View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by UUM Repository

Second International Conference on Computer and Network Technology

Comparison of attribute selection methods for web texts categorization

Rizauddin Saian
Fakulti Sains Komputer dan Matematik,
Universiti Teknologi MARA Perlis,
02600 Arau, Perlis, Malaysia.
e-mail: rizauddin@perlis.uitm.edu.my

Ku Ruhana Ku-Mahamud
College of Arts and Sciences,
Universiti Utara Malaysia,
06010 Sintok, Kedah, Malaysia.
e-mail: ruhana@uum.edu.my

***Abstract*—This paper presents a study on the performance of attribute selection methods to be used with Ant-Miner algorithm for web text categorization. The new generated data set by each attribute selection method was classified with Ant-Miner to see the performance in terms of predictive accuracy and the number of rules generated. The results of classification were also compared to C4.5 algorithm.**

***Keywords*—web mining; classification, attribute selection; machine learning.**

I. INTRODUCTION

According to a web survey by Netcraft [1], there are about 206 millions web sites in January 2010, and thus contributing to the enormous size of information on the web. As a consequence, information searching from the web is not an easy task.

Web directories are web sites that list other web sites according to category and subcategory. Categorizing web documents into web directories could facilitate information retrieval. Before committing a search, the user will select a specific category, such as “Arts”, “Business”, “Computers”, or “Sports”, resulting in more accurate and related information to the search engine results. DMOZ Open Directory Project (ODP) (http://www.dmoz.org) is an example of web directory which are mainly constructed by a huge number of human editors. However, as the web is constantly changing and expanding, this manual approach will sooner became less effective. Hence, due to the enormous size of the web, it would be good to have an automatic classier that will categorized web pages, to help the development of a web directory.

An intelligent computer web document classification algorithm could assist in building the web directories. Classification is a data mining task of assigning objects to one of several predefined categories. It is an important task in many information management and retrieval tasks on the web. Examples include helping Web spider to focus crawl, improve the quality of Web search, and assisting in the development of Web directories.

One of the emerging algorithm to classify web text documents is the Ant-Miner. Ant-Miner [2], an ant colony algorithm variant for learning classification rules accuracies was used to categorize web texts in a study done by Holden & Freitas (2004). Holden & Freitas had found that the results of using Ant-Miner was comparable to C5.0. Ant-Miner performed better than the C5.0 in term of knowledge comprehensibility. C5.0 is a commercial data mining tool originated from C4.5 [3] for discovering patterns that define categories, assembling them into classifiers, and using them to make predictions.

In web text categorization problem, attributes are the words that occur in the web pages or are also called documents. The number of attributes could be tens or even hundreds of thousands, even for small size web pages. According to the study conducted by Yang & Pederson, the text categorization performance will be improved by removing up to 98% of the attributes [4].

Attribute selection is done by searching the space of attributes for a set of attributes that would best predict the class. Several search methods such as Best-First, Exhaustive Search, Genetic Search [5], Greedy Stepwise, Race Search [6] and Random Search [7] are constantly used by researchers for attribute selection activities. The function for each search method is as listed in Tab. I.

TABLE I. SEARCH METHODS FOR ATTRIBUTE SELECTION.

Search Methods	Function
Best-First	Searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility.
Exhaustive Search	Performs an exhaustive search through the space of attribute subsets starting from the empty set of attributes.
Genetic Search	Search using simple genetic algorithm.
Greedy Stepwise	Searches the space of attribute subsets by greedy hill climbing augmented without a backtracking facility.
Race Search	Using race search methodology.
Random Search	Search randomly.

Evaluation method such as Correlation-based attribute subset selection [8], Classifier-based attribute subset selection and Consistency subset [7] were used by many researchers to select the attributes in the process of evaluating the attributes subset found by search method. The functions for each evaluation method are shown in Tab. II.

In this study, the performance of three attributes selection evaluation methods against Ant-Miner and C4.5 has been undertaken. Predictive accuracy and number of rules were used as the matrix to measure performance of the attribute selection methods. Section II describes the method that has been used in this study and the experimental design. Experimental results are presented in section III followed by conclusion in section IV.

