Machine learning model for automation of soil texture classification

K. Radhika* and D. Madhavi Latha1

Department of Information Technology, Chaitanya Bharathi Institute of Technology, Hyderabad-500 075, Telangana, India.

Received: 11-06-2018 Accepted: 12-01-2019 DOI: 10.18805/IJARe.A-5053

ABSTRACT

Soil formation is a long term process and diverse soils are formed in different localities due to various soil forming factors over the landscape. Soil classification plays critical role in various aspects of agricultural engineering. Physico-chemical parameters play an important role in soil classification. In this paper, we present a comprehensive classification model for soil texture classification by using Linear Discriminant Analysis (LDA). We took the Physico-chemical properties of the soil, which include soil moisture, temperature, electrical conductivity, pH, organic carbon, available nitrogen, available phosphorus and potassium as independent variables, while the soil type was taken as the dependent variable. Feature selection is employed using Boruta algorithm. The performance of the proposed classification model is evaluated and expressed in terms of overall accuracy and kappa coefficient. Results show that the average prediction accuracy and kappa coefficient of the proposed model are 96.3% and 0.944 respectively, indicating that the model can be used effectively for soil classification for a set of suitable dependent variables.

Key words: Accuracy, Boruta, Classification, Kappa coefficient, Linear discriminant analysis, Physico-chemical factors, Soil texture.

INTRODUCTION

Soil is composed of organic matter, minerals, gases, liquids, and organisms that provides substratum for plant and animal life. Soil classification plays critical role in crop management, soil improvement, land consolidation, management of drainage, soil erosion and irrigation (Bhaskar B.P. et al., 2015; Dickson A.A et al., 2002). The physical and chemical parameters measured from real-time field samples are the most influential features in soil characterization. Physical parameters such as moisture and temperature are related to the organization of the particles and pores, reflecting effects on root growth, speed of plant emergence and water infiltration. The chemical parameters such as pH, organic carbon, available nitrogen, phosphorus and potassium determine the nutrient availability, presence of other organisms and also mobility of contaminants.

Soil texture is useful for determining productivity of soil and an important parameter in its management (Ram. et al., 2013). Sandy Loam (SL) soils are loose, friable and can easily handle tillage operations. They facilitate drainage, aeration and allow rapid evaporation and percolation. Sandy soils have very little water holding capacity. Such soils cannot stand drought and unsuitable for dry farming. Sandy soils are poor store houses of plant nutrients and contain low organic matter. Leaching of applied nutrients is very high. In this type of soil few crops such as potato, groundnut and cucumbers can be grown (Savalia et al. 2010). Sandy Clay Loam (SCL) soils are difficult to till and require much skill in handling. SCL soils are exceedingly sticky when moist and when dry become very hard and difficult to break. They have fine pores, and are poor in drainage and aeration. They have a high water holding capacity and poor percolation, which usually results in water logging. They are generally very fertile soils, in regard to plant nutrient content. Crops such as rice, jute, sugarcane can be grown very successfully in these soils. Silt Clay Loam (SICL) soils, comprise mainly of intermediate sized particles are fairly well drained and hold more moisture than sandy soils. These soils are potentially fertile as they hold nutrients bound to the clay minerals in the soil. Plants with thicker roots, such as corn, alfalfa, and sugar beets grow well in these soils.

Standard statistical analysis techniques for soil classification are both time consuming and expensive. Machine learning is a field of computer science that uses statistical algorithms to progressively improve the performance of the computer program in detecting patterns in data. Machine learning allows automation of tasks, time cycle reduction and efficient utilization of resources and can handle multidimensional and multi-variety data.

Two general approaches have been applied to predict soil taxonomic classes using machine learning. The first approach attempts to find and extract soil class-landscape relationships from existing digitized soil polygon maps when the exact locations (GPS coordinates) of soil pedon observations are unknown. As part of this approach
different methods were used for soil classification. An Artificial Neural Network (ANN) based methodology was developed (Behren et al., 2005) for digital soil mapping that is able to predict soil unit within a test area in Rhineland, Palatinate, Germany. A classification tree based method for predicting soil distribution at an unvisited location was developed (Grinand et al., 2017) using soil-landscape pattern extracted from soil map. A Random Forest based predictive model was developed (Massawe et al., 2018) to map numerically classified soil clusters of a portion of Kilombero valley, Tanzania using Digital Soil Mapping (DSM) techniques.

