Algorithm for Producing Compact Decision Trees for Enhancing Classification Accuracy in Fertilizer Recommendation of Soil

Navneet
Maharshi Dayanand University,
Deptt. Of Computer Science and Applications,
Rohtak Haryana,India

Nasib Singh Gill
Maharshi Dayanand University,
Deptt. Of Computer Science and Applications,
Rohtak Haryana,India

ABSTRACT
Data mining is the process of automatic classification of cases, based on data patterns obtained from a data set. Number of algorithms has been developed and implemented to extract information and discover knowledge patterns that may be useful for decision support. This paper proposes a technique that compact the decision tree increase the classification accuracy. The algorithm is developed by cascading the clustering and decision tree classification algorithm. The algorithm completes the process of two steps. In the first step, clustering is performed on training instances and in second step then the classification occurs on the clusters. A Schwartz criterion is used to get the optimal number of clusters. The algorithm is tested with the soil data set and various other online available datasets using WEKA. The simulation result shows that compact tree is formed, and the classification accuracy of the proposed algorithm is better than the classification accuracy of existing algorithms. The paper also presents the real-world application of proposed work in recommendation of fertilizers for soil dataset.

General Terms
Data mining, Decision Tree

Keywords
Data mining, c4.5, WEKA, k-mean clustering, Schwarz criteria

1. INTRODUCTION
Data mining is a technique that gives meaning to the existing data. Basic purpose of the data mining is to produce a series of patterns by analyzing the existing data. In other words, it is the process to get logical patterns from database [1]. Data mining is useful due to its applications available in various fields like machine learning, data visualization and pattern recognition.

The main goal of the data mining is the prediction and the description. Prediction data mining develops a technique to analyze the existing data set to get unknown features of interest. While the description finds the patterns to describe the data that can be understood by the user. Basically predicts unknown features and description describes new information. Some of the existing techniques of data mining are not suitable in the current world due to high-dimensionality and heterogeneity of the data [2]. This paper improves the existing technique of classification so that it can be applied to the today's world applications.

Rest of the paper is divided into four sections. Section 1 contains various data mining techniques and section two describes the material and methods used in the paper with the existing technique and its drawback. The section 3 describes the modification of the existing technique and its implementation results on various datasets. The final section gives the application of the proposed algorithm in real world for fertilizer recommendation.

2. DATA MINING TECHNIQUES
Data mining is introduced, to be get used in various fields like Machine learning, Database system etc; so its techniques are derived from these fields. Data mining techniques consists of clustering, classification, nearest neighbor and regression. These techniques are described below.

2.1 Clustering
Clustering is the process to group the similar data and to separate the dissimilar data. Clustering is useful when similar type of data is targeted. The main goal of clustering algorithm is to generate minimum numbers of clusters to describe the data. The data is grouped on the basis of the similarity matrices [3]. The number of clusters can be predefined [4] or dependent on some threshold [5].

2.2 Nearest Neighbor
Nearest Neighbor is a classification technique that is completed in two phases. First phase is the learning phase in which different instance are learned. The second phase is the classification in which class for a new point is found on the basis of the weight of the existing nearest neighbor [6]. The greater weight of the nearest neighbor increases the accuracy of classification [7].

2.3 Classification
Classification is the process to classify the data in to predefined classes by satisfying given constraints. In classification training is used to generate a model that classifies the unclassified data [3]. Various existing techniques of classification are decision tree, multilayered preceptron etc.

2.4 Regression
Regression is the process to predict the unknown target values of various datasets. In regression, several dataset of known target values are used to generate a model for predication [8]. In training the difference between the target and the predicted value is minimized. The training gets completed when we get same error in successive steps. The generated model is applied to get the unknown target value.

3 MATERIALS AND METHODS

3.1 Soil Data Set description

The basic component to grow the crops is soil. Different parameters to test the suitability of soil for any particular crop are ph, electric conductivity, organic carbon, ph ,potash [9].On the basis of these parameters classification can be applied to know the crop that can be grown on particular soil even after some treatment [10]. The amount and the type of fertilizer can also be analyzed by such classification. In this paper the classification algorithm is generated to recommend the fertilizer. We need soil dataset to apply the classification algorithm. The required data is collected from http://agriharyana.nic.in. This website contains soil databases of various districts and their blocks within Haryana region. The steps to create the dataset from data of website http://agriharyana.nic.in are as follow

1. Select few samples from the available samples on website http://agriharyana.nic.in
2. Calculate the no of samples in term of percentage by dividing the number of samples taken by total number of samples.
3. Various attributes are assigned low or medium or high on the basis of their actual value and maximum value possible.
4. Design a new attribute named class on the basis of the attributes ph, phos and potash.

