Proceeding of the International Conference on Advanced Science, Engineering and Information Technology 2011

Hotel Equatorial Bangi-Putrajaya, Malaysia, 14 - 15 January 2011

ISBN 978-983-42366-4-9

An Approach for Optimal Feature Subset Selection using a New Term Weighting Scheme and Mutual Information

Shine N Das[#], Midhun Mathew[*], Pramod K Vijayaraghavan[#]

[#]*Dept. of Computer Applications, Cochin University of Science & Technology, India*
Tel: +91 4865 232989, E-mail: shine_das@rediffmail.com

[*]*P.G. Scholar, School of Computing, SASTRA University, India*
Tel: +91 4865 263590, E-mail: midhunmunnar@gmail.com

Abstract— **With the development of the web, large numbers of documents are available on the Internet and they are growing drastically day by day. Hence automatic text categorization becomes more and more important for dealing with massive data. However the major problem of document categorization is the high dimensionality of feature space. The measures to decrease the feature dimension under not decreasing recognition effect are called the problems of feature optimum extraction or selection. Dealing with reduced relevant feature set can be more efficient and effective. The objective of feature selection is to find a subset of features that have all characteristics of the full features set. Instead Dependency among features is also important for classification. During past years, various metrics have been proposed to measure the dependency among different features. A popular approach to realize dependency is maximal relevance feature selection: selecting the features with the highest relevance to the target class. A new feature weighting scheme, we proposed have got a tremendous improvements in dimensionality reduction of the feature space. The experimental results clearly show that this integrated method works far better than the others.**

Keywords—**Feature selection, Web page Classification, Feature subset selection, Mutual Information**

I. INTRODUCTION

Web represents documents on different topics or different aspects of the same topic and introduces massive volume of online unstructured or semi-structured text with diverse information sources. This opens the question of how to effectively use such a massive Web repository to retrieve information with minimum computation time and maximum relevancy. Automatic Web classification aids in better information retrieval and knowledge utilization. Users often prefer navigating through Web catalogues of pre-classified contents as they enable them to find more relevant information in a shorter time. For the classification purpose, the representative words in the web page called as features are used rather than the entire web page. Using the full feature set is infeasible and impractical because there exist a large number of features and also many features are irrelevant, correlated, or redundant. Reduced feature set with relevant features can influence the classification accuracy. Feature selection derives a feature subset that is closer to the full features set. Classification quality depends on how close the reduced feature set is to the full feature set [1]. The rapid developments in computer science and engineering allow for data collection at an unprecedented speed and present new challenges to feature selection. Wide data sets, which have a huge number of features but relatively few instances, introduce a novel challenge to feature selection problem.

Feature selection is a data preprocessing technique commonly used on high dimensional data. Its purposes include reducing dimensionality, removing irrelevant and redundant features, reducing the amount of data needed for learning, improving algorithm's predictive accuracy, and increasing the constructed model's comprehensibility [2]. Feature-selection methods are particularly welcome in

interdisciplinary collaborations because the selected features retain the original meanings, domain experts are familiar with.

Large number of features brings disadvantages for classification problem. On one hand, increased features give difficulties to calculate, because the more data occupy large amount of memory space and require more computerization time. On the other hand, a lot of features include certainly many correlation factors respectively, which results to information repeat and waste. Therefore, we must take measures to decrease the feature dimension without affecting the document representation; feature optimum extraction or selection. The number of features needs to be constrained to reduce noise and to limit the burden on system resources.

This paper focuses on issues that have not been touched in most of the earlier works. Eventhough web documents are hyperlinked; most of the classification techniques take little advantage of the link structure. Though some of the methods take context also into account, features from all fields are weighted equally which is absolutely wrong. Features identified from different fields should be assigned with different weights according to their relevance. Majority of the existing methods are based on the assumption that attributes are purely independent. But one may depends another and can be used for a phrase query or proximity query. Many of the algorithms do not consider user feedback or relevance feedback. They are greedy in nature which may lose optimal result. Instead of treating each attribute as independent one, here dependency among features is also taken into account. One of the most popular approaches to realize dependency is considered as maximal relevance feature selection. Features with the highest relevance to the target class are selected for further processing. In this paper, we propose a novel approach for solving all of these problems. The experimental results show that this approach is comparable with other feature selection methods proved promising in this field. It clearly describes that the proposed work is enough worthy since it surpasses others in terms of accuracy while the number of features is increasing.

