Evaluation of Best First Decision Tree on Categorical Soil Survey Data for Land Capability Classification

Nirmal Kumar
National Bureau of Soil Survey and Land Use Planning
Amravati Road
Nagpur, Maharashtra - 440033

G. P. Obi Reddy
National Bureau of Soil Survey and Land Use Planning
Amravati Road
Nagpur, Maharashtra - 440033

S Chatterji
National Bureau of Soil Survey and Land Use Planning
Amravati Road
Nagpur, Maharashtra – 440033

ABSTRACT
Land capability classification (LCC) of a soil map unit is sought for sustainable use, management and conservation practices. High speed, high precision and simple generating of rules by machine learning algorithms can be utilized to construct pre-defined rules for LCC of soil map units in developing decision support systems for land use planning of an area. Decision tree (DT) is one of the most popular classification algorithms currently in machine learning and data mining. Generation of Best First Tree (BF Tree) from qualitative soil survey data for LCC reported in reconnaissance soil survey data of Wardha district, Maharashtra has been demonstrated in the present study with soil depth, slope, and erosion as attributes for LCC. A 10-fold cross validation provided accuracy of 100%. The results indicated that BF Tree algorithms had good potential in automation of LCC of soil survey data, which in turn, will help to develop decision support system to suggest suitable land use system and soil and water conservation practices.

General Terms
Data mining algorithms, Decision Tree

Keywords
Best First Decision Tree, Land Capability Classification, Information gain

1. INTRODUCTION
LCC - A qualitative system - developed by the US Department of Agriculture [1] is the most used land classification system. LCC provides information of the kind of soil, its location on the landscape, its extent, and its suitability for various uses, which is needed for conservation planning [2]. LCC includes eight classes of which, first four are suitable for cropland and the limitations on their use and necessity of conservation measures and careful management increase from I through IV. The remaining four classes, V through VIII, are unsuitable for cropland, but may be used for pasture, range, woodland, grazing, wildlife, recreation, and esthetic purposes. Within the broad classes are subclasses, which signify special limitations such as (e) erosion, (w) excess wetness, (s) problems in the rooting zone, and (c) climatic limitations. The task of LCC occurs every time a soil surveyor identifies a map unit. A large and diversified dataset have already been generated through soil surveys. A pre-defined rule set learned on these data for automatically defining the LCC of the future soil units being surveyed, will be of great help for developing decision support systems for land use planning and suggesting conservation and management practices. Machine learning and data mining techniques which gives computers the ability to learn based on the inherent characteristics of data, without being explicitly programmed [3,4] may be utilized for generating these rule sets. DT is one of the most popular classification algorithms currently in machine learning and data mining [5-7].

In their simplest form, DT classifiers successively partition the input training data into more and more homogeneous sub sets by producing optimal rules or decisions, also called nodes [7-9]. The rules or the splitting criteria at these nodes are the key to successful decision tree creation. The most frequently used splitting criteria are the information gain, the information gain ratio [10], and the Gini index [11]. Some of the most popular DT methods are ID3, C4.5 [10, 12, 13], CART [11] and BF Tree [14]. A detailed review of DT applications in agricultural and biological engineering may be found in [5] and [7]. In the field of applying DT algorithms for soil survey data, [15] evaluated ID3 DT for LCC with 12 simulated samples with soil depth, slope, and texture as attributes for LCC; however, model was not validated. ID3 DT algorithm was applied on soil survey data with an accuracy of 86.84% on 10-fold cross validation [16].

2. MATERIAL AND METHODS

2.1 Training Data Used
By considering slope, soil depth and erosion as important attributes, LCC of 38 soil series of Wardha district, Maharashtra, India, was assessed as per the procedure laid down by Soil Survey Manual by All India Soil and Land Use Survey Organization [17].

Waikato Environment for Knowledge Analysis (WEKA) – an open source data mining tool – was used for generation of BF tree rules for comparison with the one manually generated.