The dataset created from the data of website contains 1941 instance each containing 9 attributes. Each attribute can be described in following table.

Table 1 List of Attributes of Soil Data Set

Attribute Name	**Attribute Description**	**Values**
Sr. No.	Serial Number of Instance	Assigned Serial no. from 1
Block	Block Code	Actual code of Blocks
Village	Village code	Serially generated code,village code starts from 1 for each block
Ph	Ph value	1,2,3*
EC	Electric Conductivity	1,2,3*
OC	Organic carbon	1,2,3*
Potash	Potash Concentration	1,2,3*
Phosphorus	Phosphorus Concentration	1,2,3*
Class	Class depending on values of attributes	MLH,MLM,MLL,MMM,MML

***1- Low Concentration, 2—medium concentration, 3- High Concentration**

3.2 C4.5

C4.5 is a classification algorithm that produces decision tree and rules for the datasets containing categorical and the numerical attributes. It is used to predict the class on the basis of given attributes by the procedure explained in next paragraph.

Information gain i.e. entropy difference is calculated for each attribute. The attribute having maximum information gain will act as decision node. In Other words, list of attribute is divided into sub-lists on the basis of attribute having the maximum information gain. The process is repeated on all sub-list until we get the same class for each attribute of sub-list or no information gain [11]. The performance of algorithm can be analyzed on already classified list by comparing the class of each attribute to the calculated class.

3.3 K-Mean Clustering

K-mean clustering is a fast algorithm to divide the data in to k groups [4]. It firstly selects k points as the centroid of the k clusters. Then it calculates the Euclidean distance of each data point to centroid of each cluster. Data point is assigned to the cluster having minimum Euclidean distance. New centroids are evaluated by calculating the mean of each cluster data points. The process is repeated until specified iterations achieved or same centroid evaluated in successive iterations.

The main drawback of the K-Mean clustering is that k is predefined. In other words, we have to specify the number of clusters initially. If we underestimate the number of clusters, then fewer numbers of clusters may lead to forced assignment of instances to the clusters [12]. If numbers of clusters are overestimated, then one instance can appear in more than one cluster. One more problem occurs if the size of one cluster is very large as compared to other clusters than the large cluster try to cover all the instances of dataset. This drawback can be removed by cascading the clustering and the classification algorithm. Muniyandi, Amuthan, Prabakar et. al. (2012)[13] presents that cascading of the K-mean clustering and C4.5 say existing technique can remove this drawback. The existing algorithm is completed in two phases first phase is the k-mean clustering, and the second phase is the classification of the clustered instances. However, it is must create the optimal number of clusters to recover all other drawbacks. That's why this paper uses Schwarz's criterion to generate the optimal number of clusters.

4. PROPOSED TECHNIQUE

Schwarz Criterion (SC) has been used in the proposed algorithm to choose the optimal number of clusters in a given range of values according to intrinsic properties of the specified data set. Schwarz Criterion is a parametric measure of how well a given model predicts the data [14]. It represents a trade-off between the likelihood of the data for the model and the complexity of the model. Proposed algorithm explains the process of the proposed work.

4.1 Proposed Algorithm

I. Input large Dataset of soil sample.
II. Initiate K=smallest value(default k=2);
III. Apply K-means to generate number of clusters say C0, C1, C2 ,……., Cn.
IV. For i=1:n
V. Calculate the Schwarz criterion for cluster Ci by using

$$SC = -2.\ln\hat{L} + k\ln(n) \quad (1)$$

Where x = data within the cluster Ci
n= the number elements in Ci
k = the number of parameters to be estimated.
$\hat{L}$ = = The maximized value of the likelihood function of the model M i.e $\hat{L} = p(x|\hat{\theta}, M)$ where $\hat{\theta}$ are the parameter values that maximize the likelihood function.

