Web Page Classification

Prepared By
Rutu Joshi
12MCEC11

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INSTITUTE OF TECHNOLOGY
NIRMA UNIVERSITY
AHMEDABAD-382481

May 2014

Web Page Classification

Major Project

Submitted in partial fulfillment of the requirements

For the degree of

Master of Technology in Computer Science and Engineering

Prepared By
Rutu Joshi
(12MCEC11)

Guided By
Prof. Priyank Thakkar

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INSTITUTE OF TECHNOLOGY
NIRMA UNIVERSITY
AHMEDABAD-382481

May 2014

Certificate

This is to certify that the Major Project Report entitled "**Web Page Classification**" submitted by **Rutu Joshi (Roll No: 12MCEC11)**, towards the partial fulfillment of the requirements for the degree of Master of Technology in Computer Science and Engineering of Nirma University, Ahmedabad is the record of work carried out by her under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this major project, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Prof. Priyank Thakkar
Guide & Assistant Professor,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Prof. Vijay Ukani
Associate Professor
Coordinator M.Tech - CSE
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. Sanjay Garg
Professor and Head,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr K Kotecha
Director,
Institute of Technology,
Nirma University, Ahmedabad

Undertaking for Originality of the Work

I, **Rutu Joshi**, Roll. No. **12MCEC11**, give undertaking that the Major Project entitled "**Web Page Classification**" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering** of Nirma University, Ahmedabad, is the original work carried out by me and I give assurance that no attempt of plagiarism has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

_____

Signature of Student

Date:
Place: AHMEDABAD

Endorsed by
Prof. Priyank Thakkar
(Signature of Guide)

Acknowledgements

Foremost, I would like to express my sincere gratitude to my guide **Professor Priyank Thakkar**, Professor, Computer Science Department, Institute of Technology, Nirma University, Ahmedabad for his patience, motivation, enthusiasm, and immense knowledge. His valuable guidance and continuous support has helped me throughout this work. I could not have imagined having a better advisor and guide for my project.

My deepest thank you is extended to **Prof. Vijay Ukani**, PG CSE - Coordinator, Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad for an exceptional support and continual encouragement throughout the Major Project.

It gives me an immense pleasure to thank **Dr. Sanjay Garg**, Hon'ble Head of Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his kind support and providing basic infrastructure and healthy research environment.

A special thank you is expressed wholeheartedly to **Dr K Kotecha**, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation he has extended throughout course of this work.

I would also thank the Institution, all faculty members of Computer Engineering Department, Nirma University, and Ahmedabad for their special attention and suggestions towards the project work.

The blessings of God and parents make the way for completion of Project without whom I could not have made it here. I am very much grateful to them.

- Rutu Joshi
12MCEC11

Abstract

Classification of web pages is essential for improving the quality of web search, focused crawling, development of web directories like Yahoo, ODP etc. This paper compares various classification techniques for the task of web page classification. The classification techniques compared include k nearest neighbours (KNN), Naive Bayes (NB), support vector machine (SVM), classification and regression trees (CART), random forest (RF) and particle swarm optimization (PSO).Impact of using different representations of web pages is also studied. The different representations of the web pages that are used comprise Boolean, bag-of-words and term frequency and inverse document frequency (TFIDF). Experiments are performed using WebKB and R8 datasets. Accuracy and f-measure are used as the evaluation measures. Impact of feature selection on the accuracy of the classifier is moreover demonstrated.

Keywords: Web Page Classification, PSO (Particle Swarm Optimization), SVM (Support Vector Machine), KNN (K Nearest Neighbours), Naive-Bayes, CART(Class ification and Regression Trees), Random Forest

Contents

List of Figures

Chapter 1

INTRODUCTION

1.1 INTRODUCTION

The internet consists of millions of web pages corresponding to each and every search word which provides highly useful information. Search engines help users retrieve web pages related to a keyword but searching those innumerable pages is tedious. Also, web pages are dynamic and volatile in nature. There is no unique format for the web pages. Some web pages may be unstructured (text), some pages may be semi structured (HTML pages) and some pages may be structured (database). This heterogeneous format on the web presents additional challenges for classification. Hence it is important for us to find a technique which accurately classifies web pages and provide only the most relevant web pages. Classification is a data mining technique which predicts pre-defined classes for datasets. Classification is a supervised learning technique. Here the classifier is trained using the training dataset. The trained classifier then assigns class labels to the testing dataset. As stated in [1] in Web Page classification, web pages are assigned to pre-defined classes mainly according to their content .

