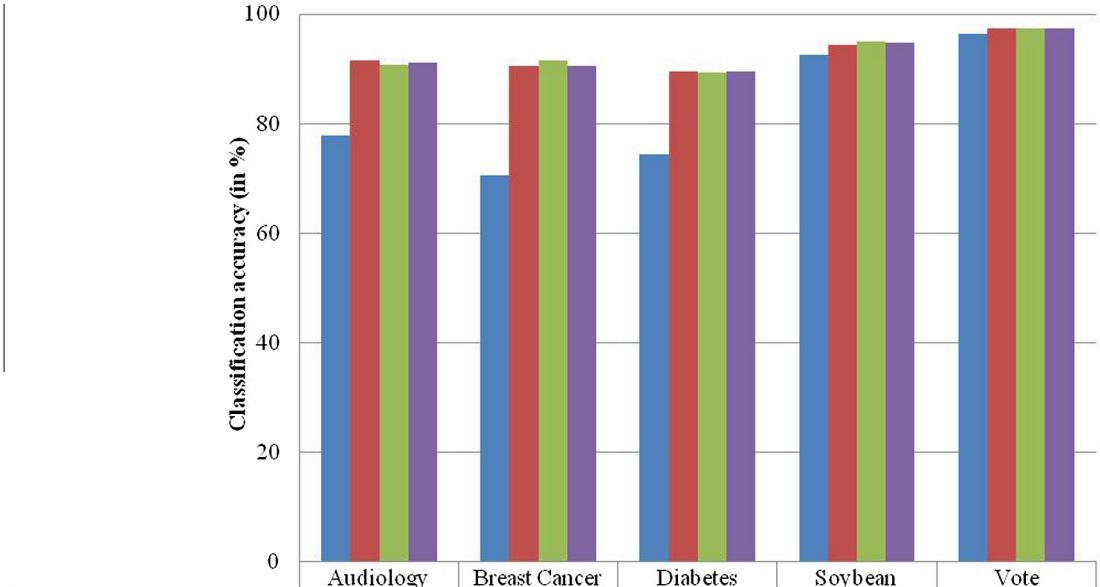
CFS attribute evaluator results in the following subset of attributes with respect to disease target class – temperature, soil-moisture, hypocotyl, stem-lesions, collar, leaf-lesions, leaf-surface, mycelia, fruiting-bodies, plant-effect from (step 1 of algorithm1). It is clear from Table 3 that groundnut disease dataset is not balanced before applying instance filter-Resample. It contains majority classes such as Alternaria leaf spot, Charcoal rot and Rust. It also contains some minority classes as Cylindrocladium black rot, Yellow mold and Zonate leaf spot. The function of instance filter-Resample is to create a random subsample of the groundnut disease dataset and balance its class distributions. The disease classes are made uniform after applying instance filter-Resample (step 2 of algorithm1) on the result of (step1 of algorithm1) as shown in Table 3.

In (step 3 of algorithm1) Random Forest algorithm is applied on the result of previous step. The resultant classification accuracy is obtained from (step 4 of algorithm 1). Similarly the performance of improved-RFC approach is observed for SU and Gain Ratio attribute evaluator methods (steps 1 to 4 of algorithm 1). It is clear from Fig. 2 that improved-RFC approach performs better than Random Forest algorithm for groundnut disease dataset. The classification accuracy obtained for Random Forest algorithm is 80.20%. Improved-RFC approach with (CFS, SU and Gain Ratio) shows a greater increase in disease classification accuracy as 97.80%.

It is also apparent from Fig. 2 that the other performance metrics – F-measure, sensitivity, specificity and ROC also show considerable rise after using improved-RFC approach on groundnut disease dataset as compared to Random Forest algorithm. This proves the adequacy of improved-RFC approach for groundnut disease diagnosis problem as compared to Random Forest algorithm. The experimental results show that classification using improved-RFC approach enhances the diagnosis of groundnut diseases. In order to prove the efficiency of improved-RFC approach, we made use of Audiology, Breast Cancer, Diabetes, Soybean and Vote multi-class datasets as benchmarking studies besides multi-class groundnut disease dataset.

