Machine Learning Classification Techniques:

A Comparative Study

Archana Chaudhary[1], Savita Kolhe[2] & Raj Kamal[3]

[1&3]SCS & IT, DAVV,   [2]Soybean Research, (ICAR)

E-mail : archana_227@rediffmail.com, savita_dakhane@yahoo.com, dr_rajkamal@hotmail**.com**

Abstract – **Machine learning is the study of computer algorithms that improve automatically with experience. In other words it is the ability of the computer program to acquire or develop new knowledge or skills from examples for optimising the performance of a computer or a mobile device. In this paper we apply machine learning techniques Bayes network, Logistic Regression, Decision Stump, J48, Random Forest, Random Tree and REPtree to build classifier models and compare them. In this study machine learning techniques are applied to agriculture data and for each classifier model correctly classified instances, incorrectly classified instances, model build time and kappa statistics are computed. The reported test results describe the applicability and effectiveness of the classification approach.**

Keywords – C4.5, Decision Stump, GSM, REPtree, 3G+, J48.

I. INTRODUCTION

Intelligent systems learn by improving through experience [1]. This learning process is not only restricted to humans but spreads across many fields including machine learning, psychology, neuroscience, education, computational linguistics, economics and bioinformatics The field of Machine learning deals with developing programs that learn from past data and is also a branch of data processing. Machine learning includes the stream in which machines learn for knowledge gain or understanding of some concept or skill by studying the instruction or from experience [2]. Machine learning techniques consist of formulation of programs that imitate some of the facets of human mind that helps us to solve highly complicated problems at a very good speed [3]. Thus, machine learning has great potential in improving the efficiency and accuracy of decisions drawn by intelligent computer programs. Machine learning includes mainly concept learning and classification learning. Classification is the most widely used Machine learning technique that involves separating the data into different segments which are non-overlapping. Hence classification is the process of finding a set of models that describe and distinguish class label of the data object [5]. Machine learning field is also useful for mobile devices such as Smart phones, smartcards and sensors, handheld and automotive computing systems [4]. Mobile Technology has fostered development with the help of increasing mobile terminals (e.g. computers, mobile computers, mobile phones, Pocket PC, PDA) and mobile networks (GSM, 3G+, wireless networks, Bluetooth etc.).Machine learning techniques like C4.5, Naïve Bayesian, Decision trees etc are helpful for mobile devices. Machine learning applications for mobile devices include Sensor based activity recognition, Mobile text categorization, Malware detection on mobile devices, Language understanding etc.

II. METHODS

A. Bayes Network Classfier

Bayesian networks are a powerful probabilistic representation, and they are used for classification purposes [2]. Bayesian networks are also called belief networks and belong to the group of probabilistic graphical models .These graphical structures are used for knowledge representation of an uncertain domain. In this network, each node in the graph represents a random variable, where as the edges between the nodes represent probabilistic dependencies among the corresponding random variables. These conditional dependencies in the graph are often estimated by using known statistical and computational methods. The Bayesian classifier learns from training data the conditional probability of each attribute Bi given the class label X [6]. Classification is done by applying

Bayes rule to calculate the probability of X given the particular instances of B1…..Bn and then predicting the class that has highest posterior probability. The naive Bayes classifier is a Bayesian network where the class has no parents and each attribute has the class as its sole parent [6]. Hence these networks have principles from graph theory, probability theory, computer science, and statistics.

B. *Logistic Regression*

The logistic regression is a type of regression model that is used for predicting the result of a categorical (a variable that can have a limited number of categories) dependent variable based on one or more predictor variables. In other words logistic regression measures the relationship between a categorical dependent variable and usually a continuous independent variable (or several), by converting the dependent variable to probability scores. Hence a Logistic Regression model is used to determine the impact of multiple independent variables presented simultaneously to predict membership of one or other of the two dependent variable categories. This model is useful in determining Protein function which is further used to predict protein-protein interaction [7].It is also useful in predicting customer retention, Forecasting stock performance, spam filtering and a variety of classification tasks. There are some basic assumptions for this model [8]. These are

- Logistic regression does not consider a linear relationship between the dependent and independent variables.
- The dependent variable must be a dichotomy (2 categories).
- The independent variables need not be interval, nor normally distributed, nor linearly related, nor of equal variance within each group.
- Larger samples are required in this model.

