Performance Analysis of Decision Tree Algorithm C5.0 using Heavy metal contamination in agricultural soil at Coimbatore

R.Beulah, Dr. M. Punithavalli

Abstract—The quality of agricultural soil is significant to human health; however soil contagion is a severe problem in India. Polycyclic aromatic hydrocarbons (PAHs) have been found to be among the major soil contaminants in India. PAH derivatives could be more lethal but their dimensions in soils are exceptionally restricted. Human activities, such as land use change, cause brutal land dilapidation in many ecosystems around the globe with prospective impacts on soil processes. Cadmium (Cd), Lead(Pb), Copper(Cu), Zinc(Zc), Nickel(Ni) infectivity of top soil and victual crops is a everywhere ecological quandary that has resulted from unrestrained industrialization, indefensible urbanization and exhaustive agricultural practices. Being a noxious element and heavy metals poses high threats to soil quality, food safety, and human health. The objective of this document is to execute research work in agriculture which is utilize the current carry out of data m ining, in order to un ravel a v ariety of pertinent problems. This work presents a system, which uses data mining technique C5.0 in order to predict the category of the analyzed soil datasets.

Index Terms— Artificial Neural Network (ANN), Polycyclic Aromatic Hydrocarbons (PAHs)

1 INTRODUCTION

Agricultural soils may also be enhanced by cadmium which causes it's accretion in plants and facade a probable peril to human being physical condition [3]. Also elevated concentrations of cadmium in soil encompass detrimental effects on environment as it enters the victuals fetter [1]. The largest amount widespread grave metal contaminations establish in nature were Cd, Cr, Cu, Hg, Pb and Ni [4]. Ecological pollution by these heavy metals has develop into widespread due to manufacturing activities and when they go in prominent level in the surroundings, they are exceptionally wrapped up by ancestry and translocated to shoot, foremost to impaired metabolism and concentrated growth. Heavy metal contamination in soil consequences in decreased soil microbial commotion, soil opulence and yield fatalities [2].

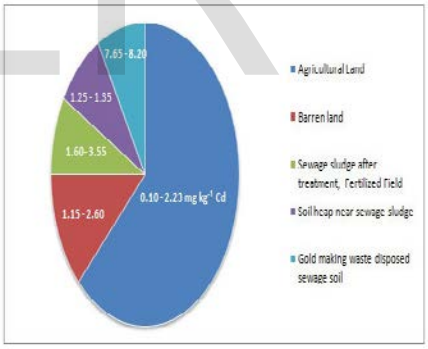

Fig 1. Cadmium concentrations (ppm) of different land(soil) type in sewage disposal areas of Coimbatore.

1.1 STUDY AREA LOCATION

The Corporation of Coimbatore is the cram area preferred which is positioned in the Southern part of the neck of land at a longitude of 77°04′00″ and 76°54′00″ and latitude of 11°03′30″ and 10°57′30″. The total geological area of the city is 105.5 sq.km[4].

Soil dilapidation is measured as one of the foremost causes of languish productivity development. Soil degradation refers to the processes, primarily human induced, by which soil declines in quality and is thus made less fit for a specific purpose, such as crop production. The main causes of soil degradation are erosion (by water or wind), compaction, salinization[5], nutrient depletion (due to a decline in organic matter content, leaching, extraction by plant roots without adequate replacement), contamination and, soil sealing (e.g. by urbanisation, road construction). In addition, problem soils refer to soils with unfavourable characteristics created by natural, long-term soil-forming processes, to yet suppressing productivity[3].

2. LITERATURE SURVEY

From the study article [6] to assess the recital of three data mining methods-Decision Trees (DT), Random Forest (RF) and Artificial Neural Network (ANN) with R Software to calculate soil map units from the Guppi-macacu watershed in Rio de janeiro state Brazil.

Researchers provided the outcome indicated that geology and respite were the factors scheming the spatial allocation of As, and living being factors particularly anthropogenic behavior were the factors calculating the spatial allotment of in the study region [7].

In this study of the calculation apparatus proved that constituent sorption by spectrally energetic Fe-oxide and soil inside of the soil was the most important method by which the spectrally undistinguished Pb and Zn ions can be predicted. The spatial patterns of predicted noxious rudiments showed that FD-ELM had the most parallel with those maps obtained by interpolating deliberate standards. In excess of all, it is accomplished that reflectance spectroscopy collective with the ELM algorithm is a hasty, reasonably priced and perfect tool for tortuous assessment of Pb and Zn and mapping their spatial sharing in dumpsite soils of Sarcheshmeh copper mine [7].