2.2 Best First Tree Algorithm
In BF tree learners the “best” node is expanded first as compared to standard DT learners such as C4.5 and CART which expand nodes in depth-first order [14]. The “best” node is the node whose split leads to maximum reduction of impurity (e.g. Gini index or information gain) among all nodes available for splitting. The resulting tree will be the same when fully grown; just the order in which it is built is different. BF tree constructs binary trees, i.e., each internal node has exactly two outgoing edges. The tree growing method attempts to maximize within-node homogeneity. The extent to which a node does not represent a homogenous subset of cases is an indication of impurity. For example, a terminal node in which all cases have the same value for the dependent variable is a homogenous node that requires no further splitting because it is “pure.” The impurity measures for nominal dependent variables are entropy-based definition of information gain and gini index. The measure used in this

study is information gain. Entropy characterizes the purity of any sample set. If the target attribute can take on *v* different values, then the entropy of set *(S)* relative to this v-wise classification is defined as

$$Entropy(S) = \sum_{i=1}^{v} -p_i log_2 p_i \qquad (1)$$

where p_i is the proportion of *S* belonging to class i.

Information gain is the expected reduction in entropy caused by splitting the training data set according to this attribute. More precisely, the information gain, *Gain(S, A)* of an attribute A, relative to a collection of examples S, is defined as

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values\ (A)} \frac{|S_v|}{|S|} Entropy(S_v) \qquad (2)$$

where *Values(A)* is the set of all possible values for attribute *A*. and S_v is the subset of *S* for which attribute *A* has value *v* $(i.e., S_v = (S \in S | A(S) = v|))$.

The tree ceases to grow when all instances belong to a single value of a target feature or when best information gain is not greater than zero.

2.3 Accuracy Assessment

In machine learning methods, such as the DT, the classification accuracy is often predicted by stratified 10-fold cross-validation [18-20]. In the process, the whole dataset is split into 10 parts. Nine parts of the dataset is used for learning and 1 for testing. This procedure is repeated 10 times so that every part of the dataset is used for both training and testing (of course one at each time). Afterwards, the overall accuracy parameters were calculated as means from the evaluation of the individual cross-validation subset. A 10-fold cross validation was applied in BF tree model.

3. RESULTS AND DISCUSSION

3.1 Induction of BF Tree

The LCC of soil series of the study area ranges from IIIs to Vies (Table 1). The same training dataset was used to evaluate BF tree in assessment of LCC.

Table 1: Soil series description and attributes for LCC

Soil series	Depth	Slope	Erosion	Capability class
Kolona	d5	d	e2	IIIes
Karanja	d5	d	e2	IIIes
Nagjhari	d5	d	e2	IIIes
Nijampur	d5	b	e1	IIIs
Pachod	d5	b	e2	IIIse
Vagholi	d5	b	e2	IIIse
Thar	d4	c	e2	IIIse
Anjangaon	d4	c	e2	IIIse
Takli	d5	c	e2	IIIse
Arvi	d4	b	e2	IIIse
Yakamba	d4	b	e2	IIIse
Chamla	d4	b	e1	IIIs
Sirasgaon	d3	c	e3	IVes
Talani	d2	c	e3	IVes
Panthargavda	d3	d	e3	VIes
Parsodi	d2	e	e3	VIes
Hridi	d4	c	e3	IIIse
Chanakpur	d3	b	e2	IVs
Wadner	d2	b	e2	IVs
Pardi	d2	b	e2	IVs
Lakhandevi	d3	e	e3	VIes
Mahakali	d3	e	e3	VIes
Karanii	d3	e	e3	VIes
Ashti	d3	d	e3	VIes
Kinala	d2	e	e3	VIes
Sewagram	d2	d	e3	VIes
Madni	d2	d	e3	VIes
Hewan	d5	b	e1	IIIs
Waigaon	d4	b	e2	IIIse
Karla	d4	b	e2	IIIse
Bothali	d5	b	e2	IIIse
Kondhali	d5	b	e2	IIIse
Wardha	d5	b	e2	IIIse
Lasanpur	d5	b	e2	IIIse
Malalpur	d5	b	e2	IIIse
Malakpur	d5	c	e2	IIIse
Sirpur	d5	c	e2	IIIse
Talegaon	d5	b	e2	IIIse

