

AN APPROACH TO CONSTRUCT DECISION TREE USING SLIQ AND KNN FOR LAND GRADING SYSTEM

Kamlesh Kumar Joshi, Pawan Patidar

M.Tech Scholar, Assistant Professor

Laxmi Narain College of Technology, Indore (M.P) India

k3g.kamlesh@gmail.com

ABSTRACT

India is a nation of farmers where most population of country is dependent on the crops and agriculture. But poor land quality and their composition affect the performance of crops production. Therefore according to the soil chemical composition, their categorization required. In order to design an appropriate, efficient and accurate classifier the decision tree algorithm is selected for data modeling. Therefore first using the soil composition a dataset is prepared and classification algorithm is optimized for accurate classification. The classification algorithm first utilized SLIQ decision tree for preparing the decision tree than KNN algorithm is applied on decision tree to extract the rules. These rules are optimized further for reducing the number of comparisons during classification. Additionally for justifying the proposed rule based solution the presented data model is compared with the traditional ID3 and SLIQ algorithm. The proposed is implemented using visual studio development technology and the performances of the algorithms are compared with traditional algorithms based on the various qualities of service parameters such as memory consumption, training time, search time and accuracy. According to the obtained results the previous algorithm shows 80% accuracy rate compared to which implemented algorithm shows 90% of accuracy.

Keywords: Classification, SLIQ, ID3, KNN, performance

I. INTRODUCTION

India is a developing country and new inventions where techniques are developed for promoting business, science and research. In this country all most 60% peoples are dependent on the agriculture and crop productions. That is directly dependent on soil quality and their productivity. In this presented work, the soil quality and their categorization techniques are investigated, which can help farmers of the specific regions for getting benefits and can improve the crop productions. This chapter provides an overview of the presented study and the document organization. Data mining is technique or application which is used for analysis of large data sets and establishes useful classification patterns in input training data sets. In this process the similarity between the given data samples are calculated and using this calculated relationships the classification and categorization task are performed. Various techniques of data analysis have used including statistical machine learning, natural trees and other analysis methods for biological and agricultural research studies. For analysis of agricultural data sets having various data mining techniques may yield outcomes useful for researchers in the field of agriculture. Both commercial and research centers had developed data mining software applications which have various methodologies. These methods have been utilized for industrial, commercial and scientific purposes [1].

Various techniques of data analysis including statistical machine learning, natural trees and other analysis tools are used for biological and agricultural research studies. This research resolute whether techniques of data mining could be used to classify soils that analyse large soil profile experimental datasets. The research intended to

start whether techniques of data mining can be used to analyse other classification methods by finding whether meaningful pattern exists across various soil profiles characterized at various research sites. The set of data has been gathered from soil surveys at various agricultural areas in India. The research has utilized working data collected from many commonly occurring soil types in order to categorized soils and correlations between a numbers of properties of soil. Soil profiles classification and characteristics of chemicals are used for preparing the soil grading data set. The analysis of these agricultural data sets with data mining techniques may help to researchers in the soil sciences and agricultural chemistry [1].

The overall aim of the research is to develop and design a data model by which soil compositions are identified for appropriate crop production and their proper utilization in agriculture or construction domains. Therefore an optimum classifier is helpful for soil properties classification and compares the different classifiers and the performance analysis. That advantages to agriculture, soil management and environment.

II. BACKGROUND

The most common data mining algorithms and decision support systems are neural networks, decision trees and logistic regression decision trees. Among these classification algorithms decision tree algorithms is the most commonly used because of it is easy to understand and cheap to implement. It provides a modeling technique that is easy for human to understand and simplifies the classification practice.

Decision tree classifiers are used successfully in many diverse areas such as radar signal classification, character

identification, remote sensing, speech recognition, medical diagnosis, and expert systems, to name only some. Perhaps, the significant feature of

Decision tree classifiers are their capability to break down a complex decision-making process into a collection of simpler decisions, thus providing a solution which is often easier to interpret.

