IJCTA, 9(40), 2016, pp. 33-41

Classification of Soil type in Salem District Using J48 Algorithm

N. Hemageetha* and G.M. Nasira**

Abstract : Data mining in agriculture is a novel research field. Analysis of soil is a major component of soil resource management in agriculture. In this paper various soil types are analyzed using Data mining clustering and classification techniques for Salem district. Compared with other classifiers, clustered J48 classifier produce high accuracy result. Soil type classification will help the farmer to cultivate suitable crops in a particular type of soil.

Keywords : Agricultural Soil, Classification, Clustering, Data mining.

1. INTRODUCTION

Data mining is the process of finding out the interesting patterns and knowledge from large amount of data. Various data analysis techniques are available for agricultural researches [1]. Soil analysis is useful for farmers to determine the type of crops to be cultivated in a particular type of soil. The main aim of this paper is to analyze the soil types for cultivation of crops in Salem district. Salem is one of the biggest districts in Tamil Nadu. This district has nine Taluks, twenty Blocks, Three hundred and seventy six Panchayats and six hundred and thirty one revenue villages. The geographical area of Salem district is 5205.30 sq. kilometers. It is located at 11.669437°N 78.140865°E, at an average elevation of 278 m (912 ft) [2]. This city is surrounded by hills. The soils of Salem District can be classified as Red soil, Black soil, Alluvial soil and Loamy soil. Salem district receives major rainfall from the South-West Monsoon followed by North-East monsoon. Salem district receives maximum rainfall through northeast monsoon [3].

This paper focuses on classification of the soil types based on available nutrients for Salem district in Tamil Nadu.

Section II presents Literature review. Section III presents a few Classification and clustering algorithms, Section IV presents Experiment And Results finally Section V presents Conclusion And Future Scope.

2. LITERATRE STUDY

The study about the previous work related to data mining in agriculture was analyzed in [4]. Among the various areas of agricultural research as crop cultivation, market analysis, vegetable price prediction [5] [6][7] and classification of soil plays an important role[8]. By analyzing the soil, guidance can be given on the type and amount of fertilizers that can be used in that soil. Soil consists of various nutrients present in organic and mineral forms. Soil testing is an important component of sustainable nutrient management in agriculture. Nitrogen, phosphorus, potassium are called major or macro nutrient because they are important for plants to grow healthy. macronutrients and micronutrients availability is affected by soil pH

* Research Scholar Department of computer Science Periyar University, Salem, Tamil Nadu *geekani2010@gmail.com*

** Assistant Professor and Head Department of Computer Application Chikkanna Govt. Arts College Tirupur, Tamil Nadu *nasiragm99@yahoo.com*

value. Salem District soil has many Macro nutrients such as Nitrogen, Phosphorus, Potassium and Micro nutrients such as Iron, Manganese, Zinc [9]. Table I shows that the Nutrient rating of pH, EC, N, P and K [9] of the soil.

Table 1

Nutrient rating of pH, EC, N, P and K

Level / *Parameter*	*Low*	*Medium*	*High*
pH	<6.5 (Acidic)	6.5 – 7.5 (Neutal)	>7.5 (Alkaline)
OC %	<0.5	0.5 – 0.75	>0.75
Ec(*ds* m^{-1})	<1.0 (Harmless)	1.0-3.0 (Injurious)	>3.0 (Critical)
N(Kg ha^{-1})	<280	280 – 450	>450
P(Kg ha^{-1})	<11	11 – 22	>22
K(Kg ha^{-1})	<118	118 – 280	>280

Soil can be classified with the help of the available Nutrients. Generally pH value of Soil falls between 6.5 and 8.5 which is suitable for most of the common crops growth [10]. In [11] soil pH level is analyzed. 7.8 to 9.4 is the pH range of soil in Salem District. In and around the Major Soil types available in Salem District are as follows:

Alluvial soil : Alluvial soil is found mostly in the deltas and flood plains. It is a major soil type found in India (Biswas and Mukherjee, 2005) Alluvial soil is also rich in potassium and has high fertility.

Black soil : The disintegration of lava rocks forms the Black soil. Nutrients like calcium, potassium and magnesium are rich in black soil.

Red soil : Red soil gets its red colour due to the presence of iron oxide. The disintegration of metamorphic and igneous rocks forms the Red Soil.

Loamy soil : Loamy soil is composed of sand,silt and a smaller amount of clay ("Loamy soil", Wikipedia). Loamy soil is important for most garden plants. Not only loamy soil holds plenty of moisture but also drains well so that sufficient air can reach the roots. The main aim of this paper is to classify various soil types in Salem District with the help of data mining techniques.