From this paper decision trees are cheering in the selection of SQ (Soil Quality) indicators. In addition, as well as morphological properties in the calculation of key soil properties such as Ks seem capable. VSA (Visual Soil assessment) could deliver morphological rejoinder variables for predicting other soil properties and emergent SQ frameworks (agricultural interest) more accomplished of representing structural dynamic [8].

3. METHODOLOGY

3.1 TYPES OF DECISION TREES

Decision trees used in data mining are essentially of two types:

Classification tree in which scrutiny is when the predicted conclusion is the set to which the data fit in. For instance ending of loan application as safe or risky.

Regression tree in which analysis is when the predicted outcome can be considered a real number. For example population of a state Both the classification and regression trees have similarities as well as differences, such as procedure used to determine where to split. There are various decision trees algorithms namely ID3 (Iterative Dichotomiser 3), C4.5, CART (Classification and Regression Tree), CHAID (CHi-squared Automatic Interaction Detector), MARS. Out of these, we will be discussing the more popular ones which are ID3, C4.5, CART.

3.2 ID3 (ITERATIVE DICHOTOMISER)

ID3 is an algorithm developed by Ross Quinlan used to generate a decision tree from a dataset. To construct a decision tree, ID3 uses a top-down, greedy search through the given sets, where each attribute at every tree node is tested to select the attribute that is best for classification of a given set. Therefore, the attribute with the highest information gain can be selected as the test attribute of the current node. ID3 is based on Occam's razor. In this algorithm, small decision trees are preferred over the larger ones. However, it does not always construct the smallest tree and is therefore, a heuristic algorithm. For building a decision tree model, ID3 only accepts categorical attributes. Accurate results are not given by ID3 when there is noise and when it is serially implemented. Therefore data is preprocessed before constructing a decision tree.

For constructing a decision tree information gain is calculated for each and every attribute and attribute with the highest information gain becomes the root node. The rest possible values are denoted by arcs. After that, all the outcome instances that are possible are examined whether they belong to the same class or not. For the instances of the same class, a single name class is used to denote otherwise the instances are classified on the basis of splitting attribute.

3.3 Advantages of ID3

The training data is used to create understandable prediction rules.

1. It builds the fastest as well as a short tree.
2. ID3 searches the whole dataset to create the whole tree.
3. It finds the leaf nodes thus enabling the test data to be pruned and reducing the number of tests.

The calculation time of ID3 is the linear function of the product of the characteristic number and node number.

3.4 Disadvantages of ID3

1. For a diminutive illustration, data possibly will be over-fitted or over-classified.
2. For creation a resolution, only one attribute is tested at an instant thus consuming a lot of time.
3. Classifying the continuous data may prove to be expensive in terms of computation, as many trees have to be generated to see where to break the continuum.
4. One disadvantage of ID3 is that when given a large number of input values, it is overly sensitive to features with a large number of values

4. C4.5

C4.5 is an algorithm used to produce a decision tree which was also developed by Ross uinlan. It is an addition of Quinlan's ID3 algorithm. C4.5 produce decision trees which can be used for categorization and consequently C4.5 is repeatedly referred to as arithmetical classifier. It is enhanced than the ID3 algorithm for the reason that it deals with both uninterrupted and distinct attributes and also with the lost values and pruning trees behind production. C5.0 is the saleable descendant of C4.5 because it is a assortment more rapidly, more remembrance competent and used for edifice slighter decision trees C4.5 performs by evasion a tree pruning method. This leads to the arrangement of slighter trees, more uncomplicated regulations and produces more spontaneous elucidation.

C4.5 follows three steps in hierarchy growth :

1. For splitting of clear-cut attributes, C4.5 follows the related approach to ID3 algorithms. Unremitting attributes for eternity produce dual splits.

Selecting attribute with the uppermost expand ratio.

These steps are continually practical to new tree undergrowth and development of the tree is stopped after scrutiny of stop condition. In sequence achieve partiality the attribute with additional amount of values. Thus, C4.5 uses grow Ratio which is a not as much of unfair assortment principle.

4.1 Advantages of C4.5

1. C4.5 is simple to execute.
2. C4.5 assembles models that can be without difficulty interprets.
3. It can be able to handle together categorical and nonstop values.
4. It cans transaction with sound and contract with lost value attributes.

4.2 Disadvantages of C4.5

1. A diminutive disparity in information can guide to dissimilar decision trees at what time using C4.5.
2. For a tiny training set, C4.5 does not work very well.