Decision tree is one of the classification methods and supervised learning algorithms, supervised learning algorithm (like classification) is chosen to unsupervised learning algorithm (like clustering) because its prior knowledge of the class labels of data records makes feature/attribute selection easy and this leads to good prediction/classification accuracy. The early decision tree algorithms are C4.5 and ID3, improved version of ID3 is C4.5, it introduced some new methods and functions, such as adopting information gain ratio, disposing of the continuous attributes, validating the model by k-fold cross-validation, and so on mentioned in [3, 4]. It has been broadly applied in information extraction from remote sensing image, disaster weather forecasting, correlation analysis of environmental variables, and so on.

During the process of training, we must find the most efficient way to split a set of cases (records) into two child nodes in the decision tree algorithm. The most common methods are entropy and gini, for evaluating the splits.

The information gain ratio is the base for choosing the split attribute of the decision tree in C4.5, for the root node of the decision tree, the attribute that has the maximum gain ratio will be chosen. When we want to construct the decision tree for the training set t , which will be divided into n subset in accordance with the gain ratio calculated. If all the classifications of the tuple contained in sub-set t_i are of the same group, the node will become a leaf node of decision tree and stop splitting. The other sub-set of t which does not satisfy this condition mentioned above will be split recursively and to construct the branch for the tree as described above, until all the tuple contained in the subset belongs to the same category. After generating a decision tree, we can extract the rules from the tree, and to classify the new data set.

III. ALGORITHM STUDY

This section represents the algorithm we had studied and the basics of the previous algorithm including ID3 as well as SLIQ. In addition of that, it also includes the enhancement and contributions on exiting classification algorithm.

ID3 DECISION TREE BASICS

ID3 is a simple algorithm of decision tree learning developed by Ross Quinlan. The primary fundamental of ID3 algorithm is to construct the decision tree by exposing a top-down, greedy search through the

specified sets to test each attribute at every tree node. In order to choose the attribute that is most useful attributes information gain is used.

To find an optimal way to classify a learning set, what we require to do is to reduce the questions asked (i.e. minimizing the depth of the tree). Thus, we require some function which can compute which questions provide the most balanced splitting. The information gain metric is like a function.

Entropy

In order to define information gain accurately, we need discussing entropy first. Let's assume, without loss of simplification, that the resulting decision tree categorizes instances into two categories, we'll call them p (positive) and n (negative)

Given a set s , containing these positive and negative targets, the entropy of s associated to this boolean classification is:

$$Entropy(S) = -P(Positive) \log_2 P(Positive) - P(negative) \log_2 P(negative)$$

P (positive): proportion of positive examples in s

P (negative): proportion of negative examples in s

Information gain

As we declared before, to reduce the decision tree depth, when we traverse the tree path, optimal attribute has to be selected if we need splitting the tree node, attribute with the most entropy reduction is the best choice which we can easily imply.

We describe information gain as the expected decrease of entropy related to specified attribute when splitting a decision tree node.

The information gain, $gain(s, a)$ of an attribute a ,

$$Gain(S, A) = Entropy(s) - \sum_{n=1}^v \frac{S_v}{S} \times Entropy(S_v)$$

Notion of gain to rank attributes can be used and where at every node is located the attribute with highest gain among the attributes not yet considered in the path from the root is used to build decision tree.

Supervised learning in quest (SLIQ) algorithm

SLIQ is a classifier of decision tree, that can take both numerical and categorical attributes it builds compact and accurate trees. Pre-sorting technique is used in the tree growing phase and an inexpensive pruning algorithm. It is appropriate for classification of huge disk-resident datasets, separately of the number of classes, attributes and records [20].

Tree building

Make tree (training data *t*)

Partition (*t*)

Partition (data *s*)

If (all points in *s* are in the same class)

Then return;

Evaluate splits for each attribute *a*;

Use best split to separation *s* into *s1* and *s2*;

Partition (*s1*);

Partition (*s2*);

The *gini*-index is used to evaluate the “goodness” of the alternative splits for an attribute

If a data set *t* contains examples from *n* classes, *gini(t)* is given as

$$gini(T) = 1 - \sum P_j^2$$

Where *p_j* is the relative frequency of class *j* in *t*. After splitting *t* into two subset *t1* and *t2* the gin index of the split data is defined as

$$gini(T)_{split} = \frac{|T1|}{|T|} gini(T1) + \frac{|T2|}{|T|} gini(T2)$$

The first technique implemented by SLIQ is a scheme that eliminates the need to sort data at each node it creates a separate list for each attribute of the training data. A separate list, called *class list*, is produced for the class labels attached to the examples. SLIQ requires that the *class list* and (only) one *attribute list* could be kept in the memory at any time.

