Selection of Important Features for Optimizing Crop Yield Prediction

Maya Gopal P S, VIT University, Chennai, India
Bhargavi R, School of Computing Science and Engineering, VIT University, Chennai, India

ABSTRACT

In agriculture, crop yield prediction is critical. Crop yield depends on various features including geographic, climate and biological. This research article discusses five Feature Selection (FS) algorithms namely Sequential Forward FS, Sequential Backward Elimination FS, Correlation based FS, Random Forest Variable Importance and the Variance Inflation Factor algorithm for feature selection. Data used for the analysis was drawn from secondary sources of the Tamil Nadu state Agriculture Department for a period of 30 years. 75% of data was used for training and 25% data was used for testing. The performance of the feature selection algorithms are evaluated by Multiple Linear Regression. RMSE, MAE, R and RR MSE metrics are calculated for the feature selection algorithms. The adjusted R2 was used to find the optimum feature subset. Also, the time complexity of the algorithms was considered for the computation. The selected features are applied to Multilinear regression, Artificial Neural Network and M5Prime. MLR gives 85% of accuracy by using the features which are selected by SFFS algorithm.

KEYWORDS

Artificial Neural Network, Feature Selection Algorithm, M5Prime, Model Validation, Multiple Linear Regression, Performance Metrics

1. INTRODUCTION & RELATED WORK

Data mining is a process of discovering previously unknown and potentially interesting patterns in large datasets (Frawley et al., 1991). The data mining process includes fixing the problem, understanding the data, preparing the data, applying the right techniques to build the models, interpreting the results and use the data into action. Now-a-days, intelligent data mining and knowledge discovery by artificial neural network and feature selection algorithms have become the important revolutionary concepts in prediction and modelling (Roddick et al., 2001, Schuize et al., 2005). Data set may contain redundant information that does not directly impact the predictions, and it may contain highly correlated attributes. The data sets are typically not gaining any new information by including all the attributes. In data mining, feature selection algorithms are useful for identifying irrelevant attributes to be excluded from the dataset (Che et al., 2017, Kotu et al., 2015). Feature selection in predictive analytics refers to the process of identifying few most important features or attributes that are essential in building a model for an accurate prediction. Efficient predictive models can improve the quality of the decision making. Feature selection optimizes the performance of the data mining algorithm and makes it easier for the analyst to interpret the outcome of the modeling. This procedure can reduce not only the cost of recognition by reducing the number of features to be collected, but in some cases it can also provide a better classification of prediction accuracy due to finite sample size effects (Jain et
al., 1982). This strategy aims at further reducing the number of features. Adding feature selection to the analytical process has several benefits: it simplifies and narrows down the scope of the features that are essential in building a predictive model, to minimize the computational time and memory requirements using the feature selection algorithms (Pal and Foody, 2010). The focus can be directed to a subset of predictors which are very essential.

Researchers work with different feature selection models to optimize their data sets. Automated feature selection for every algorithm with the conventional approach of stepwise regression for feature selection (Alvarez, 2009). Gonzalez- Sanchez et al. (2014), performed an exhaustive search of the feature selection algorithms. H. Liu et al. (1996) proposed a consistency based feature selection mechanism to evaluate the worth of a subset of the attributes by the level of consistency in the class values when the training instances are projected onto the subset of attributes. In this model the consistency of any subset can never be lower than that of the full set of attributes. M. Hall (1999) proposed a correlation based approach to feature selection in different datasets and demonstrated how it can be applied to both classification and regression problems for machine learning. Karimi et al. (2013) presented a hybrid feature selection methods by combining symmetric uncertainty measure and gain measure. Both measures for each feature-class correlation were calculated first and then rank feature according to average score value. High ranked feature greater than a threshold values was selected. They evaluated their system using knowledge discovery data dataset and Naïve Bayes algorithm. Correlation based method, Gain Ratio method and Information Gain method were used by Chaudhary et al. (2013) and presented the performance evaluation of three feature selection methods with optimized Naïve Bayes is performed on mobile device. Zhang et al. (2005) performed a principal components analysis to transform data and used stepwise feature selection for multiple linear regression (MLR). In most experiments conducted, researchers collect data that are supposedly related to the phenomenon of interest, given resource and/or time constraints on the collection and analysis of data. The oriented collection of data means that these kinds of datasets have only pre-approved features. Feature selection can enhance model quality by discarding unwanted features or simply decreasing the model and computational complexity by keeping the most important features with an example (Ruß et al., 2010).

The major research on agricultural management is the development of cost effective methods for predicting the crop yield with the available parameters like irrigation details, fertiliser details and temperature (Bocca et al., 2016; Zeynep Özcan et al., 2017). In recent years, the data analysts are working towards developing predictive models in the form of expert systems to improve agricultural yield, considering the environmental features, soil quality, irrigation and land usage (Aggarwal, 1995; Bagley et al., 2012; De wit et al., 1987; Safa et al., 2015; Patricio Grassinina et al., 2015). The crop yield is classified into long term (years ahead), medium term (months to years ahead) and short term (weeks to months ahead). Long term paddy crop is focuses on this reserach work. Recent research focused on applying the machine learning and data mining techniques to predict the yield (Majumdar et al., 2017). Several researchers used regression models for crop yield prediction to analyse its applicability with other models (Zhang et al., 2005; Ji et al., 2007; Alvarez et al., 2009; Gonzalez-Sanchez et al., 2014; Matsumura et al., 2014). Researchers compared the classical statistical models against Artificial Neural Networks, regression trees and support vector regression for better crop yield models (Drummond et al., 2003; Fortin et al., 2011; Ruß, 2010; Felipe et al., 2016).

The aim of this research work is to identify important paddy field conditions (features) using feature selection algorithms for providing a comprehensive view about paddy crop yield. Understanding the importance of features among a large dataset of features can play a key role in improving the yield under field conditions. Our study investigates the behaviour of five feature selection algorithms with sixteen features and the outcome is given as input to multiple linear regression model, artificial neural network and M5Prime to find the accuracy.
Related Content

The Effect Degree Analysis of Human Activities on Regional Groundwater Level Based on Variable Fuzzy Optimization Model
www.igi-global.com/article/the-effect-degree-analysis-of-human-activities-on-regional-groundwater-level-based-on-variable-fuzzy-optimization-model/128851?camid=4v1a

Presowing Processing of Seeds in Magnetic Field
www.igi-global.com/chapter/presowing-processing-of-seeds-in-magnetic-field/201354?camid=4v1a
Innovative Means and Technologies Ensuring Electrical Safety of Agricultural Production

Decision Support for Agricultural Consultants With Semantic Data Federation