4.3. *Case studies for testing purposes*

Audiology, Breast Cancer, Diabetes, Soybean and Vote are multi-class datasets from UCI machine learning repository and the description of each dataset is shown in Table 2. The



	Audiology	Breast Cancer	Diabetes	Soybean	Vote
Random Forest algorithm	77.87	70.62	74.34	92.53	96.32
improved-RFC (CFS)	91.59	90.55	89.45	94.28	97.47
improved-RFC (SU)	90.70	91.60	89.32	95.02	97.47
improved-RFC (Gain Ratio)	91.15	90.55	89.45	94.72	97.47

Fig. 3 – Classification accuracy rates of Random Forest algorithm and improved-RFC approach for benchmark datasets.

Table 4 – The performance index values for Random Forest algorithm and improved-RFC approach using benchmark datasets.

Benchmark datasets	Performance indices	Random forest algorithm	Improved-RFC approach		
			CFS	SU	Gain ratio
Audiology	F-measure	0.751	**0.911**	0.904	0.907
	Sensitivity	0.779	**0.916**	0.907	0.912
Breast Cancer	F-measure	0.694	0.902	**0.913**	0.902
	Sensitivity	0.706	0.906	**0.916**	0.906
Diabetes	F-measure	0.737	**0.893**	0.892	**0.893**
	Sensitivity	0.743	**0.895**	0.893	**0.895**
Soybean	F-measure	0.926	0.941	**0.949**	0.947
	Sensitivity	0.925	0.943	**0.950**	0.947
Vote	F-measure	0.963	**0.975**	**0.975**	**0.975**
	Sensitivity	0.963	**0.975**	**0.975**	**0.975**

Bold values in signify maximum increase obtained in performance index value.

experiments are conducted for each dataset, as it is realized for multi-class groundnut disease dataset. The performance of improved-RFC approach is tested with the help of three performance metrics – (i) classification accuracy, (ii) F-measure (iii) sensitivity [3,12,33]. It is an important observation from Fig. 3 that the greatest increase in classification accuracy using improved-RFC approach is 13.72% (CFS), 20.98% (SU), 15.11% (CFS and Gain Ratio), 2.49% (SU), 1.15% (CFS, SU and Gain Ratio) for Audiology, Breast Cancer, Diabetes, Soybean and Vote multi-class datasets. Significant rise in F-measure and sensitivity values in Table 4 also indicate that improved-RFC approach outperforms Random Forest algorithm.

5. Conclusions

The paper discusses an improved-RFC approach for enhancement of classification accuracy of Random Forest algorithm for multi-class datasets. The improved-RFC approach is effectively applied to groundnut disease diagnosis multi-class classification problem. Improved-RFC approach shows superior performance as compared to Random Forest algorithm. The improved-RFC approach with CFS, SU and Gain Ratio shows increase in disease classification accuracy as 97.80% as compared to Random Forest algorithm with disease classification accuracy as 80.20%. The performance of improved-RFC approach is also tested for classification accuracy, F-measure and sensitivity values with 10-fold cross validation on five benchmark datasets from UCI machine learning repository. The results for these datasets on these performance measures confirm that the improved-RFC approach shows better performance as compared to Random Forest algorithm. Therefore it is concluded that improved-RFC approach is a good substitute in dealing with computer-aided diagnosis and multi-class classification problems.

References

[1] Agriculture Research Institute. Annual report of entomology work. Hyderabad, India: Agriculture Research Institute; 2009 (in English).

[2] Arribas JI, Sánchez-Ferrero GV, Ruiz-Ruiz G, Gómez-Gil J. Leaf classification in sunflower crops by computer vision and neural networks. Comput Electron Agric 2011;78(1):9–18.