978-0-7695-4042-9/10 $26.00 © 2010 IEEE
DOI 10.1109/ICCNT.2010.13

TABLE II. ATTRIBUTE EVALUATION METHODS FOR ATTRIBUTE SELECTION.

Evaluation Methods	Function
Correlation-based	Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.
Classifier-based	Use classifier to evaluate attribute set.
Consistency subset	Measure consistency in class values for a chosen subset of attributes.

II. METHOD AND EXPERIMENTAL SETTING

This study uses the pre-classified web pages of four Universities [9] data sets. The data set contains 8282 manually classified web pages. The categories (class) are student, faculty, staff, department, course and project. For this project, only two categories were chosen, which are student and course, leaving only 2571 web pages. Fig. 1 depicts the whole process of the rules generation.

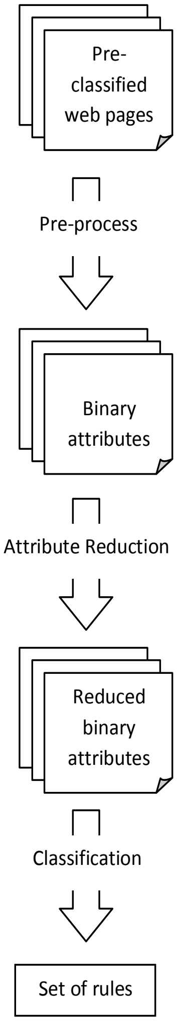

Figure 1: The process of generating rules

From those documents, a set of binary attributes (words) was extracted, leaving the html tags, numbers, stop words and punctuations (pre-process). Stop words are words that give very little contribution or none to the meaning of the text. Examples of stop words are “the”, “and”, “he”. For each attribute, a value of 1 will be given if the attribute occurs in the document and 0 otherwise.

The words “car” and “cars” came from the same word root “car”. Instead of keeping both word “car” and “cars”, the number of words will be reduced using stemming algorithm if only the root word is used. Therefore, each word extracted was stemmed using Porter Stemming algorithm [10]. According to Hull [11], there is no difference between the stemmers in terms of average performance. However, Porter's algorithm is the most common algorithm for stemming English. Attributes that occur less than 100 in the whole set of documents were also removed, and for each category, only 20 attributes was chosen.

Three attribute evaluation methods with seven search methods were used to reduce the number of attributes. The evaluation methods are Correlation-based attribute subset selection [8], Classifier-based attribute subset selection and Consistency subset [7]. The search methods used include Best-First, Exhaustive Search, Genetic Search [5], Greedy Stepwise, Race Search [6] and Random Search [7]. The attribute selection was done using Weka software [12] which generates a number of new sets of data. Tab. III shows the number of attribute selected for each attribute selection run.

Finally, for each sets generated by the attribute selection, classification of documents was performed. The results were compared with C4.5 algorithm.

A *k*-fold cross validation procedure (Fig. 2) was used to measure the accuracy of the discovered rules. The experiment consists of *k* folds (iterations), where each fold uses a different set of data as a test set. Ten different set of training and test data were generated randomly for each fold.

To evaluate the rules generated by each fold, at the end of each fold run, statistics such as predictive accuracy of the rule set and the number of rules in the rule set are calculated. The average predictive accuracy in the test set over the ten iterations of the cross validation procedure is reported in the next section.

The parameters for the Ant-Miner are the same as were used by Parpinelli [2], which are as follows:

1) *Number of ants: 3000*
2) *Minimum number of cases per rule: 10*
3) *Maximum number of uncovered cases: 10*
4) *Number of identical for convergence: 10*

TABLE III. THE NUMBERS OF ATTRIBUTES GENERATED BY VARIOUS ATTRIBUTE SELECTION METHODS.

