IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Analysis of WEKA Classification Algorithms of Data Mining for Large Data Sets

Birinder Singh Sarao
Assistant Professor, Mata Gujri College, Fatehgarh Sahib

Abstract
Data Mining has an important role in handling large data sets for multiple applications used in problem solving. It involves in extracting insights, predictions, patterns and trends from different sources like data warehouses having large datasets. For the extraction, it uses various techniques like classification, association, clustering and regression. Concept of Data Mining is being used in many areas nowadays like healthcare, business and marketing, Security, Banking and Finance, Education and many more. This paper covers the classification techniques of data mining. WEKA tool has certain classifiers, which are compared on the basis of their performance on data sets.

Keywords: Classification, Random Forest, J48, LMT, WEKA

Introduction

Knowledge Discovery in Database(KDD) is a process of recovering useful knowledge or extracting patterns from data through Data Mining [1]. It involves following steps like selection, pre-processing, transformation, Data Mining and Evaluation. KDD uses different Data Mining techniques like Clustering, classification, association, regression etc. Various tools and algorithms are used for these data mining different techniques on large datasets as per the pattern or trend requirement [2]. This paper shows the comparison of various techniques of classification used in Data Mining for extracting knowledge from large datasets by considering certain parameters. WEKA support various data mining tasks which includes preprocessing of data, classification, association, clustering etc. It provides flexibility for using existing methods on new data sets.

Classification Techniques

Classification is supervised learning technique which is used to find out that a new observation belongs to which category depending on labelled training data sets. Following steps are followed in classification: -

1. Training Data set is created
2. Class and Class attributes are detected
3. Identification of useful attributes for classification
4. In training set, a model is acquired using training examples.
5. Using Model for classification of unknown data samples

Random Forest Algorithm

Random Forest is a blending learning method which uses multiple decision trees for better accuracy. It works on the strategy where a strong learner is achieved from the combination of multiple decision trees (weak learners). Multiple bootstrapped datasets are generated from original dataset by bagging technique i.e through random sampling. One decision tree is trained from each dataset. Random feature selection is used for selecting random subset of features from each spilt in the decision tree [1,3,4]. This selection feature helps to reduce correlation between trees which improves the overall performance.

Steps performed in Random Forest

Step1: Input the training dataset D with samples m and features a.
Step2: Create n bootstrap datasets by sampling from D with replacement as D1, D2, D3…. Dn
Step3: Train a decision tree on each Di. Subset of features are randomly selected at each spilt and the best spilt is selected among the features.
Step4: Each tree selects a class and the majority one becomes the final prediction. The mean of all tree outputs is taken.

J48 Classifier

J48 classifier is an improved and implemented version of C4.5 algorithm. It built decision trees based on entropy and information gain. J48 is used for building decision trees and it is an extension of ID3[2,5].

It works on the concept that each attribute of the data can be spilt into smaller subsets for making decisions. Attribute with maximum normalized information gain is used for decision making and the output is represented in decision tree [6,7].

Weka tool for J48 offers options which are associated with tree pruning. Pruning can be used for summarizing in potential overfitting. The objective is progressive generalization of a decision tree until it gains an equilibrium of flexibility and accuracy [3,8]. The steps involved in this algorithm are

1. If the instances fit to similar class, the leaf is labelled with similar class
2. Potential data is assumed for each attribute and from the test on the attribute, the gain on the data is taken.
3. Based on the current selection parameters, best attribute is selected.

Logistic Model Tree(LMT)

A Logistic Model Tree is a hybrid model that combines decision tree learning with logistic regression at the leaves [4]. It is used for classification tasks, particularly when you want to combine the interpretability of decision trees with the probabilistic and linear modelling capabilities of logistic regression [6,8].

Steps performed in Logistic Model Tree(LMT)

1. It builds a decision tree similar to CART or C4.5, recursively partitioning the data.
2. Logistic Regression at Leaves: Instead of assigning a class label to each leaf, LMT fits a logistic regression model to the data in that leaf.
3. Splitting Criterion: Often uses log-likelihood or entropy-based metrics to decide splits.
4. Model at Each Node (Optional): Some versions also allow logistic regression at internal nodes and use LogitBoost for fitting.

