ISSN 2250-3153

Data mining plays a key role in soil data analysis of Warangal region

Velide Phani kumar[*] and Lakshmi Velide[**]

[*] TELE 9 Technologies Limited, Hyderabad, Andhra Pradesh, India
[**] Department of Biotechnology, Gokaraju Rangaraju Institute of Engineering and technology, Kukatpally, Hyderabad, Andhra Pradesh, India.

Abstract- The advancement in computers provided large amount of data.The task is to analyse the input data and obtain the required data which can be done by various data mining techniques. Present work focusses on analysis of soil profile data from various locations of Warangal Region. Naive Bayes, J48(C4.5) and JRip Algorithms were used to analyse the data JRip reported to be simple, efficient classifier of soil data. The selected soil attributes were Nitrogen, Phosphorus, Calcium, Magnesium, Sulphur, Iron, Zinc, Potassium, PH and Humus. The attributes were predicted by linear regression, least median square and simple regression. Even though all regressions provided almost equal results least median Square depicts better results. The attributes nitrogen, phosphorus and sulphur were determined in less time by linear regression and showed accurate results. As large data of a long period for different locations were analysed by data mining, it reduces the time consuming process in soil analysis by traditional methods. Thus the results can be used by the researchers to suggest suitable crop to a particular region, season and also can recommend required fertiliser based on deficit elements.

Index Terms- Soil data, Warangal, Attributes, Naïve Bayes, J48(C4.5), JRip, Datamining.

I. INTRODUCTION

Data mining is the process of analysing data from different perspectives and summarizing it into useful information - information that can be used for industrial, commercial and scientific purposes. As such the process of data mining involves sorting through large amounts of data and discovering patterns in the data (Witten and Eibe, 2005).Agricultural and biological research studies have used various techniques of data analysis including, natural trees, statistical machine learning and other analysis methods (Cunningham and Holmes, 1999).The recent advances in data mining technologies are successfully applied to the management of natural resources also. Armstrong *et al*., (2007) have reported that data mining empowers farmers in selection of site specific crop varieties by studying soil fertility. The principle objective of classification of soil is for predicting the engineering properties and behaviour of soil finally dictating the choices for use. Laboratory and various statistical techniques are time consuming and highly expensive, efficient techniques can be developed for solving complex soil data sets using data mining to improve the effectiveness and accuracy of the classification of large soil data sets [Kumar and Kannathasan,2011). Verheyen *et al*., (2001) have studied the soil characteristics by k-means approach and GPS based data mining techniques. Tripati *et al.,* (2006) have reported the weather changes by using SVMs. K-means method and K-nearest neighbour method are useful to study pollution in atmosphere, weather variables and precipitation levels(Jorquera *et al.,* 2001; Rajagopalan and Lall, 1999). Soil tests are generally performed to study the nutrient content, contamination and deficiencies in soil that to be remedied (Wikipedia,2013). Soil testing laboratories either government or private sector provides various protocols for soil analysis and literature regarding the soil characteristics. Based on the data describes the soil composition and also recommends suitable fertilisers based on the data. It also helps farmers to supply the suitable fertiliser for suitable crop for a particular season.

The present work has taken up to analyse the soil data of Warangal district by various data mining techniques and the outcome obtained has been used by researchers and also farmers for selection of suitable crop and fertiliser to that soil.

II. METHODOLOGY

Data collection

The present research has been done by collecting soil data from three agriculture areas of Warangal district, Andhra Pradesh, India. The research has utilised commonly occurring available seven soil type data for correlation and comparison. The data has provided by Agriculture department, ANGRAU, Warangal. The data contains ten attributes and 1500 instances. Soil characteristic classification is important as it gives detail study of soil qualitatively and quantitatively. Traditional classification involves tables, flowcharts etc., as it is manual approach it takes lot of time. Hence quick, reliable, automated rule based proposed method is selected to study and classify soils based on fertility. Rules like facts, concepts and theories were collected from soil testing lab. Soil testing labs classify soils as very high fertile, high fertile, median fertile, low fertile and very low fertile. The collected data had analysed by selected automated method and observed the fertility level. The results used by the experts to suggest suitable crops and also fertiliser to compensate the deficit elements of that soil. The following algorithms were used to study the collected soil data.

Naïve Bayes

Naive Bayes classifier assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. It can be trained very efficiently in a supervised learning. In many practical

applications, parameter estimation for naive Bayes models uses the method of maximum likelihood. An advantage of naive Bayes is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) (Wikipedia, 2013).

J48 (C4.5)

It is an open source algorithm in Weka data mining tool. A decision tree can be generated from the input data by C4.5 proramme. It is an algorithm used to generate a decision tree and is an extension of Quinlan's earlier ID3 Algorithm. The decision trees generated by this can be used for classification and so referred to as statistical classifier (Wikipedia, 2013).

JRip

This class implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William Cohen (1995) as an optimized version of IREP. It is based in association rules with reduced error pruning (REP), a very common and effective technique found in decision tree algorithms (Wikipedia, 2013).

Attribute prediction

The attributes were predicted by Linear regression, Least Median square and Simple regression. Linear regression is the first type of regression analysis, considers numerical prediction. If data has any nonlinearity it cannot be applied and in such cases median square techniques are used. As the median regression techniques have high computational costs they are not used for simple practical problems (Witten and Eibe, 2005).