C. *Decision Tree*

A decision tree is a classifier model that works with recursive partition of the instance space. It is used to represent a supervised learning approach. It is a simple graphic structure where non-terminal nodes represent tests on one or more attributes and terminal nodes give decision outcomes. This tree consists of one root, branches, internal nodes and leaves. This tree is drawn from left to right or beginning from the top root to downward nodes, so that it is easy to draw it. In this tree each internal node may grow out two or more branches. Each node corresponds with a certain feature or characteristic or feature and the branches correspond with a range of values or decision outcomes [9]. The major benefits of using a decision tree are:

- It is a simple model that helps in decision making.
- It is relatively easy to interpret and understand.
- It can be easily converted into a set of production rules.
- It can classify both categorical and numerical data but the resultant attribute is categorical.
- It requires no prior assumptions about the nature of the data [9].

The Decision tree techniques that we have used in this study are:

The REPTree

REPTree is a quick decision tree learner that designs a decision/regression tree using information gain as the basis of splitting. It prunes the tree using reduced error pruning. The reduced error pruning [10] is a method that checks for each internal node, whether replacing it with the most frequent class does not reduce the tree's overall accuracy. In this case, the respective node is pruned. The procedure continues until any further pruning would decrease the accuracy. In this way REPTree is an efficient technique.

The Random Tree

A random tree is a tree drawn at random from a collection of possible trees. It is known as a Random tree because each tree in the set of trees has an equal opportunity of being sampled. It implies that with 'm' random features at each node, a random tree is a tree drawn at random from a set of possible trees. These trees offer great efficiency as these trees lead to more accurate models [9]. Random tree models are used extensively in the field of Machine Learning since recent years.

The C 4.5 tree (J48)

This tree creates a decision tree based upon the attribute values of the available training data [11]. This tree works on identification of attribute that discriminates various instances most clearly. J48 is slightly modified C4.5 in WEKA. This classification technique generates a decision tree for a given set of data by recursive partitioning of data. This tree follows a Depth-first strategy. It considers all the possible tests that can split the set of data items and selects a test that gives the best information gain [9].

The Decision Stump

A decision stump is basically a single-level decision tree where the split at the root level is based on a specific attribute/value pair. It is a decision tree with only one internal node (the root) that is connected to the terminal nodes (its leaves). A decision stump makes a prediction based on the value of just a single input feature. Sometimes they are also called 1-rules [13]. Boosting is an approach to machine learning that is based on the idea of creating a predictor with a high level of accuracy by combining many weak and inaccurate rules. Whereas bagging predictors is a way of generating several versions of a predictor and using these to get an aggregated predictor. Decision stumps are used as important components of base learner modules, in machine learning techniques used for boosting and bagging.

The Random Forest

Random forest is a collection of un pruned classification or regression trees, induced from bootstrap samples of the training data, using random feature selection in the tree induction process. Prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble. Random forest generally exhibits a substantial performance improvement over the single tree classifier such as CART and C4.5.

III. DATA SAMPLES IN WEKA

Weka was designed for the purpose of processing agricultural data, motivated by the importance of this application area in New Zealand. The machine learning methods and data engineering capability of Weka made it popular for all forms of machine learning problems. The applications of Weka in the field of bioinformatics include plant genotype discrimination, automated protein annotation, probe selection for gene expression arrays, classifying gene expression profiles and extracting rules from them. Text mining is another important field of application of Weka. This workbench has been used in Text mining field to automatically extract important key phrases from text, for document categorization and word sense disambiguation. The most important feature of any machine learning problem domain, is the type of data it takes. Most learning techniques that are applied to different problem domains assume that the data are presented in a simple attribute-value format. Broadly there are nominal, linear and tree structured attributes in Weka. In WEKA most simple form of attribute is nominal which represents attribute-value pair. Weka also supports linear attributes that are totally ordered and tree-structured attributes that form a hierarchy and are partially ordered[12].The nominal and linear types of attributes are shown by the Fig. "(a)" and Fig. "(b)" below.