Depth classes: d2 – shallow (25-50 cm), d3 – moderately shallow (50- 75 cm), d4 – Moderately deep (75 – 100 cm), d5 – deep (100 – 150 cm)
Slope classes: b – very gently sloping (1-3%), c – Gently sloping (3-8%), d – Moderately sloping (8-15%), e – Moderately steep (15-30%)
Erosion classes: e1 – Slight erosion, e2 – Moderate erosion, e3 – Severe erosion

Entropy at the root node is calculated as:

$$entropy(S) = -\frac{3}{38}log_2\frac{3}{38} - \frac{2}{38}log_2\frac{2}{38} - \frac{18}{38}log_2\frac{18}{38} - \frac{3}{38}log_2\frac{3}{38} - \frac{3}{38}log_2\frac{3}{38} - \frac{9}{38}log_2\frac{9}{38} = 2.094$$

Since, the BF tree algorithm finds binary splits to split a node if the attribute *erosion* is considered, it has 3 possible splitting subsets (Figure 1).
In case of the first split in the figure 2, the entropy values for two successor nodes are thus

*entropy(e1) = [-3/3*log(3/3,2)] = 0*
*entropy(!e1)=[-3/35*log(3/35,2)-3/35*log(3/35,2)-2/35*log(2/35,2)-9/35*log(9/35,2)-18/35*log(18/35,2)]= 1.84*

the information and information gain for a split erosion e1:!e1 is calculated as

*info(erosion=e1:!e1)=3/38*info(e1)+35/38*info(!e1) = 1.695*
info gain(erosion=e1:!e1)= entropy(S)- info(erosion=e1:!e1) = 2.094 – 1.695= 0.398

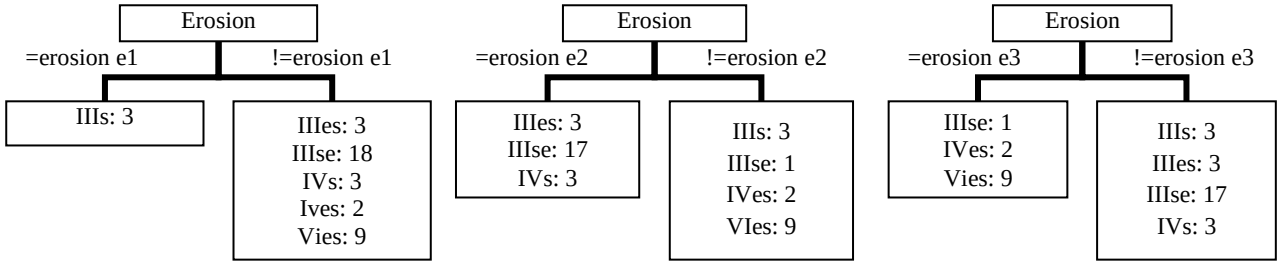

Fig 1: Possible binary splits for erosion and their class distribution at the root node

Similarly, info gains for all possible binary splits for the three attributes were calculated (Table 2). Attribute *depth* (binary split *d2-3:d4-5*) having maximum info gain (0.949), is the choice for splitting the root node.

Table 2: Possible binary splits and info gain at the root node

Binary Splits	info	Info gain
Possible binary splits for depth		
d2 : !d2	1.765	0.329
d3 : !d3	1.765	0.329
d4 :!d4	1.881	0.213
d5 :!d5	1.641	0.453
d2-3 : d4-5	**1.144**	**0.949**
d2-4 : d3-5	2.015	0.079
d2-5 : d3-4	2.015	0.079
Possible binary splits for slope		
b :!b	1.531	0.563
c :!c	1.786	0.308
d :!d	1.639	0.454
e :!e	1.767	0.327
bc : de	1.194	0.900
be : cd	1.796	0.298
bd : ce	1.837	0.257
Possible binary splits for erosion		
e1 :!e1	0.600	0.398
e2 :!e2	0.482	0.821
e3 :!e3	0.492	0.753

Child node of *depth d2-3 is* split by *slope b-c:d-e* having maximum info gain (0.940) at this node (Table 3). In the same way child node of *depth d4-5 is* split by *slope b-c:d* having maximum info gain (0.544) at this node (Table 4). Node *slope d-e* under the parent node *depth d2-3* and *slope d* under parent node *depth d4-5* is truncated as all instances belong to a single value of a target feature. Info gain at the child node *slope b-c* (0.971) under parent node *depth d2-3* (Table 5) suggests that *slope b:c is* the next split for this node. Similarly, child node *slope b-c* under parent node *depth d4-5 is* split by *erosion e2-3:e1* having maximum info gain of 0.592 (Table 6). No further splitting occurs since the improvement in the gini gain comes down to zero in all child nodes.