III. STATISTICAL ANALYSIS

Each analysis was replicated thrice by the above said three classifiers which were used to compare and evaluate the soil data based on false positive rate, true positive rate, accuracy and time.

IV. RESULTS AND DISCUSSION

Results and discussion: Table I gives details about various attributes selected for soil data analysis. Results show that JRip model is the best classifier for soil sample data analysis in comparison with other models (Table II).Table III shows that the relative absolute error is almost equal in both linear and least median regression analysis. There is no much variation in the correlation coefficient obtained in both the predicted algorithms but the time taken to build linear regression model is 61.7% less than the least median regression. This shows that the computational cost used by linear regression is much lower than the least median square. Median regression techniques have high computational costs (Witten and Eibe 2005). Though least median square model produces better results the accuracy of linear regression and least median regression are almost equal. The attributes Nitrogen, Phosphorus and sulphur were determined by linear regression technique in lesser time and gave accurate results. These predictions helps to find the attributes without carrying traditional chemical tests and thus time could be saved and reliable information obtained (Table IV).

V. CONCLUSION

Thus in conclusion the given soil data analysis was done by different algorithms. It can also be concluded that better results were obtained by least median square than linear regression analysis and classification of soil by JRip proved to be simple classifier and gave best result by constructing decision tree.

VI. ACKNOWLEDGEMENT

The authors would like to thank Agriculture department, Warangal for providing the soil data.

REFERENCES

[1] "C4.5(J48)", Wikipedia, March 2013.
[2] "JRip",Wikipedia,March 2013
[3] "Naïve Bayes", Wikipedia, March 2013.
[4] "Soil test", Wikipedia,March 2013.
[5] A.Kumar and N.Kannathasan, "A survey on data mining and pattern recognition techniques for soil data mining," *Int.J.Comp.Sci.Issues*, Vol.8,Issue 3,2011, pp. 422-428.
[6] B.Rajagopalan and U.Lall, "A K -nearest neighbour simulator for daily precipitation and other weather variables",*Water resource research*, Vol.35, Issue 10, 1999, 3089-3101.
[7] H.Jorquera,R.Perez,A.Cipriano and G.Acuna, " Short term forecasting of air pollution episodes", In Environmental modelling,P.Zannetti (ED),WIT Press, UK,2001.
[8] I.Witten and F.Eibe,"Data mining practical machine learning tools and techniques",2nd ed,Sanfrancisco: Morgan Kaufmann series in data management systems,2005.
[9] K.Verheyen, D.Adriaens, M.Hermy and S.Deckers "High resolution continuous soil classification using morphological soil profile descriptions",*Geoderma*,Vol.101,Issues 3-4, 2001, pp.31-48.
[10] L.Armstrong, D.Diepeveen and R.Maddern, "The application of data mining techniques to characterise agricultural soil profiles" Paper presented at sixth Australian data mining conference,Conferences in research and practice in information, Australia, Vol.70,2007, pp.81-96.
[11] S.J.Cunningham and G.Holmes, "Developing innovative applications in agriculture using data mining",In the proceedings of the south east Asia,Regional computer confederation conference, Newzealand,1999.
[12] S.Tripati, V.V.Srinivas and R.S.Nagundiah, "Downscaling of precipitation for climate change scenarios:Asupport vector machine approach",*J.Hydrology*,Vol.330,Issues 3-4, 2006, pp. 621-640.
[13] W.Cohen, "Fast effective rule induction",In 12th International conference on machine learning, California, USA, 1995, pp.115-123

AUTHORS

First Author – Velide Phani kumar, TELE 9 Technologies Limited, Hyderabad,Andhra Pradesh,India., Email: phani.velide@hotmail.com, Mobile: +91-9704044407
Second Author – Lakshmi Velide, Department of Biotechnology, Gokaraju Rangaraju Institute of Engineering and technology, Kukatpally, Hyderabad, Andhra Pradesh, India., Email: lakshmi.velide@gmail.com Mobile: +91-9866950998.

Table I: Attributes selected for soil data analysis

Attribute	Expanded form
N	Nitrogen
S	Sulphur
Ca	Calcium
Mg	Magnesium
Zn	Zinc
Fe	Iron
P	Phosphorus
K	Potassium
pH	Acidity or Basicity of a component
Hs	Humus

Table II: Comparison of different classifiers

Classifier	Naive Bayes	J48	JRip
Correctly Classified Instances	932	2108	2225
Incorrectly Classified Instances	1468	292	175
Accuracy	38.74	87.06	92.53
Mean Absolute error	0.426	0.0412	0.0316

Table III: Comparison of Regression analysis

Algorithm	Linear Regression	Least median Regression
Time for model building (Seconds)	2.85	12.86
Relative absolute error (%)	12.18	12.73
Correlation coefficient	0.9528	0.9746

Table IV: Comparison of actual and predicted values

Actual value by soil testing	Predicted values by linear regression	Error
11.4	11.81	0.41
5.6	5.34	-0.26
8.5	7.62	-0.88
8.1	7.79	-0.31
4.6	4.662	0.062
6.2	6.62	0.42
6.6	6.45	-0.15
2.5	3.2	0.7
7	7.5	0.5
16.8	17.4	0.6