A Comparative Study of Classification Algorithms for Spam Email Data Analysis

Aman Kumar Sharma

Assistant Professor
Department of Computer Science
Himachal Pradesh University, Shimla, India.

Suruchi Sahni

Department of Computer Science
Himachal Pradesh University, Shimla, India.

***Abstract-* In recent years email has become one of the fastest and most economical means of communication. However increase of email users has resulted in the dramatic increase of spam emails during the past few years. Data mining -classification algorithms are used to categorize the email as spam or non-spam. In this paper, we conducted experiment in the WEKA environment by using four algorithms namely ID3, J48, Simple CART and Alternating Decision Tree on the spam email dataset and later the four algorithms were compared in terms of classification accuracy. According to our simulation results the J48 classifier outperforms the ID3, CART and ADTree in terms of classification accuracy**.

Keywords: classification accuracy, ID3, CART, ADTree, J48, WEKA.

I. INTRODUCTION

Nowadays email has become one of the quickest and most inexpensive means of communication. However popularity of email has further increased spam mails during the past years. Data mining classification algorithms are used to categorize the email as spam or non spam. In this paper, we conducted experiment in the WEKA environment by using four algorithms namely ID3, J48 which is the Java implementation of C4.5 version 8, Simple Classification And Regression Tree (CART) and Alternating Decision Tree (ADTree) on the spam email dataset and then the four algorithms were compared. There are few works that compare some of the classification algorithms. Abdelghani Bellaachia, Erhan Guven [1] have used WEKA and have investigated three data mining techniques- the Naïve Bayes, the back-propagated neural network and the C4.5 decision tree algorithms and concluded that C4.5 algorithm has a much better performance than the other two techniques based on their research. In their paper the issues, algorithms and techniques for the problem of breast cancer survivability prediction in SEER database have been discussed and resolved. The paper takes into consideration the Survival Time Recode (STR), the Vital Status Recode (VSR) and Cause of Death (COD).Their future plans are to include records with missing data which they think might increase the performance since the size of the data set will increase considerably. Finally they want to predict survival time by discretizing it in terms of one year of certain cancer data such as respiratory cancer where the survivability is seriously low.

My Chau Tu, Dongil Shin, Dongkyoo Shin [2] in their research have used WEKA and have employed three algorithms to perform classification task to identify the heart disease of a patient which are decision tree C4.5 algorithm, bagging with decision tree C4.5 and bagging with Naïve Bayes. They have used 10-fold cross validation to compute confusion matrix of each model and then evaluate the performance by using precision, recall, Fmeasure and ROC space. They have concluded in their research that bagging algorithms, especially the bagging with Naïve Bayes, showed the best performance. They believe that their results will make clinical application more accessible which shall further provide great help in healing CAD. Their future improvement plans are- firstly they believe that bagging with decision tree and bagging with Naïve Bayes which are quite simple can be used with some more options which can lead to higher results. Secondly, since bagging approach leads to models that are difficult to analyze so they aim at developing a better bagging modeling technique so that the automatically generated knowledge will be easier to understand, and also provide physicians with a novel point of view on the given problem, and may reveal new interrelations and regularities.

In [3] Liangxiao Jiang, Harry Zhang, Zhihua Cai, and Jiang Su have used WEKA and have presented a new learning algorithm called One Dependence Augmented Naive Bayes (simply ODANB). Their aim was to develop a new algorithm to improve Naïve Bayes' performance not only on classification measured by accuracy but also on ranking measured by AUC. They experimentally tested their algorithm, using the whole 36 UCI datasets recommended by Weka and compared it to NB, SBC and TAN. The experimental results showed that their algorithm outperforms all the other algorithms significantly in yielding accurate ranking at the same time outperforms all the other algorithms slightly in terms of classification accuracy.