VI. Apply K-mean on Ci Clusters for k=2 say generated Clusters are Ci1 and Ci2

VII. Calculate the SC for Clusters Ci1 and Ci2 by using

$$SC1 = -2.\ln\hat{L} + 2 * k\ln(n) \quad (2)$$

Here, the number of parameters get doubled due to two cluster.

VIII. If SC>SC1 then n=n+1 i.e. new model preferred.

IX. Ci=Ci1 and Cn=Ci2

X. i=i-1

XI. End if

XII. End

XIII. For i=1:n

XIV. Compute Class frequency in Ci say it cli

XV. If cli=1then create tree with one node

XVI. Else

XVII. For each attribute in attribute list calculate gain ratio by using

$$\text{gain} = \text{info}(T) - \sum_{i-1}^{c} \frac{|T_i|}{|T|} \times \text{info}(T_i) \quad (3)$$

Where T is the total set of cases and c denotes classes.

XVIII. The attribute having maximum gain say N

XIX. If N is continuous then find threshold that denotes the greatest value of the whole training set.

XX. Create a Node and classify the data on basis of attribute N and remove the attribute from attribute list and go to step XIV.

XXI. End

5. RESULTS AND ANALYSIS

The simulation of the proposed algorithm is done using the WEKA. Figure 1, 2, 3 represents the decision tree generated by the proposed algorithm. Figure 2 represents cascaded k-mean + C4.5 algorithm [13] and figure 3 represents decision tree of C4.5 algorithm [15].

