A Comparative Analysis of Classification Algorithms on Landsat Datasets

J.Saranya[#1], Dr.M.Mohamed Sathik[#2], Dr.S.Shajun Nisha[#3]
[#1]*M.phil Research Scholar, PG & Research Department of Computer Science, Sadakathullah Appa College, Tirunelveli, Tamilnadu, India*
[#2]*Principal, Sadakathullah Appa College, Tirunelveli, Tamilnadu, India*
[#3]*Assistant Professor & Head, PG & Research Department of Computer Science, Sadakathullah Appa College, Tirunelveli, Tamilnadu, India*

Abstract

Data mining involves the arithmetic process to find patterns from large data sets. Classification, one of the main domains of data mining, it involves known structure assumption to apply to a new dataset and predict its class. There are various classification algorithms being used to classify landsat data sets. A classification method involves probability, decision tree, neural network, nearest neighbour, boolean and fuzzy logic, kernel-based etc. In this paper, we apply two classification algorithms on landsat datasets. The datasets have been selected based on their attributes. Results have been conversation using some performance evaluation measures like precision, accuracy, F-measure, Kappa statistics, mean absolute error, relative absolute error, ROC Area etc. Comparative analysis has been done through the performance evaluation measures of accuracy, precision, and F-measure. We specify features and conditions of the classification algorithms for the landsat datasets.

Keywords - *Classification, Data mining, J48, Logistic Model Tree, Landsat..*

I. INTRODUCTION

Agriculture is the main source income for many people especially for rural areas. In this classification of crops is the better yield in agriculture. So a system needs to correctly classify the crop with the help of data mining techniques. As a result accurate information of crop types is important for public and private sectors. To classify field level crop types appropriate pre processed satellite data are required. Land sat data is available for both current and earlier periods. Large amount of landsat agricultural information is made available by various government organizations, for agricultural purposes. Data mining [7, 8] is a significant method to extract information from data. There are various domains of data mining like classification, clustering, anomaly detection, association rule mining, regression, pattern mining, summarization etc.

Landsat data is widely used for regional, local and continental scales. Common Land Unit(CLU) is used to aggregate field level information based on the landsat data. The aggregated data is used for crop classification. In addition advanced landsat products such as surface reflectance are available from the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) and the Landsat Surface Reflectance Code for Landsat 5, 7, and 8. This paper describes a better evaluation performance of crop classification system that is targeted to several countries and the region of the crop is corn, btcorn and soybeans.

A Literature Review

D Ramesh et al. [1], compares the results of multiple linear regression and Density based cluster technique. The data were compared in the specific region i.e. East Godavari district of Andhra Pradesh in India. Multiple Linear Regression is applied on existing data but the results obtained are analysed and examined using Density based clustering technique.

Raorane et al. [2], proposed that several changes in the weather can be analysed by Support Vector Machine (SVM is capable of classifying data samples in two disjoint clusters) and also K-means method is used to leading the pollution in atmosphere. Data mining techniques are used to monitor the wine fermentation.

Tanvi Sharma & Anand Sharma. [3], proposed to focuses on the application of different data mining classification techniques using different machine learning tools such as WEKA and Rapid miner over the public healthcare dataset for analysing the health care system. The accuracy of every data mining classification technique is used as a standard for performance measure. The best technique for particular data set is chosen based on highest level of accuracy.

B. Motivation Justifications

The motivation behind this paper is to explore data mining techniques, which are suitable for solving agricultural problems. A huge amount of landsat records are stored in databases. This database can be utilised for research purposes. Lots of research findings have been done on the agriculture field. Hence it justify that the classification of landsat dataset with J48 and LMT is suitable for this research.

C. Outline Of The Proposed Work

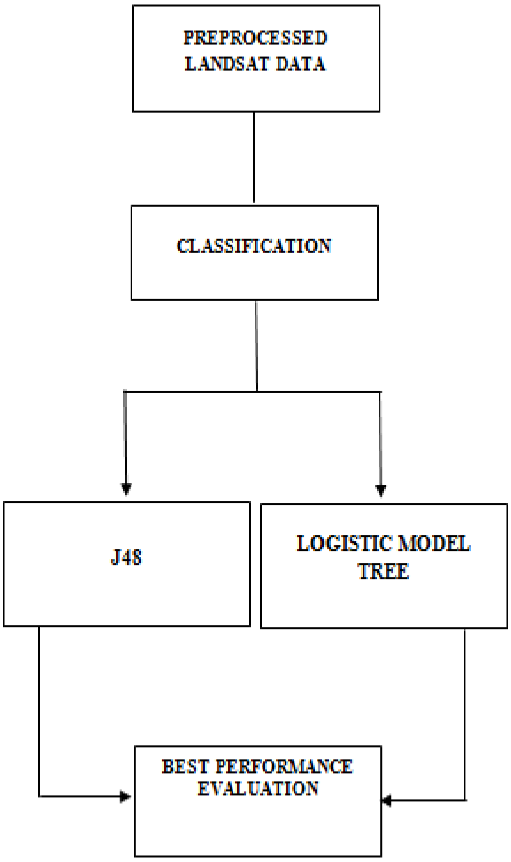

Fig 1 : Outline of the Proposed Work

D. Organization Of The Paper

The remaining paper is organized as follows: In Section II describes the classification methods. Section III Display the Experimental results, and in section IV conclusion is placed.