In [4] Patil BM, Joshi RC, Toshniwal D, Biradar S. discusses about the prediction of burn patient survivability is a difficult problem to investigate till present times. In present study a prediction Model for patients with burns was built, and its capability to accurately predict the survivability was assessed. We have compared different data mining techniques to asses the performance of various algorithms based on the different measures used in the analysis of information pertaining to medical domain. Obtained results were evaluated for correctness with the help of registered medical practitioners. The dataset was collected from SRT (Swami Ramanand Tirth) Hospital in India, which is one of the Asia's largest rural hospitals. Dataset contains records of 180 patients mainly suffering from burn injuries collected during period from the year 2002 to 2006. Features contain patients' age, sex and percentage of burn received for eight different parts of the body. Prediction models were developed through rigorous comparative study of important and relevant data mining classification techniques namely, navie bayes, decision tree, support vector machine and back propagation. Performance comparison was also carried out for measuring unbiased estimate of the prediction models using 10-fold cross-validation method. Using the analysis of obtained results, Navie bayes is the best predictor with an accuracy of 97.78% on the holdout samples, further, both the decision tree and support vector machine (SVM) techniques demonstrated an accuracy of 96.12%, and back propagation technique resulted in achieving accuracy of 95%.

In another work Zhang H. and Su J [5] compared the ranking performance of NB and DT (C4.4) classifiers. The experiments conducted with using 15 dataset from UCI data repository [6]. According to the experimental results NB algorithm outperforms the C4.4 algorithm in 8 datasets, ties in 3 datasets and loses in 4dataset. The average AUC of NB is 90.36% which is substantially higher than the average 85.25% of C4.4.Considering these results, authors argue that NB performs well in ranking, just as it does in classification.

In [7] Hu H., Li J., Plank A., Wang H. and Daggard G. in their paper discussed about the rapid development of DNA Microarray technology. Many classification methods have been used for Microarray classification. SVMs, decision trees, Bagging, Boosting and Random Forest are commonly used methods. In this paper they conducted experimental comparison of LibSVMs, C4.5, BaggingC4.5, AdaBoostingC4.5, and Random Forest on seven Microarray cancer data sets. The experimental results show that all ensemble methods outperform C4.5. The experimental results also show that all five methods benefit from data preprocessing, including gene selection and discretization, in classification accuracy. In addition to comparing the average accuracies of ten-fold cross validation tests on seven data sets, we use two statistical tests to validate findings. We observe that Wilcoxon signed rank test is better than sign test for such purpose.

II. ID3, CART, J48, ADTree

The following section describes ID3, CART, J48, ADTree decision tree algorithm briefly.

ID3

ID3 is a simple decision tree learning algorithm developed by Ross Quinlan (1983). In 1975, in a book Quinlan first presented ID3. The Concept Learning System (CLS) algorithm is the basis for the ID3 algorithm. By adding a feature selection heuristic ID3 improves on CLS. The attributes of the training instances are searched through by ID3 and the attribute that best separates the given examples is extracted by it. ID3 stops if the attribute perfectly classifies the training sets; otherwise it recursively operates on the n, where n is the number of possible values of an attribute of the partitioned subsets to get their "best" attribute. A greedy search strategy is used by the algorithm, the best attribute is picked by it and never reconsiders earlier choices by looking back[8].

CART

CART algorithm stands for Classification And Regression Trees algorithm. It is a data exploration and prediction algorithm. In the early 1980s, CART was developed by Leo Breiman, Jerome Friedman and later joined by Richard Olshen and Charles Stone who began work with decision trees when consulting in Southern California. Classification and Regression Trees is a classification method which in order to construct decision trees uses historical data. Then in order to classify new data decision trees so obtained are used. Number of classes must be known a priori in order to use CART. CART uses so called learning sample which is a set of historical data with pre-assigned classes for all observations for building decision trees [9].