Tool and Data Set Used

WEKA is an open source software which provides tools for data pre-processing and implementation of data mining algorithms on real world mining problems [9, 10] It is a powerful tool which is used for data analysis. It supports many file formats. This paper uses .arff file format for analysis[11,12].

Dataset is taken from Central Research Establishment, Home Office Forensic Science Service, Aldermaston. It includes 214 instances and 10 attributes (also includes class attribute). All the attributes are continuously valued.

Attribute Information [9]:

1. Id number: 1 to 214
2. RI: refractive index
3. Na: Sodium, Mg: Magnesium, Al: Aluminum, Si: Silicon, K: Potassium, Ca: Calcium, Ba: Barium, Fe: Iron
4. Type of glass: (class attribute)
   building_windows_float_processed,
   building_windows_non_float_processed,
   vehicle_windows_float_processed,
   vehicle_windows_non_float_processed

Results and Discussion

As mentioned above, glass dataset contains 10 attributes and all the feature can be seen on the single window after data pre-processing. Different patterns can be generated according to the number of attributes chosen. Figure 1 shows the data pre-processing of Glass dataset with different attributes

Fig-1 Data preprocessing of all attributes of data set

Data has to be further classified. Different classifiers are used on the dataset and their performance is evaluated depending upon different parameters. Figure 2 shows the J48 classification process on the screen. Classifier output is shown where precision, recall, f-measure values are calculated. Correctly classified instances are 66%.

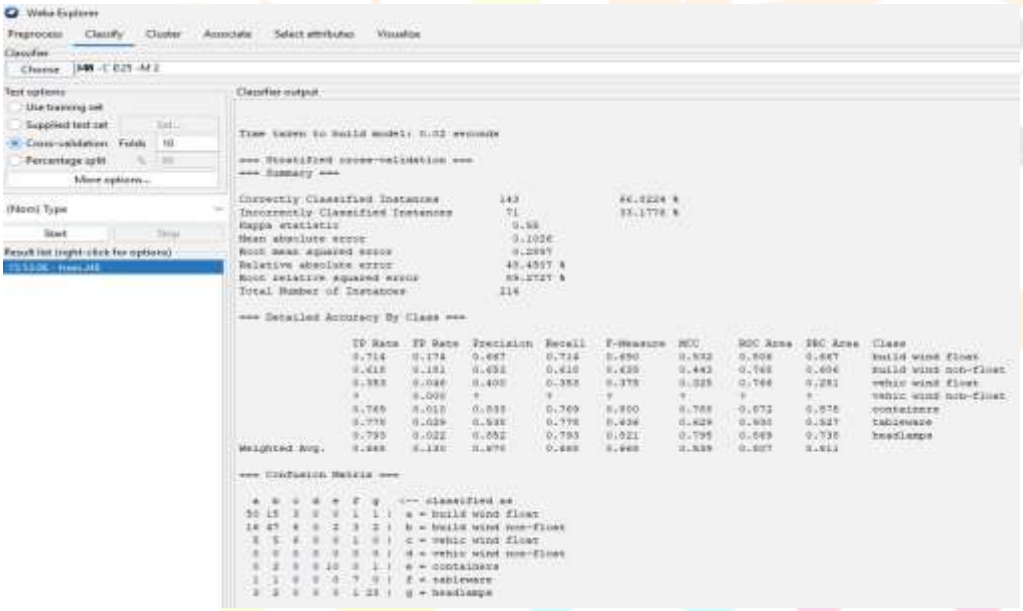

Fig-2 Processing of dataset by J48 classifier

Classification of Training set by LMT classifier is shown in Figure 3 and similarly Figure 4 shows the classification done by Random Forest Classifier. All the figures show the Correctly classified instances, Incorrectly classified instances, Precision, Recall, F-Measure.