Soil type is determine by the physico-chemical properties of soil. The pH, EC, OC, N, P, K, and soil texture are used to determine the type of the soil whether it is red,black or alluvial soil. Wide variety of crops can be grown in redsoil. Rice, Sugarcane and Banana etc are well grown in alluvial soil. Cotton, Sunflower and Pulses are grown in black soil. So the classification of soil type is very important for the farmer to cultivate the crops in appopriate soil.

3. AGRICULTURE DATAMINING ALGORITHMS

Classification of soil is very difficult to study because depending upon the fertility class of the soil, the domain experts can determine which type of crops to be cultivated in a particular soil and also determine the type of fertilizers to be used for the same. Naive Bayes and J48,Jrip, Bayesian networks classification algorithms as follows.

A. Naive bayes Classifier

British minister Bayes has been developed the Bayes theorem. In this theorem the Independent attributes and dependent attributes are two types of attributes. In this approach all the attributes are assumed to be independent to each other. Bayes theorem is based on the Naïve's assumption. In this Classifier, the value of the dependent attribute is calculated by using the values of the independent attributes.

B. J48(C4.5)

Decision trees approach are used in the data mining process. C4.5 algorithm used to generate the decision trees. Ross Quinlan developed the C4.5 algorithm. The decision trees are generated by set of labeled input data. The C4.5 algorithm is implemented using JAVA and termed as J48 classifier in Weka tool.

C. Bayesian networks

Graphical model is used in Bayesian network. Bayesian network is used to learn relationships between the items and can be used to understanding the problem domain and consequences of intervention.

D. JRip

JRip algorithm implements a propositional rule learner, Repeated incremental pruning to produce Error reduction. Jrip was proposed by William W. Cohen. It is an optimized version of IREP. The classifier divides the dataset into two sets named as growing set and pruning set, then reduces the errors in both sets and generates rules for the sets.

E. Clustering Algorithms

Grouping of a set of objects into meaningful subgroups is called Clustering. It helps the users to understand the grouping in a data set. It is an unsupervised classification technique.

F. Hierarchical Clstering

İn hierarchical clustering method, the objects are grouped into a cluster of trees. It can be further classified into divisive or agglomerative, depending on the hierarchical decomposition formed in a bottom-up or top-down fashion.

G. DBSCAN

The DBSCAN algorithm can identify clusters in large data sets with the help of the local density of database elements, using only one input parameter. The user can choose the parameter value that would be suitable for calculation. Therefore, minimal domain knowledge is required.

H. Simple K-Means

K-Means clustering is used to partition M objects into x clusters in which each object belongs to the cluster with the nearest mean. x different clusters are generated with greatest possible distinction.

