Computers and Electronics in Agriculture 124 (2016) 65–72

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

A hybrid ensemble for classification in multiclass datasets: An application to oilseed disease dataset

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

[a] School of Computer Science and IT, Devi Ahilya University, Khandwa Road, Indore 452001, Madhya Pradesh, India
[b] ICAR-Directorate of Soybean Research, Khandwa Road, Indore 452001, Madhya Pradesh, India

ARTICLE INFO

Article history:
Received 17 June 2015
Received in revised form 14 September 2015
Accepted 27 March 2016
Available online 9 April 2016

Keywords:
Machine learning
Multiclass classification
Hybrid ensemble
Oilseed disease

ABSTRACT

The paper presents a new hybrid ensemble approach consisting of a combination of machine learning algorithms, a feature ranking method and a supervised instance filter. Its aim is to improve the performance results of machine learning algorithms for multiclass classification problems. The performance of new hybrid ensemble approach is tested for its effectiveness over four standard agriculture multiclass datasets. It performs better on all these datasets. It is applied on multiclass oilseed disease dataset. It is observed that ensemble-Vote performs better than Logistic Regression and Naïve Bayes algorithms. The performance results of hybrid ensemble are compared with ensemble-Vote. The performance results prove that the new hybrid ensemble approach outperforms ensemble-Vote with improved oilseed disease classification accuracy up to 94.73%.

1. Introduction

Machine learning algorithms are useful in effective decision making in agriculture. These algorithms possess a strong capability of extracting complicated relationships that exist in the agricultural data (Rocha et al., 2010). High dimensional agricultural data requires the use of machine learning feature selection algorithms when the most explanatory or important features or attributes are to be selected from large datasets (EI-Bendary et al., 2015; Hill et al., 2014; Kundu et al., 2011; Timmermans and Hulzebosch, 1996). Machine learning classification algorithms viz. Logistic Regression and Naïve Bayes are successfully used for accurate identification of crop diseases (Phadikar et al., 2013; Sankaran et al., 2010; Gutiérrez et al., 2008; Baker and Kirk, 2007).

Soybean, groundnut and rapeseed-mustard are the three most important oilseed crops of the world. They play an important role in the oilseed economy. One of the major concerns in increasing and stabilizing the yield of oilseeds is the incidence of pests and diseases which, to a greater extent are responsible for low and unstable production of these crops. Oilseeds are susceptible to various diseases caused by bacteria, fungi, viruses, nematodes and physiological disorders. Some diseases are largely spread and cause great economic losses while others are limited in distribution and are not of much economic importance during present times, but may become major diseases in the course of time by favorable climatic conditions. Oilseed diseases considered in the present work include Alternaria leaf spot, Anthracnose, Cercospora leaf spot, Charcoal rot, Collar rot, Myrothecium leaf spot, Powdery mildew, Sclerotinia stem rot, Phyllosticta leaf spot and Rust. Crop disease diagnosis is a multiclass classification problem.

In several classification problems ensembles have proved to be effective as compared to single classification algorithm (Bolón-Canedo et al., 2012; Sun et al., 2007; Stamatatos and Widmer, 2005). Ensembles have great potential in the domain of multiclass classification. Ensemble machine learning methods have been recommended in the literature for different types of classification problems (Hsu, 2012; Kotsiantis, 2007; Dietterich, 2000; Bay, 1999; Opitz, 1999; Ting and Witten, 1999; Zheng and Webb, 1999; Ho, 1998; Breiman, 1996; Wolpert, 1992; Hansen and Salamon, 1990; Schapire, 1990).

Vote is an ensemble of Logistic Regression and Naïve Bayes algorithms in the present work. This work proposes a new hybrid ensemble approach with an aim to improve the performance results of machine learning algorithms for multiclass classification problems. The aim of the present work is also to compare proposed hybrid ensemble approach with ensemble-Vote. The proposed new hybrid ensemble approach is applied on oilseed disease diagnosis multiclass problem for accurate identification of disease(s).

* Corresponding author.
E-mail addresses: archana_scs@yahoo.in, archanaa2207@gmail.com (A. Chaudhary), savitasoham@gmail.com, savita_dakhane@yahoo.com (S. Kolhe), dr_rajkamal@hotmail.com (R. Kamal).

http://dx.doi.org/10.1016/j.compag.2016.03.026
0168-1699/

The paper is organized as follows: Section 2 describes materials and methods used in the present work. Section 3 presents new hybrid ensemble approach for multiclass classification problems. Section 4 describes results and discussion. Section 5 presents the conclusions drawn.

2. Materials and methods

The tool WEKA (Hall Mark, 2009; Witten and Frank, 2005) is used for the generation of predictive models. It is an open-source tool developed at University of Waikato, New Zealand (http://www.cs.waikato.ac.nz/ml/Weka/).

2.1. Machine learning algorithms

Logistic Regression is a popular algorithm which does regression analysis and fits a sigmoid curve. It is selected in the design of hybrid ensemble because it is a flexible estimator with low variance and is less susceptible to over-fitting. It is a simple maximum-probability estimator that approximates the probability of every class (given the features examined) and selects one with the largest value (Hill et al., 2014; Silva et al., 2013). In case of oilseed disease diagnosis problem, it will estimate the probability of each disease class and select the class with maximum probability as an answer.

