I J C E INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING
July-December 2011, Volume 3, Number 2, pp. 89–95

APPLYING CLASSIFICATION TECHNIQUES IN DATA MINING IN AGRICULTURAL LAND SOIL

Prem Kumar Chandrakar[1], Sanjay Kumar[2] and Dewashish Mukherjee[3]

Abstract: Data Mining Techniques have led over various methods to gain knowledge from vast amount of data. There are the various research tools available for the large amount of data. Data mining has more technique to analyze the large amount of data like various classification algorithms. We have studied various data mining research tools available for analyzing the large amount of data. We have studied the WEKA data mining tools and various data mining classification algorithms like Bayesian classification Algorithm, Rule based classification and classification by Decision tree. This dissertation aim to study various research tools available for the comparison for large amount of data. They are applied to soil science database and found out relationship between them. A large data set of soil database is extracted from the Department of Soil Science, Indira Gandhi Krishi Vishwavidyalaya, Raipur, Chhattisgarh, India. The database contains measurement of soil profile data from locations of Raipur districts. The data belongs to the Raipur area of Chhattisgarh region. Comparison was made between Naïve Base Classification and most effective techniques. The database is the measurement of soil profile data from location of Raipur, Chhattisgarh in India. The outcome of the research may have many benefits, to agriculture, soil management and environmental. Data mining software application includes various methodologies that have been developed by both commercial and research centers. These techniques have been used for industrial, commercial and, scientific purpose. For example data mining has been used to analyze large datasets and established useful classification and patterns in the data sets. Agriculture and Biological research studied have been used for various techniques of data analysis including, natural trees, statistical machine learning in other analysis methods. This research aim to access data mining techniques and apply them to a soil science database to establish if meaningful relationship can be found. The overall aim of the research is to classify the soils using classification techniques based on texture of soil profile.

Keywords: Naïve bayes, Soil Profile, Bayesian Statistics, Soil Database, Classification, classifiers.trees.Id3, classifiers.trees.J48, classifiers.trees.Decision, classifiers.trees.BFTree, classifiers.trees.RandomTree, Rules.DecisionTable

1. INTRODUCTION

Data Mining Software application includes various methodologies that have been developed by both commercial and research centre. These techniques have been used for industrial, commercial and scientific purposes. For example, data mining has been used to analyze large data sets and establish useful classification and patterns in the data sets. Agricultural and biological research studies have been used for various techniques of data analysis including, natural trees, statistical machine learning and other analysis methods. This research aimed to assess data mining techniques and apply them to a soil science database to establish if meaningful relationships can be found. The data set has been assembled from soil surveys at agriculture area located in Raipur, Chhattisgarh, India. The soils studies which have been conducted by the Department of Soil Science Indira Gandhi Krishi Vishwavidyalaya, Raipur, Chhattisgarh, India. The database contains measurements of soil profile data from locations of Raipur District Chhattisgarh. Provide a vast amount of information on the classification of soil profiles and chemical characteristics. The analysis of such soil data sets is difficult given the complex relationships between large numbers of variables collected from each geographical location. Soil data is best when data obtained is more efficient, so that a larger number of samples are analyzed at lower costs, in less time and with higher accuracy. The current process to assess soil data uses standard statistical procedures to interpret the soil profile data sets. The use

[1,2,3] SoS in Computer Science and IT, Pt. Ravishankar Shukla University, Raipur, Chhattisgarh, India
[1,2,3] premchandrakar@gmail.com, sanraipur@rediffmail.com, dewashishmukherjee@gmail.com

of standard statistical analysis techniques is both time consuming and expensive. If alternative techniques can be found to improve this process, an improvement in the management of these soil environments may result. It is envisaged that the information gained from this research will contribute to the improvement and maintenance of soils and the agricultural environment of Soil Science. The research has a number of potential benefits to the Soil Science. However, the analysis and interpretation of a large data set is problematic. The overall aim of the research is to classify the soils using classification techniques based on texture of soil profiles.