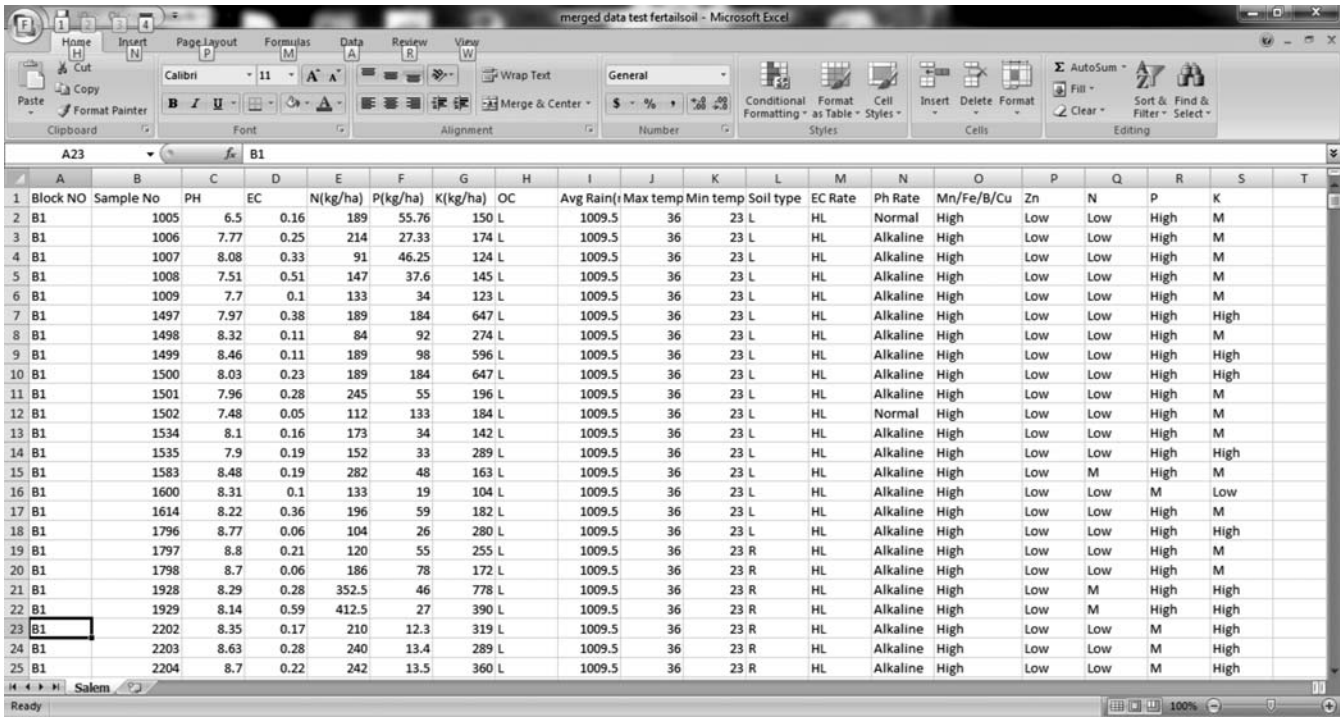
4. EXPERIMENTS AND RESULTS

A. Dataset Collection

Salem district Soil Dataset was collected from Krishi Vigyan Kendra (Farm Science Centre) Tamil Nadu Agricultural University, Santhiyur, Salem. This Dataset has 17 attributes Sample no., Block no, Soil Type ,pH value, Electric conductivity(EC), Organic Carbon(OC), Phosphorous (P), Potassium(K), Nitrogen (N), Avg Rain(mm), Max temp, Min temp, EC Rate, Ph Rate, N rate, P rate, K rate and the total of 701 instances from 792 instances of soil samples.

B. Data Formatting

The WEKA 3.6.13 (Waikato Environment for Knowledge Analysis) workbench is a state of art for machine learning algorithms and data pre-processing tools. It is the open source software for Data Mining. The Excel sheets are converted into .CSV file format which to be accessed in WEKA. From the data base collected out of 792 instances only 701 instances have been considered for our proposed methodology. The instances which had missing attributes values, noisy data and miss match. They are filtered using WEKA filters. The Preprocessed Soil data set is shown in the Figure 1.



merged data test fertailsoil - Microsoft Excel

	Block NO	Sample No	PH	EC	N(kg/ha)	P(kg/ha)	K(kg/ha)	OC	Avg Rain(r	Max temp	Min temp	Soil type	EC Rate	Ph Rate	Mn/Fe/B/Cu	Zn	N	P	K
2	B1	1005	6.5	0.16	189	55.76	150	L	1009.5	36	23	L	HL	Normal	High	Low	Low	High	M
3	B1	1006	7.77	0.25	214	27.33	174	L	1009.5	36	23	L	HL	Alkaline	High	Low	Low	High	M
4	B1	1007	8.08	0.33	91	46.25	124	L	1009.5	36	23	L	HL	Alkaline	High	Low	Low	High	M
5	B1	1008	7.51	0.51	147	37.6	145	L	1009.5	36	23	L	HL	Alkaline	High	Low	Low	High	M
6	B1	1009	7.7	0.1	133	34	123	L	1009.5	36	23	L	HL	Alkaline	High	Low	Low	High	M
7	B1	1497	7.97	0.38	189	184	647	L	1009.5	36	23	L	HL	Alkaline	High	Low	Low	High	High
8	B1	1498	8.32	0.11	84	92	274	L	1009.5	36	23	L	HL	Alkaline	High	Low	Low	High	M
9	B1	1499	8.46	0.11	189	98	596	L	1009.5	36	23	L	HL	Alkaline	High	Low	Low	High	High
10	B1	1500	8.03	0.23	189	184	647	L	1009.5	36	23	L	HL	Alkaline	High	Low	Low	High	High
11	B1	1501	7.96	0.28	245	55	196	L	1009.5	36	23	L	HL	Alkaline	High	Low	Low	High	M
12	B1	1502	7.48	0.05	112	133	184	L	1009.5	36	23	L	HL	Normal	High	Low	Low	High	M
13	B1	1534	8.1	0.16	173	34	142	L	1009.5	36	23	L	HL	Alkaline	High	Low	Low	High	M
14	B1	1535	7.9	0.19	152	33	289	L	1009.5	36	23	L	HL	Alkaline	High	Low	Low	High	High
15	B1	1583	8.48	0.19	282	48	163	L	1009.5	36	23	L	HL	Alkaline	High	Low	M	High	M
16	B1	1600	8.31	0.1	133	19	104	L	1009.5	36	23	L	HL	Alkaline	High	Low	Low	M	Low
17	B1	1614	8.22	0.36	196	59	182	L	1009.5	36	23	L	HL	Alkaline	High	Low	Low	High	M
18	B1	1796	8.77	0.06	104	26	280	L	1009.5	36	23	L	HL	Alkaline	High	Low	Low	High	High
19	B1	1797	8.8	0.21	120	55	255	L	1009.5	36	23	R	HL	Alkaline	High	Low	Low	High	M
20	B1	1798	8.7	0.06	186	78	172	L	1009.5	36	23	R	HL	Alkaline	High	Low	Low	High	M
21	B1	1928	8.29	0.28	352.5	46	778	L	1009.5	36	23	R	HL	Alkaline	High	Low	M	High	High
22	B1	1929	8.14	0.59	412.5	27	390	L	1009.5	36	23	R	HL	Alkaline	High	Low	M	High	High
23	B1	2202	8.35	0.17	210	12.3	319	L	1009.5	36	23	R	HL	Alkaline	High	Low	Low	M	High
24	B1	2203	8.63	0.28	240	13.4	289	L	1009.5	36	23	R	HL	Alkaline	High	Low	Low	M	High
25	B1	2204	8.7	0.22	242	13.5	360	L	1009.5	36	23	R	HL	Alkaline	High	Low	Low	M	High

