Performance of Classification Algorithms for Prediction of Crop Cultivation by Reducing the Dimensionality of a Data Set

I. Rajeshwari, K. Shyamala

Abstract*: Agriculture plays an important role in Indian economy. Maximizing the crop productivity in one of the main tasks that farmers are facing in their day to day life. They are also lacking in the basic knowledge of nutrient content of soil and selection of crops those best suits their soil thereby improving the crop productivity. In this work, the dataset has been taken from the soil test centres of Dindigul district, Tamilnadu. The parameters are the 12 various nutrients present in the soil samples collected from the different regions of Dindigul district. Using PCA, the dataset has been reduced to 8 parameters. Data Mining classification techniques like decision tree, KNN, Kernal SVM, Linear SVM, Logistic regression, Naive Bayes and Random forest are deployed on the original and dimensionality reduced datasets to predict the crops to be cultivated based on the availability of soil nutrient in the datasets. The performance of the algorithms are analysed based on certain metrics like Accuracy score, Cohen's Kappa, Precision, Recall And F-Measures, Hamming Loss, Explained Variance Score, Mean Absolute Error, Mean Squared Error and Mean Squared Logarithmic Error. The Confusion matrix and Classification report are used for analysis. The Decision Tree is found to be the best algorithm for the soil datasets and dimensionality reduction does not affect the prediction.

***Keywords* : Classification, Data Mining, Decision tree, Soil nutrient.**

I. INTRODUCTION

Agriculture plays a vital role in Indian Economy. Farmers are lacking in basic knowledge about their soils and crops to be cultivated to maximize the production. Crop cultivation mainly depends on the soil, weather and water. Soil is the main source of nutrients for crops. Hence maintenance of soil health is important for sustaining its productivity over long run[1]. Soil testing gives a brief idea about soil structure and mineral compositions rations and must be done frequently to analyze the vitality of soils during the seasonal changes. Indian Government has established soil test centres at every district head quarters [2]. The soil test provides information about the various chemical properties such as EC, pH, along with micro nutrients N, P, K and macro nutrients Zn, B, Cu, Fe [3]. Proper soil nutrition will prevent nutrient-related plant stress and crop losses through pests, diseases, and poor post-harvest quality. From the soil, the roots of a plant receive nutrients which help it grow. In this paper, the soil nutrient data set is subject to dimensionality reduction and data mining classification techniques are applied over both the original and dimensionality reduced data sets to predict the crops to be cultivated. The main aim of the paper is to determine the classification algorithm suitability for the soil data set, thereby assist farmers in crop selection for cultivation and does dimensionality reduction affects the performance of the classification algorithms.

The entire dataset tuples are split into two sets., viz., a training set and test set. The training sets are randomly sampled from the dataset. The remaining tuples form the test set. The classifier is built in the learning step by the classification model by analysing the training set. Prediction of the class labels are done in the classification step. The predictive accuracy of a classifier is estimated from the test set. The percentage of test sets that are correctly classified by the classifier is the accuracy of a classifier. The best way to achieve higher accuracy is to test out different algorithms. By cross-validation, the best one can be selected. Classification algorithms like decision tree, KNN, Kernal SVM, Linear SVM, Logistic regression, Naive Bayes and Random forest are deployed on the soil datasets and their performance are analysed based on certain metrics like Accuracy score, Cohen's Kappa, Precision, Recall And F-Measures, Hamming Loss, Explained Variance Score, Mean Absolute Error, Mean Squared Error and Mean Squared Logarithmic Error. The Confusion matrix and Classification report are used for analysis. The main problem is comparative study of these classification algorithms using soil dataset before and after dimensionality reduction.

Data analysis process and the classification problems become tedious with increase in the dataset dimensionality, that may result in sparse data which makes both the supervised and unsupervised learning task a bigger problem, which is called as the 'curse of dimensionality' [4]. More the number of features lower the classification accuracy. More computation cost and more memory usage happen with many classification algorithm due to high dimensional data. Also, according to Tan et al., a simplified easily understandable model and visualization techniques can be achieved when the attribute space is reduced [5].

Revised Manuscript Received on November 15, 2019

Mrs. I.Rajeshwari,, Associate Professor of Computer Science, Queen Mary's College, Chenna, Tamilnadu, India. Email: rajeshwari_i@yahoo.com

Dr. K. Shyamala, Associate Professor of Computer Science, Dr. Ambedkar Govt. Arts College, Vyasarpadi, Chennai, Tamilnadu, India. Email: shyamalakannan2000@gmail.com

Principle component analysis (PCA) is used to reduce the dimension by looking at the linear dependencies between the dataset attributes and strongest patterns in the data. It reduces the attribute space by having less needed information about the original dataset.