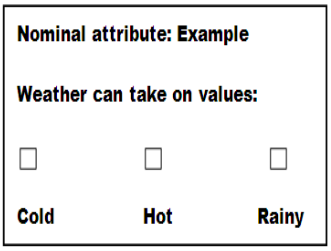

Fig . a

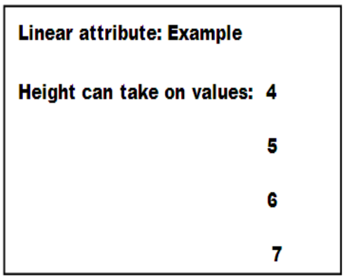

Fig. b

Similarly a tree-structured attribute is a structured attribute. This attribute type is shown by the fig. "(c)" below.

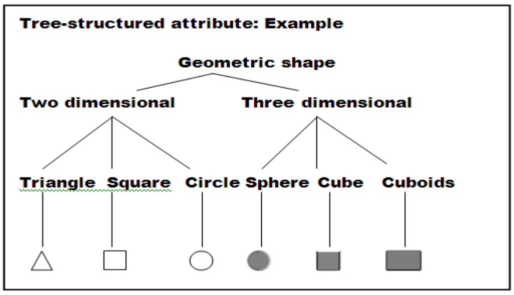

Fig. c

Weka also offers string attribute, date attribute and numeric attribute. An attribute vector or a sample object in Weka describes situations that involve relations between objects. The pictorial representation of an attribute vector is shown by the Fig. "(d)" below

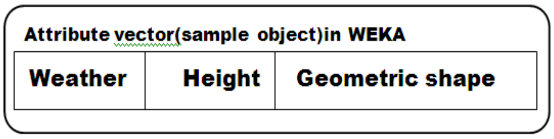

Fig. d

In this study, all data are analyzed thereafter mined with the aim of WEKA and are saved in ARFF

(Attribute Relation File Format) format (WEKA's data format) [14]. ARFF format consists of special tags in order to differentiate between attributes, values and names of the given data. In this work the dataset chosen was a large Soybean dataset that contains 683 instances. There are a total of 15 classes and 35 attributes in this dataset. This dataset consists of information related to Soybean.

IV. COMPARATIVE RESULTS

The Soybean data samples were used for machine learning classification techniques. The Machine learning techniques Bayes network, Logistic regression, J48, Random Forest, Decision Stump, Random tree, REPtree were used for simulation. Here we split our original dataset of 683 samples into 66% for training purpose and remaining 34% for testing purpose. Weka incorporates k-fold cross-validation, in which the original sample is randomly partitioned into k subsamples. Further from these k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining (k – 1) subsamples are used as training dataset. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. These k results from the folds are then averaged to produce estimation. In this work we have used 10-fold cross validation. Cohen's kappa coefficient is a statistical measure of degree of agreement between different raters related to same observations. These variations can be measured in a situation in which two or more independent observers or entities are evaluating the same thing. If the two observers randomly assign their ratings, then they may sometimes agree just by chance. Kappa gives us a numerical rating of the degree to which this agreement occurs. It is a measure of this difference, standardized to lie on a -1 to 1 scale, where 1 is perfect agreement, 0 implies expected by chance, and negative values indicate agreement lesser than chance, i.e. it indicates potential disagreement between the observing entities. The results of application of these techniques are shown in table I below.

Table I Classification results for each examined technique.