Evaluation Methods	Search Method	Number of Attributes
Classifier-based	Best-First	8
	Exhaustive Search	8
	Genetic Search	8

Evaluation Methods	Search Method	Number of Attributes
	Greedy Stepwise	6
	Race Search	6
	Random Search	9
Correlation-based	Best-First	16
	Exhaustive Search	16
	Genetic Search	16
	Greedy Stepwise	16
	Random Search	17
Consistency subset	Best-First	20
	Genetic Search	19
	Greedy Stepwise	20

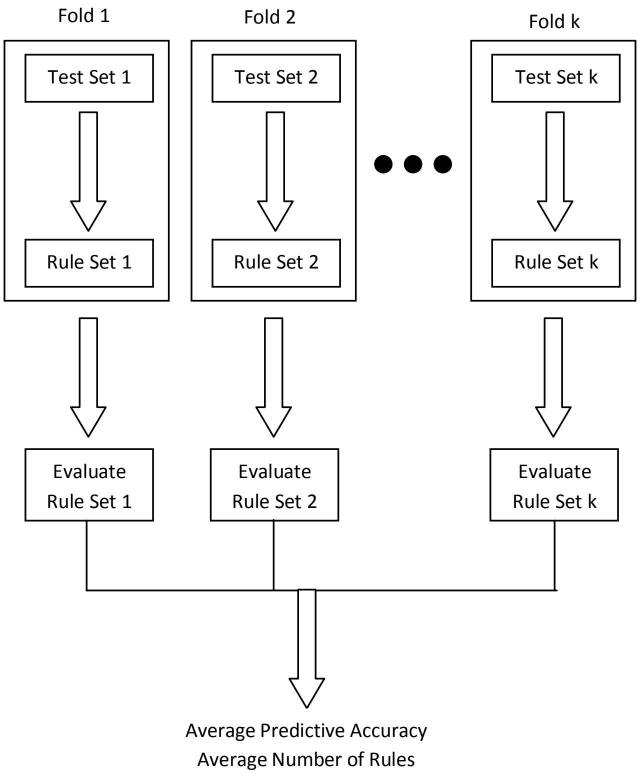

Figure 2: *k*-fold cross validation procedure

III. RESULTS

Tab. IV shows the results of the classification for each new data set created by the attribute selection in terms of average predictive accuracy, while Tab. V list out the average number of rules generated by each classifier for different attribute selection methods. The number after the "±" sign is the standard deviation.

TABLE IV. COMPARISON BETWEEN C4.5 AND ANT-MINER FOR AVERAGE PREDICTIVE ACCURACY.

Evaluation Methods	Search Method	Predictive Accuracy	
		C4.5 (%)	Ant-Miner (%)
Classifier-based	Best-First	92.29 ± 1.63	88.77 ± 1.66
	Exhaustive Search	92.29 ± 1.63	88.77 ± 1.66
	Genetic Search	92.29 ± 1.63	88.77 ± 1.66
	Greedy Stepwise	92.38 ± 1.63	89.94 ± 0.86
	Race Search	92.38 ± 1.63	89.94 ± 0.86
	Random Search	92.31 ± 1.62	89.63 ± 0.87
Correlation-based	Best-First	93.74 ± 1.58	91.30 ± 1.45
	Exhaustive Search	93.74 ± 1.58	91.30 ± 1.45
	Genetic Search	93.74 ± 1.58	91.30 ± 1.45
	Greedy Stepwise	93.74 ± 1.58	91.30 ± 1.45
	Random Search	93.92 ± 1.49	93.36 ± 0.67
Consistency subset	Best-First	94.03 ± 1.42	88.96 ± 1.38
	Genetic Search	93.89 ± 1.41	88.80 ± 1.67
	Greedy Stepwise	94.03 ± 1.42	91.02 ± 0.99

TABLE V. COMPARISON BETWEEN C4.5 AND ANT-MINER FOR AVERAGE NUMBER OF RULES.