The organization of this paper is as follows. In Section 2, the literature review conducted is described and the study to be done is introduced. Section 3 describes about PCA, the dataset collection, its preprocessing activities and the classification techniques used. Section 4 discusses the results of the classification algorithms for predicting the crop to be cultivated. Finally, section 5 includes conclusion and future work.

II. LITERATURE REVIEW

Howley et al. [6] have applied PCA on the original high dimensional spectral dataset and applied the classification techniques like C4.5, RIPPER, linear SVM, RBF SVM, k-NN and linear regression. They have come to the conclusion that combining PCA and classification techniques improves the classification accuracy for high dimensional dataset. They have used NIPALS method [7] and computed the needed n Principal Components(PCs). When more PC is included, C4.5 and RIPPER have resulted in the same error, whereas linear SVM, RBF SVM, k-NN and linear regression have resulted in little bit smaller error.

Mohini et al. [8] has stressed that the attributes those result in increased effectiveness without any compromise in performance can alone be retained through dimension reduction which is to be carried out prior to classification.

Popelinsky et. al. [9] has applied PCA and used C5.0, instance-based learner and Naïve Bayes on the dimensionality reduced dataset. Two runs were conducted, viz., the first n PCs added with the original attribute; and, the PCs replaced them. With the first run, all algorithms resulted in improved classification accuracy and with the second run, Naïve Bayes algorithm resulted in increased accuracy.

Gansterer et al. [10] have investigated the benefits of dimensionality reduction in the context of latent semantic indexing for e-mail spam detection by deploying Vector space model(VSM) and latent semantic indexing (LSI). One of the problems they have dealt with is reducing the feature set and analyzing the classification performance. They have concluded that in the spam filtering, large amount of problem reduction is possible without any degradation in the performance of the classifier.

Priya et al. [11] had applied dimensionality reduction on the Pima Indian Type II Diabetes dataset and studied the performance of Naïve Bayes on it based on classification accuracy, False Negatives, False Positives, True Negatives and True Positives. It was concluded that some feature subsets have negative impact while some have positive impact on the performance of the classifier; and by finding the optimal feature subsets, the performance of the classifier can be enhanced.

From the review, it is found that most of the authors have anlaysed the performance of classification algorithm based on accuracy on the dimensionally reduced dataset and the original dataset. In this study, more techniques are applied on a single dataset and analysed on more metrics. Classification algorithms like decision tree, KNN, Kernal SVM, Linear SVM, Logistic regression, Naive Bayes and Random forest are deployed on the soil datasets and their performance are analysed based on certain metrics like Accuracy score, Cohen's Kappa, Precision, Recall And F-Measures, Hamming Loss, Explained Variance Score, Mean Absolute Error, Mean Squared Error and Mean Squared Logarithmic Error using PYHTON. The Confusion matrix and Classification report are used for analysis.

III. RESEARCH METHODOLOGY

Data collection, cleaning and coding:

For this work, data was collected from the soil test centers in Dindigul district. Standard methods were adopted in the soil testing centers for the identification of the soil nutrients, viz., EC, pH, N, P, K, Fe, Zn, Cu, B, OC, S and Mn. Samples from various villages covering almost the entire district were collected. Nearly 44000 samples were collected. The collected data was preprocessed by removing unwanted data, noisy data and blank data. Extreme values in each attributes are treated as noisy data. The data was cleaned by removing all the abbreviations and converted to numerical form. 35700 samples from the processed data were taken. The data was initially coded in EXCEL and finally converted to comma delimited .csv.

Principal Component Analysis:

All 12 parameters are vital for plant growth and yield. Dimensionality reduction can be applied for speeding up the classification process. One of the tools used for identifying the principal component and reducing the dimensionality is PCA. PYTHON was used to analyze PCA for the soil data. Component identification is done with the cumulative percentage of variance. Components that has up to 80% of variance aare considered and hence the first eight components: pH, N,P,K, S, Zn, Fe and Mn can be taken as principal soil components. With the existing features, in both the original and the dimensionality reduced datasets, a target class was added which gives the crops that can be cultivated. The class label is given in the number form.

The various classification algorithms were applied on both the resultant datasets with 0.1% of the data taken as test data.

Decision tree:

Decision tree is a tree structure classifier where, each internal node represents a test on an attribute, each branch represents an outcome of a test and each leaf node represents a class [12]. Based on some chosen properties, this algorithm categorizes the population. This algorithm based on entropy was applied on the dataset and the decision tree was obtained.