Machine Learning Classification Technique	Correctly classified	Incorrectly classified	Kappa statistic	Time taken (in sec)
Bayes Network	93.26% (637)	6.73% (46)	.9263	.06s
Logistic Regression	93.85% (641)	6.14% (42)	.9326	.02s
J48	91.50% (625)	8.49% (58)	.9068	.06s
Random Forest	90.48% (618)	9.51% (65)	.8956	.16s
Decision Stump	27.96% (191)	72.03% (492)	.1942	.02s
Random Tree	84.04% (574)	15.95% (109)	.8250	.02s
REPtree	84.77% (579)	15.22% (104)	.8326	.08s

It is clear from the table I above that Logistic Regression and Bayes network outperform other techniques for the given dataset. It is also clear from the table I above that the Kappa rating of Bayes Network and Logistic Regression are high as compared to other learning techniques. Further the build time taken by Random Forest is significantly higher than other Machine learning techniques. The decision stump does not perform well in classification and is with the lowest kappa statistic.

V. CONCLUSION

This paper gives a brief comparative study on the performance of different machine learning techniques. These techniques are simulated in Weka. In this study these techniques are applied on Soybean dataset. We have used 66% split and 10 fold cross validation. It is clear from our study that Logistic Regression and Bayes network outperform other techniques for the given dataset. Furthermore the decision stump does not perform well in classification and is with the lowest kappa statistic.

VI. REFERENCES

[1] J. Thornton, "Techniques In Computational Learning", Chapman and Hall, London, 1992.

[2] Mohd Fauzi bin Othman, Thomas Moh Shan Yau, "Comparison of Different Classification Techniques Using WEKA for Breast Cancer", Biomed 06, IFMBE Proceedings 15, pp. 520-523, Springer, 2007.

[3] Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, "Rotation, scale, and translation resilient public watermarking for images," IEEE Trans. Image Process., vol. 10, no. 5, pp. 767-782, May 2001.

[4] Schank, R, Dynamic Memory: A Theory of Reminding and Learning in Computers and People. Cambridge University Press, Cambridge, 1982.

[5] Archana Chaudhary,Savita Kolhe, Rajkamal, "Machine Learning Techniques for Mobile Intelligent Systems: A Study",IEEE International conference on Wireless and Optical Communications Networks, ISBN 978-1-4673-1988-1, 2012

[6] Sandhya Joshi,P. Deepa Shenoy,Venugopal K R,L .M. Patnaik, “Evaluation of different stages of Dementia employing Neuropsychological and Machine learning techniques”,IEEE ICAC,2009.

[7] Bouckaert, R.R. , “Properties of Bayesian network Learning Algorithms”, In R. Lopex De Mantaras & D. Poole (Eds.), In Press of Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence (pp. 102-109). San Francisco, CA, 1994.

[8] Qingshan Ni, Zheng-Zhi Wang, Qingjuan Han, Gangguo Li, Xiaomin Wang, Guangyun Wang, “Using Logistic Regression Method to Predict Protein Function from Protein-Protein Interaction Data”, ICBBE 2009. 3rd International Conference on Bioinformatics and Biomedical Engineering, E-ISBN 978-1-4244-2902-8,2009.

[9] http://www.uk.sagepub.com.

[10] Yongheng Zhao, Yanxia Zhang, “Comparison of decision tree methods for finding active objects”, Advances of Space Research, 2007.

[11] J. Ross Quinlan, "Generating Production Rules from Decision Trees. Proceedings of the Tenth International Joint Conference on Artificial Intelligence (pp. 304–307), 1987.

[12] Robert J. McQueena, Stephen R. Gamer, Craig G. Nevill-Manning,Ian H. Witten , “Applying machine learning to agricultural data” , Computers and Electronics in Agriculture 12 (1995) 275-293,Elsevier,1995.

[13] Wayne Iba and Langley Pat, “ Induction of One-Level Decision Trees”, Proceedings of the Ninth International Conference on Machine Learning, Aberdeen, Scotland, 1–3 July 1992, San Francisco, CA: Morgan Kaufmann, pp. 233–240,1992.

[14] www.weka.com

◈◈◈