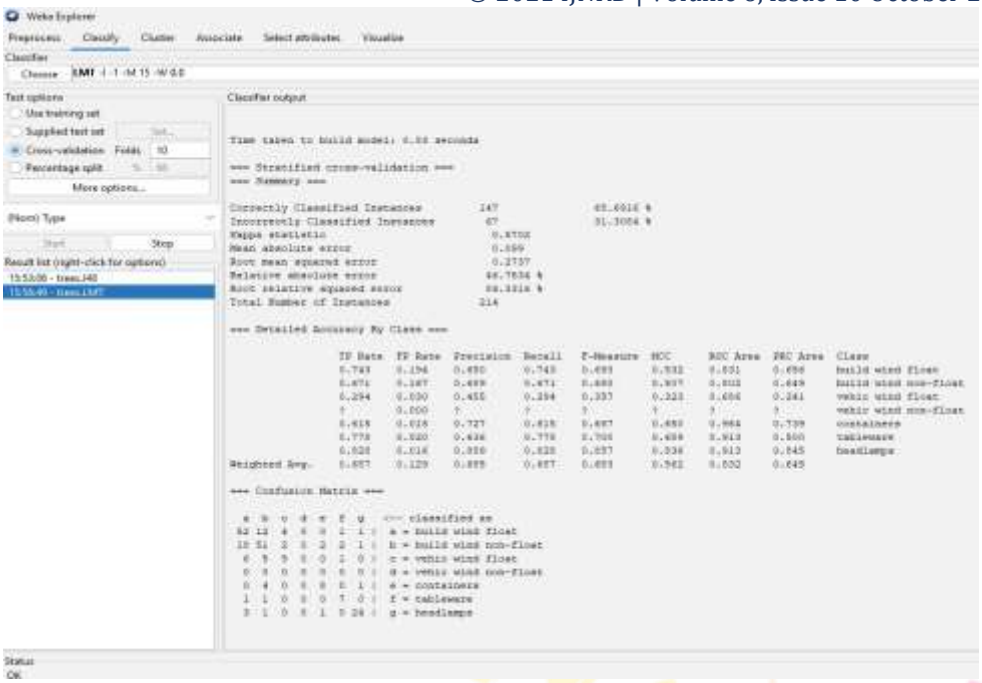

Fig-3 Processing of dataset by LMT classifier

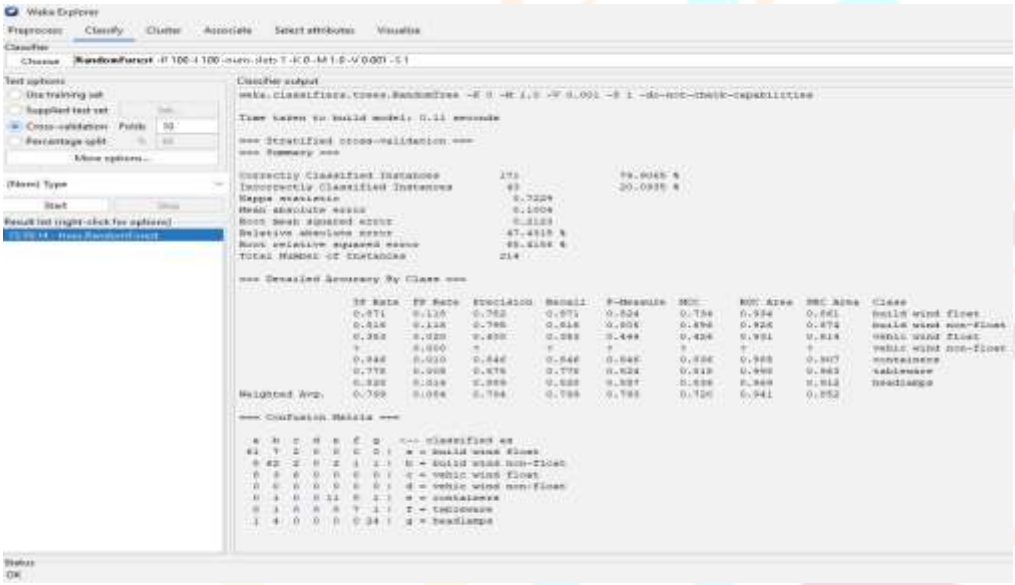

Fig-4 Processing of dataset by Random Forest Classifier

F-measure, precision, and recall are key evaluation metrics used in classification problems, especially for imbalanced datasets or information retrieval tasks (like search engines, spam detection, etc.) [12,13]

Recall is the section of classified examples as class X, among all examples which truly have class X i.e how much portion of capture class.