2. DATA MINING

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data Mining refers to extracting or "mining" knowledge from large amounts of data that may be stored or online Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction [1]. Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. Some observers consider data mining to be just one step in a larger process known as knowledge discovery in databases (KDD). Other steps in the KDD process, in progressive order include data cleaning, data integration, data selection, data transformation, (data mining), pattern evaluation, and knowledge presentation [2].

3. VARIOUS TECHNIQUES OF DATA MINING

Data mining represents the integration of several fields, including machine learning, database systems, data visualization, statistics, and information theory. The knowledge discovery in databases is a complex process, which covers many interrelated steps. These steps are connected through the feedback loop based on the results obtained in the discovery. A user ultimately controls the parameters of the process [3]. Some of the steps can be outlined as the following:-

- Learning the application domain, which may include studying prior knowledge, analyzing the goals of the mining process, and building concept hierarchies.
- Data cleaning and data preprocessing, which included removal of noise and outliers if appropriate, unifying different formats of the data and different data sources, handling missing data fields, and sampling.
- Construction of data warehouses for On-Line Analytical Processing (OLAP), which includes aggregating the data and materializing views.
- Choosing the data mining algorithm that the data will be analyzed with. It includes the decision on the purpose of the model and rules derived by the data mining process. The algorithms for data mining may include clustering, classification, association, predication, etc.
- Selection of the data items, dimensions, attributes, and measures used in the mining process.
- Data reduction and projection, which includes finding good data representation, reduction of data dimensions, focusing on relevant items of data and relevant attributes.
- Pattern extraction, which includes finding rules, patterns, and models through data mining algorithms.
- Interestingness analysis, which might filter out the uninteresting patterns.
- Visualization of the discovered knowledge through tables, rules, graphs, charts, maps, diagrams etc.
- Application of the discovered knowledge through documenting it, reporting it to the users, taking actions based on the discovered rules, and resolving conflict with previously believed information [4].

4. ACM CLASSIFICATION NUMBER

H.2 [DATABASE MANAGEMENT]

H.2.8 [Database Applications]

5. PROBLEM DEFINITION AND RELATED WORKS

In this Section, we first define the conceptual problem of soil science and relationship between soil science and data mining classification technique and related works in this areas.

5.1 Conceptual Problem of Soil

The analysis of such soil data sets is difficult given the complex relationships between large numbers of variables collected for each geographical location. Soil data is best when data obtained is more efficient, so that a larger number of samples are analysed at lower costs, in less time and with higher accuracy.

The problem definition of our work is:-

(a) The first problem is to find accuracy of classifier.

(b) Second one is find best classifier method from soil data.

6. DATA MINING CLASSIFICATION

Classification is the process of automatically creating a model of classes from a set of records that contain class labels. Popular classification techniques include decision trees, neural networks, k-nearest neighbor, and Naive Bayesian classifier etc [5].

6.1 Bayesian Classification

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers [6].

6.2 Rule-Based classification

Where the learned model is represented as a set of IF-THEN rules. We first examine how such rules are used for classification. We then study ways in which they can be generated, either from a decision tree or directly from the training data using a sequential covering algorithm [7].

6.3 Classification by Decision Tree

Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label [8].

7. SOIL SCIENCE

Soil science is the study of soil as a natural resource on the surface of the earth including soil formation, classification and mapping; physical, chemical, biological, and fertility properties of soils.

7.1 Classification of Soil Science

The most common engineering classification system for soils in North America is the Unified Soil Classification System (USCS). The USCS has three major classification groups: (1) coarsegrained soils (e.g. sands and gravels); (2) fine-grained soils (e.g. silts and clays); and (3) highly organic soils (referred to as "peat"). The USCS further subdivides the three major soil classes for clarification [9] [10].