SVM:

The unique feature of SVM is, it chooses the decision boundary in such a way that the distance from the nearest data points of all the classes are maximum.

It finds the most optimal decision boundary, ie., it finds the maximum margin from the nearest points of all the classes [13]. Support vectors are the nearest points from the decision boundary that maximize the distance between the decision boundary and the points. Linear SVM and C-Support Vector (Sigmoid Kernal) classifications are implemented here.

kNN:

The kNN is a non-parametric approach. Here, in a dataset, a k samples nearer to unknown samples are grouped together (e.g., based on distance functions), from which the average of the response variables are calculated and the class labels of the unknown samples are found (i.e., the class attributes of the k nearest neighbor) [12]. In this study k is taken as 5.

Logistic regression:

Linear regression is a statistical model used to find the linear relationship between two or more variables—a dependent variable and independent variable(s), ie., when one (or more) independent variables increases (or decreases), the dependent variable also increases (or decreases) [14]. For classification, Logistic regression is a linear model, but not for regression. In this study, logistic function is used to model the probabilities describing the possible outcomes of a single trial.

Naive Bayes:

The naive Bayesian classifier is a supervised statistical classifier which can deal with any number of features or classes [15]. Here, Bayes theorem is used to calculate the most likely class label of the new instance.

Random forest:

A random forest is a collection of unpruned decision trees which fits lot of decision tree classifiers. It is often used when the training data sets and input variables are large. It improves the predictive accuracy by averaging and controls over-fitting[16]. Here, each tree is built by replacing the sample drawn from the training set. When the tree is constructed, to split a node, the best split among a random subset of the features is chosen. Here the random forest algorithm is applied on the dataset with 5 as the number of trees.

IV. RESULT AND DISCUSSIONS

The algorithms were executed on both datasets in .CSV (Comma Separated format) files using PYTHON and the various metrics were analysed. Table I. shows accuracy on training set, accuracy on test set, mean accuracy and standard deviation resulted by the various algorithms applied on the datasets.As can be seen from the table I, the Decision Tree algorithm has cent percent accuracy both on the training and test set. The algorithms following in the line are Random forest, kNN and NaiveBayes. The accuracy score function shows the accuracy of correct prediction, which is either a fraction (default) or the number (normalize=False). If the entire set is predicted correctly, then the accuracy is 1.0; otherwise it is less than 1. The accuracy score function shows that the Decision Tree algorithm has the correct predictions, followed by Random forest, kNN and NaiveBayes. Decision Tree has the highest mean accuracy followed by Random forest, kNN and NaiveBayes. Standard deviation accuracy should be a minimum one. The table I. shows that the Decision tree algorithm results in the minimum standard deviation accuracy, i.e., its results is more accurate prediction followed by linear SVM, kNN and Logistic regression.

When comparing the performance based on accuracies between the original and dimensionality reduced datasets, there is no significant change with the decision tree and random forest algorithm. For kNN, NaiveBayes and Kernal SVM, the performance improves when the dimensionality is reduced, whereas for linear SVM and logistic regression the performance decreases when dimensionality is reduced.

Table II. shows the average precision, average recall, average f1-score, explained variance score and Cohen Kappa Score for the various algorithms applied on the datasets. Precision is used to check the wrong labelling of negative sample and recall is used to find all the positive samples [17]. The F1-measure (f1-score) gives the weighted harmonic mean of the precision and recall. Cohen Kappa Score expresses the level of agreement between two annotators on a classification problem. From the table 4.2, it is clear that Decision Tree has the highest score is all metrics, followed by random forest, kNN and Naïve Bayes. Cohen Kappa score indicates that decision tree algorithm has the best agreement. When comparing the performance between the original and dimensionality reduced datasets, there is no change with the decision tree and no significant change with the random forest algorithm. For kNN, NaiveBayes and Kernal SVM, the performance improves when the dimensionality is reduced, whereas for linear SVM and logistic regression the performance decreases when dimensionality is reduced.

The table III shows the Mean Absolute Error, Mean Squared Error, Mean Squared Logarithmic Error and Hamming loss of the algorithms applied on the dataset. It is clear that decision tree has null error rate, followed by random forest and kNN. The decision tree algorithm results in no hamming loss followed by random forest.When comparing the performance between the original and dimensionality reduced datasets, there is no change with the decision tree. For kNN, NaiveBayes and Kernal SVM, the performance improves when the dimensionality is reduced, whereas for linear SVM and logistic regression the performance degrades when dimensionality is reduced and no slight reduction in performance with the random forest algorithm.