Naïve Bayes is a well-known classification algorithm (Kotsiantis, 2007; Wu et al., 2007). It calculates approximately the conditional probability of every class given the observation, selecting a class with the largest posterior probability as an answer (Silva et al., 2013). It is used in the design of hybrid ensemble because it requires minimum storage space in both the training and classification phases to store the probabilities, hence it is a suitable algorithm for high dimensional multiclass datasets like oilseed disease dataset.

Ensemble algorithms are the methods that create a set of base classifiers to merge and then classify new data samples by casting a vote on their predictions. The ensemble learning consists of two phases. The first phase consists of creation of the base classifiers and the second phase consists of voting task. **Vote** is an ensemble algorithm or meta-classifier for merging predictions from multiple machine learning algorithms (Namsrai et al., 2013; Kotsiantis, 2007). The combined prediction is determined by a combination rule (Bauer and Kohavi, 1999; Kittler, 1998; Battiti and Colla, 1994).

The rationale of using ensemble-Vote in the proposed hybrid ensemble is that it is robust to noise and random errors of classification. It attains considerably greater classification accuracy trained on high dimensional datasets as compared to single machine learning algorithm (Namsrai et al., 2013; Kotsiantis, 2007; Bauer and Kohavi, 1999; Battiti and Colla, 1994). Machine learning algorithms Logistic Regression and Naïve Bayes are combined using ensemble-Vote with a combination rule, in the proposed new hybrid ensemble. The ensemble proposed is intended to offer better classification accuracy as compared to ensemble-Vote.

The feature selection phase, also called as attribute selection or feature ranking is applied to datasets for choosing a subset or ranking of relevant features. **Gain Ratio** (Hall and Smith, 1998) is a popular feature ranking algorithm. The purpose of using Gain Ratio in the hybrid ensemble is that it is an enhancement of Information Gain which resolves the issue of bias toward features with a larger set of values (Ibrahim et al., 2012). Gain Ratio is applied to a variety of classification problems (Silva et al., 2013; Shouman et al., 2011; Danger et al., 2010; Yen and Mike Chu, 2007).

Real-world multiclass datasets, such as oilseed disease dataset, have non-uniform class distribution. This non-uniformity of class distribution considerably influences the performance of a classification algorithm in training phase. A supervised instance filter transforms instances so that they are classified in the context of prediction. Filtering techniques like resampling or **Resample** or simple random sampling is successful in scaling up the classification accuracy attained by machine learning algorithms (Özçift, 2011). In sampling there are two ways for making random selection. The samples are chosen: (i) with substitution (ii) without substitution. The difference among the two ways is that if a sample is chosen more than once, the sampling strategy used is with substitution. The imbalanced distribution of oilseed disease classes makes the dataset appropriate to test the consequence of resampling strategy. Therefore we have used an instance filter - Resample (with substitution) to rescale class distribution of oilseed disease dataset so that resulting class distribution is uniform.

2.2. Datasets

The performance of hybrid ensemble approach is tested on standard agriculture datasets. After testing hybrid ensemble successfully on standard datasets, it is applied on high dimensional multiclass oilseed disease dataset.

2.2.1. Standard datasets

Three real standard datasets from UCI machine learning repository (Frank and Asuncion, 2010) and one from WEKA repository (http://www.cs.waikato.ac.nz/ml/weka/datasets.html) are used for the purpose of testing the proposed hybrid ensemble approach. The description of these standard agriculture datasets is shown in Table 1.

2.2.2. Oilseed disease multiclass dataset

The oilseed disease dataset is created from different sources (Gupta and Chauhan, 2005; Ghewande et al., 2002; Bartaria et al., 2001; Hartman et al., 1999; Michalski et al., 1983) by considering disease symptoms and plant-part(s) affected. There are 13,360 instances in oilseed disease dataset with no missing attribute values. There are 10 classes in our dataset. All attributes are of the type nominal. There are 22 disease influencing attributes, one attribute as the name of oilseed crop and one target class of oilseed diseases as shown in Table 2.

2.3. Performance evaluation indices

The performance of hybrid ensemble approach is evaluated with the help of performance indices viz. classification accuracy, specificity, sensitivity, Receiver Operating Characteristics (ROC), F-measure, Kappa Statistics (KS) and precision (Azar et al., 2014; Özçift,, 2011). The main formulations of these indices are:

$$\text{Classification accuracy } (in\%) = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \times 100 \quad (1)$$

$$\text{Sensitivity} = \frac{T_P}{T_P + F_N} \quad (2)$$

Table 1
Description of standard agriculture datasets.

Datasets	Classes	Attribute types	Instances	No. of attributes
Soybean	19-Class	Nominal	683	35
Iris plants	3-Class	Real	150	4
Mushroom	2-Class	Nominal	8124	22
Grub-damage	4-Class	Real & nominal	155	8

Table 2
Description of oilseed disease dataset.

