ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 8, October 2013

Comparative Study on Classification Meta Algorithms

Dr. S. Vijayarani[1] Mrs. M. Muthulakshmi[2]

Assistant Professor, Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore, Tamilnadu, India[1]

M.Phil Research Scholar, Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore, Tamilnadu, India[2]

ABSTRACT: Data mining is one of the most important research areas in the field of computer science. Data mining techniques are used for extracting the hidden knowledge from the large databases. There are various research domains in data mining such as image mining, text mining, sequential pattern mining, web mining, and so on. The purpose of text mining is to process unstructured information, extract meaningful numeric indices from the text and thus make the information contained in the text accessible to the various data mining algorithms. There are various methods in text mining such as information retrieval, document similarity, information extraction, clustering, classification, and so on. Searching of similar documents has an important role in text mining and document management. Classification is one of the main tasks in document similarity. It is used to classify the documents based on their category. In this research work, we have analyzed the performance of three Meta classification algorithms namely Attribute Selected Classifier, Filtered Classifier and LogitBoost. These algorithms are used for classifying computer files based on their extension. For example – pdf, txt, doc, ppt, xls and so on. The performances of Meta algorithms are analyzed by applying performance factors such as classification accuracy and error rate. From the experimental results, it is analyzed that LogitBoost performs better than other algorithms.

Keywords: Data mining, Text mining, Classification, AttributeSelectedClassifier, Filtered Classifier, LogitBoost.

I. INTRODUCTION

Text mining or knowledge discovery from text (KDT) deals with the machine supported analysis of text. It uses methods from information retrieval, information extraction and natural language processing (NLP) and also connects them with the algorithms and methods of Knowledge discovery of data, data mining, machine learning and statistics. Current research in the area of text mining tackles problems of text representation, classification, clustering, or the search and modelling of hidden patterns. [5]

Text mining is used to describe the application of data mining techniques to automated discovery of useful or interesting knowledge from unstructured or semi-structured text. Text mining is the procedure of synthesizing the information by analysing the relations, the patterns, and the procedures among textual data semi-structured or unstructured text. Text mining, sometimes alternately referred to as text data mining refers to the process of deriving high-quality information from text. High quality information is typically derived through the divining of patterns and trends through means such as statistical pattern learning. [6] Text mining involves the process of structuring the input text (usually analyzing, along with the addition of some derived linguistic features and the removal of others and subsequent insertion into a database) deriving patterns within the structured data and finally evaluation and interpretation of the output.

Some of the important applications of text-mining include Enterprise Business Intelligence, Data Mining Competitive Intelligence, E-Discovery, National Security, Intelligence Scientific discovery especially Life Sciences, Records Management, Search or Information Access and Social media monitoring. [13] Some of the technologies that have been developed and can be used in the text mining process are information extraction, concept linkage, summarization, categorization, clustering, topic tracking, information visualization and question answering.

The rest of this paper is organized as follows. Section 2 discusses the review of literature. Section 3 describes the classification Meta techniques and the various algorithms used for classification. Experimental results are analyzed in Section 4 and Conclusion are given in Section 5.

II. LITERATURE REVIEW

P. Kalaiselvi et al [7] discussed the performance of the different classifier methods like Bagging, Dagging, Decorate, Multi Class Classifier, and MultiboostAB are compared. Bagging is best algorithm to finding the accuracy than other algorithms. In this experiment Robot Navigation datasets are used and the classification accuracy and time is calculated by 10-fold validation methods. In future the same experiments will conduct with different datasets instead of multiple dataset, MULTICLASS and combine few ensembles with the different base classifier to study how the ensemblers combined with the base classifiers boost the performance accuracy.

Nikita Bhatt et al [10] discussed the different approaches of Meta learning based on dataset characteristics provides a system that automatically provides ranking of the classifiers by considering different characteristics of datasets and different characteristics of classifiers after the generation of the Meta Knowledge Base, Ranking is provided based on Adjusted Ration of Ratio (ARR) or accuracy or time that helps non-experts in algorithm selection task.