There are three knowledge discovery domains that are applicable to web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web structure mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. Finally, web usage mining is the process of extracting interesting patterns in web access logs.

1.2 DEFINITION

In web page classification, a web page is assigned to one or more pre-defined classes mainly according to its content. Here classification is a supervised learning technique.

Here the classifier is trained using the training dataset. The trained classifier then assigns class labels to the testing dataset. Depending on the number of classes, classification can be divided into binary and multi-class classification. In binary classification, one of the two classes are assigned to the instance while in multi-class classification, one of the many (more than two) classes are assigned to the instance Also, depending on the number of classes that an instance may be assigned to, classification can be divided into single-label and multi-label classification. In single-label classification, an instance is assigned one class while in multi-label classification, an instance may be assigned more than one class.

1.3 OBJECTIVE

Classification of web pages is essential for improving the quality of web search, focused crawling, development of web directories like Yahoo, ODP etc. Also, web pages are volatile and dynamic in nature. There is no unique format for the web pages. Some web pages may be unstructured (text), some pages may be semi structured (HTML pages) and some pages may be structured (database). This hetereogeneous format on the web presents additional challenges for classification. At present, classical methods of text document classification are not appropriate for web document classification. It is very necessary to classify web pages using appropriate method. Hence, the objective here is to address the Web Page Classification problem by the Particle Swarm Optimization technique.

1.4 THESIS ORGANIZATION

The rest of the thesis is organized as follows:

Chapter 2, Literatue Survey, focusses on the work related to web page classification. The papers related to web page clasification techniques are studied extensively and are explained in this chapter.

In Chapter 3, Classification Methods, different classifiers for web page classification are discussed. The classification techniques are PSO (Particle Swarm Optimization), SVM (Support Vector Machine), KNN (K Nearest Neighbours), Naive-Bayes, CART(Class ification and Regression Trees) and Random Forest.

Chapter 4 presents Implementation Methodology. R8 and WebKB dataset are represented in boolean, TFIDF and bag-of-words format. All the classification techniques are experimented and the results are evaluated using accuracy and f-measure.

Chapter 5 includes conclusions based on the results that were obtained after applying classification techniques on datasets.

Chapter 2

LITERATURE SURVEY

Classification of web pages using various techniques was extensively studied. M. A. Nayak in paper [2] discussed 4 classification techniques namely decision trees, k-nearest neighbour, SVM and naive bayes. The paper focussed on obtaining accurate system results. When decision tree gave accurate result, bayesian network did not and vice versa due to their different operational profiles. Since many methods of web page classification were proposed, no clear conclusion about the best method was obtained.

In paper [3], the best results were acquired using SVM with linear kernel function (followed by method of k-nearest neighbours) and term frequency (TF) document model with attribute selection by mutual information score. Here special attention was laid on treating with short documents which often occurred on the internet.

Aixin Sun et. Al in [4] concentrated on the effects of using context features (text, title and anchor words) in web classification using SVM classifiers. Experimental results showed that SVM based web classification methods performed very well on the WebKB data set even using the text components only. Also context features consisting of title components and anchor words improved the classification accuracy significantly. However, the method without using anchor words could not deliver consistently good classification for all the dataset classes.

Paper [5] compared the SVM performance using 4 different kernel functions performance. Experimental results showed that anova kernel function yielded the best result of these 4 kernel functions. The LSA-SVM, BPN and WVSVM were also compared. The experiment demonstrated that WVSVM yielded better accuracy even with a small data set. When a smaller category had training data, WVSVM could still categorize web pages with acceptable accuracy.

Yong Zhang et. al in [6] described that even with a small dataset LS-SVM yielded better accuracy with faster speed and reduced runtime of the algorithm. Even if the smaller category had less training data, the LS-SVM was still able to categorize web

pages with acceptable accuracy.

Xue et. al in [7] concluded that combination of different feature had better classification performance than others. Also composite of plain text and html structure information had better classification performance. Some researchers believed that the GAUSSIAN kernel function had always better classification performance. Inspite of the wide usage of SVM some problems were still to be researched. Also selection of kernel function lacked theoretical support. The kernel function selection is hard lacking of theoretical support. To improve the performance of classification or reduce the complexity, there is a need to analyze the characters of specific corpora to find an effective representation of features and select an appropriate classifier.