Attribute number	Attribute description	Possible values of attributes	Assigned values
1.	Oilseed-crop	Rapeseed-mustard, Soybean, Groundnut	1–3
2.	Temperature	Normal, lower-than-normal, greater-than-normal	1–3
3.	Soil moisture	High, normal, low	1–3
4.	Relative-humidity	High, normal, low	1–3
5.	Severity	Minor, severe	1–2
6.	Leaves	Normal, abnormal	1–2
7.	Leaf-symptoms	Angular, black-dots, blighting, brown, chlorotic, circular, circular-brown-with-yellow-margin, concentric-rings, crinkling, dark-brown, dark-brown-with-purple-margin, dirty-white, epiphyllous, floury, grey-brown-with-minute-lesion, grey-spots-with-chlorotic-halo, greyish-lesion, grayish-green, green-yellow-islands, irregular, large-and-wavy, leathery-and-dark, light-brown, light-brown-to-dark-brown, light-purple, marginal-and-apical, marginal-zonate-irregular, necrotic, orange-pustules, reddish-brown-to-purple, pale-green-to-dull-grey, small, tan-reddish-brown-pustules, tan-yellow-pustules, water-soaked, white-creamy-yellow-pustules, white-powdery, white-sporodochia, wilting, yellow-halos, does-not-apply	1–41
8.	Seed	Normal, abnormal	1–2
9.	Seed-symptoms	Black-discoloration, blighted, damping-off, dark-lesions, decay, discolored, drooping, irregular, pink-to-dark-purple-discoloration, purple-stain, reddish-brown-discoloration, reddish-purple-to-reddish-black, rotten, shriveled, small, sunken, wilting, does-not-apply	1–18
10.	Pod	Normal, abnormal	1–2
11.	Pod-symptoms	Black-sporodochia, circular, circular-to-linear, dark-brown, decay, dirty-white, elongated, floury, reddish-purple-to-reddish-black, rotten, sunken, white-powdery, yellow-to-brown-discoloration, does-not-apply	1–14
12.	Stem	Normal, abnormal	1–2
13.	Stem-symptoms	Black-and-sooty, circular, circular-brown-with-yellow-margin, circular-to-linear, dark-brown, dirty-white, distortion, dull-gray-to-dark-brown, elongated, floury, grey-to-brown, grey-to-white, irregular, light-brown-discoloration, necrotic, purple-brown-border, reddish-dark-brown, reddish-purple-to-reddish-black, shredded, sunken, swollen, tan-white, water-soaked, white-pustules, white-powdery, wilting, does-not-apply	1–27
14.	Root	Normal, abnormal	1–2
15.	Root-symptoms	Black-and-rotten, light-brown-discoloration, shredded, does-not-apply	1–4
16.	Collar	Normal, abnormal	1–2
17.	Collar-symptoms	Dark-brown, decay, light-brown, shredded, water-soaked, does-not-apply	1–6
18.	Leaf-surface	Upper, lower	1–2
19.	Mycelia	White-cottony-mats, white, white-fluffy, does-not-apply	1–4
20.	Sclerotia	Black, globose-to-subglobose, minute, mustard-sized, reddish-brown-to-dark-brown, silvery-white-to-light-black, does-not-apply	1–7
21.	Fruiting-bodies	Black, black-with-concentric-rings, minute-spherical, does-not-apply	1–4
22.	Plant-effect	Drying, normal, wilting, withering, premature-ripening, stunted-growth	1–6
23.	Leaf-defoliation	Present, absent	1–2
24.	Target class	Anthracnose, Alternaria leaf spot, Cercospora leaf spot, Charcoal rot, Collar rot, Myrothecium leaf spot, Phyllosticta leaf spot, Powdery mildew, Rust and Sclerotinia stem rot or blight	1–10

$$\text{Precision} = \frac{T_P}{T_P + F_P} \tag{3}$$

$$\text{Specificity} = \frac{T_N}{F_P + T_N} \tag{4}$$

In the Eqns. above T_P and T_N represent the number of true positives and true negatives, F_P and F_N signify the number of false positives and false negatives for classification. Another performance metric *F*-measure is a weighted average of precision and recall.

$$F\text{-measure} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{5}$$

KS is used to measure the agreement between forecasted and experimental values of a dataset, while correcting the agreement that occurs by chance. If KS value for any machine learning algorithm approaches near to 1, its performance is considered to be appreciable and is less driven by chance (Azar et al., 2014). It is a recommended metric for estimation purposes and it is calculated by

$$\text{KS} = \frac{P_{TA} - P_{HY}}{1 - P_{HY}} \tag{6}$$

where P_{TA} denotes total agreement probability, and P_{HY} represents the hypothetical probability of chance agreement.

ROC curves are also found useful to evaluate the performance of a disease diagnosis test (Azar et al., 2014). It has sufficient information for clarity and improving the performance of any machine learning algorithm. It offers a trade-off between sensitivity and specificity. Therefore, in the present work ROC is also observed when hybrid ensemble is applied on oilseed disease dataset.

3. The proposed hybrid ensemble approach

The hybrid ensemble design is based upon the principle that combining the results of multiple machine learning algorithms is superior to the result of single algorithm.

3.1. Algorithm of hybrid ensemble

Hybrid ensemble approach uses hybridization at two levels. First level of hybridization is achieved by combining Gain Ratio feature ranking and supervised instance filter- Resample methods. Second level of hybridization is achieved by merging the predictions of

Logistic Regression and Naïve Bayes using ensemble-Vote. The pseudo-code of hybrid ensemble approach is shown in Algorithm 1.