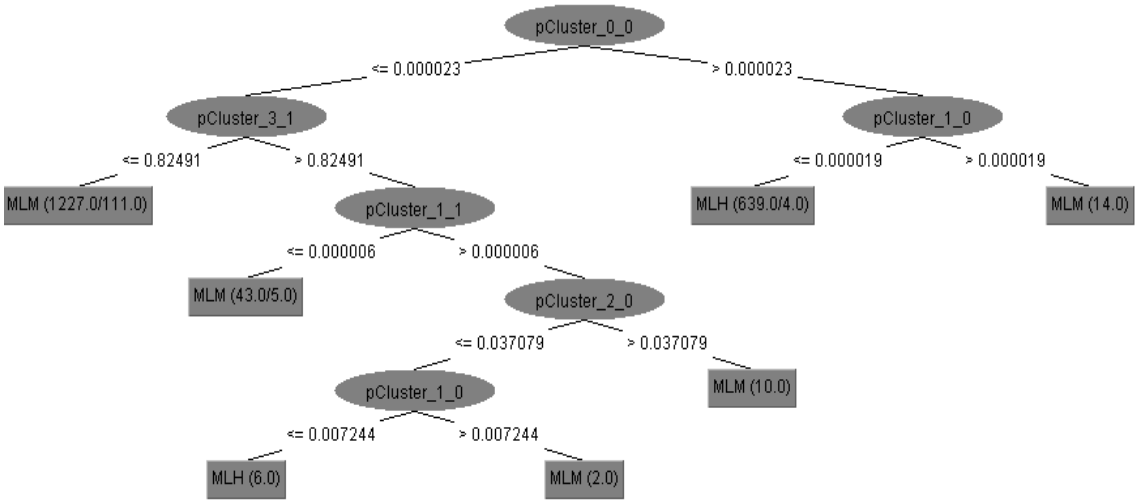

Figure 1. Decision Tree of Proposed Algorithm

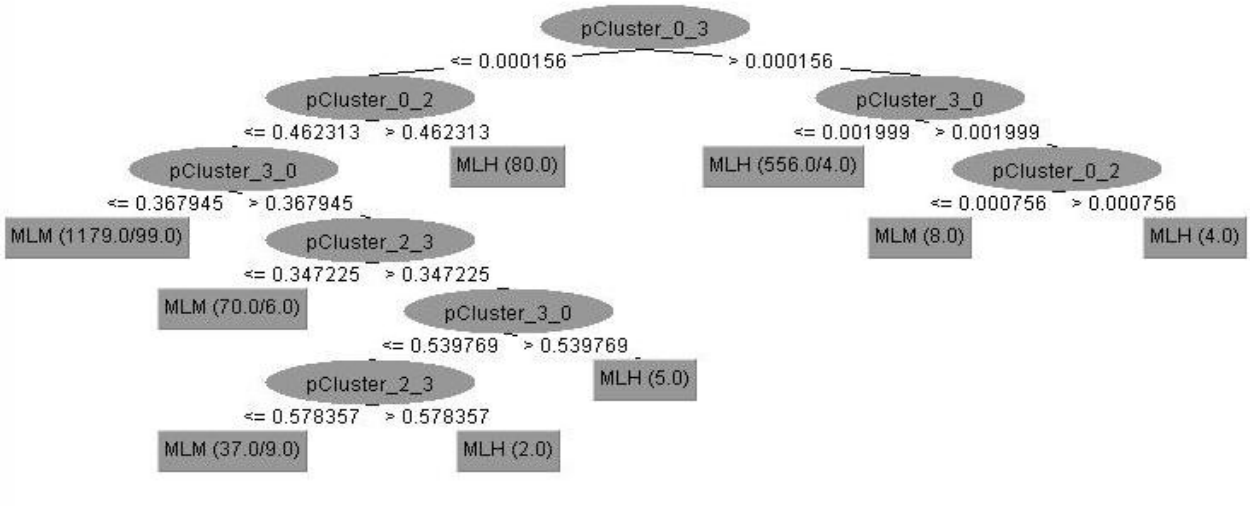

Figure 2: Decision Tree of K-Mean and C4.5 Algorithm

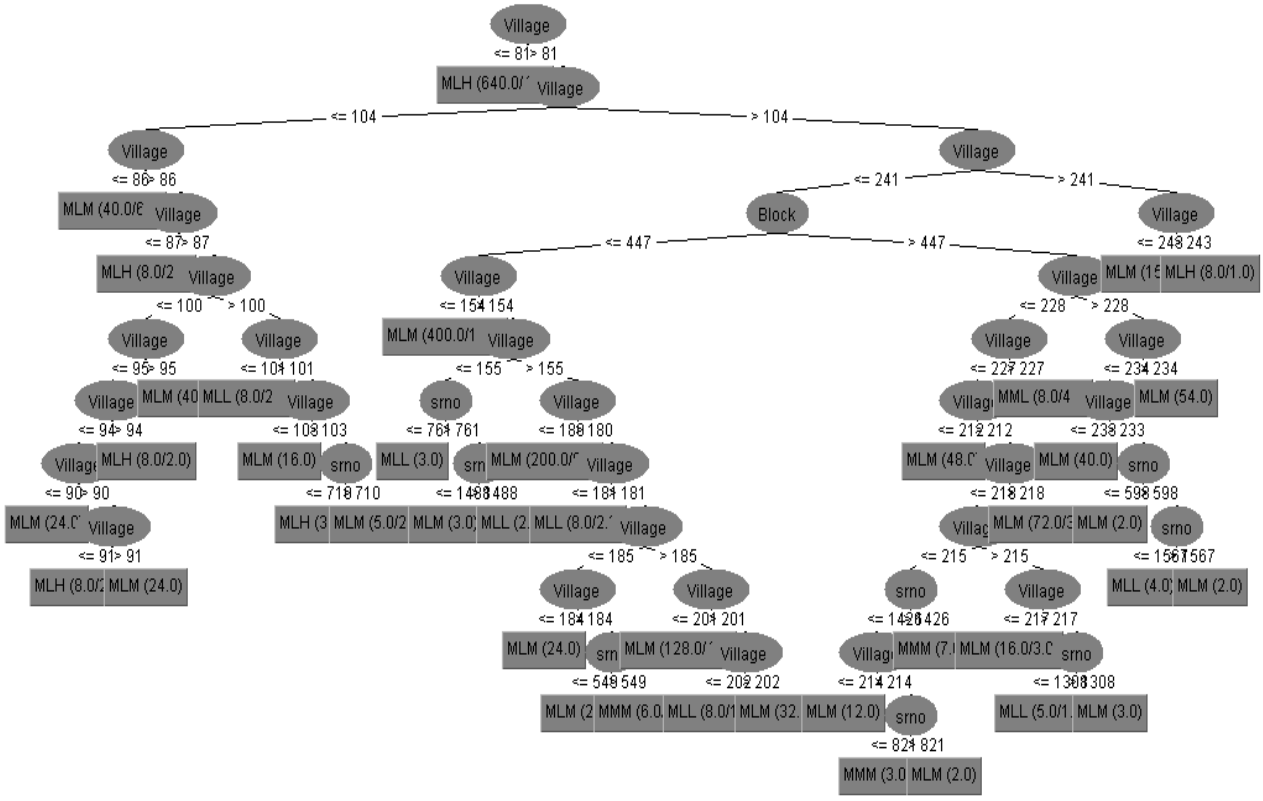

Figure 3: Decision Tree of C4.5 Algorithm

The performance of the algorithm can be measured by using various parameters like TP rate, FP rate, recall, etc. True positive rate (TP rate) is the number of instance belongs to same class as specified by the algorithm divided by the total number of instance. False-positive rate (FP rate) is the number of instance doesn't belong to the class specified by the algorithm divided by the total number of instances. Precision is the probability that randomly selected instance is correctly classified that can be given as $\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \times \ 100\%$ Recall is the average of probabilities of all instances within dataset. $\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \times \ 100\%$ .F-measure is mean of precision and the recall can be given as $\text{F} - \text{measure} = \frac{2*\text{precision}*\text{recall}}{\text{precision+recall}}$.

Table 2: Comparison of Various algorithms using various parameters

Algorithm	Size of tree	Number of leaves	Classification accuracy	TP rate	FP rate	Precision	recall	F-measure
J48(C4.5)	81	41	94.17	0.942	0.07	0.928	0.942	0.932
Existing	17	9	95.6	0.946	0.067	0.934	0.946	0.937
Proposed	13	7	97.17	0.951	0.065	0.941	0.951	0.941