Pfahringer et al [14] presented a novel meta-feature generation method in the context of meta-learning, which is based on procedures that compare the performance of individual base learners in a one-to-one manner. In addition to these new meta-features, a new meta-learner called Approximate Ranking Tree Forests (ART Forests) that performs very competitively when compared with several state-of-the-art meta-learners. The experimental results are based on a large collection of datasets and show that the proposed new techniques can improve the overall performance of meta-learning for algorithm ranking significantly. A main point in this approach is that each performance figure of any base learner for any specific dataset is generated by optimizing the parameters of the base learner separately for each dataset.

Artur Ferreira et al [3] presented an overview of boosting algorithms to build ensembles of classifiers. The basic boosting technique and its variants are addressed and compared for supervised learning. The extension of these techniques for semi-supervised learning is also addressed. For face detection, boosting algorithms have been the most effective of all those developed so far, achieving the best results.

III. METHODOLOGY

Text classification is one of the important research issues in the field of text mining where the documents are classified with supervised knowledge. In this research work, computer files can be classified based on their extension. For Example – pdf, doc, ppt, xls and so on. The main objective of this research work is to find the best classification algorithm among Attribute Selected Classifier, Filtered Classifier and LogitBoost. The methodology of the research work is as follows:

1. Dataset – Computer Files can be collected from the system hard disk.

2. Classification Meta Algorithms
   - Attribute Selected Classifier
   - Filtered Classifier
   - LogitBoost

3. Performance factors
    - Classification accuracy
    - Error rate

4. Best Technique among classification Meta algorithms
    - LogitBoost

A. DATASET

A synthetic dataset can be collected from the computer systems which are stored in the hard disk. This dataset contains 9000 instances and four attributes namely file name, file size, extension and file path. Weka data mining tool is used for analyzing the performance of the classification algorithms.

B. CLASSIFICATION META ALGORITHMS

Classification is an important data mining technique with broad applications. It is used to classify each item in a set of data into one of predefined set of classes or groups. Classification algorithm plays an important role in document classification. There are various Meta classification algorithms such as AttributeSelectedClassifier, Bagging, Decorate, Vote, FilteredClassifier, LogitBoost, END, Dagging, Rotation Forest, and so on. In this research work, we have analyzed three Classification Meta Algorithms. The algorithms are namely AttributeSelectedClassifier, Filtered Classifier and LogitBoost.

C. ATTRIBUTE SELECTED CLASSIFIER

Dimensionality of training and test data is reduced by attribute selection before being passed on to a classifier. Some of the important options in attribute selected classifier are as follows

- Classifier -- The base classifier to be used.

- Debug -- If set to true, classifier may output additional info to the console.
- Evaluator -- Set the attribute evaluator to use. It is used during the attribute selection phase before the classifier is invoked.
- Search -- Set the search method**.** This method is used during the attribute selection phase before the classifier is invoked.

D. FILTERED CLASSIFIER

This Class is used for running an arbitrary classifier on data that has been passed through an arbitrary filter. Similar to classifier, the structure of the filter is based exclusively on the training data and test instances will be processed by the filter without changing their structure. Some of the important options in Filtered classifier are as follows