Algorithm 1. Hybrid ensemble.

Input: $D_T = \{x_1, x_2, \ldots x_n\}$ // Training dataset which contains a set of training instances and their associated class labels.
N = Number of machine learning algorithms selected for classification.
Output: Classification prediction P.
Method:
step 1. Obtain suitable ranking of features of D_T by applying Gain Ratio over D_T.
step 2. Apply supervised instance filter-Resample (with substitution) on the result of (step 1) and obtain $D_{gain\text{-}ratio\text{-}resample}$.
(This step is optional and is used only in case of non-uniform datasets).
step 3. For each i from 1 to N do
i. Apply machine learning classification algorithm$_i$ on the attributes of $D_{gain\text{-}ratio\text{-}resample}$.
ii. Obtain classification prediction P_i from machine learning classification algorithm$_i$.
step 4. Apply ensemble-Vote with a combination rule for merging the predictions $P_1 \ldots P_i$.
step 5. Obtain classification prediction P.

3.2. Design of hybrid ensemble

The proposed hybrid ensemble approach is used for multiclass classification tasks. The design of hybrid ensemble approach is shown in Fig. 1. First we select the multiclass dataset for classification. Gain Ratio feature selection results in ranking of the features of multiclass dataset (step1 of Algorithm 1). After applying Gain Ratio feature ranking algorithm, supervised instance filter-Resample is applied for balancing the class distribution (step 2 of Algorithm 1).

The use of sampling through Resample is optional in this approach. If the dataset already has uniform class distribution, then there is no need of using Resample. Now we choose machine learning algorithm Logistic Regression (step 3(i) of Algorithm 1) and obtain classification prediction – P_1 (step 3(ii) of Algorithm 1). Next we select Naïve Bayes algorithm (step 3(i) of Algorithm 1) and obtain classification prediction – P_2 (step 3(ii) of Algorithm 1). Subsequently the predictions from Logistic Regression (P_1) and Naïve Bayes (P_2) are combined using ensemble-Vote with a combination rule (step 4 of Algorithm 1). Finally the classification prediction of hybrid ensemble is obtained (step 5 of Algorithm 1). Consequently we examine the performance of hybrid ensemble.

3.2.1. Combination rules

The choice of a suitable combination rule or fusion method in ensemble design can further enhance the performance of the hybrid ensemble. In case of oilseed disease diagnosis dataset, the probability rules estimate the disease prediction with the assumption that all the disease classes are a priori equi-probable and the product will be dominated by the expert decision outcome which provides the maximum (maximum probability) or least (minimum probability) support for a particular hypothesis – occurrence of a disease. The product of probabilities rule estimates the probability of occurrence of a disease by merging the posterior probabilities produced by individual classification method with the help of product of probabilities. The average of probabilities rule assigns an instance to a disease class whose average of posterior probabilities is the maximum. Majority voting calculates the votes received for a hypothesis – occurrence of a particular disease from the individual classification methods. The disease class which obtains greatest number of votes is then chosen as the majority decision outcome.

An important observation is that the minimum probability rule performs better when there are no estimation errors or missing values in the dataset. Maximum probability and majority voting rules are less sensitive to estimation errors (Kittler, 1998). A comparison of combination rules – maximum probability, minimum probability, product of probabilities, average of probabilities and majority voting is performed on all standard datasets considered in the present work and the rule selected is shown in Table 3. The combination rule or fusion method that performs the best in forming hybrid ensemble is shown as '✓' and other rules(s) as '✗' in Table 3.

3.3. Functionality of hybrid ensemble

The functionality of hybrid ensemble is shown in this section with the help of standard agriculture dataset- Soybean. Prior to the demonstration of functionality of hybrid ensemble, we observe that for Soybean dataset the ensemble-Vote performs better than Logistic Regression and Naïve Bayes individual classification algorithms. Gain Ratio feature ranking method is applied on Soybean dataset to rank the attributes in order of their ranks (step 1 of Algorithm 1). In (step 2 of Algorithm 1) supervised instance filter-Resample is applied on the result obtained from the previous step. We combine the classification predictions of algorithms Logistic Regression (step 3(ii) of Algorithm 1) and Naïve Bayes (step 3(ii) of Algorithm 1) using ensemble-Vote with a combination rule-maximum probability or average of probabilities (step 4 of Algorithm 1). The resultant disease classification accuracy of hybrid ensemble obtained is observed as 95.46% (step 5 of Algorithm 1) which is greater than ensemble-Vote (94.43%).

4. Results and discussion

Ten-fold cross validation has been successfully used for evaluating the performance of a machine learning algorithm(s) as it offers reliable approximates for classification accuracy on each classification task (Arora and Jain, 2014; Azar et al., 2014; Baldi et al., 2000). The experiments conducted for evaluating the performance of hybrid ensemble are performed using 10-fold cross validation strategy.