The rest of this paper is organized as follows. Section II describes the related works. Section III gives the details of the proposed work. Section IV analyses the experimental results to compare with other feature selection methods. In the last section, we give the conclusion and future works.

II. RELATED WORKS

Existing methodologies make use of different combinations of different methods promising in the field of feature selection in classification. Preprocessing is a common step done by all of these methodologies. So far, lots of selection methods have been proposed to identify salient features which briefly reviews only on filter model feature selection methods. A large number of studies on feature selection have focused on non-text domains. These studies typically deal with much lower dimensionality.

A number of feature selection techniques were described in the TC literature, while [3] found document frequency (DF), information gain (IG) and χ^2(Chi-square) to be the most effective (reducing the feature set by 90-98% with no performance penalty, or even a small performance increase due to removal of noise). It is also observed that contrary to a popular belief in information retrieval that common terms are less informative, document frequency, which prefers frequent terms (except for stop words), was found to be quite effective for document categorization. Their comparative study of feature selection method in statistical learning of text categorisation focused on aggressive dimensionality reduction evaluated five methods. This suggests that the DF thresholding is not just an ad hoc approach to improve efficiency, but a reliable measure for selecting informative features.

An algorithm for feature selection which approximates Optimal Feature Selection model is presented in [4] and it is proved to have good efficiency and scalability which in some cases could lead to only slight accuracy gains since it does not take advantage of the induction algorithm's properties. Information Gain (IG), an information theoretic measure, was used to rank [5] the features so that a threshold could be established above which the features were selected for the reduced set of features. But feature selection is done in a single step which does not undergo any optimization. In this, they used only the content words and ignored other features such as HTML tags and links. M. Lan et. al. [6] proposed a term weighting method called tf*rf, and compared their method using the traditional SVM, with other term weighting methods, i.e. (tf.x2, tf.ig, tf.or), on two widely used data sets. The experimental results showed that methods based on information theory, i.e. (tf.x2, tf.ig, tf.or), perform poorly if compared with their proposed term-weighted method in terms of accuracy.

To evaluate the significance of features, many measurements (e.g., distance, Gini index, χ^2-test and dependency) have been introduced [7, 8]. Among them, distance discriminant is a straightforward one. As an illustration, Relief, which is introduced by Kira et. al. [9] and later enhanced by Kononenko [10], typically belongs to this kind. In Relief, the relevant weight of feature is measured by Euclidean distance between instances, and this weight co-reflects its discriminative ability to different classes. A feature has higher weight if it has the same value for instances within the same class and different values to other instances. Relief randomly picks out an instance from training dataset and then calculates distances between the instance and its nearest neighbors from the same and opposite class, respectively. These distance values are later used to update relevance scores of features [9]. To further improve the efficiency or robustness, several variations of Relief have been investigated recently [11]. For instance, Liu et.al, [12] chose instances by

selective sampling, rather than random one, which does not exploit data characteristics.

III. PROPOSED WORK

This paper provides a novel and efficient approach for optimal feature subset selection by feature pruning and dependency analysis. Our objective is to find how to select good features from the entire feature space. Then, a two-stage feature selection algorithm is proposed by combining a different term weighting approach (for content, URL, heading, title, anchor text and information in the meta-tags) and wrapper model feature selection method. This allows selecting a compact set of superior features *S* with *m* features, which jointly have the largest dependency on the target class *c* at very low cost.

A wrapper is a feature selector [13] that convolves with an automatic classifier (we use Naïve Bayes classifier), with the direct goal to minimize the classification error of the particular classifier. Usually, wrappers can yield high classification accuracy for a particular classifier at the cost of high computational complexity and less generalization of the selected features on other classifiers. This is different from existing methods, which does not optimize the classification error directly.