Table 3: Possible binary splits and info gain at the child node *depth d2-3*

Binary Splits	info	info gain
Possible binary splits for slope		
b : !b	0.198	1.089
c : !c	0.256	1.031
d : !d	0.391	0.896
e : !e	0.362	0.925
bc : de	**0.128**	**1.159**
bd : ce	0.340	0.947
be : cd	0.346	0.941
Possible binary splits for depth		
d2 : d3	0.466	0.821
Possible binary splits for erosion		
e2 : e3	0.198	1.089

Table 4: Possible binary splits and info gain at the child node *depth d4-5*

Binary Splits	info	info gain
Possible binary splits for slope		
b : cd	0.502	0.559
c : bd	0.593	0.468
d : bc	0.327	0.734
Possible binary splits for depth		
d4 : d5	0.619	0.442
Possible binary splits for erosion		
e1 : !e1	0.327	0.734
e2 : !e2	0.406	0.655
e3 : !e3	0.659	0.402

Table 5: Possible binary splits and info gain at the child node *slope b-c* under node *depth d2-3*

Binary Splits	info	info gain
Possible binary splits for slope		
b : !b	0.000	0.971
Possible binary splits for depth		
d2 : d3	0.125	0.846
Possible binary splits for erosion		
e2 : e3	0.000	0.971

Table 6: Possible binary splits and info gain at the child node *slope b-c* under node *depth d4-5*

Binary Splits	info	info gain
Possible binary splits for erosion		
e1 : !e1	0.000	0.592
e2 : !e2	0.085	0.506
e3 : !e3	0.321	0.271
Possible binary splits for slope		
b : c	0.285	0.307
Possible binary splits for depth		
d4 : d5	0.326	0.265

The structure of the tree manually developed is found to be similar to that of WEKA (Figure 2). The accuracy of BF tree algorithm was assessed through a 10 - fold cross validation method. The confusion matrix has been shown in table 2. The analysis shows that BF tree algorithm found to be 100% accurate with a kappa coefficient of 1.

4. CONCLUSIONS

The result indicates that DT algorithms have immense potential over the traditional procedures in LCC for their fast and easy rules generation. Explicit rules could be formulated with better accuracy for classifying complex soil-site data acquired over diversified landscapes. The large size training dataset with higher variability may produce a robust DT model for automation of LCC and their incorporation in decision support systems to suggest suitable land use system and soil and water conservation practices.