Table – I: Performance of the algorithms on both the data sets based on various accuracies

ALGORITHM	ACCURACY ON TRAINING SET		ACCURACY ON TEST SET		ACCURACIES (MEAN)		ACCURACIES (STD DEV.)	
	DRDS	ODS	DRDS	ODS	DRDS	ODS	DRDS	ODS

DT	1	1	1	1	0.99993	0.99993	0.00013	0.00013
KerSVM	0.735	0.648	0.713	0.64	0.72616	0.67769	0.00949	0.02437
Knn	0.992	0.892	0.907	0.867	0.9014	0.86305	0.0045	0.00454
LinSVM	0.532	0.565	0.517	0.551	0.53309	0.56637	0.00326	0.00369
LogREG	0.519	0.585	0.504	0.575	0.51928	0.58387	0.00457	0.00578
NaiveBayes	0.907	0.878	0.906	0.874	0.9044	0.87732	0.01112	0.01286
RF	1	1	0.985	0.989	0.98626	0.98873	0.00207	0.00236

*ODS – ORIGINAL DATASET DRDS – DIMENSIONALITY REDUCED DATA SET

Table –II: Avg. precision, avg. recall, avg. f1-score, explained variance score and Cohen Kappa score for the algorithms

ALGORITHM	AVG PRE		AVG REC		AVG F1 SCORE		EXPLAINED VARIANCE SCORE		COHEN KAPPA SCORE	
	DRDS	**ODS**	**DRDS**	**ODS**	**DRDS**	**ODS**	**DRDS**	**ODS**	**DRDS**	**ODS**
DT	1	1	1	1	1	1	1	1	1	1
KerSVM	0.73	0.67	0.71	0.64	0.72	0.65	0.89223	0.69432	0.65687	0.57115
kNN	0.91	0.87	0.91	0.87	0.9	0.86	0.96947	0.95091	0.88795	0.83891
LinSVM	0.44	0.53	0.52	0.55	0.41	0.46	0.8129	0.8197	0.38008	0.43128
LogREG	0.43	0.59	0.5	0.57	0.41	0.51	0.7914	0.82259	0.36137	0.46281
NaiveBayes	0.91	0.88	0.91	0.87	0.9	0.86	0.84178	0.60428	0.88722	0.84857
RF	0.98	0.99	0.98	0.99	0.98	0.99	0.99522	0.99574	0.98168	0.98727

Table – III: MAE, MSE, MSLE and Hamming Loss for the algorithms

ALGORITHM	MEAN ABSOLUTE ERROR		MEAN SQUARED ERROR		MEAN SQUARED LOG ERROR		HAMMING LOSS	
	DRDS	**ODS**	**DRDS**	**ODS**	**DRDS**	**ODS**	**DRDS**	**ODS**
DT	0	0	0	0	0	0	0	0
KerSVM	0.39376	0.68839	0.69959	2.02035	0.02673	0.06627	0.28734	0.3596
kNN	0.11296	0.1714	0.19847	0.32001	0.00715	0.01277	0.09317	0.1335
LinSVM	0.72423	0.67214	1.34522	1.26251	0.04067	0.03694	0.48282	0.44866
LogREG	0.74832	0.63125	1.39283	1.17868	0.04603	0.0368	0.49589	0.42813
NaiveBayes	0.22031	0.42046	1.03921	2.65534	0.02481	0.06906	0.09373	0.12565
RF	0.01979	0.01531	0.03099	0.02763	0.00131	0.00084	0.01531	0.01064

V. CONCLUSION AND FUTURE WORK

This study has conducted a comparison between the performance of classification algorithms like decision tree, KNN, Kernal SVM, Linear SVM, Logistic regression, Naive Bayes and Random forest on the original soil dataset and dimensionality reduced dataset using PYTHON. The algorithms are compared on various metrics. It is concluded that Decision tree is found suitable for this dataset with little compromise in the execution time for predicting the crops to be cultivated. The other algorithms that top in the rank are random forest and kNN. When execution time is given importance, Naïve Bayes can be used. There is no significant difference in the performance of the algorithms when dimensionality reduction is applied. The work can be extended further by performing dimensionality reduction using other methods.