4.1. Performance analysis of hybrid ensemble on standard agriculture datasets

The performance of hybrid ensemble approach is tested on four standard agriculture datasets viz. Soybean, Iris Plants, Mushroom and Grub-damage using three performance indices – classification accuracy or accuracy, F-measure and precision (Azar et al., 2014; Özçift,, 2011). It is an important observation that ensemble-Vote performs better than Logistic Regression and Naïve Bayes algorithms for standard agriculture datasets in the context. Table 4 shows the performance observations for ensemble-Vote and hybrid ensemble for indices classification accuracy or accuracy (in%), *F*-measure and precision as weighted average. It is clear from Table 4 that the hybrid ensemble approach performs better than ensemble-Vote.

Substantial rise in classification accuracy is observed for Grub-damage dataset. Applying ensemble-Vote on Grub-damage dataset results in classification accuracy as 47.74% and after applying

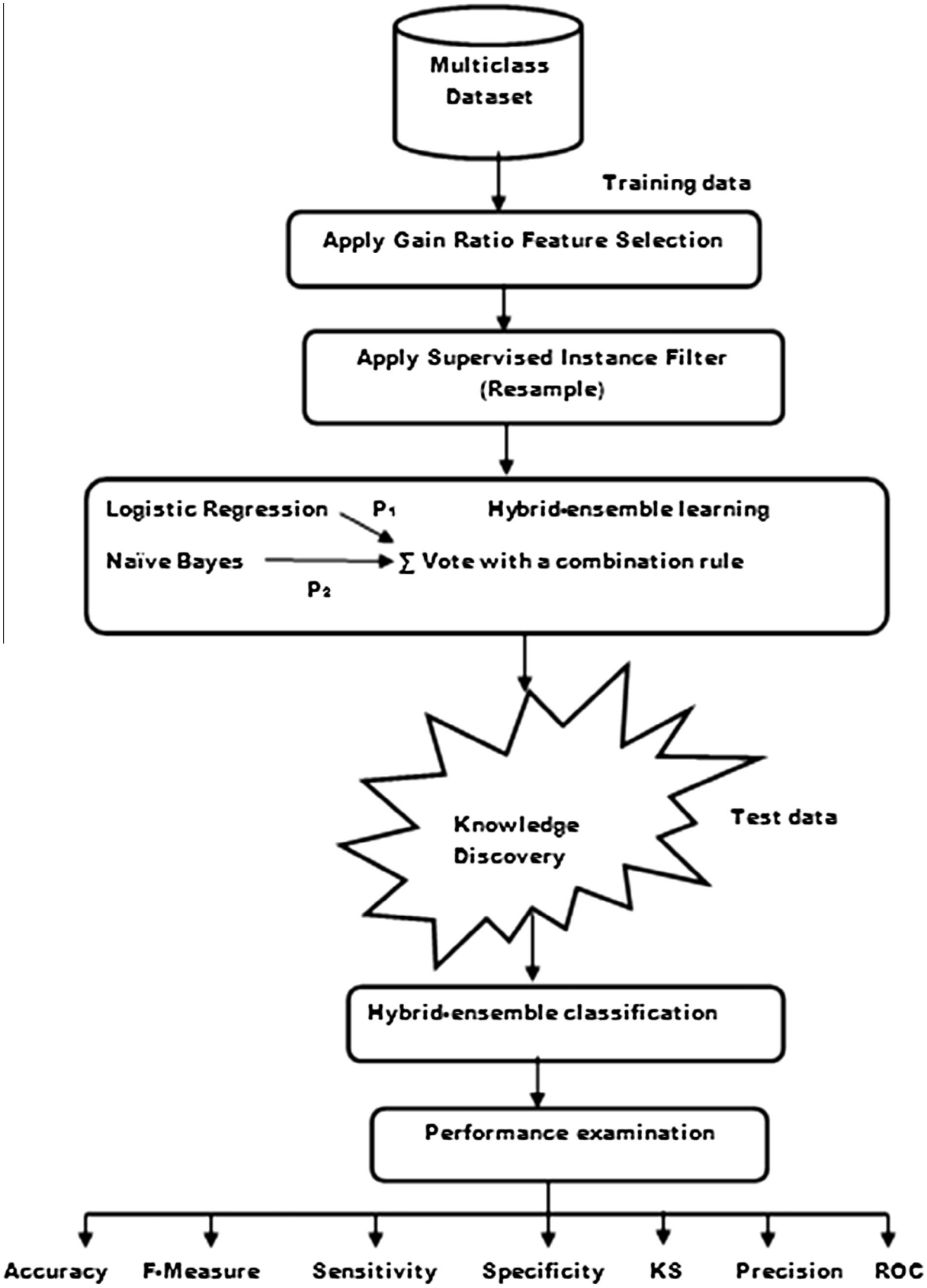


Fig. 1. Design of hybrid ensemble approach.

Table 3
Combination rules selected for standard datasets.

Standard agriculture datasets				
Combination rule	Soybean	Iris plants	Mushroom	Grub-damage
Maximum probability	✓	✓	✗	✗
Minimum probability	✗	✓	✓	✓
Product of probabilities	✗	✓	✗	✗
Average of probabilities	✓	✓	✗	✗
Majority voting	✗	✗	✗	✗

hybrid ensemble approach (steps 1–5 of Algorithm 1) the classification accuracy significantly increases to 71.61%. The accuracy results shown in Table 4 confirm that the proposed hybrid ensemble performs better than ensemble-Vote. *F*-measure and precision values shown in Table 4 also show significant increase with hybrid ensemble as compared to ensemble-Vote.