Figure 1: Preprocessed Soil Data Set

The framework of soil type classification is given in Figure 2.

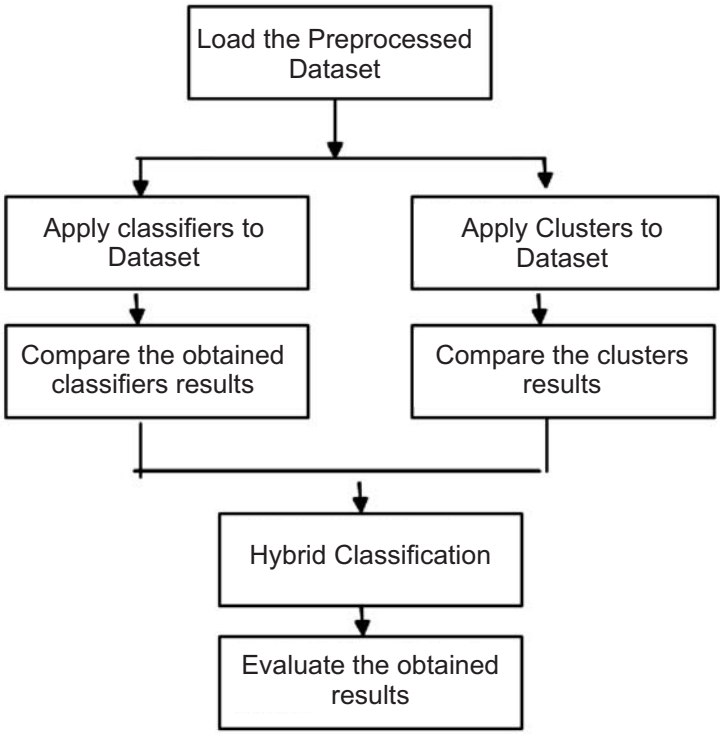


Figure 2: Framework of Soil type Classification

5. RESULTS AND DISCUSSION

A. Soil type Classification

Soil type is classified based on the available parameters. 701 samples were processed through four classifiers (Naive Bayes, J48, Bayesian networks, Jrip) for obtaining better Accuracy and less Error rate Other existing classifiers used among various researchers were also compiled with soil data which lead to less accuracy and more Error rate [12]. The Weka Screen shot of J48 classifier is given in Figure 3.