IV. IMPLEMENTATION

This section includes the implementation strategy and the system implementation technology. In addition of that it also includes the implemented classes, functions and utilized reference classes which are used in system design. Finally the gui navigation options for navigating the presented system for land classification.

PROPOSED ALGORITHM

In order to advance the classification accuracy the give algorithm steps are works as:

1. Call SLIQ algorithm
2. Build SLIQ rules
3. For each rule in rule set
4. Find confidence (rule);

$$5. confidence = 1 - \sum_{i=0}^n \sqrt{x_i^2 - y_i^2}$$

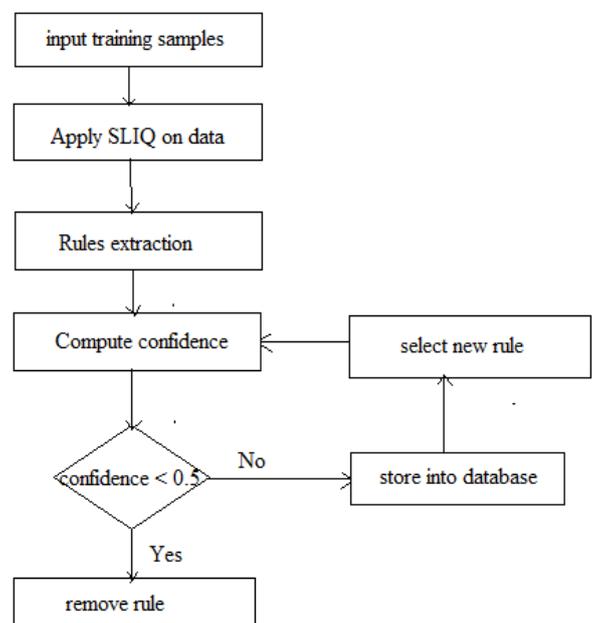
6. If confidence < .5

7. Remove rule;

8. End if

9. End for

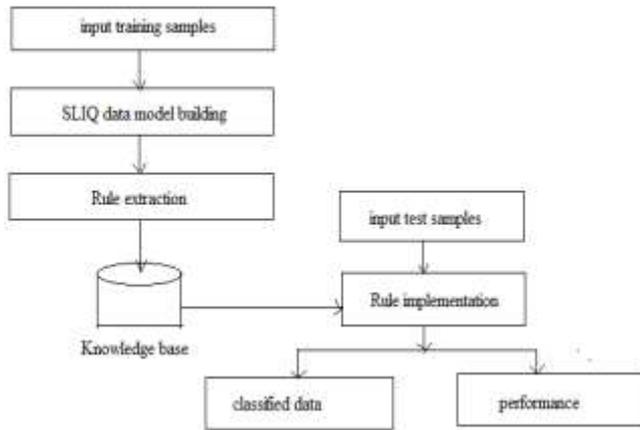
The above given process consumes the SLIQ algorithm after data processing using SLIQ algorithm a set of rules are generated which is reduced to improve the accuracy of the proposed data model, therefore each rules are evaluated using KNN algorithm for finding the confidence. Confidence is a value between 0-1 and less the distance having higher confidence. Thus if a rule set having confidence level below then .5 are removed from the rule set and it may be a weak set of rules. The entire learning and rule reduction process is given using figure 4.1. Here the given confidence values indicate the amount of attributes are matched with the training input datasets in terms of distance.