5. CART

CART (Classification and Regression Trees). It was introduced by Breiman in 1984. CART algorithm put up both classification and regression trees. The classification tree is making by CART by the binary splitting of the attribute. Gini Index is used as selecting the splitting attribute. The CART is also used for regression analysis with the help of regression tree. The regression feature of CART can be used in forecasting a dependent variable given a set of predictor variable over a given period of time. CART has an average speed of processing and supports both continuous and nominal attribute data.

5.1 Advantages of CART

1. CART able to handle lost values automatically with substitute splits.
2. Uses any arrangement of uninterrupted/separate variables.
3. CART automatically performs variable collection.
4. CART can launch communications between variables.
5. CART does not vary according to the

monotonic alteration of extrapolative variable

5.2 Disadvantages of CART

1. CART may well include unbalanced decision trees.
2. CART splits simply by single variable
3. It is Non-parametric.

TABLE 1
SNAPSHOT OF SAMPLE DATASETS OF AGRICULTURAL SOIL

pH value	EC	C	P	K	Fe	Zn	Cu	B	Mn
6	2.7	15	85	1035	13	0.6	0.2	3	14
4	2.9	15	43	551	12	0.9	0.1	2	12
6	1.6	10	45	707	15	1.1	0.4	2	5
7	0.1	5	17	883	1	2	1.7	2	1
7	0.2	9	4	836	2	3.1	1.5	2	2
8	2.4	12	43	545	15	1.3	0.4	2	5
8	2.7	20	95	246	6	0.5	0.1	3	11
7	0.4	2	19	852	1	1.5	1.4	2	3
7	2	20	62	650	3	0.9	0.1	0	9
6	2.5	13	40	334	14	1.2	0.4	2	8
8	2.4	18	23	736	14	1.2	0.3	2	7
7	0.4	2	6	963	6	2.9	1.2	1	3

PH-potential of Hydrogen, EC-Electrical Conductivity, C-Carbon, P- Phosphorus, K- Potassium, Fe-Iron, Zn- Zinc, Cu-Copper, B-Boron, Me- Magnesium

6. Proposed Methodology

6.1 C5.0

C5.0 algorithm was made further memory efficient by creating its successor which is called C5.0. It is successor which is called C5.0. It is used to make even smaller decision trees. A set of training examples are required where in each example can be seen as a pair an input object and its corresponding output value class. The algorithm builds a classifier by analyzing the training set, such that it correctly classifies both test and training example types of Decision trees. It is enhanced decision tree version of C4.5. It constructs the hierarchy representation using Information grow calculate. In this model, the decision tree can contain any number of branches. While generating the tree, splitting the element is based on the utmost information gain. Information gain is the development of multiplying the prospect of the class period the log of that class possibility. Entropy is the measure of contamination of an attribute. Information gain is generated based on the calculation of entropy values for parent and child attributes. This progression is done iteratively until there is in no need for additional opening. An advantage of C5.0 is its applicability for big datasets, less remembrance expenditure, faster and gives supports to boosting. Disadvantage is, it does not exertion well with diminutive training samples.

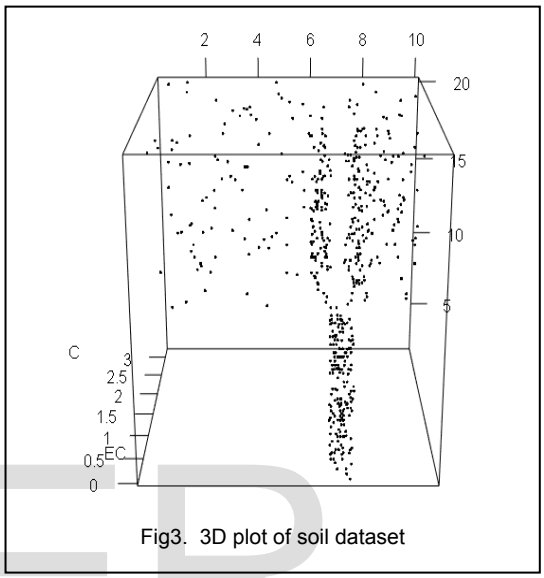

Fig3. 3D plot of soil dataset

7. OUTCOME AND DISCUSSIONS

C5.0 algorithm necessitates correct samples to assemble the decision tree model. Sample should be adjoining to the unique dataset, which is provided as the contribution for the classification algorithm. If the model is truthful then a perfect model is built these algorithms. So we need superior example techniques to extract samples from the unusual dataset. In this exertion used C5.0 algorithm for heavy metal contagion in agricultural soil. Usually, Dataset may encompass either balanced data or unbalanced data. If data set has balanced data then its module are balanced in the target attribute.