Recall= True Positive(TP)/ (True Positive(TP) + False negative(FN))

Precision is the proportion of true positive predictions out of all positive predictions made by the model.

Precision=True Positive/(True Positive(TP) +False Positive(FP))

F-Measure: The harmonic mean of precision and recall. It is an important measure as it gives equal importance to precision and recall.

F-measure=2*recall*precision/precision + recall

Table-1 shows the comparison of various classifiers where test mode cross validation is considered for 5 folds and as well as for 10 folds.

Table-1 Different Classifiers Overall Evaluation Summary

Classifiers	Test Mode (Cross Validation)	Precision	Recall	F-Measure	Correctly Classified Instances	Incorrectly Classified Instances
J48	5 Folds	0.664	0.654	0.657	140	74
	10 Folds	0.670	0.668	0.668	143	71
LMT	5 Folds	0.689	0.682	0.682	146	68
	10 Folds	0.685	0.687	0.683	147	67
Random Forest	5 Folds	0.793	0.794	0.791	170	44
	10 Folds	0.794	0.799	0.793	171	43

Conclusion

As the same dataset is applied on different classifier with cross validation of different folds. Precision, Recall and F-measure values are high in case of Random Forest Classifier. It is observed that Correctly Classified Instances are 79% in case of Random Forest classifier. So the accuracy is high in RF classifier than J48 Classifier and LMT Classifier. Different other algorithms can also be applied from WEKA tool as per the prerequisite of the problem to get speedy and more accurate results

References

1. Minyechil Alehegn, R. J. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. International Research Journal of Engineering and Technology (IRJET).
2. Nagaparameshwara chary, S. D. (2017). A Survey on Comparative Analysis of Decision Tree Algorithms in Data Mining. International Conference on Innovative Applications in Engineering and Information Technology(ICIAEIT-2017).
3. Milos Ilic, Petar Spalevic and Mladen Veinovic, Wejdan Saed Alatresh, "Students' success prediction using Weka tool", in INFOTEH JAHORINA Vol. 15, March 2016.
4. J.V. Kanu Patel, (2014, February). Comparison of Different Classification algorithms on iris datasets using Weka. International Journal of Advance Engineering and Research Development(IJAERD), 1(1)
5. V. Rao and J. Sachdev, "A machine learning approach to classify news articles based on location," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, 2017, pp. 863-867
6. Shakya, D. S. (2020, March 26). Analysis of Artificial Intelligence based Image Classification Techniques. Journal of Innovative Image Processing, 2(1), 44–54.
7. Lakshmi Devasena, C. 2014. Competency Assessment between JRip and Partial Decision Tree Classifiers for Credit Risk Estimation. International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4 (5), (May – 2014), 164-173.
8. Lakshmi Devasena, C. 2014. Effectiveness Assessment between Sequential Minimal Optimization and Logistic Classifiers for Credit Risk Prediction. International Journal of Application or Innovation in Engineering & Management, Volume3, Issue 4, (April 2014), 55 - 63.
9. storm.cis.fordham.edu/~gweiss/data-mining/weka-data/glass.arff
10. Payal P.Dhakate, Suvarna Patil, K. Rajeswari, Deepa Abin, "Preprocessing and Classification in WEKA Using Different Classifier", Int. Journal of Engineering Research and Applications, Vol. 4, Issue 8(Version 5), August 2014, pp- 91-93
11. Sunita B. Aher, Lobo L.M.R.J., "COMPARATIVE STUDY OF CLASSIFICATION ALGORITHMS", International Journal of Information Technology and Knowledge Management, July-December 2012, Volume 5, No. 2, pp. 239-243
12. Deepali Kharche, K. Rajeswari, Deepa Abin, "COMPARISON OF DIFFERENT DATASETS USING VARIOUS CLASSIFICATION TECHNIQUES WITH WEKA", International Journal of Computer Science and Mobile Computing, Vol.3 Issue.4, April- 2014, pp 389-393
13. S. S. Aman, Kumar Sharma, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis," International Journal on Computer Science and Engineering, vol. 3, no. 5, (2011). Pp 1890-1895.