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