Evaluation Methods	Search Method	Number of Rules	
		C4.5 (%)	Ant-Miner (%)
Classifier-based	Best-First	8.27 ± 0.94	8.70 ± 0.21
	Exhaustive Search	8.27 ± 0.94	8.70 ± 0.21
	Genetic Search	8.27 ± 0.94	8.70 ± 0.21
	Greedy Stepwise	7.00 ± 0.00	7.70 ± 0.15
	Race Search	7.00 ± 0.00	7.70 ± 0.15
	Random Search	7.4 ± 0.85	9.10 ± 0.46
Correlation-based	Best-First	18.93 ± 2.61	7.40 ± 0.27
	Exhaustive Search	18.93 ± 2.61	7.40 ± 0.27
	Genetic Search	18.93 ± 2.61	7.40 ± 0.27
	Greedy Stepwise	18.93 ± 2.61	7.40 ± 0.27
	Random Search	19.83 ± 2.45	7.00 ± 0.15
Consistency subset	Best-First	20.20 ± 1.63	7.30 ± 0.33
	Genetic Search	20.94 ± 1.88	8.00 ± 0.26
	Greedy Stepwise	20.20 ± 1.63	7.80 ± 0.33

Tab. IV and Tab. V show that Correlation-based evaluation with Random Search is the best attribute selection method for Ant-Miner, with the highest predictive accuracy and lowest number of rules. As for the C4.5 case, it seems like C4.5 performs slightly better (with Consistency subset evaluation method) than Ant-Miner (with Correlation-based attribute subset selection) in terms of predictive accuracy. However, the number of rules generated is more than double the number of rules generated by Ant-Miner. As for the number of attributes is concerned, it shows that the lesser the number of attributes slightly decrease the predictive accuracy, but increase the knowledge comprehension (the number of rules). However, reducing too many attributes like in the Classifier-based attribute subset selection case would slightly reduce the predictive accuracy.

Eventually, Correlation-based evaluation with Random Search attribute selection, which contributes the best predictive accuracy as well as the number of rules for Ant-Miner, does not provide the best number of attributes removed. The number of attributes selected for this attribute selection method is 100% which is higher than the one generated by Classifier-based attribute subset selection. Even though predictive accuracy depends on the attributes selected, the choice of classifiers also plays a critical role. Classifier will perform differently for different domain of data sets.

IV. CONCLUSION

The best attribute selection method for web texts categorization is the combination of Correlation-based evaluation with Random Search as the search method. However, this attribute selection method will not give the best performance in attributes reduction. Using Classifier-based attribute subset selection will reduce more attributes, but sacrifice the performance of the classifier. It is also found that Ant-Miner performed better than C4.5 for web texts categorization.

For the future project, it is suggested to test the performance of attribute selection on higher dimension of web data sets, with more categories, since this project only focus on two categories. Higher dimension of data sets may cause higher dimension of attributes. On the other hand, a study on reducing the size of attributes dimension could also be done in related to the linguistic relationship to generalized words, as a manual preliminary step before performing the attribute selection method.

REFERENCES

[1] "January 2010 Web Server Survey - Netcraft," Netcraft, Jan. 2010.

[2] R.S. Parpinelli, H.S. Lopes, and A.A. Freitas, "Data mining with an ant colony optimization algorithm," Evolutionary Computation, IEEE Transactions on, vol. 6, 2002, pp. 321–332.

[3] R.J. Quinlan, C4.5: programs for machine learning, San Mateo, CA: Morgan Kaufmann Publishers Inc., 1993.

[4] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization," Proceedings of ICML-97, 14th International Conference on Machine Learning, Morgan Kaufmann Publishers, San Francisco, US, 1997, pp. 420-412.

[5] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989.

[6] A.W. Moore and M.S. Lee, "Efficient Algorithms for Minimizing Cross Validation Error," ICML, 1994, pp. 190-198.

[7] H. Liu and R. Setiono, "A Probabilistic Approach to Feature Selection - A Filter Solution," ICML, 1996, pp. 319-327.

[8] M.A. Hall, "Correlation-based Feature Subset Selection for Machine Learning," University of Waikato, 1998.

[9] "CMU World Wide Knowledge Base (Web->KB) Project Web Site," CMU, 1998.

[10] M.F. Porter, "An algorithm for suffix stripping," Program, vol. 14, 1980, pp. 130–137.

[11] D.A. Hull, "Stemming algorithms: A case study for detailed evaluation," Journal of the American Society for Information Science, vol. 47, Dec. 1998, pp. 70–84.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "The WEKA data mining software: an update," SIGKDD Explor. Newsl., vol. 11, 2009, pp. 10–18.