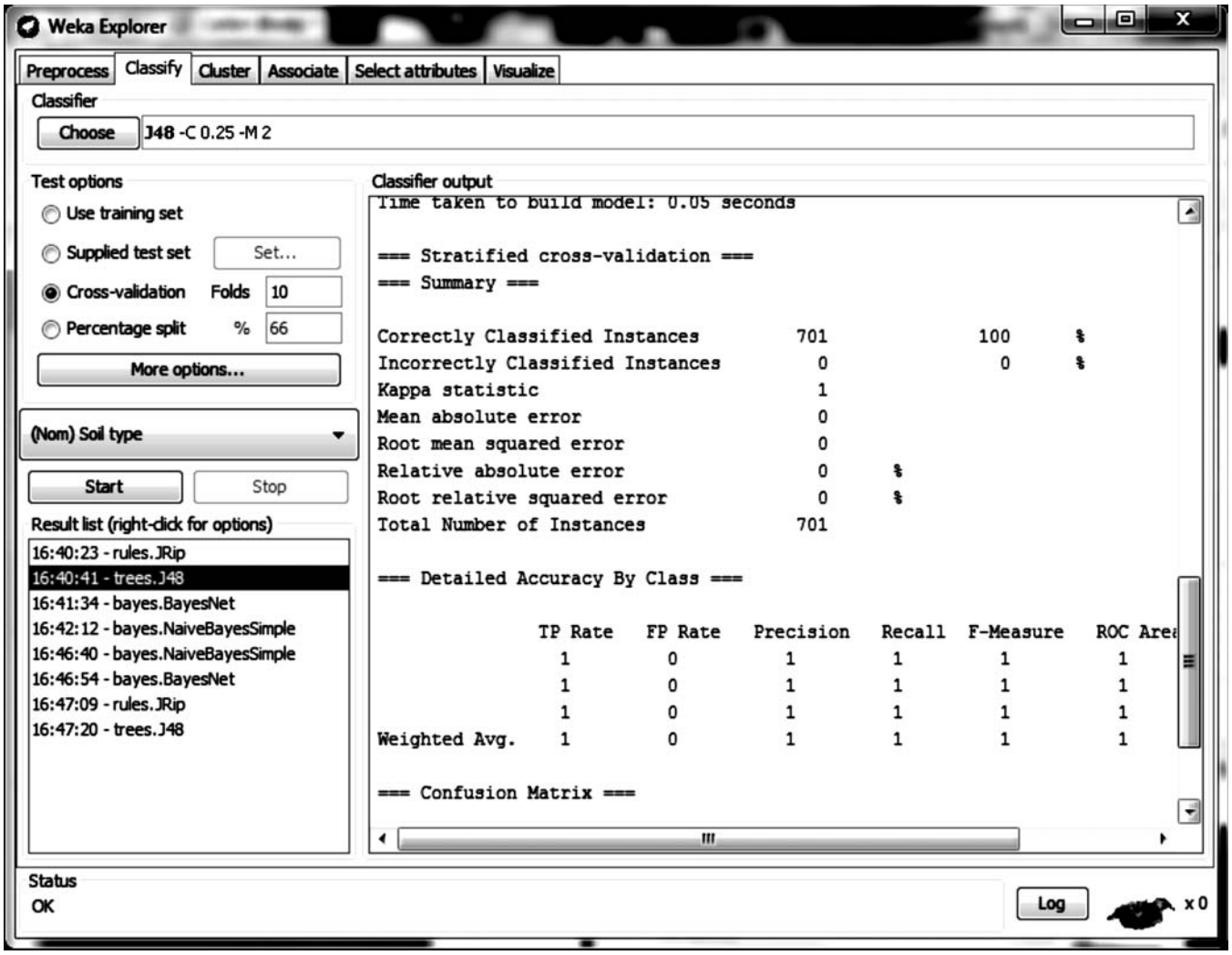


Figure 3: Weka Screen Shot of J48 Classifier for Soil type

The results of all the classifiers are compared based on its Accuracy and Mean absolute Error Rates. The J48 classifier shows the best results compared with other classifiers. Classifiers results for Soil type are compared which are given in Table 2.

Table 2

Comparison Results of clasifier soil type

S.No	*Classifier*	*Naive bayes Simple*	*Bayes Net*	*J48*	*Jrip*
1.	Correctly Classified Instances	499	489	673	658
2.	Incorrectly Classified Instances	202	212	28	43
3.	Accuracy	71.18%	69.76%	96.00 %	93.87%
4.	Kappa Statistics	0.52	0.55	0.93	0.90
5.	Mean Absolute Error	0.17	0.16	0.02	0.04

The above table shows that the accuracy of J48 result is high compared with Naive bayes, Jrip and Bayes Net. The Mean absolute Error Rates is less compared with other classifiers. The kappa statistics are compared on the basis of Tenfold cross-validation. Out of 701 instances J48 classify 673 instances correctly, its Accuracy is 96.00% which is high compared with other classifiers. The number of instances currently classified and incorrectly classified are given in the figure 4.