REFERENCES

1. E.Manjula and S.Djodiltachoumy, "Data Mining Technique To Analyze Soil Nutrients Based On Hybrid Classification", DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4794 Volume 8, No. 8, September-October 2017 International Journal of Advanced Research in Computer Science,ISSN No. 0976-5697.
2. B.V.RamaKrishna and Dr.B.Satyanarayana, "Agriculture Soil Test Report Data Mining for Cultivation Advisory", International Journal of Computer Application (2250-1797), March-April 2016, Vol. 6 – No.2
3. Ashish Kumar Dogra and Tanuj Wala, "A Comparative Study of Selected Classification Algorithms of Data Mining", IJCSMC, Vol. 4, Issue. 6, June 2015, pg.220 – 229, ISSN 2320–088X
4. Warren Buckler Powell. Approximate Dynamic Programming: Solving the Curses of Dimensionality. Wiley-Interscience, 1st edition, 2007.
5. Tan H. Witten and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2 edition, 2005.
6. Tom Howley, Michael G. Madden, Marie-Louise O'Connell, and Alan G. Ryder. The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. Knowledge Based Systems, 19(5):363–370, 2006.
7. Paul Geladi and Bruce R. Kowalski. Partial least-squares regression: A tutorial. Analytica Chimica Acta, 185:1–17, 1986.
8. Mohini D Patil and Dr. Shirish S. Sane, "Effective Classification after Dimension Reduction: A Comparative Study", International Journal of Scientific and Research Publications, Volume 4, Issue 7, July 2014 1 ISSN 2250-3153
9. Lubomir Popelinsky." Combining the principal components method with different learning algorithms". In Proceedings of the ECML/PKDD2001 IDDM Workshop, 2001.
10. Wilfried N. Gansterer, Andreas G. K. Janecek, and Robert Neumayer. Spam filtering based on latent semantic indexing. In Micheal W. Berry and Malu Castellanos, editors, Survey of Text Mining II - Clustering, Classification, and Retrieval, volume 2, pages 165–185. Springer, 2008.

11. Priya Mohan and Ilango Paramasivam, A Study on Impact of Dimensionality Reduction on Naïve Bayes Classifier, Indian Journal of Science and Technology, Vol 10(20), DOI: 10.17485/ijst/2017/v10i20/101599, May 2017
12. "Comparison of data mining classification algorithms for breast cancer prediction", Chintan Shah, Anjali G. Jivani, https://www.researchgate.net/publication/269270867, Conference Paper • July 2013 DOI: 10.1109/ICCCNT.2013.6726477.
13. "Comparison of Different Classification Algorithms for Fault Detection and Fault Isolation in Complex Systems", Marcel Jung, Octavian Niculita, Zakwan Skaf, ScienceDirect, 6th International Conference on Through-life Engineering Services, TESConf 2017, 7-8, November 2017, Bremen, Germany, www.elsevier.com.
14. "Performance Analysis of Classification Algorithms in Predicting Diabetes", Meraj Nabi, Abdul Wahid, and Pradeep Kumar , International Journal of Advanced Research in Computer Science, Volume 8, No. 3, March – April 2017, ISSN No. 0976-5697
15. "Classification Algorithms for Data Mining: A Survey", Raj Kumar, Dr. Rajesh Verma, International Journal of Innovations in Engineering and Technology (IJIET), Vol. 1 Issue 2 August 2012 ISSN: 2319 – 1058
16. "Predicting yield of the crop using machine learning Algorithm", P.Priya, .Muthaiah & M.Balamurugan, International Journal Of Engineering Sciences & Research Technology (IJESRT), April 2018, ISSN: 2277-9655.
17. "A Comparison Study between Data Mining Algorithms over Classification Techniques in Squid Dataset", Fartash Haghani, Payam Hassany Shariat Panahy, Nasim Khanahmadliravi, Seyed Ahmad. Mousavi, International Journal of Artificial Intelligence, ISSN 0974-0635; Int. J. Artif. Intell. Atumun (October) 2012 , Volume 9, Number A12, Copyright © 2012 by IJAI (CESER Publications).

AUTHORS PROFILE

Mrs. I.Rajeshwari, has completed her Masters and M.Phil in Computer science. She is now pursing her Ph.D. in the University of Madras. Her areas of interest are Data Mining, Computer Networks and algorithms.

Dr. K. Shyamala is working as an Associate Professor in the PG and Research Department of Computer Science, Dr. Ambedkar Govt. Arts College, Vyasarpadi, Chennai,
Tamilnadu, India. She has her Masters degree, M.Phil and Ph.D. in Computer Science. She has 29 years of teaching and research experience. Six candidates have completed Ph.D. under her guidance. She has authored numerous books, published 62 research articles and conducted several conferences. She has also chaired sessions in International conferences. She has served as program committee member and chairman for Board of Studies in various colleges and universities. Her area of specialization includes Data Mining, WBAN, Agent Based Computing and Advanced Computer Networks.