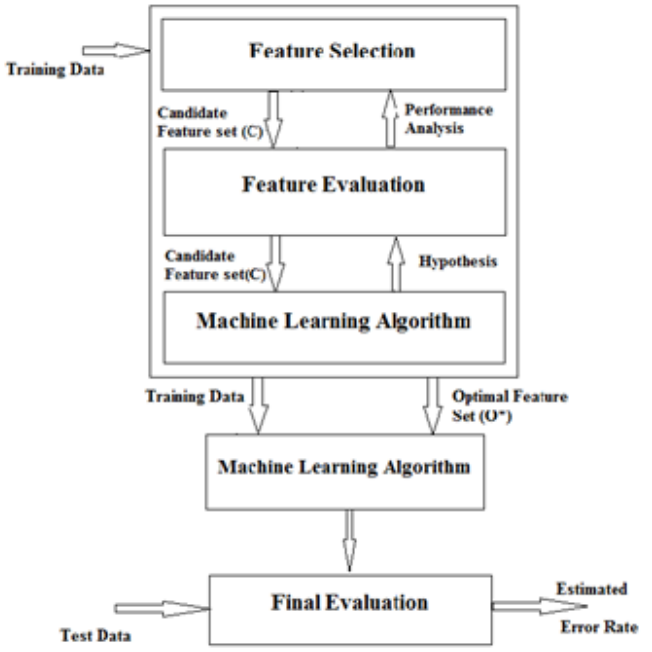

Fig. 1. Optimal Feature Subset Selection

Proposed term weighting approach is entirely different from others, since most of the existing approaches use the same weight for all the features. In this paper, we present a different weighting scheme which is purely based on the field where the term is present. This scheme is given in Table I. After applying this weighting scheme proportional to the number of occurrences of that particular feature, we select the features which are having a score above a threshold value. This threshold value is dynamically varying according to the length of the document or maximum weight of the terms. Feature set thus selected is called candidate set C.

TABLE I
PROPOSED WEIGHTING SCHEME

Term Field	Weight
Content	1
URL by n-grams	2
Heading	2
Title	2
Anchor Text – To the same web site	1
Anchor Text – To a different web site	0.5
Keywords	3
Description	3

Each feature in C is analyzed with its Error Rate (ER) to decide whether it can be included in optimal feature set O*. If ER increases with a feature, it indicates that that particular feature is irrelevant and it can be pruned.

A. *Proposed Algorithm*

Algorithm: Feature_Selection
Input: A Web page, Web_Document
Output: Optimal Feature set, O*
Remarks: C, Candidate feature set

```
Feature_Selection (Web_Document)
        Input_Document= Pre_Processing(Web_Document);
        C=Candidate_Feature_Selection (Input_Document);
        O*= Optimal_Feature_Selection(C);
        return O*;
```

Algorithm: Candidate_Feature_Selection
Input: Pre-processed document, Input_Document
Output: C
Remarks: w_i, weight of i^{th} feature

```
Candidate_Feature_Selection (Input_Document)
        F←Full feature set(Input_Document);
        for all f_i Є F
                w_i←Weight_Scheme(f_i);
        W←∑w_i;
        for all i, 1 ≤ i ≤ |F|
                w_i←Normalize(w_i, W);
        T←Thresholding(W);
        C← ϕ;
        for all i, 1 ≤ i ≤ |F|
                if w_i>T
                        C← C ∪ f_i;
        return C;
```

```
Algorithm:      Optimal_Feature_Selection
Input:          C
Output:         O*
Remarks:        ER, Error Rate during classification
Optimal_Feature_Selection(C)
        Sort(C);
        O ← ϕ;
        Set ER with O as arbitrarily high;
        repeat until C = ϕ
                first ← Top(C);
                C ← C – first;
                Calculate ER with O ⋃ first;
                if ER with O ⋃ first  < ER with O
                        O ← O ⋃ first;
        O*=mRMR(O);
        return O*;
```

Rather than treating each attribute as independent one, dependency among features is also analyzed for better results. One of the most popular approaches to realize dependency is maximal relevance feature selection: selecting the features with the highest relevance to the target class c. Relevance is usually characterized in terms of correlation or Mutual Information (MI), of which MI is one of the widely used measures to define dependency of features. In this paper, we focus MI based feature selection method that can be applied for optimal feature subset selection as a combination of mRMR and wrapper model.