7.1 Sample Code C5.0

```
>data1<- read.csv(file.choose();header=T)
>data1
>head(data1)
>str(data1)
>table(data1$EC)
>head(data1)
```

```
>set.seed(9850)
>q<-runif(nrow(data1))
>data1r<-data1[order(q),]
>str[data1r]
>install.packages("C50")
>library(C50)
```

8. RESULT

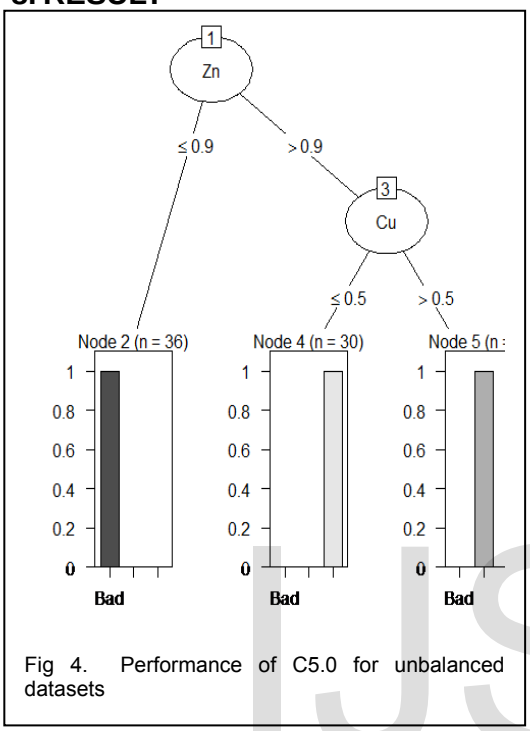

Fig 4. Performance of C5.0 for unbalanced datasets

9. CONCLUSION

Chemical remediation of the metal-polluted soils concerning use of lime, phosphatic fertilizers, and oxides of Mn and Fe in the middle of others leftovers, the lucrative choice. Phytoremediation, although looking good-looking, is too sluggish. The majority effectual approach will include being the one of adopting defensive actions rather than for the cure of the metal-polluted soils, for the reason that the later will not be either cost-effective (or) virtually practicable, for that soil scientist must dedicate a fair share of their research attempt to ecological quality problems. To increase an enhanced understanding of how soils might be used and nevertheless protected in waste management efforts, soil scientists must dedicate a fair divide of their research endeavor to environmental eminence problems.

REFERENCES

[1] Tóth, G., Hermann, T., Da Silva, M. R., & Montanarella, L. (2016). Heavy metals in agricultural soils of the European Union with implications for food safety. Environment International, 88, 299–309.

[2] Taghizadeh-Mehrjardi, R., Nabiollahi, K., & Kerry, R. (2016). Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. Geoderma, 266, 98–110.

[3] SOLGI, E., & PARMAH, J. (2015). Analysis and assessment of nickel and chromium pollution in soils around Baghejar Chromite Mine of Sabzevar Ophiolite Belt, Northeastern Iran. Transactions of Nonferrous Metals Society of China, 25(7), 2380–2387.

[4] Snakin, V. V, Krechetov, P. P., Kuzovnikova, I. O., Alyabina, A. F., Stepichev, A. V, & Gurov, A. F. (1996). The system of assessment of soil degradation. Soil Technology, 8 (1996), 331–343.

[5] Nicolova, M., Spasova, I., Georgiev, P., & Groudev, S. (2016). Microbial removal of toxic metals from a heavily polluted soil. Journal of Geochemical Exploration.

[6] Krasilnikov, P., Makarov, O., Alyabina, I., & Nachtergaele, F. (2016). Assessing soil degradation in northern Eurasia. Geoderma Regional, 7(1), 1–10.

[7] Argyraki, A., Kelepertzis, E., Botsou, F., Paraskevopoulou, V., Katsikis, I., & Trigoni, M. (2017). Environmental availability of trace elements (Pb, Cd, Zn, Cu) in soil from urban, suburban, rural and mining areas of Attica, Hellas. Journal of Geochemical Exploration.

[8] Vibha, L., Vardhan, G. M. H., Prashanth, S. J., Shenoy, P. D., Venugopal, K. R., & Patnaik, L. M. (2007). A hybrid clustering and classification technique for soil data mining. IET-UK International Conference on Information and Communication Technology in Electrical Sciences (ICTES 2007), (Ictes), 1090–1095.

[9] Bindraban, P. S., van der Velde, M., Ye, L., van den Berg, M., Materechera, S., Kiba, D. I., … van Lynden, G. (2012). Assessing the impact of soil degradation on food production. Current Opinion in Environmental Sustainability, 4(5), 478–488.