International Journal of Pure and Applied Mathematics
Volume 118 No. 22 2018, 139-144
ISSN: 1314-3395 (on-line version)
url: http://acadpubl.eu/hub
Special Issue

ijpam.eu

FEATURE SELECTION FOR YIELD PREDICTION USING BORUTA ALGORITHM

[1]Maya Gopal P.S, [2]Bhargavi R
[1,2]School of Computing Science and Engineering, VIT University, Chennai, India
[1]mayagopal.ps2016@vitstudent.ac.in

Abstract: Feature selection is one of the important tasks in the data analytic research where the datasets have large number of variables. These applications include crop yield prediction, irrigation management, gene expression array analysis, and combinatorial chemistry. The objective of variable selection is three-fold: improving the prediction performance, providing effective predictors, and a better understanding of the underlying process that generated the data. Appropriate features give better prediction accuracy. This research paper focuses on feature selection using Boruta algorithm for the yield prediction. The algorithm is designed as a wrapper around a Random Forest classification algorithm. The selected features are given as the input of Multiple Linear Regression (MLR) for prediction accuracy. Analysis on the MLR model reveals the direct relationship of yield with crop area, number of open wells, and maximum temperature and an inverse relationship with canal length, number of tanks and the nitrogen fertilizer resulting at 84% accuracy.

Keywords: Boruta algorithm,MLR, Random Forest classification , yield prediction.

1. Introduction and Related work

Data mining is the process of discovering previously unknown and potentially interesting patterns in large datasets. Many data sets have large number of features [Deogun et al 1997]. Data sets may contain various redundant information which do not impact the prediction, also the data set may contain relatively correlated attributes. There are siginificant drawbacks in considering large number of features for any mining or learning purpose. Some of the disadvanges include a) increase in the processing time of the algorithms,b) utilization of too many resources and c) simply difficult in maintanence [Kohavi R, John GH (1997)]. Another important issue is many machine learning algorithms give less accuracy when the number of features are significantly higher than optimal. The selection of the optimal feature set giving best possible prediction results is desirable for practical reasons. These issuses have been deeply studied and there are plenty of algorithms which were developed to reduce feature set to a manageable size. Finding all important attributes, instead of only the non-redundant ones, may be very useful in itself. In particular, this is necessary when one is interested in understanding mechanisms related to the subject of interest, instead of merely building a black box predictive model. When dealing with crop yield prediction, identification of all important features which are related to production is necessary for complete understanding of the process, whereas a minimal-optimal set of features might be more useful [Nilsson et al 2007]. The all-relevant problem of feature selection is more difficult than usual minimal-optimal one. The degradation of the prediction accuracy, upon removal of the feature from the feature set, is sufficient to declare that a particular feature important. But lack of this effect is not sufficient to declare it as unimportant. Therefore, it needs another criterion for declaring variables important or unimportant. Filtering methods cannot be used for declaring variables important or otherwise becuase of lack of direct correlation between given features. The decision is not a proof that this feature is not important in conjunction with the other features [Guyon et al 2003]. The wrapper algorithms are computationally more demanding than filters. The research paper focuses on the implementation of the Boruta algorithm developed by R development team for finding all important features for crop yield prediction.

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., performed an exhaustive search of the feature selection algorithms. H. Liu et al. 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 his model the consistency of any subset can never be lower than that of the full set of attributes. M. Hall 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 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 methods were used by A. Chaudhary, et. al. and presented the performance evaluation of three feature selection methods with optimized Naïve Bayes is performed on mobile device. . Zhang et al. performed a Principal Components Analysis to transform data and used stepwise feature selection for 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 bogus features or simply decreasing the model and computational complexity by keeping the most important features with an example[Ruß et al 2010].

2. Boruta algorithm

Boruta algorithm is a wrapper built around the random forest classification algorithm implemented in the R package random forest [Liaw et al 2002]. The random forest classification algorithm is relatively quick, can usually be run without tuning of parameters and gives a numerical estimate of the feature importance. It is an ensemble method in which classification is performed by voting of multiple unbiased weak classifiers or decision trees. These trees are independently developed on different bagging samples of the training set. The important measure of an attribute is obtained as the loss of accuracy of classification caused by the random permutation of attribute values between objects. It is computed separately for all trees in the forest which use a given attribute for classification. Then, the average and standard deviation of the accuracy loss are computed. The Z score computed by dividing the average loss by its standard deviation can be used as the important measure. In Boruta, Z score has the important measure since it takes into account the fluctuations of the mean accuracy loss among trees in the forest. For each attribute, the corresponding `shadow' attribute is created, and its values are obtained by shuffling values of the original attribute across objects. Then a classification is performed using all attributes of this extended system and importance of all attributes has been computed. The importance of a shadow attribute can be non zero only due to random fluctuations. Thus, the set of important shadow attributes is used as a reference for deciding which attributes are truly important. Moreover, it is dependent on the particular realization of shadow attributes. Therefore, it needs to repeat the re-shuffling procedure to obtain statistically valid results.