7.2 Classification of Soil in Chhattisgarh

The state of Chhattisgarh is endowed with varying soil type, surplus manpower and favorable agro-ecological conditions through which prosperity in the agriculture sector can be attained. It covers an area of 13.51 and gross cropped 5.8 MH, representing 4.1 and 4.07 per cent of the total geographical and gross cropped area, respectively of the country. The state is endowed with unique geographical settings. It has hilly region in the north of one hand and on the other, it hasvplateau area in the south and the plain area is spread over between these two zones. On the basicvof rainfall, temperature, soil type and topography of the land, the state can be broadly divided-vinto three appropriate zones. 1. Northern Hills 2. Chhattisgarh plains and 3. Bastar plateau. The climate of Chhattisgarh state, in general, is sub-humid with an average rainfall of about 1400 mm. The day time temperature during peak summer season is usually very high in the entire area varying from 430 C at Raigarh to 380 C at Bastar in the second-fortnight of May. The monsoon sets in around 10 th June in the southen most tip of Bastar region and finally extends over the entire area by 25th June [11].

8. RESEARCH METHODOLOGY

Existing parameter to analyzed physical properties of soil science. In this section we explain the existing parameter to analyzed physical properties of soil science.

8.1 Correctly Classified Instance

If the problem is a multi-class one (i.e. more than two classes) then AUC is calculated for each class in turn by treating all other classes as the negative class. It is possible to achieve a high AUC on one class while the overall classification accuracy is somewhat lower. Another possibility is due to the precision that is used to output stuff. Classification accuracy is output to four decimal places (and is a percentage between 0 and 100) while AUC is output to three decimal places (and is a number between 0 and 1)[12].

8.1 Incorrectly Classified Instance

Incorrectly classified instances refers to the case where the instances are used as test data and again are the most important statistics here for our purposes [13].

8.2 Kappa Statistic

Inter observer variation can be measured in any situation in which two or more independent observers are evaluating the same thing [14].

8.3 Mean Absolute Error

The mean absolute error is a quantity used to measure how close forecasts or predictions are to the eventual outcomes [15].

8.4 Root Mean Square Error

The root mean square deviation (RMSD) or root mean square error (RMSE) is a frequently-used measure of the differences between values predicted by a model or an estimator and the values actually observed from the thing being modeled or estimated [16].

8.5 Root Relative Square Error

The relative absolute error is very similar to the relative squared error in the sense that it is also relative to a simple predictor, which is just the average of the actual values. In this case, though, the error is just the total absolute error instead of the total squared error [17].

8.6 Soil Data Collection

The dataset was collected as part of a survey by soil data of Raipur district and included a large amount of information from different location within the Chhattisgarh. This information was collected from various locations where a pit was dug and samples taken. The samples were then sent for chemical and physical analysis at the agricultural laboratories in Department of Soil Science Indira Gandhi Krishi Vishwavidyalaya , Raipur ,Chhattisgarh.

8.7 Data Mining Process

The data mining process was conducted in accordance with the results of the statistical analysis. The following steps are a general outline of the procedure that allowed a cluster analysis to be conducted on the dataset.

8.8 Data Collection Cleaning and Checking

Relevant data was selected from a subset of the Soil science database.

8.9 Data Formatting

The data was formatted into an Excel format from the Access database, based on the ten soil types and relevant related fields. The data was then copied into a single Excel spread sheet. The Excel spread sheet (ESS) was then formatted to replace any null or missing values in the soil data set to allow coding for the file in the next phase.

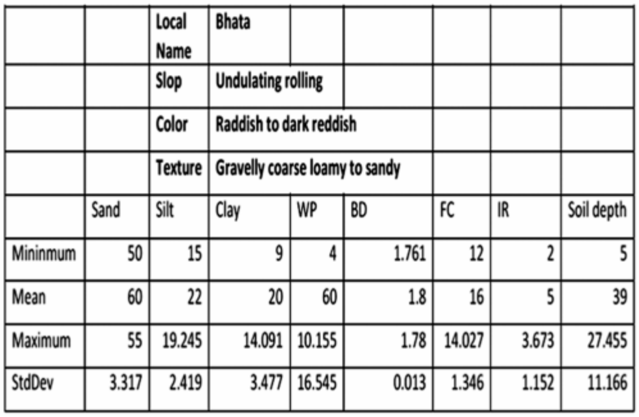
8.10 Data Coding

The soil data set was then converted into a comma delimited (CSV) format file for the ESS. This file was then saved and opened using a text editor. The text editor was used to format and code the data into the type that will allow the data mining techniques and programs to be applied to it. The coding was formatted so that the input will recognize names of the attributes, the type of value of each attribute and the range of all attributes. Coding was then conducted to allow the machine learning algorithms to be applied to the soil data set to provide relevant outcomes that were required in the research[18].