Given two random variables x and y, their mutual information is defined in terms of their probabilistic density functions p(x), p(y) and p(x,y) :

$$I(x;y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy. \quad \text{.....(1)}$$

In Max-Relevance, the selected features x_i are required, individually, to have the largest mutual information $I(x_i;c)$ with the target class c, reflecting the largest dependency on the target class. The top m features in the descent ordering of $I(x_i;c)$, are often selected as the m features. The purpose of feature selection is to find a feature set S with m features $\{x_i\}$, which jointly have the largest dependency on the target class c. Max- Dependency:

$$\max D(S,c), \quad D = I(\{x_i, i = 1, \ldots, m\}; c). \quad \text{.....(2)}$$

Max-Relevance is to search features which approximates D(S,c) with the mean value of all mutual information values between individual feature x_i and class c:

$$\max D(S,c), \quad D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i;c). \quad \text{.....(3)}$$

When two features highly depend on each other, the respective class-discriminative power would not change much if one of them were removed. Therefore, the following minimal redundancy condition can be added to select mutually exclusive features

$$\min R(S), \quad R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j). \quad \text{.....(4)}$$

The criterion combining max-relevancy and min-redundancy constraints is called "minimal-redundancy-maximal-relevance" (mRMR) and a simplest form to optimize D and R simultaneously:

$$\max \Phi(D,R), \Phi = D - R. \quad \text{.....(5)}$$

We combine mRMR with our wrapper model feature selection to obtain a low cost, high relevant, superior set of features which significantly improve the classifier accuracy in automatic web page classification.

IV. EXPERIMENTAL RESULTS

To ascertain the validity of the proposed measure, we performed the experiments of automatic web page categorization and the obtained results using the proposed measure were compared with those using other commonly used measures. To validate performance fairly, 16 benchmark datasets were adopted in our simulation experiments. These datasets are all available from the UCI Machine Learning Repository available from [14]. Since these datasets may embody missing values, they would be processed during the preprocessing phases. For missing values, we replaced them with the most frequently used values. In simulation experiments, datasets were firstly fed into different feature selectors, which will generate different feature subsets from the same dataset. Since the number of features chosen by these selectors is different, we chose the same quantity of features for the sake of impartiality and the selected features were arranged in a descending order according to their priorities. After that, datasets with newly selected features were passed to external learning algorithms to assess classification performance. Currently, various outstanding learning algorithms are available. In our experiments, a popular classifier, namely NBC (Naive Bayes Classifier), is chosen to test prediction capability of the selected subset. The reason to choose it is because of its relatively high efficiency. NBC utilizes Bayes formula to distinguish which label an instance belongs to. Moreover, the conditional probability distribution of any given class satisfies normal distribution. Many experiments have demonstrated that NB classifier has good performance compared with others on various real datasets.

The experimental platform was Weka, which is an excellent tool in data mining and brings together many machine learning algorithms under a common frame work. To achieve impartial results, ten 10-fold cross validations had been adopted for each algorithm-dataset combinations while verifying classification capability. This is to say, for each dataset before and after feature selection, we run classification algorithm on it 10 times and at each time, a 10-fold cross validation was used, and the final results were their average values.

TABLE II
DATA SETS FOR OUR EXPERIMENTS

Sl. No.	Dataset	No. of Instances	No. of features
1	Annealing	798	38
2	Audiology (Standardized)	226	69
3	Breast Cancer Wisconsin (Diagnostic)	569	32
4	Census-Income (KDD)	299285	40
5	Congressional Voting Records	435	16
6	Connect-4	67557	42
7	Covertype	581012	54
8	Cylinder Bands	512	39
9	Dermatology	366	33
10	Flags	194	30
11	Heart Disease	303	75
12	Image Segmentation	2310	19
13	Internet Advertisements	3279	1558
14	KDD Cup 1999 Data	4000000	42
15	Meta-data	528	22
16	Statlog (German Credit Data)	1000	20

Details of datasets used in our experiments are given in Table II. Table III shows the comparison of our proposed algorithm with Information Gain (IG), Term Frequency (TF) and Gini Index (GI) algorithms in terms of accuracy. It is clearly observable that our method works far better than the others. One may also observe that our proposed method clearly surpasses others in many cases.

TABLE III
A COMPARISON OF ACCURACIES OF CLASSIFICATION WHILE USING DIFFERENT FEATURE SELECTION ALGORITHMS ON 16 DATA SETS. BOLD VALUE REPRESENTS THE MAXIMUM ONE.