The Boruta algorithm consists of following steps:

1. Extend the information system by including copies of all the shadow attributes.
2. Shuffle the added attributes with the original attribute to remove their correlations with the response.
3. Run a random forest classifier on the extended information system and calculate the Z score value.
4. Find the maximum Z score among shadow attributes (MZSA), and then assign a hit to every attribute that scored better than MZSA.
5. For each attribute with undetermined, perform a two-sided test of equality with the MZSA.
6. Find the attributes which have significantly lower importance than MZSA as `unimportant' and remove that attributes from the information system.
7. Find the attributes which have significantly higher importance than MZSA as `important'.
8. Remove all shadow attributes from the information system.
9. Repeat the process until the importance is assigned for all the attributes.

The time complexity of the algorithm is O (P.N), where P and N are respectively the numbers of attributes and objects.

3. Multiple Linear Regression model

Multiple Linear Regression is one of the statistical models that specifies how one set of features, called dependent features, functionally depend on another set of features, called independent features. It has been applied most frequently for crop yield prediction because production (yield) is normally dependent on number of parameters such as production area, irrigation, fertilizers, and weather parameters. Here, yield is the dependent feature (response feature) and other parameters are independent features .

MLR model is described by [Gonzalez- Sanchez et al.]

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_{ik} + \varepsilon_i$$

where, k is the number of feature, x_{ij} is the i^{th} observation of the feature, x_j, β_1, β_2, $\beta_{3,...}$ β_k are the regression coefficients and ε_i is the error term or residual. The above equation can be written as $Y = \sum X\beta + \varepsilon$.

4. MLR Model for Crop yield Prediction

The regression equation defines the given regression model with six independent features. Here, PD is the dependent feature and AH, OW, Tmax, TK, CL and NF are the independent features. The slope of variables indicates that, for a given value of the particular variable estimate to decrease or increase of predicted value.

$$PD = -0.061 - 0.166CL - 0.244TK + 0.216OW + 2971.376AH + 144.980T\max - 242.483NF$$

5. Results and discussion

The agricultural data set consists of 745 objects and 14 attributes. Among the 14 attributes two of them are rejected, one after the initial round 2, and one during the final round. The remaining attributes are indicated as confirmed. Boruta performed 77 random Forest runs in 38 secs. The confirmed important 12 attributes are Area, Canals Length, K, N, P, Open Well, Tanks, Tube Well, Solar Radiation, Average Temperature, maximum temperature and seed rate. Two attributes confirmed as unimportant are Rainfall and minimum Temperature. Figure 1 shows the Z scores variability among attributes during the Boruta run.