II. METHODOLODY

In this paper j48 and LMT classification methods are used to find the best accuracy of crop Classification.

A. Dataset

Dataset consists of different attributes that have complex relationships between dynamic variables.

B. Classification Algorithms

1) J48

A predictive machine-learning model which decides the target value of a new sample based on different attribute values of the available data is J48 decision tree. The different attributes denoted by the internal nodes of a decision tree, the branches between the nodes tell us the possible values that these attributes can have in the experimental samples, while the terminal nodes tell us the final value of the dependent variable

2) LMT

A classification model with an associated supervised training algorithm that combines logistic prediction and decision tree learning is logistic model tree (LMT)[4] . Logistic model trees uses a decision tree that has linear regression models at its authorization to provide a section wise linear regression model.

III EXPERIMENTAL RESULT

B. Performance Metrics

1) True Positive Rate

A true positive is an outcome where the model correctly predicts the positive tuples. It measures the percentage of actual positives that are correctly identified.

True Positive Rate $= \frac{TP}{TP+FN}$

2) False Positive Rate

The false positive rate is ratio between the numbers of negative events wrongly classified as positive.

False Positive Rate $= \frac{TP}{FP+TN}$

3) Precision

It is the proportion of instances that are truly of a class divided by the total instances classified as that class.

Precision $= \frac{TP}{TP+FP}$

4) Recall

It is the proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate)

Recall $= \frac{TP}{TP+FN}$

5) F-Measure

A combined measure for precision and recall is calculated as

F-measure$= \frac{2*precion*recall}{precision+recall}$

6) Matthews Correlation Coefficient

MCC has a range of values from -1 to 1 where -1 indicates completely wrong binary classifier while 1 indicates a completely a correct binary classifier. Using the MCC allows one to gauge how well their classification model/function is performing.

MCC = $\frac{TP*TN-FP*FN}{[(TP+FP)*(FN+TN)*(FP+TN)*(TP+FN)]\hat{}(\frac{1}{2})}$

7) Precision Recall Curve

The PRC is called as Precision recall characteristics curve. It is a comparison of two operating characteristics (PPV and sensitivity) as the criteria. A PRC curve is a graphical plot which explain the performance of binary classifiers as its discrimination threshold is varied.

8) Receiver Operating Characteristics

ROC is connection of two operating characteristics TPR and FPR. A receiver operating characteristic curve is a graphical action which analyses the performance of a classified as its partiality threshold is varied. It is a completion of plotting the true positive rate vs. false positive rate at varied threshold settings.

9) Mean Absolute Error

Mean Absolute Error is the common difference between the Original Values and the Predicted Values. It gives us the capacity how far the predictions were from the actual output. They don't give us any plan of the direction of the error. It under predicting the data or over predicting the data. Mathematically, it is written as

Mean Absolute Error = $\frac{1}{N}\sum_{j=1}^{N}|y_j - \hat{y_j}|$

10) Mean Squared Error

Mean Squared Error(MSE) is similar to Mean Absolute Error, the only difference being that MSE takes the fair and square difference between the original values and the predicted values. The advantage of MSE being that it is easier to compute the gradient, whereas Mean Absolute Error requires complicated linear programming tools to compute the gradient. As, we take square of the error, the effect of larger errors become more pronounced then smaller error, hence the model can now focus more on the larger errors.

Mean Squared Error = $\frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y_j})^2$

11) Kappa Statistic

Kappa Statistic analyze the accurateness of the system with the random system. The Measurement of Observer Agreement for Categorical Data, is an experimental probability of agreement and is a theoretical expected probability of agreement under a baseline constraint for appropriate set [6].

Kappa= (total accuracy- random accuracy)/(1- random accuracy)

C. Performance Evaluation

TABLE I
Detailed Accuracy for J48 Algorithm

	CORN	BT CORN	SOY BEAN	WEIGHTED AVERAGE
TP RATE	0.978	0.949	0.885	0.939
FP RATE	0.035	0.047	0.016	0.037
PRECISION	0.865	0.963	0.947	0.941
RECALL	0.978	0.949	0.885	0.939
F-MEASURE	0.918	0.956	0.915	0.939
MCC	0.901	0.901	0.889	0.898
ROC	0.992	0.990	0.988	0.990
PRC	0.949	0.990	0.964	0.976