The second approach attempts to construct soil class-landscape relationships from soil pedon observations made by field sampling at known locations. As part of this approach Conditional hypercube sampling and random forest modelling were used (Rad et al., 2014) for mapping soil taxonomy groups for around 85,000 ha in Golestan province, Iran. The updated soil map using this technique was 13.4% more accurate than existing conventional soil series map. Kriging with external drift, cokriging, regression Kriging and Residual Maximum Likelihood- Empirical Best linear Unbiased Predictor(REML-EBULP) were used (Hengl et al., 2007) for prediction of soil properties for making use of the ancillary variables as covariates. Logistic regression and multinomial logistic regression were used (Jafari et al., 2014) to produce soil class maps in the Zarand region of Southeast Iran. Digital soil mapping model based on remote sensing data from the Moderate resolution Imaging Spectroradiometer (MODIS) platform was developed (Vagen et al., 2016) for Africa based on balanced field survey data. The choice of these approaches largely depends on the availability of soil pedon observations with known locations.

The objective of this study is to develop and evaluate a machine learning model using LDA to identify soil texture based on physico-chemical parameters. The relative importance of these parameters is then investigated. The current study is based on the real time data set having soil samples collected from six different locations of Hyderabad region of Telangana in different climatic conditions. The selection of these localities was based on the variability in the habitat features. The observed results from this investigation provide useful information for efficiently mapping the physico-chemical characteristics of soil and soil texture.

**MATERIALS AND METHODS**

**Study Site:** The current research was undertaken to characterize the soils of Hyderabad under Telangana region in order to assess their land capability, irrigability and suitability for various crops and also to assess soil fertility constraints of the area. Hyderabad is geographically located in the northern part of Deccan plateau, in southern part of India on the banks of Musi River. Hyderabad has an altitude of about 542m above the sea level and has a unique combination of tropical wet and dry climate that borders on a hot semi-arid climate. This region remains fairly warm through most parts of the year and does not receive much rainfall in the monsoon. During summer months, mercury goes as high as 42°C and while in winters, the minimum temperature may come down to as low as 12°C. Thus for most parts of the year, the climate of Hyderabad remains fairly moderate. The city’s soil type is mainly red sandy with some area of black cotton soil. Hyderabad falls in the seismic zone-I and is seismically least exposed to earthquakes.

The original city of Hyderabad was founded on the banks of river Musi and has grown over centuries on both the banks of the river. Due to indiscriminate urbanization and lack of planning, the river has become a receptacle for all the untreated domestic and industrial waste waters coming out of Hyderabad and therefore has become polluted. Numerous efforts to clean it have failed. A manmade lake known as the HussainSagar Lake separates the two cities, which in the beginning was a reservoir of drinking water. Later in 1920s two more drinking water tanks were constructed, Himayath Sagar and Osman Sagar, both taking water from Musi river upstream.

**Sample collection:** Soil samples from six different locations of Hyderabad (Table 1) at different parts of the year are considered for the present study. The Soil samples were collected at regular intervals with 30 days duration between each sample. The soil samples were collected with the help of clean sterilized scalpel from each spot which were free from gravels and debris and collected in sterilized polythene bags and were labelled with reference to sampling sites with date of collection. These soil samples were sent to soil testing laboratory at Agricultural University, Rajendranagar, Hyderabad for the estimation of Physico-chemical parameters such as Percentage of moisture, pH, Temperature, Electrical conductivity(EC), Organic carbon(OC), Available Nitrogen(N), Available Phosphorus(P), Available Potassium(K).

Estimation of physico-chemical parameters: The Physico-chemical parameters of all the collected samples are estimated by using standardized methods (Table 2) in the soil testing laboratory at Agricultural University, Rajendranagar. The percentage of soil moisture was estimated on the same day by gravimetric method (ElGamal, 2016). The Physico-chemical parameters were estimated using soil-landscape pattern extracted from soil map. A Random Forest based predictive model was developed (Massawe et al., 2018) to map numerically classified soil clusters of a portion of Kilombero valley, Tanzania using Digital Soil Mapping (DSM) techniques.

| Place  | Area of Sample          |
|--------|-------------------------|
| OU     | Rhizosphere of Grass    |
| CBIT   | Rhizosphere of Grass    |
| Balapur | Rhizosphere of Paddy  |
| Meerpet | Rhizosphere of Palak |
| Khanapur | Rhizosphere of Red sorrel |
| Ashoknagar | Polluted soil |

**Table 1: Sampling Sites.**
Table 2: Estimation of Physico-chemical parameters.

| Parameter          | Method                                                |
|--------------------|-------------------------------------------------------|
| pH                 | Potentiometric method (Jackson, 1977)                 |
| Electrical conductivity | Conductivity bridge method (Jackson, 1977)           |
| Organic Carbon     | Walkley and Black’s oxidation method (Walkley and Black, 1934) |
| Total Nitrogen     | Subbaiah and Asija method (Subbaiah and Asija, 1956) |
| Available Phosphorus | Olsen’s extraction method (Olsen, 1954)              |
| Available Potassium | Flame photometer method (Muhr et al., 1965)          |
| Soil Moisture      | Gravimetric method of moisture estimation (ElGamal, 1995) |
| Temperature        | -                                                     |

The pH values of the soils were measured using standard digital pH meter. Electrical conductivity of the soil was estimated using digital electric conductivity meter method (Jackson, 1977) in 1:2 soil and water ratio. For estimation of organic carbon, Oxidation method (Walkley and Black, 1934) was used. Available Nitrogen was estimated by Subbaiah and Asija method (1956). Available Phosphorus was estimated by Olsen et al. (1954) and available potassium by Muhr et al. (1965) using Flame Photometer.