[3] Azar AT, Elshazly HI, Hassanien AE, Elkorany AM. A random forest classifier for lymph diseases. Comput Methods Prog Biomed 2014;113(2):465–73.

[4] Baker KM, Kirk WW. Comparative analysis of models integrating synoptic forecast data into potato late blight risk estimate systems. Comput Electron Agric 2007;57(1):23–32.

[5] Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification and overview. Bioinformatics 2000;16(5):412–24.

[6] Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. Feature selection and classification in multiple class datasets: an application to KDD Cup 99 dataset. Expert Syst Appl 2011;38(5):5947–57.

[7] Breiman L. Random forests. Mach Learning 2001;45(1):5–32.

[8] Butler DR, Wadia KDR, Reddy RK, Das ND, Johnson B, Meena K, Krishnamurty K, Sreeenivas B, Srivastava NN. A weather-based scheme to advise on limited chemical control of groundnut leaf spot diseases in India. Exp Agric 2000;36 (4):469–78.

[9] Butler DR, Wadia KDR, Jadav DR. Effects of leaf wetness and temperature on late leaf spot infection of groundnut. Plant Pathol 1994;43(1):112–20.

[10] Cutler DR, Edwards Jr TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ. Random forests for classification in ecology. Ecology 2007;88(11):2783–92.

[11] Filters Das S. wrappers and a boosting-based hybrid for feature selection. Proc. ICML'01 Proceedings of the 2001 international conference on machine learning, Williamstown, USA. p. 74–81.

[12] Farid DM, Zhang L, Rahman CM, Hossain MA, Strachan R. Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. Expert Syst Appl 2014;41(4):1937–46.

[13] University of California Irvine. UCI Machine Learning repository. Link: http://archive.ics.uci.edu/ml/.2010

[14] Ghewande MP, Desai S, Basu MS. Diagnosis and management of major diseases of groundnut. Bulletin. Junagadh, Gujarat, India: National Research Centre for Groundnut; 2002. p. 1–36.

[15] Ghimire B, Rogan J, Miller J. Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic. Remote Sens Lett 2010;1(1):45–54.

[16] Gislason PO, Benediktsson JA, Sveinsson JR. Random forests for land cover classification. Pattern Recogn Lett 2006;27 (4):294–300.

[17] Glezakos TJ, Moschopoulou G, Tsiligiridis TA, Kintzios S, Yialouris CP. Plant virus identification based on neural