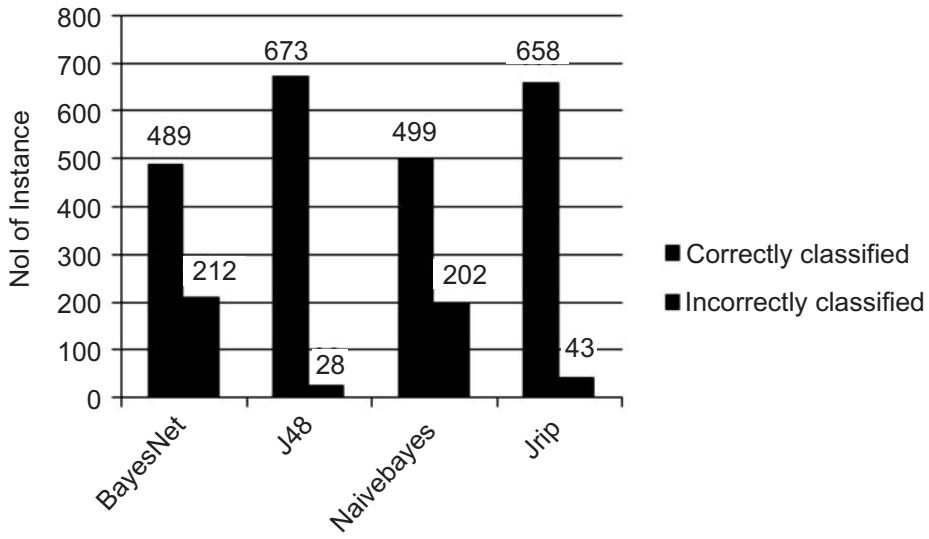


Figure 4: The number of instances classified for soil type

This shows that accuracy of J48 is more compared with Jrip. Naive bayes,Jrip and BayesNet. The Mean absolute Error Rates and the kappa statistics are compared on the basis of Tenfold cross-validation.

B. Soil type Clustering

In this paper the Soil type is clustered based on the available parameters. Simple K-Means, Hierarchical, and DBSCAN clusters have been generated and compared the results based on the number of clusters created and time taken to built the model. Figure5 shows the K-Mean WEKA Screen shot.

