Comparison of Machine Learning Algorithms to Classify Web Pages

Ansam A. AbdulHussien
Lecturer, Department of continuous education
University of information technology and communication
Baghdad, Iraq

***Abstract*—The 'World Wide Web', or simply the web, represents one of the largest sources of information in the world. We can say that any topic we think about is probably finding it's on the web. Web information comes in different forms and types such as text documents, images and videos. However, extracting useful information, without the help of some web tools, is not an easy process. Here comes the role of web mining, which provides the tools that help us to extract useful knowledge from data on the internet. Many researchers focus on the issue of web pages classification technology that provides high accuracy. In this paper, several 'supervised learning algorithms' evaluation to determining the predefined categories among web documents. We use machine learning algorithms 'Artificial Neural Networks (ANN)', 'Random Forest (RF)', 'AdaBoost' to perform a behavior comparison on the web pages classifications problem.**

Keywords—Web page classification; artificial neural networks; random forest; adaboost

I. Introduction

In the computer world, data represent an interesting area. It is constantly increasing and expanding exponentially, and it is important for us to find useful information from this massive data. The overall process of analyzing data to find understandable and useful information is called data mining. In the last few years, most enterprise-owned data have been stored in structured data stores such as relational databases [1] These data are easily accessible for exploratory purposes using several data mining techniques. However, the nature of the data has changed dramatically since the advent of the Internet, which has characteristics that make it different from structured data Some of these characteristics are The huge volume on the web and growing exponentially, The web contains various types and formats of data . This includes structured data, such as the table, semi-structured data such as XML documents, unstructured data such as text on web pages, and multimedia data such as images and movies, the incompatibility of information on the Internet urge the researchers from around the world are involved in building web content. As a result. We may find pages with similar or identical content.

In addition, web data have hyperlinks, which means that web pages are linked together so that anyone can navigate through pages within the same site or across different sites that make the Information noise. The reasons for this are two issues. First, the typical web page usually contains much information such as the main body of the page, links, ads, and much more. Thus, the page does not have a specific structure. Second, there is no qualitative control over information, meaning that anyone can upload content on the web, regardless of its quality, finally, A large portion of the content on the web is considered dynamic, meaning that the information is updated frequently and continuously. For example, weather information is updated continuously.

All these characteristics make data extraction on the web more challenging while giving us opportunities to discover useful and valuable knowledge from the web. Because of a wide range of data types, the traditional techniques to extraction data have become inadequate. This has led to the crystallization of a need to develop new techniques and algorithms aimed at data mining on the Internet. The rest of this research paper is structured as follows. In Section 2, Related work is discussed; Section 3 describes Machine Learning Algorithms. Data Collection and preprocessing is discussed in Section 4, Section 5 explained Entropy Term Weighting Schema, Experimental Results and Discussion are shown in Section 6. Concluding remarks are given in Section 7.

II. Related Work

WPC techniques use concepts from many fields like Information filtering and retrieval, Artificial Intelligence, Text mining, Machine learning techniques and so on. In the machine-learning model, a classifier was given training with already classified examples and it learns the rules for classification during this training phase. Then this classifier is used for classification of the new pages. Mach work has been previously begun on WPC Mention some of them:

I. Anagnostopoulos, et al. [2] suggested a system to identify and categorize web pages, based on information filtering. The system is a three layer Probabilistic NN (PNN) having biases and radial basis neurons in the middle layer and competitive neurons in the output layer. This is an eCommerce area study domain. Thus, PNN hopes to identify eCommerce web pages to classify them to respective type based on a framework describing commercial transactions fundamental transactions on the web.

In the same direction, Feng Shen, et al. in [3] proposed a new deep learning based text classification model to solve the problem of Chinese web text categorization of dimension reduction by use away way to learn the data feature from massive data is to use deep learning NN structure. Deep learning network has the excellent feature learning ability. It can combine objects of low-level features to form advanced abstract representations of the object which will be more suitable for classification.