networks with evolutionary preprocessing. Comput Electron Agric 2010;70(2):263–75.
[18] Guo L, Chehata N, Mallet C, Boukir S. Relevance of airborne lidar and multispectral image data for urban scene classification using random forests. ISPRS J Photogrammetry Remote Sens 2011;66(1):56–66.
[19] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explorations 2009;11(1):10–8.
[20] Hall MA. Correlation-based feature selection for machine learning Ph.D. thesis. Hamilton, New Zealand: University of Waikato; 1999.
[21] Hall MA, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. IEEE Trans Knowl Data Eng 2003;15(6):1437–47.
[22] Chen XW, Liu M. Prediction of protein–protein interactions using random decision forest framework. Bioinformatics 2006;21(24):4394–400.
[23] Ibrahim HE, Badr SM, Shahee MA. Adaptive layered approach using machine learning techniques with gain ratio for intrusion detection systems. Int J Comput Appl 2012;56 (7):10–6.
[24] Izge UA, Mohammed ZH, Goni A. Levels of variability in groundnut (Arachis hypogaea L.) to cercospora leaf spot disease - implication for selection. Afr J Agric Res 2007;2 (4):182–6.
[25] Jhorar OP, Mavi HS, Dhiman JS. A graphical technique for short-range forecast of early and late leaf spot disease of groundnut. J Res Punjab Agric Univ 1987;24(4):607–12.
[26] Kim T, Heald CW. Inducing inference rules for the classification of bovine mastitis. Comput Electron Agric 1999;23(1):27–42.
[27] Kolhe S, Kamal Raj, Saini HS, Gupta GK. A web-based intelligent disease-diagnosis system using a new fuzzy-logic based approach for drawing the inferences in crops. Comput Electron Agric 2011;76(1):16–27.
[28] Kumari V, Gowda MVC, Tasiwal V, Pandey MK, Bhat RS, Mallikarjuna N, Upadhyaya HD, Varshney RK. Diversification of primary gene pool through introgression of resistance to foliar diseases from synthetic amphidiploids to cultivated groundnut (Arachis hypogaea L.). Crop J 2014;2(2-3):110–9.
[29] Lacroix R, Wade KM, Kok R, Hayes JF. Prediction of cow performance with a connectionist model. Trans ASAE 1995;38:1573–9.
[30] McQueen RJ, Garner SR, Nevill-Manning CG, Witten IH. Applying machine learning to agricultural data. Comput Electron Agric 1995;12(4):275–93.
[31] Mitchell RS, Sherlock RA, Smith LA. An investigation into the use of machine learning for determining oestrus in cows. Comput Electron Agric 1996;15(3):195–213.
[32] Mitra SK, Pathak PK. The nature of simple random sampling. Ann Stat 1984;12:1536–42.
[33] O¨ zc- ift A. Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. Comput Biol Med 2011;41(5):265–71.
[34] Phadikar S, Sil J, Das AK. Rice diseases classification using feature selection and rule generation techniques. Comput Electron Agric 2013;90(C):76–85.
[35] Pietersma D, Lacroix R, Lefebvre D, Wade KM. Performance analysis for machine-learning experiments using small data sets. Comput Electron Agric 2003;38(1):1–17.
[36] Ramírez J, Górriz JM, Segovia F, Chaves R, Salas-Gonzalez D, López M, Álvarez I, Padilla P. Computer aided diagnosis system for the Alzheimer's disease based on partial least squares and random forest SPECT image classification. Neurosci Lett 2010;472:99–103.
[37] Seera M, Lim CP. A hybrid intelligent system for medical data classification. Expert Syst Appl 2014;41(5):2239–49.
[38] Singh F, Oswalt DL. Major diseases of groundnut. ICRISAT-Skill development series No.6, Patancheru, Andhra Pradesh, India 1992. 50-36.
[39] Titapiccolo JI, Ferrario M, Cerutti S, Barbieri C, Mari F, Gatti E, Signorini MG. Artificial intelligence models to stratify cardiovascular risk in incident hemodialysis patients. Expert Syst Appl 2013;40(11):4679–86.
[40] Tripoliti EE, Fotiadis DI, Argyropoulou M, Manis G. A six stage approach for the diagnosis of the Alzheimer's disease based on fMRI data. J Biomed Inform 2010;43(2):307–20.
[41] Tripathy AK, Adinarayana J, Vijayalakshmi K, Merchant SN, Desai UB, Ninomiya S, Hirafuji M, Kiura T. Knowledge discovery and leaf spot dynamics of groundnut crop through wireless sensor network and data mining techniques. Comput Electron Agric 2014;107:104–14.
[42] Weiss GM, Provost F. Learning when training data are costly: the effect of class distribution on tree induction. J Artif Intell Res 2003;19:315–54.
[43] Chaudhary A, Kolhe S, Kamal Raj. A hybrid ensemble for classification in multiclass datasets: an application to oilseed disease dataset. Comput Electron Agric 2016;124:65–72.
[44] Kolhe S, Kamal Raj, Saini HS, Gupta GK. A fuzzy-logic based on-line disease diagnosis system for soybean. Soybean Res 2009;7:73–81.
[45] Chaudhary A, Kolhe S, Kamal Raj. Machine learning classification techniques: a comparative study. Int J Adv Comput Theory Eng 2013;2(4):2319–526.