Sl. No	IG	Prop. Algm	TF	GI
1	96.17	**97.88**	91.81	95.36
2	74.41	**76.51**	74.05	74.42
3	**73.50**	73.24	71.04	70.98
4	95.21	**95.42**	95.31	95.09
5	70.57	**74.21**	70.15	70.56
6	83.54	82.45	83.01	82.12
7	94.25	**95.06**	93.45	92.17
8	93.21	**94.08**	92.47	93.88
9	87.16	87.06	86.95	**87.21**
10	95.23	**96.54**	96.00	95.87
11	82.65	81.36	**83.65**	81.32
12	74.39	75.35	**75.91**	74.30
13	89.99	90.04	88.36	**90.27**
14	**92.54**	91.56	90.42	90.48
15	82.65	**84.26**	81.54	81.56
16	89.65	**90.04**	88.24	89.69
Average	85.95	**86.57**	85.15	85.33

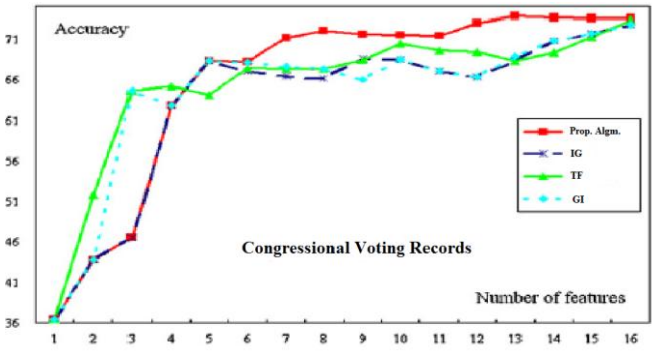

Fig. 2(a) Congressional Voting Records

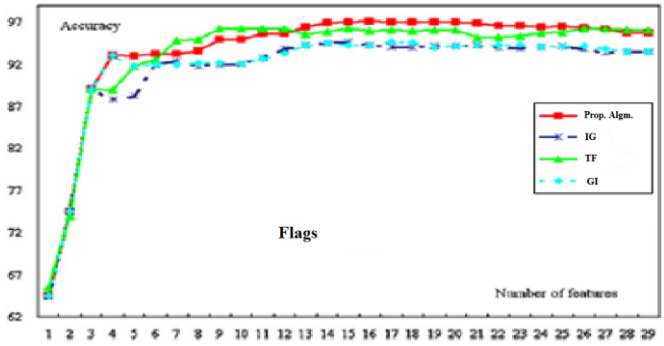

Fig. 2(b) Flags

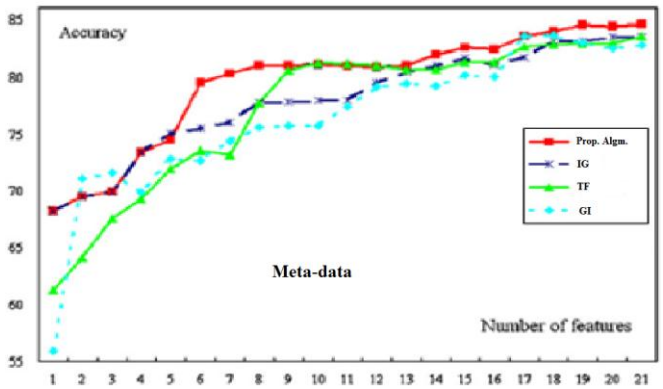

Fig. 2(c) Meta-data

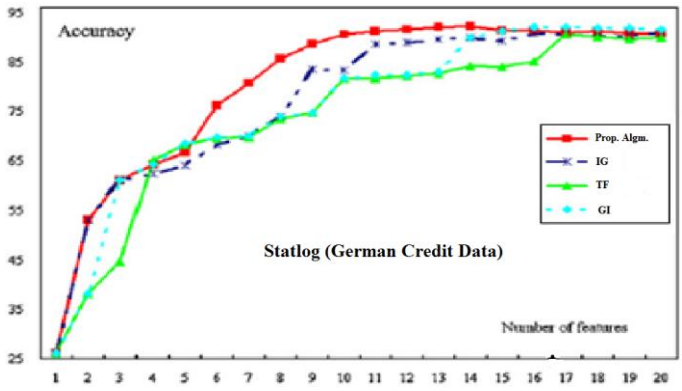

Fig. 2(d) Statlog (German Credit Data)

Fig 2 shows the graphical representation of comparison of these methods with some particular datasets. It shows accuracy vs. no. of features for (a) Congressional Voting Records (b) Flags (c) Meta-data and (d) Statlog (German Credit Data). From the view of average performance, we can infer that proposed method is superior to other selectors.