TABLE II
Detailed Accuracy for LMT Algorithm

	CORN	BT CORN	SOY BEAN	WEIGHTED AVERAGE
TP RATE	0.913	0.986	0.984	0.971
FP RATE	0.000	0.019	0.057	0.017
PRECISION	1.000	0.986	0.923	0.973
RECALL	0.913	0.986	0.923	0.973
F-MEASURE	0.955	0.986	0.952	0.971
MCC	0.946	0.967	0.937	0.955
ROC	0.996	0.997	0.996	0.997
PRC	0.986	0.998	0.987	0.993

TABLE III
Performance Error of Classification Algorithms

PERFORMANCE OF ALGORITHM	J48	LMT
MEAN ABSOLUTE ERROR	0.0637	0.1141
ROOT MEAN SQUARED ERROR	0.1784	0.1821
RELATIVE ABSOLUTE ERROR	16.2859%	29.1954%
ROOT RELATIVE SQUARED ERROR	40.3892%	41.2281%

KAPPA STATISTICS	0.8962	0.9511

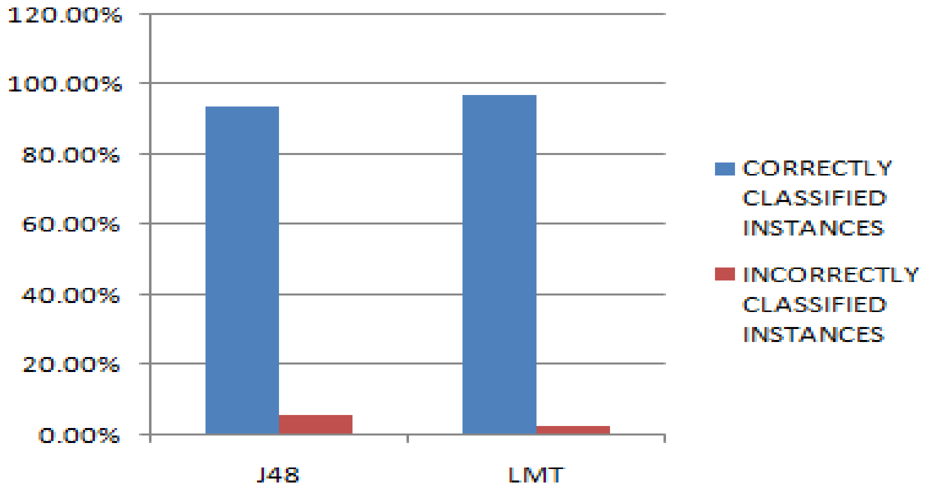

Fig 2 : Accuracy of Classification Algorithms

IV. CONCLUSION

This research focuses on finding the best algorithm between J48, LMT to enhance the classification of landsat dataset. The algorithms are used to classify the crops in certain countries with their production. From the experimental results the J48 classifiers have minimum error rate than the LMT classifiers. By analysing the experimentation results of the landsat dataset, it is concluded that LMT algorithm has produced the best classification performance than the J48 algorithm and has slightly difference performance.

ACKNOWLEDGEMENT

I would like to thank Dr.M.Mohamed Sathik for providing his valuable inputs to improve this paper. I convey my fruitful thanks to Dr.S.Shajun Nisha for their valuable suggestions and support for necessary resources to accomplish this research.

REFERENCES

[1] D Ramesh, and B Vishnu Vardhan, “Analysis Of Crop Yield Prediction Using Data Mining Techniques”, IJRET, 2015.

[2] Raorane A.A, “Data Mining: An effective tool for yield estimation in the agricultural sector “, 2012.

[3] Tanvi Sharma, Anand Sharma & Vibhakar Mansotra “Performance Analysis of Data Mining Classification Techniques on Public Health Care Data, ” International Journal of Innovative Research in Computer and Communication Engineering , 2016

[4] N. Landwehr, M. Hall, and E. Frank, “Logistic model trees, Mach. Learn.,” Springer, 2005.

[5] Anshul Goyal and RajniMehta,” Performance Comparison of Naive Bayes and J48 Classification Algorithms,” *IJAER* , 2012.

[6] Muhammad Alghobiri, “A Comparative Analysis of Classification Algorithmson Diverse Datasets, ” *Engineering, Technology & Applied Science Research, 2018.*

[7] N.M. Ramos, J.M. Delgado, R.M.Almeida, M.L. Simoes and S.Manuel, “Application of Data Mining Techniques in the Analysis of Indoor Hygrothermal Conditions”, Springer, 2015.

[8] B.Bakhshinategh, O.R. Zaiane, S. Elatia, D. Ippercial. “Educational Data Mining Applicatins and tasks: A Survey of the last 10 years”, Education and Information Technologies, 2018.

[9] Kathija and Dr. S.Shajun Nisha., “Breast Cancer Data Classification Using SVM and Naïve Bayes Techniques., ” International Journal of Innovative Research in Computer and Communication Engineering, 2016.

[10] A.Ameer Rashed Khan, Dr.S.Shajun Nisha and Dr.M.Mohamed Sathik., “Clustering Techniques For Mushroom Dataset.,” International Research Journal of Engineering and Technology (IRJET), 2018.