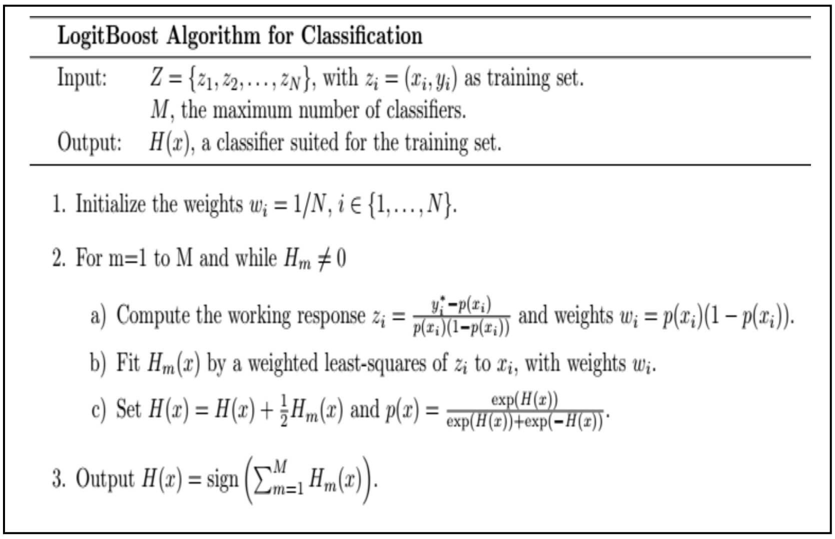
- Classifier -- The base classifier to be used.

- Debug -- If set to true, classifier may output additional info to the console.

- Filter -- The filter to be used.

E. LOGITBOOST

LogitBoost algorithm is an extension of Adaboost algorithm. It replaces the exponential loss of Adaboost algorithm to conditional Bernoulli likelihood loss. This Class is used for performing additive logistic regression. This class performs classification using a regression scheme as the base learner, and can handle multiclass problems.



```
LogitBoost Algorithm for Classification
Input:  $Z = \{z_1, z_2, \ldots, z_N\}$, with $z_i = (x_i, y_i)$ as training set.
        $M$, the maximum number of classifiers.
Output: $H(x)$, a classifier suited for the training set.

1. Initialize the weights $w_i = 1/N, i \in \{1, \ldots, N\}$.
2. For m=1 to M and while $H_m \neq 0$
    a) Compute the working response $z_i = \frac{y_i^* - p(x_i)}{p(x_i)(1-p(x_i))}$ and weights $w_i = p(x_i)(1 - p(x_i))$.
    b) Fit $H_m(x)$ by a weighted least-squares of $z_i$ to $x_i$, with weights $w_i$.
    c) Set $H(x) = H(x) + \frac{1}{2}H_m(x)$ and $p(x) = \frac{\exp(H(x))}{\exp(H(x))+\exp(-H(x))}$.
3. Output $H(x) = \text{sign}\left(\sum_{m=1}^{M} H_m(x)\right)$.
```

IV. EXPERIMENTAL RESULTS

A. Accuracy And Error Rate

There are various measures used for classification accuracy such as true positive rate, precision, F Measure, ROC Area, and kappa Statistics. The TP Rate is the ratio of play cases predicted correctly cases to the total of positive cases. F Measure is a way of combining recall and precision scores into a single measure of performance. Precision is the proportion of relevant documents in the results returned. ROC Area is a traditional to plot the same information in a normalized form with 1-false negative rate plotted against the false positive rate

Table I
Accuracy Measures For Classification Meta Algorithms

Parameters	**Attribute Selected Classifier**	**Filtered Classifier**	**LogitBoost**
Correctly Classified Instances	95.44	97.12	97.91
Incorrectly Classified Instances	4.56	2.88	2.09
TP Rate	95.40	97.10	97.90
Precision	95.30	95.40	98.10
F Measure	94.90	96.10	97.70
ROC Area	99.00	99.80	99.90
Kappa Statistics	94.25	96.37	97.37