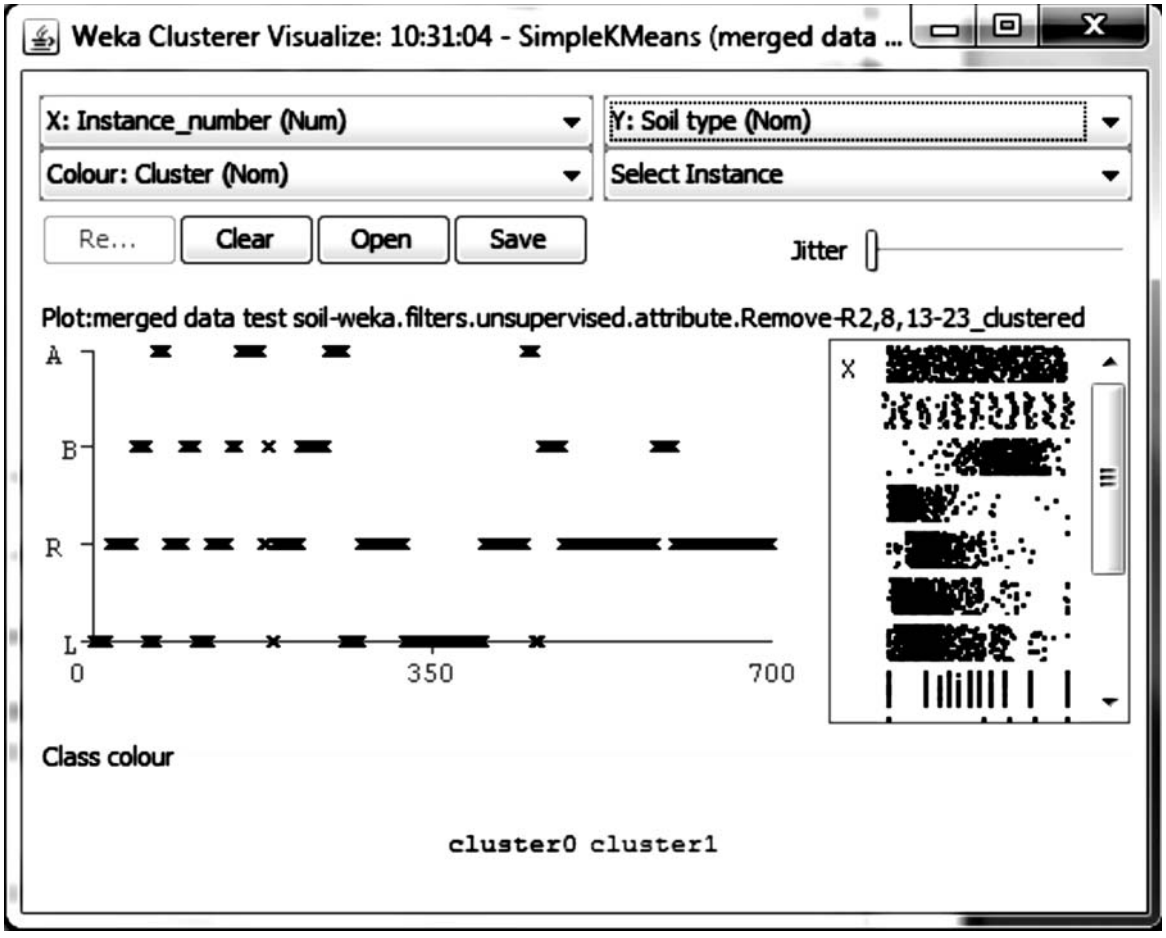


Figure 5: Cluster Visualize-Simple *k*-mean

Table 3 shows that the results compression between Simple K-Mean, Hierachical and DBSCAN based on number of clusters generated and time taken to built the model.

Table 3

Comparison of differnet clstering results

S. No	*Algotrithms*	*Number of Clusters*	*Time taken to build model*
1.	Simple K-Means	2	0.03 seconds
2.	Hierarchical	3	0.87 seconds
3.	DBSCAN	24	0.75 seconds

The result showed that Simple K-Means is the best clustering algorithm compare with other clstering algorithm for soil type grouping.

C. Hybrid Classification(Clustered J48)

The clustered J48 has 2 phases. In the first phase clusters for the dataset is generated before generating the decision tree. The soil dataset is taken and perform similarity function based on the number of instances in the data set. The Euclidean distance metric is used to find similarity between selected instances and compare the instance. This clustered phase is implemented in Weka datamining tool.

Phase 1 steps :

1. Load soil dataset for training
2. Based on the distance metric, divide the dataset into subsets.
3. Calculate the distance for all instances.
4. If a instance belongs to the same cluster and add it into new set then remove it from the un clustered data set.

In the second phase the J48 is generated from clustered data set. This is implemented in java API in weka.

Phase 2 steps :

1. Load Clustered Data set to classifier and Create a node M
2. If instances are all the same class CC then
   Return M as a leaf node with CC as the label.
3. If list of attribute is null then
   Return M as a leaf.
4. The highest information gain attribute is selected for the test-attribute
5. For each k_i of test –attribute generate a branch from M.
6. If S is null then
   Attach the leaf node with the common class in instance
   Else
   Attach the node return by tree generate function repeat the steps.

It increases the Accuracy of the classification. This is implemented by Java API in WEKA.

Java weka.classifier.tree.J48 -t "c:clusterdataset.arff"

The output of clustered J48 for soil type is shown in Figure 6.

The four soil types in Salem district are grouped into two using the k-mean clustering technique. The clustering results show that major area in the Salem District contains Red soil and Block soil. In this paper clustered J48 is generated from output of 2 clusters, 3 clusters and 4 clusters of K-mean cluster. Table 4 shows that the comparison results of J48 classifier to the Clustered J48 with number of clusters 2, 3 and 4.

Number of Leaves : 26

Size of the three : 38

Time taken to build model : 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly classified Instances	695	99.1441 %
Incorrectly Classified Instances	6	0.8559 %
Kappa statistic	0.9851	
Mean absolute error	0.0081	
Root mean square error	0.0761	
Relative absolute error	2.108 %	
Root relative squared error	17.376 %	
Total Number of Instances	701	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0. 975	0. 003	0. 975	0. 975	0. 975	0. 998	cluster0
0. 988	0	1	0. 988	0. 994	0. 993	cluster1
0. 997	0. 012	0. 99	0. 997	0. 993	0. 996	cluster2
Weighted Avg. 0. 991	0. 007	0. 991	0. 991	0. 991	0. 996	

=== Confusion Matrix ===

a	b	c	← classified as
77	0	2	a = cluster0
1	241	2	b = cluster1
1	0	377	c = cluster2