J. Jagani and et al [4] M.S. Othman and et al [5] the authors discuss the result of classifying web documents using the extraction and machine learning techniques. Six web document features have been identified which are text, meta tag and title (A), title and text (B), title (C), meta tag and title (D), meta tag (E) and text (F). The Support Vector Machine (SVM) method is used to classify the web document while four types of kernels namely: Radial Basis Function (RBF), linear, polynomial and sigmoid kernels was applied to test the accuracy of the classification.

E. Sarac and et al [6] introduced that increase in the amount of information on the Web has caused the need for accurate automated classifiers for Web pages to maintain Web directories and to increase search engine performance. Every tag and every term on each Web page can be considered as a feature there is a need for efficient methods to select the best features to reduce the feature space of the WPC problem. The aim is to apply a recent optimization technique, namely the firefly algorithm (FA) to select the best features for Web page classification problem. The firefly algorithm (FA) is a metaheuristic algorithm, inspired by the flashing behavior of fire flies. Using FA to select a subset of features and to evaluate the fitness of the selected features J48 classifier of the Weka data mining tool is employed. Another related work M.Klassen [66]and J. Jagani J. Jagani and K. patel [7] where the authors are extract useful knowledge from large web data ,and handle those data and achieve various functionalities .they going to discuss a new technique that will work on hierarchical as well as multi pass approach that is having the advantages of both multi pass and hierarchical approach by combining the benefits of both and designed a new algorithm, the discussion is based on various neural network learning algorithms that help to handle large web data as well as better classification and clustering of data with less number of errors. Self-Organizing Maps called SOM and Learning Vector Quantization known as LVQ are very constructive learning algorithms that classify and cluster the web data. MLVQ and HLVQ techniques are following a concept of multi pass in which more than one pass can be performed on the same model using different algorithms.

A. Herrouz and et al [8] authors used techniques Apriori Algorithm and implementation of Naive Bayes Classifiers. Apriori Algorithm finds interesting association or correlation relationships among a large set of data items interesting association or correlation relationships among a large set of data items, relationships among huge amounts of transaction records can help in many decision making process and use Naive Bayes Classifier to calculate probability of keywords among a large data item sets The Naive Bayes Classifier uses the maximum a posterior estimation for learning a classifier.

III. Machine Learning Algorithms

A. Artificial Neural Networks (ANN)

An interconnected set of virtual neurons created by software programs similar to the work of a biological neuron or electronic structures (electronic chips designed to simulate the work of neurons) using the mathematical model to process information based on the communicative method of computing. Neural networks generally consist of simple processing elements that do a simple job, but the overall behavior of the network is determined by the connections between these different elements called the neurons and the indicators of these elements [9]. The first suggestion of the idea of neural networks comes from the mechanical action of brain neurons that can be likened to electrical, biological networks to process the information contained in the brain.

Artificial neural networks are composed of nodes called neurons or processing elements, which are connected together to form a network of nodes. Each contact between these nodes has a set of values called weights which contribute to the determination of the values resulting from each processing element based on the input values of that element. Neural network arranged in layers of artificial cells [10]: input layer and output layer and layers between them called hidden layers. Each cell in one of these layers relates to all the neurons in the next layer and all the neurons in the layer preceding it. All connections between a neuron and another is characterized by a value called weighting, it is the importance of the connection between these two elements.

The neurons multiply each input value from the previous layer neurons with the weight of the communication by these neurons and then multiply the multiplication outcomes, The conversion is different with neuron type, the output of the transformation considers the output of the neuron which is transferred to the neurons of the next layer. The feed forward-back propagation neural network is adapted as the classifiers. The activation of these input units is propagating forward through the network, and finally, the value of the output unit determines the categorization decisions.

B. Random Forest

Random Forest [11] is the one of a Machine Learning Algorithm work as a large collection of the correlated decision tree. The random forest lies in one of those Class of Machine Learning Algorithms which does 'ensemble' classification. By Ensemble, Collective Decisions of Different Decision Trees. RF is making a prediction about the class not based on One Decision Tree, but by an (almost) Unanimous Prediction, made by Decision Trees.