J48

J48 is an open source Java implementation of the C4.5 algorithm in the WEKA data mining tool. C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is a software extension and thus improvement of the basic ID3 algorithm designed by Quinlan. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier [10]. For inducing classification rules in the form of Decision Trees from a set of given examples C4.5 algorithm was introduced by Quinlan. C4.5 is an evolution and refinement of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees, rule derivation, and so on. A set of records are given .

ADTree

An Alternating Decision Tree (ADTree) is a machine learning method for classification. The ADTree data structure and algorithm are a generalization of decision tree and have connections to boosting. ADTrees were introduced by Yoav Freund and Llew Mason. However, the algorithm as presented had several typographical errors. Clarifications and optimizations were later presented by Bernhard Pfahringer, Geoffrey Holmes and Richard Kirkby. Implementations are available in WEKA and JBoost. Original boosting algorithms typically used either decision stumps or decision trees as weak hypotheses. As an example, boosting decision stumps creates a set of T weighted decision stumps (where T is the number of boosting iterations), which then vote on the final classification according to their weights. Individual decision stumps are weighted according to their ability to classify the data. Boosting a simple learner results in an unstructured set of T hypotheses, making it difficult to infer correlations between attributes. Alternating decision trees introduce structure to the set of hypotheses by requiring that they build off a hypothesis that was produced in an earlier iteration. The resulting set of hypotheses can be visualized in a tree based on the relationship between a hypothesis and its "parent." Another important feature of boosted algorithms is that the data is given a different distribution at each iteration. Instances that are misclassified are given a larger weight while accurately classified instances are given reduced weight[11].

An alternating decision tree consists of decision nodes and prediction nodes. Decision nodes specify a predicate condition. Prediction nodes contain a single number. ADTrees always have prediction nodes as both root and leaves. An instance is classified by an ADTree by following all paths for which all decision nodes are true and summing any prediction nodes that are traversed. This is different from binary classification trees such as CART or C4.5 in which an instance follows only one path through the tree[12].

III. EXPERIMENTAL RESULTS

In this section we compare the classification accuracy results of the four decision tree algorithms namely ID3, J48, CART and ADTree. The simulations were conducted using a large spam email dataset consisting of 4601 instances having 58 attributes. The UCI dataset has been modified accordingly and used. All simulations were performed using weak machine learning environment which consists of collection of popular learning schemes that can be used for practical data mining. We list below the steps taken to achieve desired results:
Step 1. Firstly, the dataset named Spam email was taken from the UCI machine learning repository http://www1.ics.uci.edu/~mlearn/MLRepository.html. The "spam" concept is diverse: advertisements for products/web sites, make money fast schemes, chain letters, pornography. UCI Machine Learning collection of spam e-mails came from their postmaster and individuals who had filed spam. Their collection of non-spam e-mails came from filed work and personal e-mails, and hence the word 'george' and the area code '650' are indicators of non-spam. These are useful when constructing a personalized spam filter. One would either have to blind such non-spam indicators or get a very wide collection of non-spam to generate a general purpose spam filter. In this dataset the number of instances is 4601 out of which 1813 Spam which is equal 39.4% and the numbers of attributes are 58 out of which 57 are continuous and 1 has nominal class label.

Attribute Information:

The last column of 'spambase.data' denotes whether the e-mail was considered spam (1) or not (0). Most of the attributes indicate whether a particular word or character was frequently occuring in the e-mail. The run-length attributes (55-57) measure the length of sequences of consecutive capital letters. For the statistical measures of each attribute.

Here are the definitions of the attributes:

48 continuous real [0,100] attributes of type word_freq_WORD = percentage of words in the e-mail that match WORD, i.e. 100 * (number of times the WORD appears in the e-mail) /total number of words in e-mail. A "word" in this case is any string of alphanumeric characters bounded by non-alphanumeric characters or end-of-string.