9. RESULT

9.1 Statistical Results

When we applied soil data in weka data mining software first step analyzing statistical result like minimum , mean ,maximum and stander derivation. Table's show statistical result and graph between physical properties of soil and statistical values.



		Local Name	Bhata					
		Slop	Undulating rolling					
		Color	Raddish to dark reddish					
		Texture	Gravelly coarse loamy to sandy					
	Sand	Silt	Clay	WP	BD	FC	IR	Soil depth
Mininmum	50	15	9	4	1.761	12	2	5
Mean	60	22	20	60	1.8	16	5	39
Maximum	55	19.245	14.091	10.155	1.78	14.027	3.673	27.455
StdDev	3.317	2.419	3.477	16.545	0.013	1.346	1.152	11.166

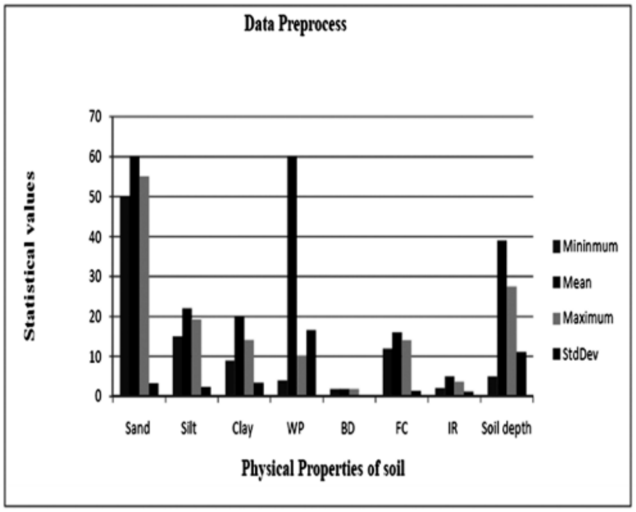


Fig. 8.1(a): Statistical Graph (Bhata)

Local Name		Dorsa						
Slop		level gently, undulating						
Color		Brownish grey						
Texture		Clay Loam						
	Sand	Silt	Clay	WP	BD	FC	IR	Soil depth
Mininmum	25	25	35	15	1.3	25	1	50
Mean	35	35	45	18	1.65	30	2	150
Maximum	30	30	38.082	16.7	1.557	27.1	1.5	113.818
StdDev	3.317	3.317	3.248	1.05	0.108	1.525	0.332	33.193

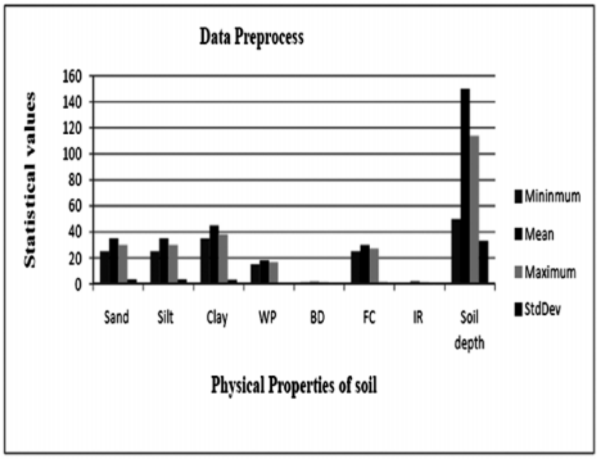


Fig. 8.1(b): Statistical Graph (Dorsa)

	Local Name	Kanhar						
	Slop	level						
	Color	Dark grey brown to black						
	Texture	Clayey						
	Sand	Silt	Clay	WP	BD	FC	IR	Soil depth
Mininmum	20	25	15	16	1.2	30	0.6	80
Mean	25	35	45	20	1.55	35	1.5	150
Maximum	22.88	29.6	30.1	17.8	1.401	32.09	1.05	102.3
StdDev	1.648	3.204	10.535	1.376	0.117	1.611	0.303	22.099

