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

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