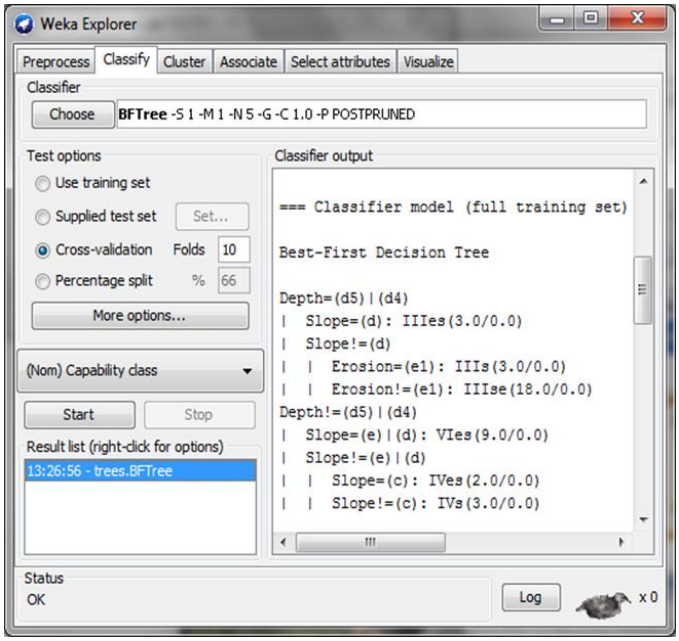

Fig. 2: Decision tree developed with BF Tree module of WEKA

Table 2: Confusion Matrix of 10-fold cross validation in WEKA

Confusion Matrix						
a	**b**	**c**	**d**	**e**	**f**	**Classified as**
3	0	0	0	0	0	a = IVs
0	**2**	0	0	0	0	b = Ives
0	0	**9**	0	0	0	c = Vies
0	0	0	**3**	0	0	d = IIIs
0	0	0	0	**18**	0	e = IIIse
0	0	0	0	0	**3**	f = IIIes

5. REFERENCES

[1] Klingebiel, A. A., Montgomery, P. H. 1961. Land capability classification. Agriculture handbook no 210. Soil conservation service, Washington D.C. US Department of Agriculture (USDA).

[2] Fenton, T. E. 2005. Land Capability Classification. In Encyclopedia of Soil Science, Second Edition. CRC press. Pp 962-964.

[3] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthuruswamy, R. (Eds.), 1996. Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park, CA, USA.

[4] Diplaris, S., Symeonidis, A. .L, Mitkas, P. A., Banos, G., Abasc, Z. 2006. A decision-tree-based alarming system for the validation of national genetic evaluations. Comput. Electron. Agric. 52, 21–35.

[5] McQueen, R. J., Garner, S. R., Nevill-Manning, C. G., Witten, I. H. 1995. Applying machine learning to agricultural data. Comput. Electron. Agric.12 (4), 275–293.

[6] Gangrade, A., Patel, R. 2009. Building privacy-preserving C4.5 decision tree classifier on multiparties. International J. Comp. Sci. Engg. 1(3), 199-205.

[7] Huang, Y., Lan, Y., Thomson S J, Fang A, Hoffmann W C, Lacey R E. 2010. Development of soft computing and applications in agricultural and biological engineering. Comput. Electron. Agric.71, 107–127.

[8] Safavian, S. R., Landgrebe, D. 1991. A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics 21, 660–674.

[9] Trépos, R., Masson, V., Cordier, M. O., Chantal, G. O., Jordy, S. M. 2012. Mining simulation data by rule induction to determine critical source areas of stream water pollution by herbicides. Comput. Electron. Agric. 86, 75–88.

[10] Quinlan, J. R. 1993. C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, California.

[11] Breiman, L., Friedman, J. H., Olshen, R. A. 1984. Classification and Regression Trees. Belmont: Wadsworth International Group. California, USA.

[12] Quinlan, J. R. 1986. Induction of decision trees. Machine Learning 1, 81-106.

[13] Quinlan, J. R. 1996. Improved use of continuous attributes in C4.5. J. Artifi. Intel. Res. 4, 77-90.

[14] Shi, Haijian. 2007. Best-first Decision Tree Learning. Masters Degree Theses. University of Waikato Masters Theses. Pages 104.

[15] Tamboli, N. M., Kamble, A. M., Metkewar, P. S. 2012. LCC Decision tree analysis using ID3. Int. J. Comp. Appl. 41(19), 19-22.

[16] Nirmal Kumar, Obi Reddy, G. P., Chatterjee, S., Dipak Sarkar (2013). An application of ID3 decision tree algorithm for land capability classification. *Agropedology*. 22(1): 35-42.

[17] AISLUS. 1971. All India Soil and Land Use Survey, Soil Survey Manual, Indian Agricultural Research Institute (IARI) Publ. New Delhi.

[18] Weiss, S. M., Kulikowski, C. A. 1991. Computer systems that learn. San Mateo, CA: Morgan Kaufman Publishers.

[19] Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Montreal, Canada. Morgan Kaufmann, San Francisco, USA.

[20] Kirchner, K., T¨olle, K. H., Krieter, J. 2006. Optimisation of the decision tree technique applied to simulated sow herd datasets. Comput. Electron. Agric. 50, 15–24.