**Feature selection:** Feature selection is the process of selecting a subset of relevant features for use in model construction. Feature selection is useful if the data contains many features that are either redundant or irrelevant, and can thus be removed without incurring much loss of information. Feature selection reduces over fitting by eliminating redundant data which causes noise. This results in improvement in accuracy and faster training. In this study, Boruta algorithm is applied on the soil data for identifying important features.

**Linear Discriminant Analysis (LDA):** Linear Discriminant Analysis (LDA) is a classification method originally developed in 1936 by R. A. Fisher. It is simple, mathematically robust and often produces models whose accuracy surpasses more complex methods. This linear classification model does not require multiple passes over the data for optimization.

**Algorithm:**
1. Fisher’s Linear Discriminant Analysis (LDA) builds \( j = \text{min}(k-1, p) \) discriminant functions that estimate Discriminant scores \( (D_{ij}) \) for each of \( i = 1, \ldots, n \) subjects (soil samples) classified into \( k \) groups, from \( p \) linearly independent predictor variables \( (x) \) as
   \[
   D_{ij} = w_{ij} X_{1i} + w_{ij} X_{2i} + \ldots + w_{ij} X_{pi}, \quad [i = 1, \ldots, n \text{ and } j = 1, \ldots, \text{min}(k-1, p)]
   \]
2. Discriminant weights \( (w_{ij}) \) are estimated by ordinary least squares so that the ratio of the variance within the \( k \) groups to the variance between the \( k \) groups is minimal.
3. Classification functions of the type \( C_j = c_{j0} + c_{j1} X_{1j} + c_{j2} X_{2j} + \ldots + c_{jp} X_{pj} \) for each of the \( j = 1, \ldots, k \) groups can therefore be constructed from the Discriminant scores.
4. The coefficients of the classification function for the \( j^{th} \) group are estimated from the within sum of squares matrices \( (W) \) of the Discriminant scores for each group and from the vector of the \( p \) Discriminant predictors means in each of the classifying groups \( (M) \) as
   \[
   C_j = W^t M \text{ with } c_{j0} = \log_{10} C_j M
   \]
5. A subject is then classified into the group for which its classification function score is higher.

**Model evaluation:** The constructed LDA based soil texture classification model is evaluated using Hold-out method in which the model is trained on the training dataset and evaluated against the test dataset. A random split of 70% of data for training and 30% of data for testing is performed. The trained model is not exposed to the test data set during training and hence, the predictions made on the test dataset are indicative of the performance of the model.

**RESULTS AND DISCUSSIONS**

**Data overview:** The data set under study consists of 72 observations and 8 variables. The soil type is a factor variable which consists of 3 categories. Scatter plots (Fig. 1) show how much one variable is affected by another. The relationship between two variables is called their correlation. The function pairs. Panels in psych package of ‘R’ language is used to create a scatter plot of matrices with bivariate scatter plots below the diagonal, histograms on the diagonal, and the Pearson correlation above the diagonal. In the upper triangle, correlation coefficients have been indicated. These variables help to separate soils.

![Bivariate Scatter Plot of the soil data set](image-url)
Relative importance of physico-chemical parameters: Boruta method was applied on the dataset to identify redundant and irrelevant features. Blue box plots in Boruta variable importance chart (Fig. 2) correspond to minimal, average and maximum Z score of a shadow attribute. Red, yellow and green box plots represent Z scores of rejected, tentative and confirmed attributes respectively. The tentative attributes will be classified as confirmed or rejected by comparing the median Z score of the attributes with the median Z score of the best shadow attribute. It is observed that the z-scores of all the eight physico-chemical parameters are greater than the maximum z-score of the shadow attribute and hence are significant in soil texture classification. Hence, there are only green box plots and no red and yellow box plots in Fig. 2 indicating that all the physico-chemical parameters considered in this study are confirmed.

Model performance: Confusion matrix is one of the most intuitive metrics used for finding the correctness and accuracy of the model. The performance of the proposed LDA based soil texture classification model is assessed based on the confusion matrices. In classification point of view, True Positive (TP) represents the number of samples which are positive and also classified as positive, False Positive (FP) represents the number of samples which are positive but classified as negative, True Negative (TN) represents the number of samples which are negative and also classified as negative and False Negative (FN) represents the number of samples which are negative but classified as positive.