V. CONCLUSION AND FUTURE WORKS

This paper proposed a novel task and also a set of hybrid approaches for finding feature subset selection. The proposed techniques aim at helping document classification based on the maximal relevancy at minimum feature set. We have also built a system based on feature weighting to extract the features using a different term weighting approach for content, URL, heading, title, anchor text and information present in the meta-tags. Instead of treating each attribute as independent one, dependency criterion is also considered – maximal relevance feature selection. In this paper, we focus MI based feature selection method that can be applied for optimal feature subset selection as a combination of mRMR and wrapper model. Thus we achieve the objective of our research. We compare its performance with other feature selection methods. The experiments show that our work has a better performance than other feature selection methods.

We believe that this work represents an important step toward this direction and it is a promising method for feature selection which contributes more for document classification. Further research works can substitute a more efficient classifier like SVM instead of NBC and can concentrate on much diverse training data. Also, more advanced techniques can be used for dependency analysis as well as relevance feedback.

REFERENCES

[1] Walid Ahmed Fouad, Amr Ahmed Badr and Ibrahim Farag Abd El-Rahman, "A Comparative Study of Web Document Classification Approaches", Proceedings of the 37th International Conference on Computers and Industrial Engineering, October 2007.

[2] Indra Mahadevan, Selvakuberan Karuppasamy and Rajaram Ramasamy, "Resource Optimization in Automatic web page classification using integrated feature selection and machine learning", International Arab Journal of e-technology.

[3] Yiming Yag & Jan O Pedersen, "A comparative study on feature selections in Text Categorization", Proceedings of the Fourteenth International Conference on Machine Learning, pp. 412 – 420, 1997.

[4] Daphne Koller & Mehran Sahami, "Toward Optimal Feature Selection" , International Conference on Machine Learning 1996.

[5] Richong Zhang, Michael Shepherd, Jack Duffy, Carolyn Wattersan, "Automatic Web Page Categorization using Principal Component Analysis", Proceedings of the 40th Hawaii International Conference on System Sciences – 2007.

[6] M. Lan, S. Y. Sung, H. B. Low and C. L. Tan, "A comparative study on term weighting schemes for text categorization", *Proceedings of the International Joint Conference on Neural Networks*, pp. 1032-1033, 2005.

[7] G. Forman, "An extensive empirical study of feature selection metrics for text classification", Journal of Machine Learning Research 3, pp. 1289–1305, 2003.

[8] H. Liu, L. Yu, "Toward integrating feature selection algorithms for classification and clustering", IEEE Transactions on Knowledge and Data Engineering 17 (4), pp. 491–502, 2005.

[9] K. Kira, L. Rendell, "A practical approach to feature selection", Proceedings of the 9th International Conference on Machine Learning, Morgan Kaufmann, Los Altos, CA, pp. 249–256, 1992.

[10] I. Kononenko, "Estimating attributes: analysis and extensions of relief", Proceedings of the 11th European Conference on Machine Learning, Springer, Berlin, pp. 171–182, 1994.

[11] M. Robnik-Sikonja, I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF", Machine Learning pp. 23–69, 2003.

[12] H. Liu, H. Motoda, A. Yu, "A selective sampling approach to active feature selection, Artificial Intelligence", 159 (1–2) 49–74, 2004.

[13] Roberto Ruiz, Jes´us S. Aguilar-Ruiz, and Jos´e C. Riquelme, "Wrapper for Ranking Feature Selection", Lecture Notes in Computer Science, 2004, Volume 3177/2004, 384-389.

[14] http://archive.ics.uci.edu/ml/datasets.html, Department of Information and Computer Science, University of California.