Learning process

Methodology

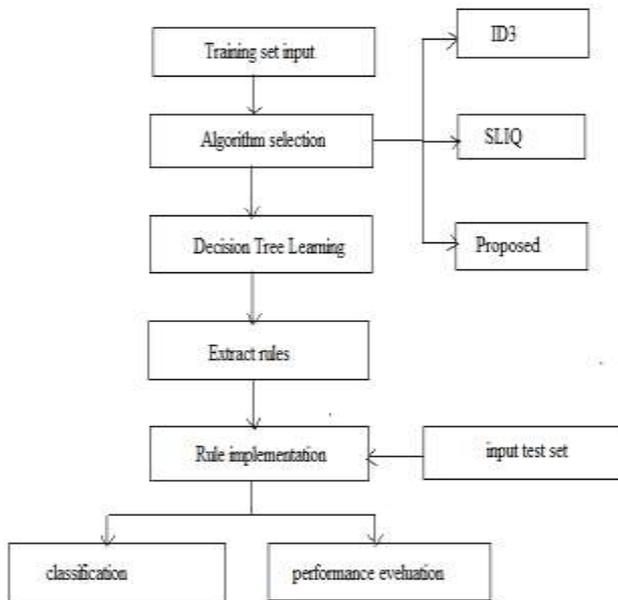
The proposed methodology is based on the decision tree assembling technique for improving the classification accuracy of SLIQ classifiers. The proposed methodology for performance improvement is discussed using Proposed Methodology



Proposed Methodology

System Architecture

This section provides the system design for demonstration of the effectiveness of the proposed classification algorithm over the traditional techniques implemented with the proposed classifier. The proposed simulation architecture is given using simulation architecture.



Simulation Architecture

V. RESULT & ANALYSIS

RESULT

In results the performance of the implemented system. Additionally the comparative outcomes are also included to justify the presented data model for agriculture land classification.

TRAINING TIME

The amount of time required to develop data model using the input training samples are known as the training time.



BUILD TIME

CLASSIFICATION TIME

The amount of time required to classify the input test set is known as the classification or search time.



SEARCH TIME

MEMORY CONSUMPTION

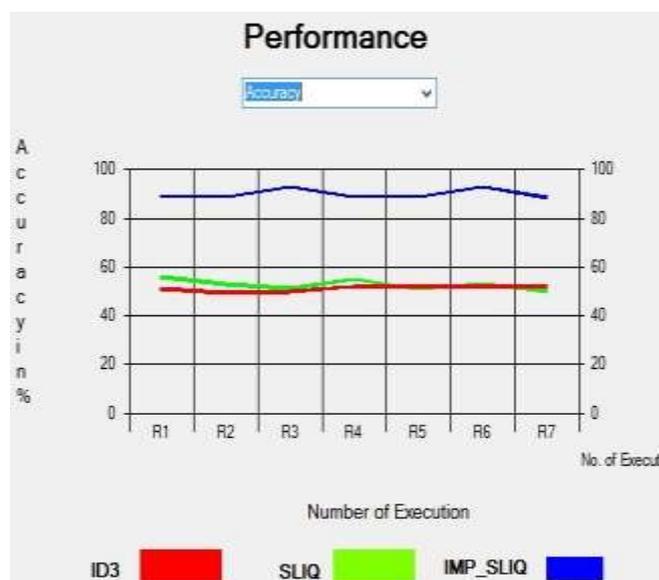
The amount of main memory required to successfully execute the implemented algorithms are known as the memory consumption of the system.



MEMORY CONSUMPTION

ACCURACY

The amount of test data correctly recognized by a classifier is known as the accuracy of the algorithm.



COMPARATIVE ACCURACY

ANALYSIS

The performance of the implemented system is analyzed on the basis of different parameters including accuracy(in %),training time(in ms),search time(in ms), and memory consumption(in kb).Table shows the performance summary.

S. No.	Algo. Parameter	ID3	SLIQ	Proposed
1	Accuracy (in %)	82.98	82.146	90.548
2	Training Time (in Ms)	0.218	0.187	0.172
3	Search Time (in Ms)	0.609	0.500	0.500
4	Memory Consumption(in Kb)	54840	55768	56412

VI. CONCLUSION

Data mining is an essential computer application for evaluation of huge data and identification of essential target patterns over data. Therefore this technique is widely accepted and utilized for automated data analysis. In this presented work the data mining techniques more specifically decision tree algorithms are investigated for efficient and accurate pattern extraction from raw data. Basically the implemented work includes three keys objectives to accomplish.