The training algorithm for random forests applies the general method of bootstrap aggregating, or bagging, Each tree is trained on a bootstrapped sample at each node of training data, the algorithm searches across a random a subset of the variables to determine a split, this procedure leads to best model performance because it decreases the variance of the model, without increasing the bias. This means that the predictions of a single tree are very sensitive to noise in its training set, To classify an input feature vector in random forests, the vector is submitted as an input to each of the trees in the forest. Each tree gives class and it is said that the trees vote for that class. In the classification time, the forest chooses the class that has the most votes.

C. AdaBoost

AdaBoost, abbreviated "Adaptive Boosting", is a "machine learning meta-algorithm" produced by Yoav Freund and Henry Martyn Robert Schapire, it is a type of "Ensemble Learning" where varied learners are employed to construct a stronger learning algorithm. AdaBoost is one of the most efficient supervised learning algorithms of the last years [12]. it has to be inspired learning theoretical developments and also provided solid theoretical foundation, very accurate prediction, great simplicity building and easily explainable modeling that proved successful in wide applications, It is used in most cases with several alternative typs of learning algorithms to enhance their performance by which combined the output of the other learning algorithms ('weak learners') into into a weighted addition that represents the ultimate output of the boosted classifier, but it is sensitive to noising data and outliers in some cases, it may be less oversensitive to the overfitting problem than other learning algorithms.

AdaBoost works by choosing a base algorithm (e.g. Decision trees) and iteratively improving it by accounting for the incorrectly classified examples in the training set.

IV. Data Collection and Preprocessing

Data collection is a process used to crawling the web page from a website, We can get updated data and maintain this document for later processing by storing in the database. In this paper, The type of data in the database is a health data includes Diseases information, the dataset used in this experiment collecting randomly from search engines.

The web pages consist varied information like nouns, stop words, navigation, image, link structure, Advertisement, and lay out all of these are unnecessary in classification in order to achieve a standard representation of all documents must be applied preprocessing.The preprocessing contains three steps (lexical analysis, string tokenizer, stop words elimination and stemming) as Fig. 1.

The lexical URL analysis technique used to enhance the classification accuracy of web classifiers, the system uses the tools of lexical analysis to reduce the amount of data by removing any Useless word and can get the Stream of useful words that could be used in the next steps of the preprocessing. Identify words in the plain text using, tokenization, which processes represented each word as a token.

After the string tokenizer process is applied the stop words, elimination is done, which means the pronouns, prepositions like “to”, “the”, etc. and conjunctions are removed from the document because these words don’t have any meaning or indications about the content.Thus affecting the quality of classification, Finally, The last part of preprocessing is stemming, Stemming is a technique used to minimize the words to their grammatical roots. The stemming process is applied to remove suffixes like “Ed”, “ing”, “ly”, etc. by removing these suffixes we can reduce the terms in the document and lessen the complexity and also it is necessary to do text analysis to make information retrieval efficient in especially in data mining applications.

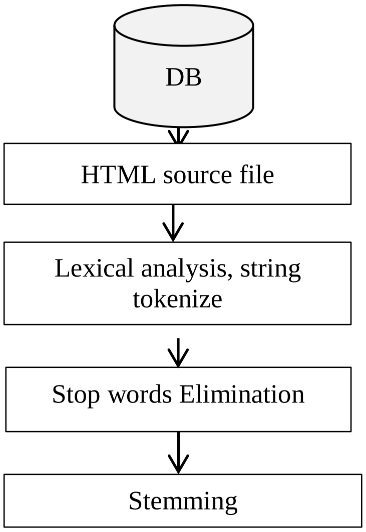

Fig. 1. Preprocessing steps for web page.