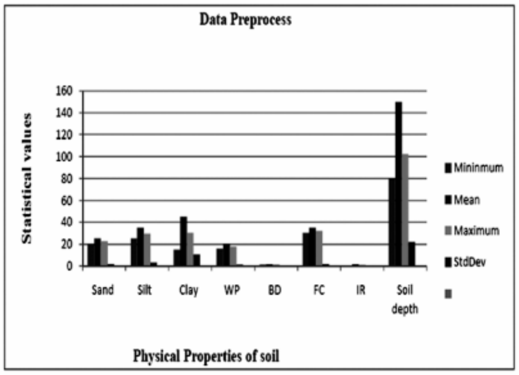


Fig. 8.1(c): Statistical Graph (Kanhar)

	Local Name	Matasi						
	Slop	Level gently , undulating						
	Color	Yellow						
	Texture	Sandy Loam						
	Sand	Silt	Clay	WP	BD	FC	IR	Soil depth
Mininmum	40	25	15	8	1.5	8	1.5	30
Mean	55	35	25	12	1.65	22	1.6	80
Maximum	47.03	30.3	20.3	9.65	1.588	13	1.546	55.3
StdDev	4.895	3.335	3.335	1.293	0.052	4.714	0.032	17.77

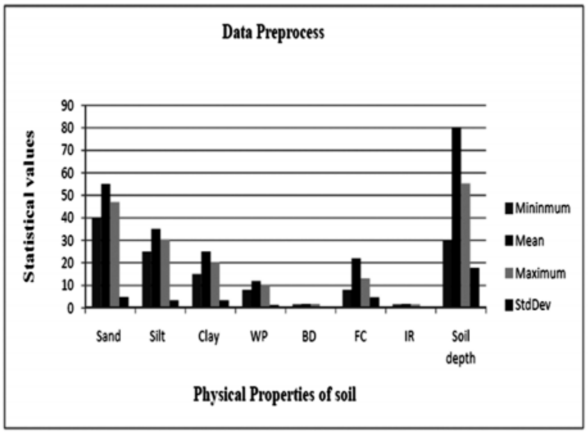
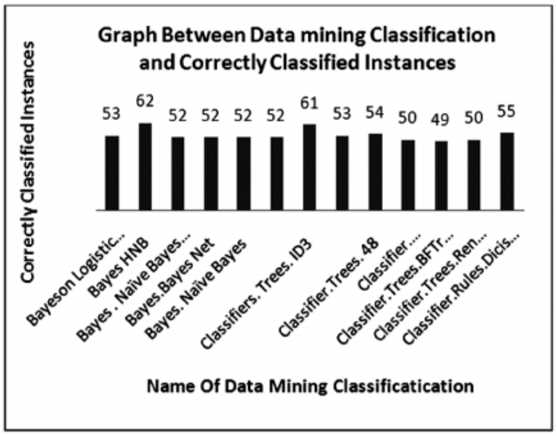


Fig. 8.1(d): Statistical Graph (Matasi)