4.2. Application of hybrid ensemble on oilseed disease multiclass dataset

Before applying hybrid ensemble it is important to note the performance of ensemble-Vote on oilseed disease dataset. Logistic Regression on oilseed disease dataset results in classification accuracy 68.88%. Naïve Bayes algorithm shows disease classification accuracy as 70.97%. Ensemble-Vote shows better performance as compared to Logistic Regression and Naïve Bayes algorithms and results in the disease classification accuracy as 71.67%.

After testing successfully hybrid ensemble approach on four standard agriculture datasets, it is applied on real-life oilseed disease multiclass dataset for accurate diagnosis of oilseed disease (s). The attributes of oilseed disease dataset after applying Gain Ratio (step 1 of Algorithm 1) are ranked for each oilseed-crop in order with respect to target class as Relative-humidity, Root-symptoms, Collar-symptoms, Plant-effect, Root, Pod, Soil moisture, Leaves, Leaf-defoliation, Mycelia, Pod-symptoms, Stem-symptoms, Sclerotia, Collar, Seed-symptoms, Fruiting-bodies, Temperature, Stem, Leaf-symptoms, Leaf-surface, Severity and Seed.

Class distributions in case of oilseed disease dataset are balanced by using Resample. The instance filter-Resample is used with substitution for maintaining uniformity of class distributions. The effect of using supervised instance filter-Resample on class distributions of oilseed disease dataset (after applying step 2 of Algorithm 1) is shown in Table 5.

Table 6 shows the performance observations for classification accuracies as observed for hybrid ensemble approach using 10-fold cross validation in comparison to ensemble-Vote. After successfully completing – steps 1–2 of Algorithm 1, we apply Logistic Regression (step 3(i) of Algorithm 1) and then Naïve Bayes (step 3(i) of Algorithm 1) on the result of previous step. By applying (step 3(ii) of Algorithm 1), increase in disease classification accuracies is observed as 90.32% (P_1) and 88.88% (P_2) as compared to Logistic Regression (68.88%) and Naïve Bayes (70.97%) respectively.

Table 4
Performance index values obtained for ensemble-Vote and hybrid ensemble using standard agriculture datasets.

Standard agriculture datasets					
Machine learning ensemble method	Performance indices	Soybean	Iris plants	Mushroom	Grub-damage
Ensemble-Vote	Accuracy	94.43	96.00	66.58	47.74
	F-measure	0.944	0.960	0.687	0.473
	Precision	0.947	0.960	0.789	0.472
Hybrid ensemble (After applying steps 1–5 of Algorithm 1)	Accuracy	**95.46**	**99.33**	**67.09**	**71.61**
	F-measure	**0.953**	**0.993**	**0.690**	**0.716**
	Precision	**0.954**	**0.993**	**0.787**	**0.718**

Bold text indicates performance improvement.

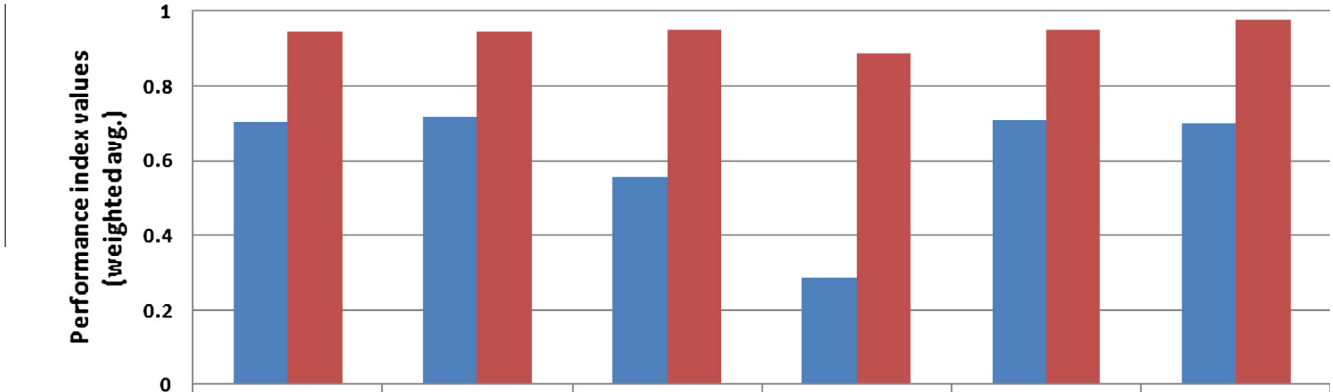
Table 5
Class distributions in oilseed disease dataset before and after sampling.

Class	Class labels of oilseed disease dataset	Before sampling	After sampling
01	Alternaria leaf spot	1680	1690
02	Anthracnose	1170	1160
03	Cercospora leaf spot	1200	1290
04	Charcoal rot	1200	990
05	Collar rot	720	720
06	Myrothecium leaf spot	1320	1220
07	Powdery mildew	1320	1350
08	Sclerotinia stem rot	950	960
09	Phyllosticta leaf spot	1400	1440
10	Rust	2400	2540

Table 6
The classification accuracies obtained for oilseed disease dataset with 10-fold cross validation.