6 continuous real [0,100] attributes of type char_freq_CHAR = percentage of characters in the e-mail that match CHAR, i.e. 100 * (number of CHAR occurences) / total characters in e-mail

1 continuous real [1,...] attribute of type capital_run_length_average = average length of uninterrupted sequences of capital letters

1 continuous integer [1,...] attribute of type capital_run_length_longest = length of longest uninterrupted sequence of capital letters

1 continuous integer [1,...] attribute of type capital_run_length_total = sum of length of uninterrupted sequences of capital letters = total number of capital letters in the e-mail

1 nominal {0,1} class attribute of type spam = denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail.
Missing Attribute Values: None
Class Distribution:

Spam	1813 (39.4%)
Non-Spam	2788 (60.6%)

Then it was later converted into the ARFF (Attribute Relation File Format) which is the WEKA data File Format. This data contained 4601 number of Instances and out of which 1831 are of Spam category (that is 39.4%) and total number of Attributes are 58 out of which 57 are continuous and 1 is nominal class label.

Step 2. Secondly, the WEKA tool version 3.6 which is available from http://www.cs.waikato.ac.nz/ml/weka was downloaded. WEKA is an open source machine learning software.

Step 3. In WEKA environment while preprocessing stage for the dataset as mentioned in the step1 attributes were discretize using supervised attribute filters. Filtering refers to the preprocessing of a data file before simulations start. Certain data mining techniques algorithm especially ID3 algorithm of decision tree technique require all data to be categorical. This involves discretisation of numeric or continuous attributes. WEKA provides a number of filters for both supervised and unsupervised learning. For this experiment we applied supervised learning since the data mining classification strategy comes under supervised learning and we have used classification tasks – Decision Tree Algorithms.

Step 4. When the attributes were discretize using supervised attribute filters for the dataset containing 4601 instances and 58 attributes then we got the WEKA window showing the graphical attribute presentation of spam email data. This is one way among many present of visualizing data using WEKA. At a time for the attribute distribution of a single selected attribute a histogram is shown by the main GUI. Within a histogram individual colours indicate individual classes our data set has two (spam and non-spam) classes. All the distributions are shown for 58 attributes.

Step 5. Then we applied the four selected classification algorithms in the WEKA environment. Over fitting was avoided by evaluating the classification algorithms using 10-fold cross-validation. The performance of each classifier was assessed with a stratified 10-fold cross-validation method. This approach has the advantage that all the data is, at some point, used for model evaluation, as opposed to simply splitting the data into testing and training sets. Instead, the data was divided into 10 equal sized fragments, each of which was in turn used as an independent test set, while the other fractions were used for training the classifier. Then 10 pair of training sets and testing sets are created. Classification error was then estimated as the average performance over the 10 test sets. The classification and feature selection tasks were performed with the help of a freely available software package WEKA, Version 3.6 (WEKA Machine Learning Project, The University of Waikato, New Zealand) and we obtained the following results:

TABLE I CLASSIFICATION ACCURACY TEST RESULTS

Instances(4601) **Algorithms**	**Correctly classified instances**	**Incorrectly classified instances**
ID3	4100(89.1111%)	433(9.411%)
J48	4268(92.7624%)	333 (7.2376%)
ADTree	4183(90.915%)	418 (9.085%)
SimpleCART	4262(92.632%)	339(7.368%)

Table I demonstrate the classification accuracy results of four classification decision tree algorithms. It is evident from the table I that J48 has the highest classification accuracy (92.7624%) where 4268 instances have been classified correctly and 333 instances have been classified incorrectly. The Second highest classification accuracy for CART algorithm is 92.632% in which 4262 instances have been classified correctly. Moreover the ADTree showed a classification accuracy of 90.915%. The ID3 algorithm results in lowest classification accuracy which is 89.111% among the four algorithms. The ID3 was also not able to classify 68 instances. So the J48 outperforms the CART, ADTree and ID3 in terms of classification accuracy.