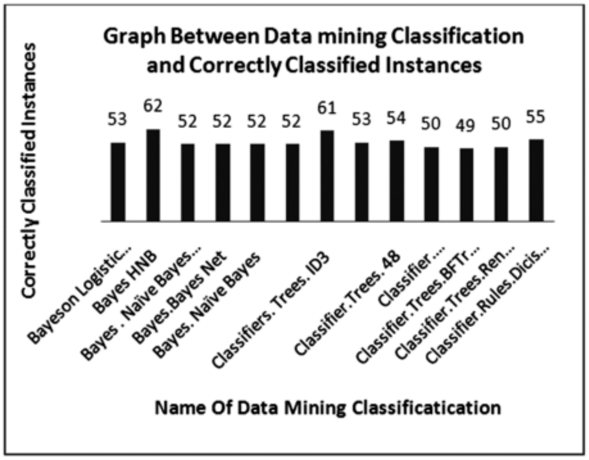
9.2 Data Mining Results

Bhata Soil Database

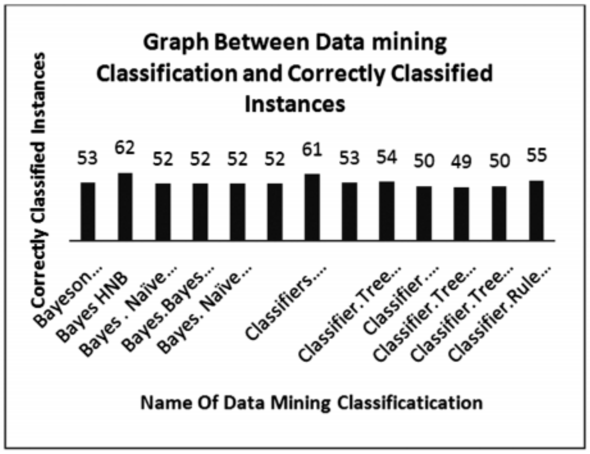
When we applied data mining classification techniques in soil data form weka data mining software . We finding results of defferent parameters like Correctly classified instance, Incorrectly classified instance, kappa statistic, Mean Absolute Error, Relative Absolute error, Root Relative Squared error. In this section we present results of parameters in tabulation form of classification techniques and parameters and also present each class accuracy results in graphical form.



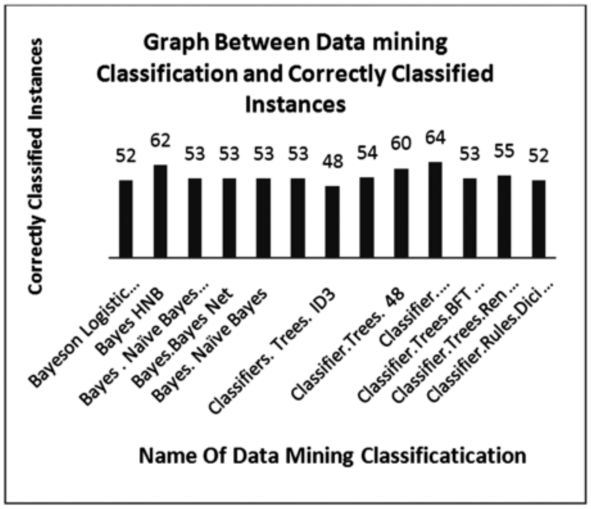
Dorsa Soil Database



Kanhar Soil Database



Matasi Soil Database



10. CONCLUSION

This problem is solved with the help of data mining classification techniques. When we take large amount of soil data as input and analyzed it in weka data mining software then error rate of various classifiers are detected, and we applied the soil data as a input data and analyzed to various data mining classification techniques such as Bayes:- Bayesian Logistic Regression, Bayes HNB, bayes.NaiveBayesSimple, Bayes.BayesNet, bayes. Naive Bayes,bayes.Naive Bayes Updateable, Tree:- classifiers. trees.Id3, classifiers.trees.NBTree, classifiers.trees.J48, classifiers.trees.Decision,classifiers.trees.BFTree, classifiers.trees.RandomTree, and Rule:- classifiers.rules. DecisionTable tesed by weka data mining tool. Analyzing to all various data mining classification techniques with soil profile test parameters like Correctly Classified Instances, Incorrectly Classified Instances, Kappa statistic, Mean absolute error, Root mean squared error, Relative absolute error, Root relative squared error. With the help of this method we find when Bayes HNB classification technique is applied to soil data set, the correctly classified instances are more classified. The Kappa statistic , Mean absolute error, Root mean squared error , Relative absolute error are less than the remaining Classifiers, like Bayesian classifier,J48.of soil profile. The time to build the Bayes HNB Classifier is less than the remaining Classifier. So, The Naive Bayes Classifier is the efficient classification technique among remaining classification techniques. Normalized Expected Cost of Bayes HNB is more accurate when compared to Bayesian Network best classifier. finally we say Bayesian Classification technique is best classification technique.

11. FUTURE WORK

The recommendations arising from this research are: That data mining techniques may be applied in the field of soil research in the future as they will provide research tools for the comparison n of large amounts of data. Data mining techniques, when applied to an agricultural soil profile, may improve the verification of valid soil profile classification.