Machine learning method	Classification accuracies (in%)
Logistic Regression	68.88
Naïve Bayes	70.97
Ensemble-Vote	**71.67**
Logistic Regression (After applying steps 1–3 (ii) of Algorithm 1)	90.32
Naïve Bayes (After applying steps 1–3 (ii) of Algorithm 1)	88.88
Hybrid ensemble (After applying steps 1–5 of Algorithm 1)	**94.73**

Bold text indicates performance improvement.



	Precision	Sensitivity	Specificity	KS	F-Measure	ROC
Ensemble-Vote	0.704	0.717	0.554	0.2857	0.708	0.701
Hybrid ensemble	0.946	0.947	0.949	0.8858	0.948	0.977

Fig. 2. Performance graph of hybrid ensemble and ensemble-Vote for identification of oilseed disease(s).

Ensemble-Vote with a combination rule (maximum probability) is used for combining the predictions (step 4 of Algorithm 1) of Logistic Regression (P_1) and Naïve Bayes (P_2). The resultant disease classification accuracy of hybrid ensemble (step 5 of Algorithm 1) significantly increases to 94.73% as shown in Table 6.

The performance indices – precision, sensitivity, specificity, KS, *F*-Measure and ROC as observed for Vote and hybrid ensemble are shown in Fig. 2.

It is clear from Fig. 2 that all the six performance indices show significant increase with hybrid ensemble approach on oilseed disease features as compared to the ensemble-Vote.

5. Conclusions

The paper proposes a new hybrid ensemble approach for improvement of classification accuracy for multiclass classification problems. It is successfully applied for accurate diagnosis of oilseed diseases. The performance of proposed hybrid ensemble is tested for classification accuracy with 10-fold cross validation on four standard agriculture datasets. The accuracy results obtained for these standard datasets prove that the hybrid ensemble approach shows better classification accuracies as compared to ensemble-Vote for all of these standard datasets. The hybrid ensemble approach is applied to oilseed disease diagnosis multiclass classification problem. The disease classification accuracy is improved up to 94.73% by applying hybrid ensemble approach as compared to ensemble-Vote which yields accuracy 71.67%. Hence it is concluded that hybrid ensemble approach might be a good alternative for accurate classification in similar multiclass classification or prediction problems.

References

Arora, A., Jain, R., 2014. Machine learning for diagnosis of soybean diseases. Soybean Res. 12, 256–262 (Special Issue - 2).

Azar, A.T., Elshazly, H.I., Hassanien, A.E., Elkorany, A.M., 2014. A random forest classifier for lymph diseases. Comput. Meth. Programs Biomed. 113 (2), 465–473.

Baker, K.M., Kirk, W.W., 2007. Comparative analysis of models integrating synoptic forecast data into potato late blight risk estimate systems. Comput. Electron. Agric. 57 (1), 23–32.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A., Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification and overview. Bioinformatics 16 (5), 412–424.

Bartaria, A.M., Shukla, A.K., Kaushik, C.D., Kumar, P.R., Singh, N.B., 2001. Major diseases of Rapeseed-Mustard and their management. Technical bulletin No. 10, National Research Centre for Rapeseed-Mustard (ICAR), Bharatpur (Rajasthan), India, pp. 1–44.

Battiti, R., Colla, A.M., 1994. Democracy in neural nets: voting schemes for classification. Neural Netw. 7 (4), 691–707.

Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach. Learn. 36 (1), 105–139.

Bay, S.D., 1999. Nearest neighbor classification from multiple feature subsets. Intell. Data Anal. 3 (3), 191–209.

Bolón-Canedo, V., Sanchez-Marono, N., Alonso-Betanzos, A., 2012. An ensemble of filters and classifiers for microarray data classification. Pattern Recogn. 45 (1), 531–539.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140.

Danger, R., Segura-Bedmar, I., Martínez, P., Rosso, P., 2010. A comparison of machine learning techniques for detection of drug target articles. J. Biomed. Inform. 43 (6), 902–913.

Dietterich, T.G., 2000. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach. Learn. 40 (2), 139–157.

EI-Bendary, N., EI-Hariri, E., Hassanien, A.E., Badr, A., 2015. Using machine learning techniques for evaluating tomato ripeness. Exp. Syst. Appl. 42 (4), 1892–1905.

Frank, A., Asuncion, A., 2010. UCI machine learning repository. <http://archive.ics.uci.edu/ml/datasets.html>, (accessed 29.05.2015).

Ghewande, M.P., Desai, S., Basu, M.S., 2002. Diagnosis and management of major diseases of groundnut. In: Bulletin, National Research Centre for Groundnut, Junagadh, Gujarat, India, pp. 1–36.

Gupta, G.K., Chauhan, G.S., 2005. Symptoms, identification and management of soybean diseases. Technical bulletin, National Research Centre for Soybean (ICAR), Indore, India, pp. 1–92.