After the preprocessing of all web pages document,the database has been created and contains all the unique words in each page. The number of unique words (25769) represents distinctive words that appeared several times in the page. Each word represents one feature vector. This feature vector contains the document terms weight. The Term weight calculated by using the entropy term weighting scheme

V. Entropy Term Weighting Scheme

A process that identifies most regular words in each class or category, as well as calculating the weights of them by implementing the term weighting scheme, Entropy method is based on a probabilistic analysis of the texts. It provides a more accurate weight. Calculate term weighting from two aspects which are local term **Ljk** weighting and global **Gk**. Entropy term weighting scheme on each term is calculated as **Ljk** x **Gk** calculate from two formulas [13]

$$L_{jk} = \begin{cases} 1 + \mathrm{Log}\, TF_{JK} & (TF_{JK} > 0) \\ 0 & (TF_{jk=0}) \end{cases} \qquad (1)$$

And

$$G_k = \frac{1 + \sum_{k=1}^{n} \frac{TF_{jk}}{F_k} \ \mathrm{Log} \frac{TF_{jk}}{F_k}}{\log n} \qquad (2)$$

(n) is the number of documents in a database and TF_(jk)Is the term frequency of each word in Doc j as The F_(k) Is a frequency of the term k in the entire document collection. Term weight input to principal components algorithm"PCA" which use to reduce the original data vectors to a small number of relevant features and it calculates the eigenvectors of the covariance matrix, after that projects the testing data onto a lower dimensional feature space which is defined by the eigenvectors

VI. Experimental Results and Discussion

The experiments were carried out to show the effectiveness of these algorithms and difference between them .The data type used is web pages that are randomly downloaded from search enginesyahoo, google, etc. and divided to eight categories, namely Respiratory, Hearts, liver, cancer, diabetes, dermatology, joint, digestive, 269 web page was used contain information about this Diseases, this dataset divided into 70% for training phase and 30% for testingIt was evaluated the classification performance using the standard information retrieval measures (F. Measures), it considers both the precision and the recall as below:

$$Precision = TP / (TP + FP) \quad (3)$$

$$Recall = TP / (TP + FN) \quad (4)$$

$$F.Measures = 2{*}TP /(2{*}TP + FP + FN) \quad (5)$$

Where the precision for a class is the number of true positives (i.e. The number of pages correctly labeled as belonging to the positive class) divided by the total number of pages labeled as belonging to the positive class (i.e. The sum of true positives and false positives which are pages incorrectly labeled as belonging to the class).

Recall in this context is defined as the number of true positives divided by the total number of pages that actually belong to the positive class (i.e. The sum of true positives and false negatives, which are pages which were not labelled as belonging to the positive class but should have been). The experiments were conducted using three algorithms and test the accuracy and efficiency for each of them; results are shown in Table 1 and Fig. 2:

TABLE I. Comparison Results of Three Classifiers

Algorithm	Precision	Recall	*F.Measures*
Random forest	92.94%	82.24%	87.26%
Artificial neural networks	90.33%	79.93%	84.82%
AdaBoost	82.6%	82.4%	81.7%

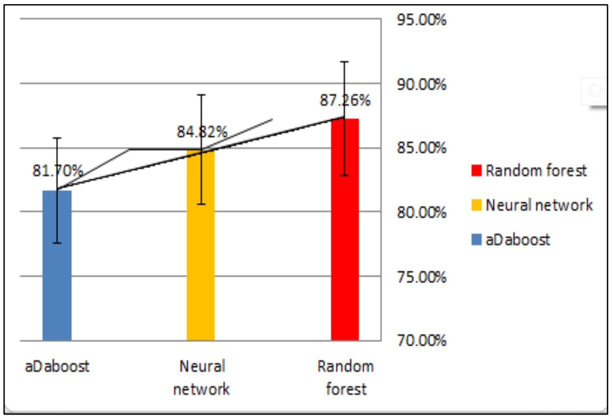

Fig. 2. Performance of RF, NN and adaBoost algorithm.

The results above show that the random forest has a higher precision than neural network and ADABOOST The value of F. Measures are also greater from Other in the classification. But the NN has the best competitive performance with ADABOOST. A bagging algorithms model rather than a boosting algorithm. A too complex model has a low bias, but large variance, while a too simple model has low variance, but large bias, both leading a high error, but two different reasons. One powerful modeling algorithm that makes good use of bagging is Random Forests. Random Forests work by training numerous decision trees each based on a different resampling of the original training data. In Random Forests the bias of the full model is equivalent to the bias of a single decision tree (which itself has high variance). By creating many of these trees, in effect a "forest", and then averaging them the variance of the final model can be greatly reduced over that of a single tree. In practice the only limitation on the size of the forest is computing time as an infinite number of trees could be trained without ever increasing bias and with a continual (if asymptotically declining) decrease in the variance.