Figure 6: Clustered J48 Result for Soil Type

Table 4

Comparison J48 results with clustered j48 results for soil type

S. No	*Classifier*	*J48*	*Clusterd J48 with 2 clusters*	*Clusterd J48 with 3 clusters*	*Clusterd J48 with 4 clusters*
1.	Correctly Classified Instances	673	700	695	680
2.	Incorrectly Classified Instances	28	1	6	21
3.	Accuracy	96.00 %	99.85%	99.14 %	97.00%
4.	Kappa Statistics	0.52	0.99	0.99	0.96
5.	Mean Absolute Error	0.17	0.00	0.01	0.02

From the Table 4 the results shows that accuracy of Clustered J48 with 2cluster is high compared with other clustered J48 classifier and ordinary J48 classifier. The Mean absolute Error Rate is less compared with other classifiers. The kappa statistics are compared on the basis of Tenfold cross-validation. Out of 701 instances J48 classify 700 instances correctly, the Accuracy is 99.86% which is high compared with other classifiers. The result shows that the clustered J48 is more Accuracy than the J48 classifier.

6. CONCLUSION

Data Mining Techniques in agriculture will help the farmers to improve the crop productivity. Various decision tree algorithms are used for classification of soil type. This paper shows that J48 gives more accurate result. Further the output of clustered soil data set is given as an input to J48 classifier. The result shows that the accuracy of clustered J48 is more than J48 classifier. The result shows that major area in Salem District contains Red soil and Black soil. So classifiation of soil type will help the farmers to cultivate the crops based on the type of soil in Salem District.

7. ACKNOWLEDGMENT

The author gratefully acknowledges the support of Dr.N.Sriram, Programme Coordinator Krishi Vigyan Kendra (Farm Science Centre), Tamil Nadu Agricultural University, Santhiyur, Salem, who was of great help in the analysis and understanding of soil nutrients and soil types.

8. REFERENCES

1. Mucherino.A, Petraq Papajorgji and P.M.Pardalos, "A survey of data mining techniques applied to agriculture". *Published online 2009 © Springer-verlag*
2. R.Santhi et at (2014) ,GIS based Soil map for salem district of Tamilnadu. TechInical Folder, TNAU,Coimbatore.
3. Eia Report For Stp – 35.0 Mld At Mankuttai,Salem City Municipal Corporation
4. Dr. G.M. Nasira , N. Hemageetha ,"Perspective on Classification Techniques in Agriculture",*International Journal of Computing Technology and Information Security* Vol.1, No.2, pp. 40-46 , December, 2011. *ISSN: 2231-1998 © 2011. www.ijctis*
5. Dr. G.M. Nasira , N. Hemageetha, "Vegetable price prediction using data mining classification technique" *Preceedings of the International Conference on pattern Recognition, Informatics and Medical Engineering (PRIME 2012), PP. 99-102 ISBN No:978-1-4673-1038-3. © 2012 IEEE.*
6. *Dr.* G.M. Nasira , N. Hemageetha, " Forecasting Model for Vegetable Price Using Back Propagation Neural Network*" International Journal of Computational Intelligence and Informatics, Vol 2, no.2 sep 2012 PP 110-115.*
7. Dr.G.M. Nasira , N. Hemageetha, " Radial Basis Function Model for Vegetable Price Prediction " Preceedings of the International Conference on pattern Recognition, *Informatics and Mobile Engineering (PRIME 2013), PP. 424 – 428 ISBN No.978-1-4673-5843-9 © 2013 IEEE.*
8. N. Hemageetha, "A Survey on appliation of Datamining Techniques to Analyze the Soil for Agricultural purpose", Proceedings of the 10th INDIACom; INDIACom-2016; IEEE Conference ID: 37465 2016 3rd International Conference on "Computing for Sustainable Global Development", ISSN 0973-7529; ISBN 978-93-80544-20-IEEE.
9. Natesan et at(2007),. Technical Bulletinon "Soil test crop response based fertilizer prescription for different soils and crops in tamil nadu", AICRP-STCR TamilNadu Agricultural University, Coimbatore.
10. Soil Testing Kit Hand Book
11. Hemageetha,, Dr. G.M. Nasira , "Analysis of the Soil data Using Classification Techniques for Agricultural Purpose "International Journal of Computer Sciences and Engineering IJCSE E-ISSN: 2347-2693 Vol-4 Issue -6 June 2016
12. Vrushali Bhyar, "comparative Analysis of Classification Techniques on soil data to predict Fertility rate for Arangabad District", International Journal of Emerging Trends and Technology in computer science , Vol 3, Issue 3, March 2014