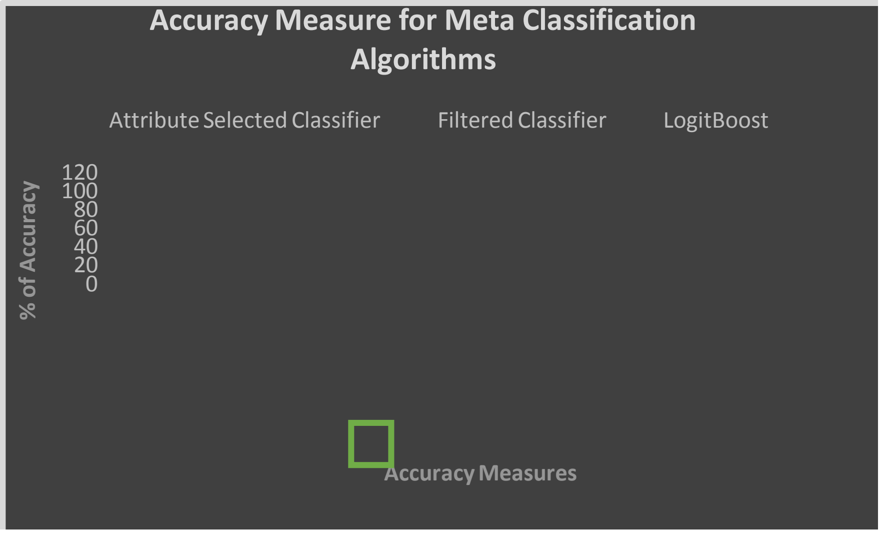


Figure 1: Accuracy Measure for Classification Meta Algorithms

From the above graph, it is analyzed that the LogitBoost algorithms performs better than the other algorithms. Therefore the LogitBoost classification algorithm performs well because it contains highest accuracy when compared to Attribute Selected Classifier and Filtered Classifier.

B. Error Rate

They are the mean absolute error (M.A.E), root mean square error (R.M.S.E), relative absolute error (R.A.E) and root relative squared error (R.R.S.R) [10]. The mean absolute error (MAE) is defined as the quantity used to measure how close predictions or forecasts are to the eventual outcomes. The root mean square error (RMSE) is defined as frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed. Relative error is a measure of the uncertainty of measurement compared to the size of the measurement. The root relative squared error is defined as a relative to what it would have been if a simple predictor had been used.

Table II
Error Rate For Classification Meta Algorithms

Algorithm	MAE	RMSE	RAE	RRSE
Attribute Selected Classifier	0.46	5.41	6.69	29.11
Filtered Classifier	0.31	3.94	4.45	21.19
LogitBoost	0.29	3.56	4.18	19.13

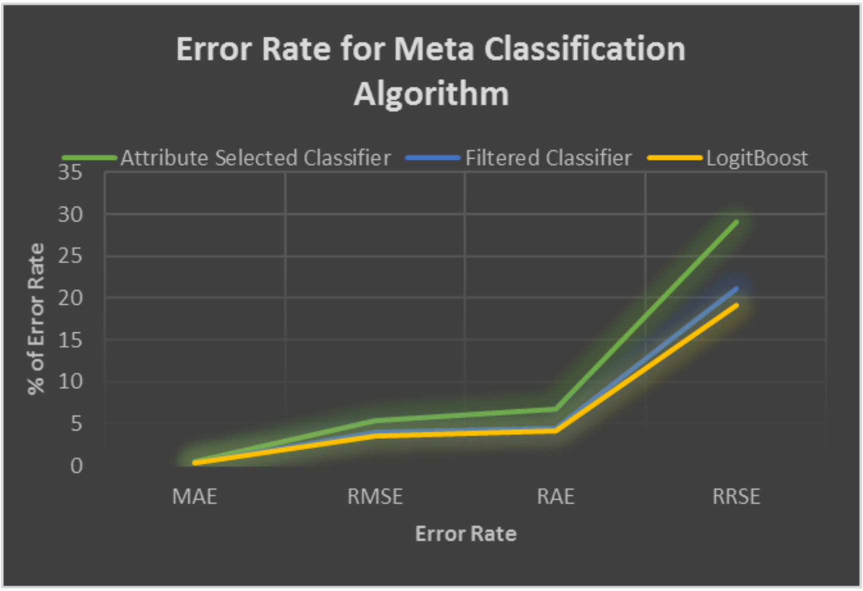


Figure 2: Error Rate for Classification Meta Algorithms

From the above graph, it is analyzed that the LogitBoost algorithms performs better than the other algorithms. Therefore the LogitBoost classification algorithm performs well because it attains lowest error rate when compared to Attribute Selected Classifier and Filtered Classifier.