AdaBoost weak learners have high bias and low variance. To reduce the bias of a large number of 'small' models with low variance by building up one learner at the top of another, the boosting ensemble tries to decrease the bias, for a little variance. While the result of random forest nearly with a neural network, both algorithms have strength and weakness points. Here we are going to focus on the positive side of random forest compared to the neural network. The RF faster, usually finishes within minutes and it Easier to train, RF has less hyper-parameter to tune, In contrast, in NN have huge numbers of parameters to choose from like the number of layers, a number of neurons in each layer, activation function, learning rate, etc.

VII. Conclusion

From the testing phase, the accuracy of RF classifier is 87.26 %, NN is 84.82% and AdaBoost is 81.7 % According these results, we found that RF can classify more accurately than the ANN and AdaBoost classifiers. The value of F1 is greater from Other when classifying the pages. But the ANN has the best performance compared with AdaBoost. Finally, RF Can handle small data most NN architectures require big data to generalize very well, and the number of documents should be a much larger from a number of features. While RF could give a proper accuracy with a small number of documents even with too many features.

References

[1] T. Bourgeois, " Information Systems for Business and Beyond", Edition, Textbook Equity, Saylor Academy, 2014.

[2] I. Anagnostopoulos, C. Anagnostopoulos, V. Loumos and E. Kayafas, “Classifying Web pages employing a probabilistic neural network”, IEE Proceedings-Software, Vol.151, No.3, June 2004, PP.139 - 150.

[3] F.Shen, X.Luo and Yi.Chen, “Text Classification Dimension Reduction Algorithm for Chinese Web Page Based on Deep Learning”, International Conference on Cyberspace Technology (CCT 2013), pp. 451 – 456, Beijing, China, 23 Nov. 2013.

[4] M.S.Othman, L.M Yusuf and J. Salim, “Web classification using extraction and machine learning techniques”, In Information Technology (ITSim), 2010 International Symposium in Vol 2, PP. 765 -770, Kuala Lumpur, 15-17 June 2010.

[5] E. Sarac and S.A.Ozel, “Web page classification using firefly optimization”, In Innovations in Intelligent Systems and Applications (INISTA), 2013 IEEE International Symposium on, PP.1-5, Albena, Bulgarian, 19-21 June 2013.

[6] M.Klassen, “A frame work for search forms classification” In Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on, PP.1029-1034 Seoul, Korea, 14-17 Oct. 2012.

[7] J. Jagani and K. patel, “An Enhanced Approach for Classification in Web Usage Mining using Neural Network Learning Algorithms for Supervised Learning”, International Journal of Computer Applications, Vol 90, No 1, March 2014, pp.25-30.

[8] K. J. Patel and K.J. Sarvakar, “Web Page Classification Using Data Mining”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, No.7, July 2013, pp. 2513- 2520.

[9] J . Heaton, " Artificial Intelligence for Humans: Deep learning and neural networks", Vol 3, Heaton Research, Incorporated, 323 pages, 2015.

[10] N. Gupta," Artificial Neural Network", Network and Complex Systems, Vol 3, no. 1, pp. 24-28, 2013.

[11] T.K Ho, "Random decision forests." In Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on, vol. 1, pp. 278-282. IEEE, 1995.

[12] Y. Freund and R. E. Schapire."A decision-theoretic generalization of on-line learning and an application to boosting." In a European conference on computational learning theory, pp. 23-37. Springer, Berlin, Heidelberg, 1995.

[13] Z.S. Lee, M. A.Maarof, A. Selamat and S. M. Shamsuddin, " Enhance Term Weighting Algorithm as Feature Selection Technique for Illicit Web Content Classification ", The 8th International Conference on Intelligent Systems Design and Applications, pp.145–150 , Kaohsiung, Taiwan ,IEEE, 26-28 Nov. 2008 .