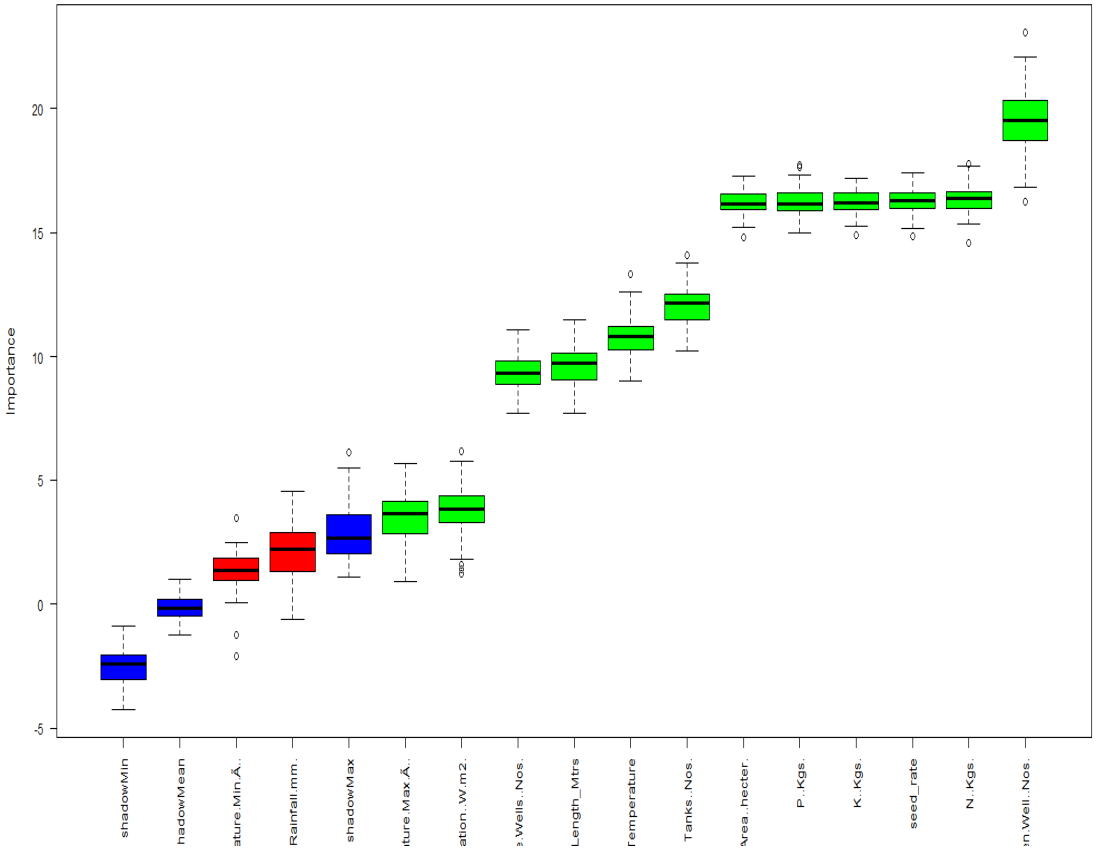

Figure 1. Boruta result plot for given(agricultural) data.

Blue boxplots correspond to minimal, average and maximum Z score of a shadow attribute. Red and green boxplots represent Z scores of respectively rejected and confirmed attributes.

The attStats function creates a data frame containing each attribute's Z score statistics and the fraction of random forest runs in which this attribute is more important than the most important shadow one is shown in table 1. After 77 iterations it is selected 12 attributes and rejected 2 attributes.

Table 1. Attributes statistics generated by attStats function

Attribute	meanImp	medianImp	minImp	maxImp	normHits	decision
Canals Length	9.65155	9.73844	7.70481	11.49111	1.00000	Confirmed
Tanks	12.06570	12.16517	10.24978	14.11364	1.00000	Confirmed
Tube Wells	9.36174	9.32223	7.70838	11.09445	1.00000	Confirmed
OpenWell	19.48032	19.50108	16.25013	23.07546	1.00000	Confirmed
Area	16.21193	16.16335	14.83029	17.26187	1.00000	Confirmed
Rainfall	2.12296	2.22054	-0.63201	4.53189	0.32586	Rejected
Average Temperature	10.77270	10.80036	9.01738	13.33703	1.00000	Confirmed
Minimum Temperature	1.22926	1.34266	-2.10880	3.46207	0.04494	Rejected
Maximum Temperature	3.51274	3.65767	0.93181	5.68089	0.69663	Confirmed
Solar Radiation	3.84015	3.82126	1.20870	6.15276	0.76405	Confirmed
Nitrogen	16.35776	16.37735	14.57802	17.74994	1.00000	Confirmed
Phosphate	16.25422	16.17145	14.96816	17.70242	1.00000	Confirmed
Potash	16.24291	16.21778	14.91014	17.16561	1.00000	Confirmed
Seed rate	16.29077	16.28856	14.86162	17.41635	1.00000	Confirmed

The selected features are fed into the MLR to find the accuracy.The accuracy has been identified based on its Adjusted R^2 value. The Adjusted R^2 shows the variance explained by the model. For the current model it gives 84% of the Adjusted R^2 value for the given agricultural dataset. From the value, it could be understood understand that the selected features are important for the yield prediction.