V. CONCLUSION

Data mining can be defined as the extraction of useful knowledge from large data repositories. Text mining is a technique which extracts information from both structured and unstructured data and also finding patterns which is novel and not known earlier. In this paper, the classification meta algorithms are used for classifying computer files which are stored in the computer. The Classification Meta algorithms include three techniques namely Attribute Selected Classifier, Filtered Classifier and LogitBoost. By analyzing the experimental results it is observed that the LogitBoost classification technique has yields better result than other techniques.

REFERENCES

[1]. Abdullah Wahbeh H, Mohammed Al-Kabi., “Comparative Assessment of the Performance of Three WEKA Text Classifiers Applied to Arabic Text”, Vol. 21, No. 1, pp. 15- 28, 2012.

[2]. Abdullah Wahbeh H, Qasem Al-Radaideh A, Mohammed Al-Kabi N, and Emad Al-ShawakfaM., “A Comparison Study between Data Mining Tools over some Classification Methods”.

[3]. Artur Ferreira., “Survey on Boosting Algorithms for Supervised and Semi-supervised Learning”.

[4]. Christophe Giraud-Carrier., “Meta learning - A Tutorial”.

[5]. Christoph Goller, Joachim Löning., Thilo Will, Werner Wolff., “Automatic Document Classification: A thorough Evaluation of various Methods”

[6]. Falguni Patel N, Neha Soni R., “Text mining: A Brief survey”, Volume-2 Number-4 Issue December-2012.

[7]. Ian Witten H, Eibe Frank, Mark Hall A., “Data Mining Practical Machine Learning Tools and Techniques”.

[8]. Kalaiselvi P, Nalini C., "A Comparative Study of Meta Classifier Algorithms on Multiple Dataset", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013.

[9]. Kaushik Raviya H, Biren Gajjar., "Performance Evaluation of Different Data Mining Classification Algorithm Using WEKA".

[10]. Mahendra Tiwari, Manu Bhai Jha, OmPrakash Yadav., "Performance analysis of Data Mining algorithms in Weka", IOSR Journal of Computer Engineering (IOSRJCE), ISSN: 2278-0661, ISBN: 2278-8727 Volume 6, Issue 3, PP 32-41, (Sep-Oct. 2012).

[11]. Nikita Bhatt, Amit Thakkar, Amit Ganatra., "A Survey & Current Research Challenges in Meta Learning Approaches based on Dataset Characteristics", Volume-2, Issue-1, March 2012

[12]. Mrs. Sayantani Ghosh, Mr. Sudipta Roy, Prof. Samir Bandyopadhyay K., "A tutorial review on Text Mining Algorithms, Vol. 1, Issue 4, June 2012.

[13]. Shaidah Jusoh, Hejab Alfawareh M., "Techniques, Applications and Challenging Issues in Text Mining", Vol. 9, Issue 6, No 2, November 2012.

[14]. Shilpa Dhanjibhai Serasiya, Neeraj Chaudhary., "Simulation of Various Classifications results using WEKA", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-1, Issue-3, August 2012.

[15]. Quan Sun, Pfahringer, "Pairwise meta-rules for better meta-learning-based algorithm ranking Machine learning", Springer US, Machine Learning, 93(1):141-161, 2013.

Dr. S. Vijayarani

She has completed MCA, M.Phil and PhD in Computer Science. She is working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy, security, bioinformatics and data streams. She has published papers in the international journals and presented research papers in international and national conferences.

Mrs. M. Muthulakshmi

She has completed M.Sc in Computer Science and Information Technology. She is currently pursuing her M.Phil in Computer Science in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of interest are data mining, text mining and semantic web mining.