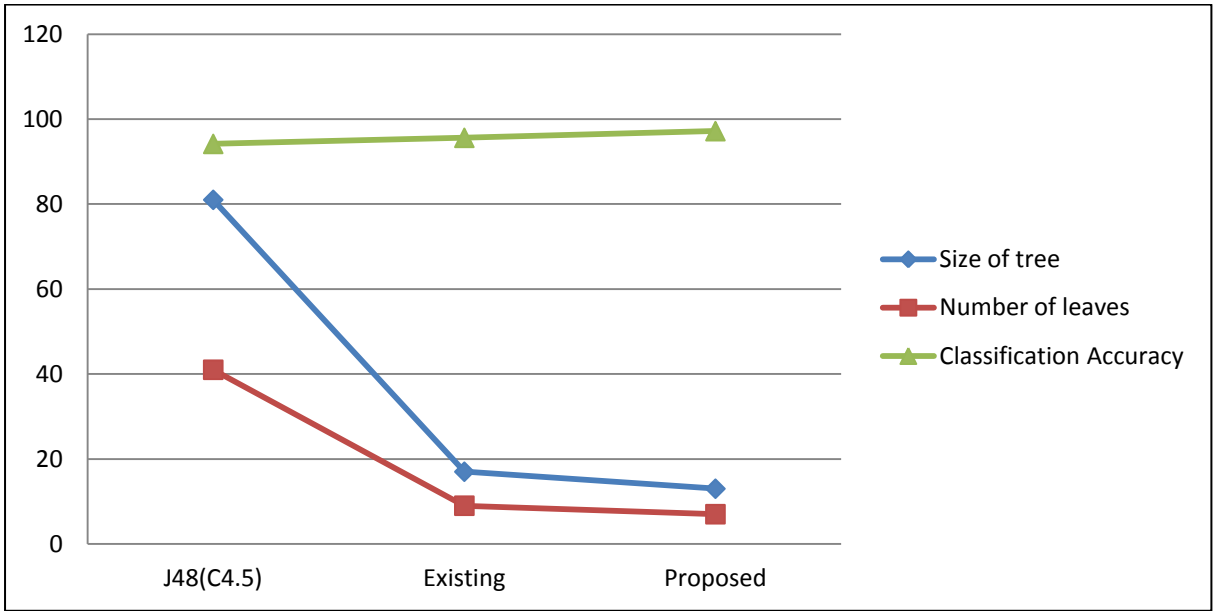

Figure 4 Size and Accuracy Comparison of tree by using various Algorithm

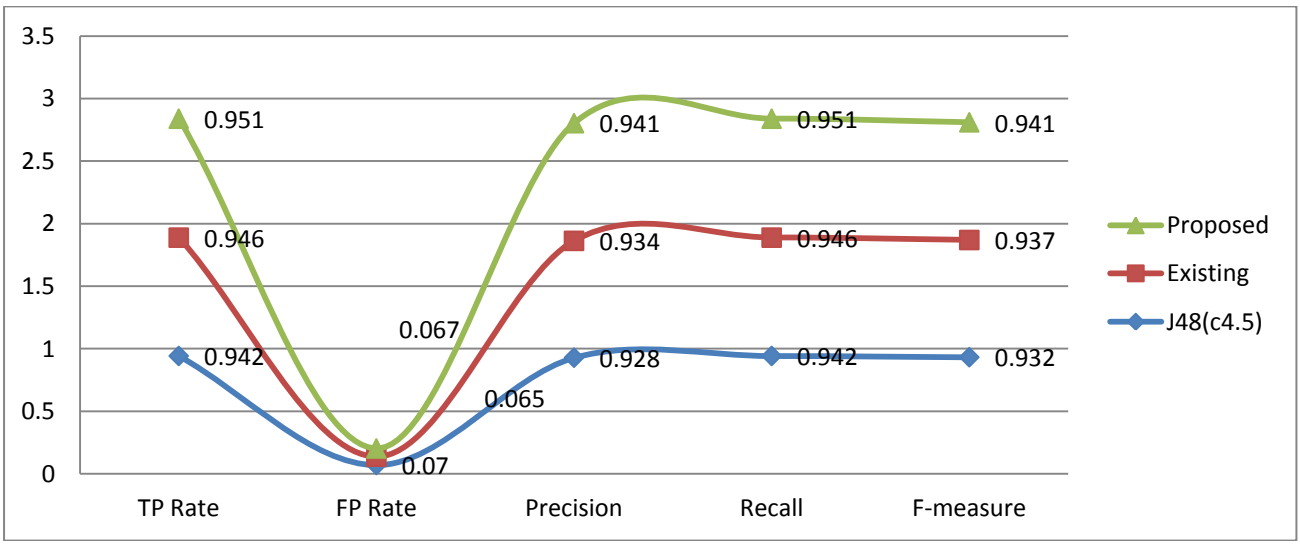

Figure 5 Comparison of proposed and existing algorithm

Table 2 shows the comparison among C4.5 (J48), cascaded k-mean + J48 and the proposed algorithm. The classification accuracy and size of the proposed algorithm is better than the other algorithms C4.5 (J48) and cascaded K mean +J48.Overall a compact tree having greater classification accuracy is generated by the proposed algorithm. Figure 3 and figure 4 shows the graphical comparison of size as well as other parameters for J48 and existing and proposed algorithm. The algorithm can be compared on other datasets. Various datasets downloaded that are available with WEKA are used to verify the performance of the proposed algorithm. The datasets used are diabetes, glass and the ionosphere. Table 3 specifies various characteristics and performance comparison of the different algorithm on these three datasets.

Table 3: Comparison of C4.5, existing, proposed algorithms on different datasets.

Data set description			Tree size			Number of leaves			Classification accuracy		
Name	**Number of instances**	**Number of attributes**	**J48**	**Existing**	**proposed**	**J48**	**Existing**	**proposed**	**J48**	**Existing**	**proposed**
Diabetes	768	9	39	17	5	20	9	3	73.82	75.32	76.80
Glass	214	10	59	55	51	30	28	26	66	68	69
Ionosphere	351	35	35	11	9	18	6	5	91.45	92.9	94.45

Table 3 shows the proposed algorithm gives compact size of tree as well as the higher classification accuracy for all datasets.

6. APPLICATION OF PROPOSED ALGORITHM IN REAL WORLD

Proposed algorithm can be used to recommend the fertilizer according to the block and village. The algorithm will classify the soil based on its property. The fertilizer for each class is recommended by scientific research drawn from http://agriharyana.nic.in. In other words, proposed work recommends different fertilizer for different blocks of Haryana state in country India for particularly crop of wheat. The result for the data set described in section 3.1 is shown in the figure below.