IV. CONCLUSIONS AND FUTURE RESEARCH

In this research we have performed the experiments in order to determine the classification accuracy of four algorithms in terms of which algorithm better determine whether a particular email is spam or not with the help of an attractive data mining tool known as WEKA. Four algorithms namely ID3, J48, Simple CART and ADTree were compared on the basis of different percentage of correctly classified instances. All these four come under the classification methods of data mining which makes a relationship between a dependent (output) variable and independent (input) variable by mapping the data points. In simple terms, classification problem refers to identifying an object as belonging to a given class for example whether a particular mail is spam or non-spam.

It is clear form the simulation results that the highest classification accuracy performance is for the J48 classifier for the spam email datasets containing 58 attributes with each 4601.

Furthermore Simple CART also showed similar results that were only slightly different from J48. ADTree and ID3 classifiers showed less accuracy as compared to the previous two mentioned. This indicates that J48 classification algorithm should be favored over Simple CART, ADTree and ID3 classifiers in the spam email application where classification accuracy performance is important. In future work we can include the extension of the simulation performed in the WEKA environment by keeping the number of instances same in a given dataset but decreasing the number of attributes and comparing the classification accuracy performance of the proposed algorithms. Moreover, other factor can also be taken for instance the time requirement to compare the accuracy of the proposed algorithms with respect to the decrease in the number of attributes while keeping the number of instances constant which we believe shall surely bring out certain important aspects about the different algorithm which can prove useful in the research field.

REFERENCES

[1] Abdelghani Bellaachia, Erhan Guven, “Predicting Breast Cancer Survivability Using Data Mining Techniques”, accessed online from www.siam.org/meetings/sdm06/workproceed/bellaachia.pdf on Dec 06, 2010.

[2] My Chau Tu, Dongil Shin, Dongkyoo Shin, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms", dasc, pp.183-187, 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009.

[3] Liangxiao Jiang, Harry Zhang, Zhihua Cai, and Jiang Su, “One Dependence Augmented Naive Bayes”, accessed online from citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.437.

[4] Patil BM, Joshi RC, Toshniwal D, Biradar S., “A New Approach: Role of Data Mining in Prediction of Survival of Burn Patients”, accessed online from www.springerlink.com/index/8pnh75n137t99892.pdf.

[5] Zhang H. and Su J., “Naive bayesian classifiers for ranking”, 15th European Conference on Machine Learning, ECML 2004. Accessed online http://www.cs.unb.ca/profs/hzhang/publications/NBRanking.pdf.

[6] UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science. Accessed Online from http://www.ics.uci.edu/~mlearn/MLRepository.html.

[7] Hu H., Li J., Plank A., Wang H. and Daggard G., “A Comparative Study of Classification Methods for Microarray Data Analysis”, In Proc. Fifth Australasian Data Mining Conference, Sydney, Australia (2006).
[8] Laurent Hyafil, RL Rivest, “Constructing Optimal Binary Decision Trees is NP-complete” , Information Processing Letters, Vol. 5, No. 1. (1976), pp. 15-1 R. Kohavi and J. R. Quinlan. Decision-tree discovery.
[9] http:// www.CART-Wikipedia, the free encyclopedia.htm accessed on 14/12/2010.
[10] http://www.c4.5-Wikipedia, the free encyclopedia.htm accessed on 16/12/2010.
[11] Bernhard Pfahringer, Geoffrey Holmes and Richard Kirkby, “Optimizing the Induction of Alternating Decision Trees”, Proceedings of the Fifth Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. 2001, pp. 477-487.
[12] http://en.wikipedia.org/wiki/Alternating_decision_tree.

AUTHORS PROFILE

Mr. Aman Kumar Sharma is working as Assistant Professor at the Department of Computer Science, Himachal Pradesh University, Shimla. He is a post graduate in computer application and perusing his Ph.D. in the area of software engineering. He has fourteen years of teaching experience at post graduate level. His research interest include component based software engineering, data mining and operating systems.

Ms. Suruchi Sahni , M.Tech. from Himachal Pradesh University, Shimla. Her area of research interest is data mining.