The derived overall performance metrics of the classification model from the confusion matrix are shown in Table 3. Accuracy is defined as the fraction of instances that are correctly classified ((TP+FN)/(TP+FP+TN+FN)). The Kappa statistic is a measure of agreement between the predictions and the actual class labels and the value ranges from 0 and 1. It is a measure of comparison of overall accuracy to the expected random chance accuracy. The higher the Kappa metric of the classifier, the better is the classifier as compared to a random chance classifier. The observed accuracy and Kappa of the classification model are 96.3% and 0.944 respectively. Precision is the ratio of correctly predicted positive observations to the total positive predicted observations (TP/(TP+FP)). Recall is defined as the ratio of correctly predicted positive observations to the total observations in the actual class (TP/(TP+FN)). F measure is the weighted average of Precision and Recall. It takes both False Positives and False Negatives into account. The overall Precision, Recall and F-measure are observed to be 0.96, 0.97 and 0.96 respectively. These results show that this model can be used more effectively to automate soil texture classification for classifying large soil datasets for facilitating improved crop and soil management.

Table 3: Overall statistics.

| Overall Performance metric | Value |
|----------------------------|-------|
| Accuracy                   | 0.963 |
| Kappa                      | 0.944 |
| Precision                  | 0.96  |
| Recall                     | 0.97  |
| F-measure                  | 0.96  |

Fig 2: Boruta Variable Importance Chart.

REFERENCES
Behrens T, Förster H, Scholten T, Steinrücken U, Spies ED, Goldschmitt M. (2005). Digital soil mapping using artificial neural networks. *Journal of Plant nutrition and Soil Science*, 168: 21-33.

Bhaskar B.P., S.V. Bobade, S.S. Gaikwad, Dipak Sarkar, S.G. Anantwar and Tapas Bhattacharyya (2015) Soil informatics for agricultural land suitability assessment in Seoni district, Madhya Pradesh, India. *Indian Journal of Agricultural Research*. 49: 315-320.

Dickson A.A., E.A. Allison-Oguru and N.O. Isirimah (2002). Fertility capability classification based land evaluation in relation to socio-economic conditions of small holder farmers in bayelsa state of nigeria, *Indian Journal of Agricultural Research*. 36: 10-16.

El-Gamal, A.D. (1995). Systematical studies on the algae isolated from some cultivated areas and laboratory studies on the effect of light, temperature and humidity on three selected soil algae. Ph.D. Thesis, Fac. Of Sci., Al-Azhar University, Cairo, Egypt.

Grinand C, Arrouays D, Laroche B, Martin MP (2008). Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. *Geoderma*. 143:80-90.

Hengl T, Toomanian N, Reuter HI, Malakouti MJ. (2007). Methods to interpolate soil categorical variables from profile observations: Lessons from Iran, *Geoderma*, 140: 417-427.

Jackson, M.L. (1977) Soil Chemical Analysis, Pentice Hall of India. Private Limited-New Delhi.

Jafari A, Khademi H, Finke PA, Van de Wauw J, Ayoubi S. (2014). Spatial prediction of soil great groups by boosted regression trees using a limited point dataset in an arid region, southeastern Iran. *Geoderma*. 232:148-163.
Massawe BH, Subburayalu SK, Kaaya AK, Winowiecki L, Slater BK (2018). Mapping numerically classified soil taxa in Kilombero Valley, Tanzania using machine learning, Geoderma, **311**: 143-148.

Muhr GR et al. (1965) Soil Testing in India, USAID, New Delhi. pp.120.

Olsen SR et al. (1954). Estimation of available phosphorus in soils by extraction with sodium bicarbonate. U.S. Dep. Agric. Washington, D.C. Circ.939.

Rad MR, Toomanian N, Khormali F, Brungard CW, Komaki CB, Bogaert P. (2014). Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. Geoderma, **232**: 97-106.

Ram R.L., P.K. Sharma and N. Ahmed (2013). Characterization and fertility assessment of soils of Markapur Mandal, Prakasam district, Andhra Pradesh for sustainable land use planning. Indian Journal of Agricultural Research, **47**: 127-133.

Savalia S.G and Gundalia J.D. (2010). Characterization and evaluation of soil-site suitability for groundnut in the soils of Uben Irrigation Command Area of Saurashtra region in Gujarat, Legume Res., **33** (2): 79 – 86.

Subbaiah, B.V. and Asija, G.L. (1956). A rapid procedure for the estimation of available nitrogen in soils. Current Sci., **25**: 259-260.

Vågen TG, Winowiecki LA, Tondoh JE, Desta LT, Gumbricht T. (2016). Mapping of soil properties and land degradation risk in Africa using MODIS reflectance, Geoderma, **263**: 216-225.

Walkely, A and Black, I.A., (1934), An examination of the Degtjareff method for determining soil organic matter and a proposed modification of the chromic acid titration method, Soil Science, **37**:29-38.