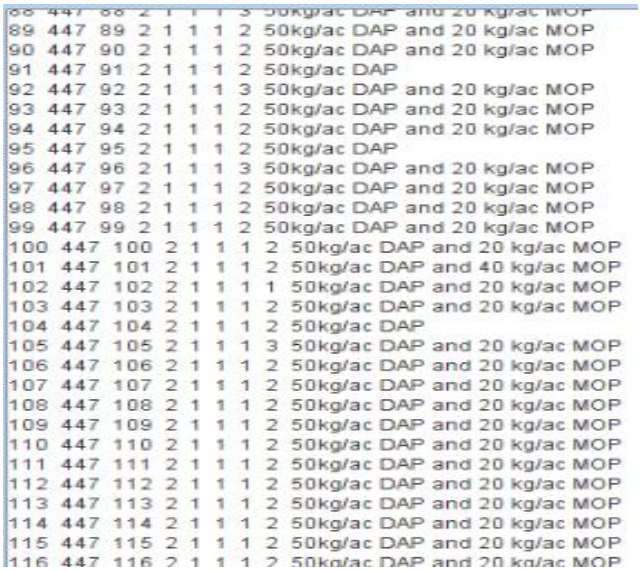

Figure 6 Result of Fertilizer Recommendation for soil Dataset

7. CONCLUSION

This paper proposes a technique that produces a compact decision tree having increased classification accuracy. The algorithm is developed by cascading the clustering and decision tree classification algorithm. The SC (Schwarz Criterion) is applied to get the optimal number of clusters, and then the C4.5 decision tree is applied to get the decision tree. The algorithm is simulated using WEKA on the soil data set and three other datasets, and the result shows improved classification accuracy and compact decision tree which results in fast and more accurate recommendation of fertilizers for soil of real world dataset. Various parameters like TP rate, FP rate, precision, recall, and f-measure are also evaluated to analyze the performance of the proposed algorithm. In future, the decision tree generating rules can be optimized and fertilizers recommendation can extended to other crops of different seasons.

8. REFERENCES

[1] Kruse ,R. , Riccia ,G. , Della et. al.,2000,Computational Intelligence in Data Mining, Springer, New York, NY, USA.

[2] Stonebraker, M., Agrawal ,R. et. al.,1993,.DBMS Research At A Crossroads: The Vienna Update. In Proc. of the 19th VLDB Conference, pp 688-692, Dublin, Ireland.

[3] Chen M.S., Han J., and Yu P.S.,December 1996. Data mining: An overview from database perspective. IEEE Transactions on Knowledge and Data Engg., 8(6): pp. 866-883.

[4] Mac Queen, J. B. ,1967, Some Methods For Classification And Analysis Of Multivariate Observations. Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, University of California Press, pp 281- 297.

[5] Hastie, T. and Tibshirani, R.,1996, Discriminant Adaptive Nearest Neighbor Classification, IEEE Transaction on Pattern Analysis and Machine Intelligence, 18(6),pp 607-616.

[6] Tan, Pang-Ning, Steinbach, Michael and Kumar Vipin,2006,.Introduction to Data Mining, Addison Wesley.

[7] Mitchell, T.M.,1997,.Machine Learning, McGraw-Hill Companies, USA.

[8] Singh, Nanhay,2012, Data Mining With Regression Technique, Journal of Information Systems and Communication ISSN: 0976-8742 & E-ISSN: 0976-8750, Volume 3, Issue 1,pp. 199-202.

[9] Methods Manual-Soil Testing in India,2011, Department of Agriculture & Cooperation Ministry of Agriculture Government of India.

[10] Gholap, Jay, et al, 2012, Soil Data Analysis Using Classification Techniques and Soil Attribute Prediction. arXiv preprint arXiv:1206. 1557.

[11] Wu, X., Kumar ,V. et. al.,2008, Top 10 algorithms in data mining, Knowledge Information System.

[12] Gaddam, R. , Shekhar et. al .,March 2007, K-Means+ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-Means Clustering and ID3 Decision Tree Learning Methods, IEEE transactions on knowledge and data engineering, vol. 19, no. 3.

[13] Muniyandi, Amuthan, Prabakar et. al. ,2012, Network Anomaly Detection by Cascading K-Means, International Conference on Communication Technology and System Design 2011,Elsevier , Procedia Engineering 30 – pp174 – 182.

[14] Cavanaugh, Joseph E., and Neath. Andrew.,1999. A. Generalizing the derivation of the Schwarz information criterion. Communications in Statistics-Theory and Methods 28.1: 49-66.

[15] Quinlan, J. R. ,1993, C4.5: Programs for Machine Learning, Morgan Kaufmann.

[16] Natural Resources Conservation Service, United States Department of Agriculture. Website:" http://soils.usda.gov/survey/geography/ssurgo/description_statsgo2.html"