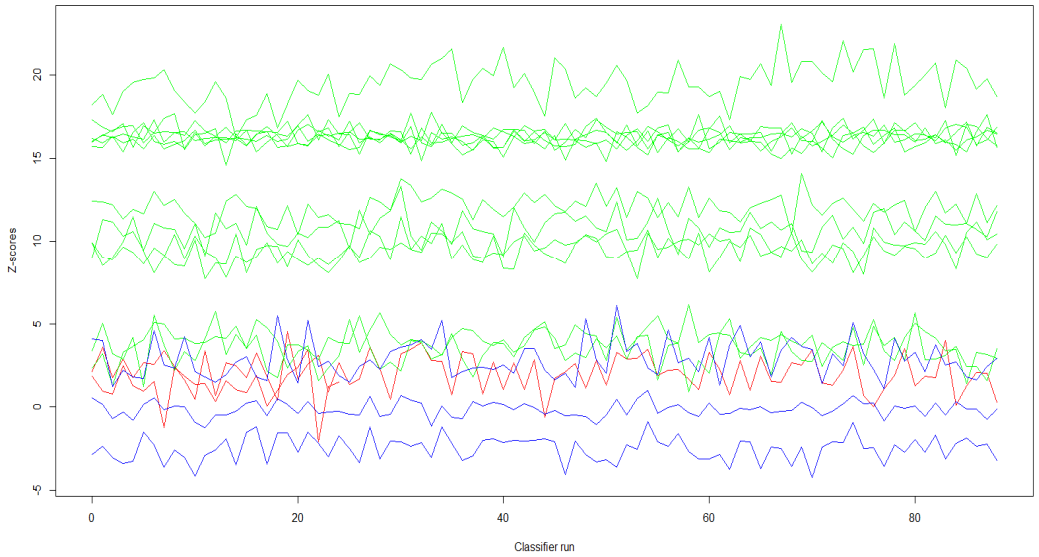

Figure 2. Z score evolution during Boruta run

Figure 2 shows the Z score evolution during Boruta run. Green lines correspond to confirmed attributes, red to rejected ones and blue to respectively minimal, average and maximal shadow attribute importance.

6. Conclusion

In this research work , Bourta algorithm has been used to select the important features for the prediction of crop yield. It has been observed that the features crop area, canal length, number of open well, number of tube well, number of tanks, maximum temperature, average temperature, fertilizers nitrogen, phopers and pottash, solar radiation and seed rate are important features. These features are given as the input of MLR model for finding accuracy. By using these features into the model has achieved 84% of accuracy.

References

[1] Alvarez, R., 2009. Predicting average regional yield and production of wheat in the argentine pampas by an artificial neural network approach. Eur. J. Agron. 30 (2), 70–77

[2] Chaudhary A., S. Kolhe and Rajkamal, “Performance Evaluation of feature selection methods for Mobile devices”, Vol. 3, Issue 6, Nov - Dec 2013, pp. 587-594.

[3] Deogun JS, Raghavan VV, Sarkar A, Sever H and others (1997), "Data mining: Trends in research and development", Lin and Cercone. Vol. 191, pp. 9.

[4] Gonzalez-Sanchez, A., Frausto-Solis, J., Ojeda-Bustamante, W., 2014. Predictive ability of machine learning methods for massive crop yield prediction. Span. J. Agric. Res. 12 (2)

[5] Guyon I, Elisseeff A (2003). "An Introduction to Variable and Feature Selection." Journal of Machine Learning Research, 3, 1157-1182.

[6] Hall M., “Feature Selection for Discrete and Numeric Class Machine Learning”, Department of Computer Science, The University of Waikato 1999.

[7] Karimi Z., M. Mansour and A. Harounabadi “Feature Ranking in Intrusion Detection Dataset using combination of filtering”, International Journal of Computer Applications,Vol. 78, September 2013.

[8] Kohavi R, John GH (1997, "Wrappers for Feature Subset Selection", Artificial Intelligence, 97, 273-324.

[9] Liaw A, Wiener M (2002). "Classification and Regression by Random Forest" R News, 2(3),18 - 22.

[10] Liu H. and R. Setiono, “A probabilistic approach to feature selection - A filter solution,” the 13th International Conference on Machine Learning, pp. 319-327, 1996.

[11] Nilsson R, Pena J, Bjorkegren J, Tegner J (2007). "Consistent Feature Selection for Pattern Recognition in Polynomial Time" The Journal of Machine Learning Research, 8, 612.

[12] R Development Core Team (2010). R: A Language and Environment for Statistical Computing.R Foundation for Statistical Computing, Vienna, Austria.

[13] Ruß G, Kruse R, 2010. Feature selection for wheat yield prediction. In: Research and development in intelligent systems XXVI (Bramer M et al., eds.), Springer-Verlag, London.

[14] Zhang, B., Valentine, I., Kemp, P., 2005. Modelling the productivity of naturalised pasture in the north island, New Zealand: a decision tree approach. Ecol. Model. 186 (3), 299–311

[15] S.V.Manikanthan and K.srividhya "An Android based secure access control using ARM and cloud computing", Published in: Electronics and Communication Systems (ICECS), 2015 2nd International Conference on 26-27 Feb. 2015,Publisher: IEEE,DOI: 10.1109/ECS.2015.7124833.

[16] T. Padmapriya and V. Saminadan, “Priority based fair resource allocation and Admission Control Technique for Multi-user Multi-class downlink Traffic in LTE-Advanced Networks”, International Journal of Advanced Research, vol.5, no.1, pp.1633-1641, January 2017.