Gutiérrez, P.A., López-Granados, F., Peña-Barragán, J.M., Jurado-Expósito, M., Hervás-Martínez, C., 2008. Logistic regression product-unit neural networks for mapping Ridolfia segetum infestations in sunflower crop using multitemporal remote sensed data. Comput. Electron. Agric. 64 (2), 293–306.

Hall Mark, E.F., 2009. The WEKA data mining software: an update. SIGKDD Explor. 11 (1), 10–18.

Hall, M.A., Smith, L.A., 1998. Practical feature subset selection for machine learning. In: Proceedings of 21st Australasian Computer Science Conference (ACSC'98), Perth, pp. 181–191.

Hansen, L.K., Salamon, P., 1990. Neural network ensembles. IEEE Trans. Pattern Anal. Mach. Intell. 12 (10), 993–1001.

Hartman, G.L., Sinclair, J.B., Rupe, J.C., 1999. Compendium of Soybean Diseases, four ed. The American Phytopathological Society. Academic press, St. Paul, Minnesota, pp. 1–100.

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. Comput. Electron. Agric. 108 (1), 250–257.

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20 (8), 832–844.

Hsu, K.-W., 2012. Hybrid ensembles of decision trees and artificial neural networks. In: IEEE international conference on Computational Intelligence and Cybernetics (CyberneticsCom), Bali, pp. 25–29.

Ibrahim, H.E., Badr, S.M., Shahee, M.A., 2012. Adaptive layered approach using machine learning techniques with gain ratio for intrusion detection systems. Int. J. Comput. Appl. 56 (7), 10–16.

Kittler, J., 1998. Combining classifiers: a theoretical framework. Pattern Anal. Appl. 1 (1), 18–27.

Kotsiantis, S.B., 2007. Supervised machine learning: a review of classification techniques. Informatica 31 (3), 249–268.

Kundu, P.K., Panchariya, P.C., Kundu, M., 2011. Classification and authentication of unknown water samples using machine learning algorithms. ISA Trans. 50 (3), 343–520.
Michalski, R., Davis, J., Visht, V., Sinclair, J., 1983. A computer-based advisory system for diagnosing soybean diseases in Illinois. Plant Dis. 67, 459–463.
Namsrai, E., Munkhdalai, T., Li, M., Shin, J.-H., Namsrai, O.-E., Ryu, K.H., 2013. A feature selection-based ensemble method for arrhythmia classification. J. Inform. Process. Syst. 9 (1), 31–40.
Özçift, A., 2011. Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. Comput. Biol. Med. 41 (5), 265–271.
Opitz, D.W., 1999. Feature selection for ensembles. In: Proceedings of the National Conference on Artificial Intelligence. Menlo Park, CA, USA, pp. 379–384.
Phadikar, S., Sil, J., Das, A.K., 2013. Rice diseases classification using feature selection and rule generation techniques. Comput. Electron. Agric. 90 (C), 76–85.
Rocha, A., Hauagge, D.C., Wainer, J., Goldenstein, S.K., 2010. Automatic fruit and vegetable classification from images. Comput. Electron. Agric. 70 (1), 96–104.
Sankaran, S., Mishra, A., Ehsani, R., Davis, C., 2010. A review of advanced techniques for detecting plant diseases. Comput. Electron. Agric. 72 (1), 1–13.
Schapire, R.E., 1990. The strength of weak learnability. J. Mach. Learn. 5 (2), 197–227.
Shouman, M., Turner, T., Stocker, R., 2011. Using decision tree for diagnosing heart disease patients. Proceedings of the Ninth Australasian Data Mining Conference 121, 23–30.
Silva, L.O.L.A., Koga, M.L., Cugnasca, C.E., Costa, A.H.R., 2013. Comparative assessment of feature selection and classification techniques for visual inspection of pot plant seedlings. Comput. Electron. Agric. 97, 47–55.
Stamatatos, E., Widmer, G., 2005. Automatic identification of music performers with learning ensembles. Artif. Intell. 165 (1), 37–56.
Sun, S., Zhang, C., Zhang, D., 2007. An experimental evaluation of ensemble methods for EEG signal classification. Pattern Recogn. Lett. 28 (15), 2157–2163.
Timmermans, A.J.M., Hulzebosch, A.A., 1996. Computer vision system for on-line sorting of pot plants using an artificial neural network classifier. Comput. Electron. Agric. 15 (1), 41–55.
Ting, K., Witten, I., 1999. Issues in stacked generalization. Artif. Intell. Res. 10 (1), 271–289.
Witten, I.H., Frank, E., 2005. Data Mining: Practical machine learning tools and techniques, second ed. Morgan Kaufman series in Data Management Systems, San Francisco, CA, USA.
Wolpert, D.H., 1992. Stacked generalization. Neural Netw. 5 (2), 241–259.
Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J., Steinberg, D., 2007. Top 10 algorithms in data mining. Knowledge Inform. Syst. 14 (1), 1–37.
Yen, E., Mike Chu, I.-W., 2007. Relaxing instance boundaries for the search of splitting points of numerical attributes in classification trees. Inf. Sci. 177 (5), 1276–1289.
Zheng, Z., Webb, G.I., 1999. Stochastic attribute selection committees. Lect. Notes Comput. Sci. 1574, 123–132.