In future we will find the best classifier for this data in terms of memory and time. Further research would look at increased dataset size to determine if this would increase the instance classification. This would create more focus on data mining and less on current statistical methods. There were a number of areas not explored by the research due to time limitations, such as the differences between the soil profile.

REFERENCES

[1] Cunningham S. J., and Holmes, “Developing Innovative Applications in Agriculture using Data Mining”. In the Proceedings of the Southeast Asia Regional Computer Confederation Conference, 1999, pp. 1-2.

[2] F. N. Afrati, A. Gionis, and H. Mannila. “Approximating a Collection of Frequent Sets”. In Proc. 2004 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD’04), pp. 12-19, Seattle, WA, Aug. 2004.

[3] Kantardzic, Mehmed (2003). “Data Mining: Concepts, Models, Methods, and Algorithms”. John Wiley & Sons. ISBN: 0471228524. OCLC 50055336.

[4] Fayyad Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). “From Data Mining to Knowledge Discovery in Databases”. Retrieved 2008-12-17.

[5] Thair Nu Phyu “Survey of Classification Techniques in Data Mining”, Proceedings of the International MultiConference of Engineers and Computer Scientists, 2009, Vol 1, IMECS 2009, March 18-20, 2009, Hong Kong.

[6] T. Joachims. “A Statistical Learning Model of Text Classification with Support Vector Machines”. In Proc. Int. 2001 ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR’01), pp. 128-136, New Orleans, LA, Sept. 2001.

[7] Andrew W. “Moore Professor School of Computer Science Carnegie Mellon University”, Naïve Bayes Classifiers, www.cs.cmu.edu/~awm awm@cs.cmu.edu 412-268-7599

[8] Neapolitan R.E., “Learning Bayesian Networks”, Prentice Hall, Upper Saddle River, NJ, 2004.

[9] N. Friedman, M. Linial, I. Nachman, D. Peer (August 2000). “Using Bayesian Networks to Analyze Expression Data”. Journal of Computational Biology, (Larchmont, New York: Mary Ann Liebert, Inc.) 7(3-4), 601-620. doi:10.1089/106652700750050961. ISSN 1066-5277. PMID 11108481.

[10] A. Ansari, S. Ansari, “The Concept of Data Mining, Its Application and Issues”, “TECHNOLOGY FORCES (Technol, Forces)”, Journal of Engineering and Sciences, January-June 2010. pp. 26-30.

[11] W.L. Buntine and T. Niblett. “A Further Comparison of Splitting Rules for Decision-tree Induction”. Machine Learning, 8:75-85, 1992.

[12] M. Ankerst, C. Elsen, M. Ester, and H.P. Kriegel. “Visual Classification: An Interactive Approach to Decision Tree Construction”. In Proc. 1999 Int. Conf. Knowledge Discovery and DataMining (KDD’99), pp. 392-396, San Diego, CA, Aug. 1999.

[13] Mckenzie N., and Ryan P. (1999). “Spatial Prediction of Soil Properties using Environmental Correlation”.

[14] P. Bhargavi et al. “Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils”, IJCSNS International Journal of Computer Science and Network Security, 9(8), August 2009, pp. 117-121.

[15] Patil, S.K., Mishra V.N., and Jaggi I.K., (1997). “Soil Testing Metthods and Fertilizer Recommendations”. IGKV, Raipur (C.G.).

[16] Qu Xiaoyu Correctly Classified Instances Calculated https://list.scms.waikato.ac.nz

[17] Qu Xiaoyu http://www.cs.waikato.ac.nz/~ml/weka/mailinglist_etiquette.html

[18] Arie Ben-David (February 2008). “Comparison of Classification Accuracy Using Cohen’s Weighted Kappa”. Expert Systems with Applications: An International Journal (Pergamon Press, Inc. Tarrytown, NY, USA) 34 (2), 825-832. doi:10.1016/j.eswa.2006.10.022.http://portal.acm.org/citation.cfmid=1322819

[19] Hyndman R., and Koehler A., (2005). “Another Look at Measures of Forecast Accuracy”.