1. Therefore first the soil properties are analysed on the basis of their chemical composition like soil ph,soil texture,soil thickness,soil erosion,organic matter etc.those properties are helps for deciding effective attributes by which the land are classified according to their appropriate use. There are only three kind of class labels are decided agriculture use, construction and none.
2. After investigation some improvements on traditional technique are performed. Therefore SLIQ algorithm and ID3 algorithm is implemented first and then the promising technique SLIQ is improved using knn classifier. Additionally the performance of classifier is evaluated. Finally for justifying the outcomes of the implemented classifier the evaluated performance is compared with the traditional classification schemes.
3. According to the evaluated performance and given performance summary the implemented classification technique provides the efficient performance and accurate classification

outcomes for soil dataset. Thus the implemented classification scheme is adoptable as compared to the traditional techniques.

VII. REFERENCES

- [1] P. Bhargavi, Dr. S. Jyothi, "Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.8, August 2009
- [2] Yoshihiro Kawamura, Shigeru Takasaki, Masashi Mizokami, "Using decision tree learning to predict the responsiveness of hepatitis C patients to drug treatment", 2012 Federation of European Biochemical Societies Published by Elsevier B.V.
- [3] Brain Decoding of fMRI Connectivity Graphs Using Decision Tree Ensembles, 978-1-4244-4126-6/10/\$25.00 ©2010 IEEE
- [4] S Keskar, R Banerjee, "Time-Recurrent HMM Decision Tree to Generate Alerts for Heart-Guard Wearable Computer", Computing in Cardiology, 2011, 2011 - ieeexplore.ieee.org
- [5] Jiawei Han and Micheline Kamber, "Data mining Concepts and Techniques", Second Edition, <http://akademik.maltepe.edu.tr/~kadiredem/772sData.Mining.Concepts.and.Technique-s.2nd.Ed.pdf>
- [6] A Comparison of Several Approaches to Missing Attribute Values in Data Mining, Jerzy W. Grzymala-Busse and Ming Hu, Springer-Verlag Berlin Heidelberg 2001, pp. 378–385,
- [7] "Data Mining - Classification & Prediction Introduction", http://www.idc-online.com/technical_references/pdfs/data_communications/Data_Mining_Classification_Prediction.pdf
- [8] Data Mining - Cluster Analysis, http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm
- [9] Mohd. Mahmood Ali, Mohd. S. Qaseem, Lakshmi Rajamani, A. Govardhan, "Extracting Useful Rules Through Improved Decision Tree Induction Using Information Entropy", International Journal of Information Sciences and Techniques (IJIST) Vol.3, No.1, January 2013
- [10] M. Jayakameswaraiah and S. Ramakrishna, "Implementation of an Improved ID3 Decision Tree Algorithm in Data Mining System", International Journal of Computer Science and Engineering Volume-2, Issue-3 E-ISSN: 2347-2693 Published: 30 March 2014
- [11] Rashmi R. Tundalwar, Prof. Manasi Kulkarni, "Web Spam Detection Using Improved Decision Tree Classification Method", International Journal of Computer Science and Information Technologies, Vol. 5 (4), 2014, 4936-4942
- [12] A. S. Galathiya, A. P. Ganatra and C. K. Bhensdadia, "Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning", International Journal of Computer Science and Information Technologies, Vol. 3 (2), 2012, 3427-3431
- [13] Hongze Qiu, Haitang Zhang, "Fuzzy SLIQ Decision Tree Based on Classification Sensitivity", I. J. Modern Education and Computer Science, 2011, 5, 18-25
- [14] Yajuan Wang, Marc Simon, Pramod Bonde, Bronwyn U. Harris, Jeffrey J. Teuteberg, Robert L. Kormos and James F. Antaki, "Prognosis of Right Ventricular Failure in Patients with Left Ventricular Assist Device Based on Decision Tree with SMOTE", Copyright (c) 2011 IEEE. Personal use is permitted
- [15] David A. N. Ussiri, Rattan Lal, "Land Management Effects on Carbon Sequestration and Soil Properties in Reclaimed Farmland of Eastern Ohio, USA", Open Journal of Soil Science, 2013, 3, 46-57