Damodaram et. al in [8] worked on trying to identify phishing website by selecting an appropriate technique. Here PSO produced more accurate classification models than Associative classifiers. Here after detecting more than 1050 websites for both its application effectiveness and its theoretical groundings, PSO was concluded as one of the most successful paradigms in network security.

PSO was used for classifying multidimensional real dataset in [9] where the parameters were set in such a way that it gave the best result. Though PSO is one of the most efficient optimization technique, its performance depends upon PSO variants and parameters(C1, C2). Also, achieving 100 percent accuracy in PSO based classifier is uncertain and time of convergehce is also uncertain.

PSO was compared with decision-tree algorithm, naive bayes classifier and K-nearest neighbour algorithm on Reuter-21578 and TREC-AP in [10]. The experimental results indicated that PSO yielded much better performance than other conventional algorithms. Hence PSO could be considered as an effective algorithm for document classification problem.

Falco et. al [11] used 3 fitness functions on 13 datasets. PSO was compared with 9 other classification techniques like Multi Layer Perceptron Artificial Neural Network (MLP), Bayes Network, Naive Bayes Tree etc. Here PSO was in 4Th position, quite close to its predecessors. Also, PSO seemed effective for 2 class problem but contrasting results were obtained for more than 2 classes. Hence no clear conclusion was inferred.

2.1 Comparison of Papers

1. Automatic Web Page Classification [3]

2. Web Classification Using Support Vector Machine [4]

3. Web page classification based on a support vector machine using a weighted vote schema [5]

4. Web Page Classification Based on SVM [7]

5. Web Page Classification Based-on A Least Square Support Vector Machine with Latent Semantic Analysis [6]

1	2	3	4	5
DATASETS				
a training set of Czech written documents	Webkb	sports news	PKU collection from the Chinese Web Page Categorization Contest held by China Computer Federation and Macroeconomic collection from the research supported in part by Chinese national key fundamental research program and Chinese national fund of natural science.	12, 684 Chinese web pages from Internet
METHODS				
Comparison of Naïve Bayes, C4.5 algorithm, SVM & K nearest neighbors	Comparison of SVM(X), SVM(T), SVM(A) AND SVM(TA) where X, T, A, TA are four kinds of web page representations, i.e., X (text only), T (text + title), A (text+anchor words) and TA (text + title + anchor words)	Comparison of WVSVM(Weighted Vote SVM), LSA-SVM & BPN(Back Propagation Network)	Comparison of Naive Bayes & SVM	Comparison of LS-SVM, SVM & KNN
RESULTS				
The best result has been acquired using Support vector machines with linear kernel function	SVM(TA) delivered the best performance as it considered both the titles and anchor words	Using Anova kernel svm	GAUSSIAN kernel function has always better classification performance	Using LS-SVM
REPRESENTATION				
Binary, tf, tfidf representation	SVM-set of words representation	-----	Representation not used because it is tricky to incorporate such representations for structural features.	-----

Chapter 3

CLASSIFICATION METHODS

3.1 kNN method

K-nearest neighbor (kNN) is a lazy learning method. Here the training dataset is not used to train the classifier. For a test instance say t, the kNN method compares t with the training dataset to find the k most similar training instances. It then returns the class which represents the maximum of the k instances of the dataset. Normally k=1 is not opted for classification due to noise and other anomalies in the dataset. Hence k=3 is chosen for knn classification. Here the similarity function is the most important factor which is chosen depending on different scenarios like Euclidean distance, Hamming distance, Cityblock distance, cosine similarity etc.

3.2 Naive Bayes Method

Naive Bayes classifier is based on Bayes' theorem. Here classification is considered as estimating the posterior probabilities of class for testing dataset.

$$P(C|x) = \frac{P(x|C)P(C)}{P(x)}$$

$$P(x|C) = P(x_1, x_2, ..., x_n|C) = \prod_{i=1}^{n} p(x_i|C)$$

where $P(C|x)$ is the posterior probability of class given attribute, $P(C)$ is the prior probability of class, $P(x|C)$ is the probability of predictor given class and $P(x)$ is the prior probability of predictor. Here for each class, probability is calculated. Thereafter the class which has the highest probability is assigned to the instance of testing dataset.

Different distributions are used for different representations in naive bayes classification. Normal distribution is fit for TFIDF representation as it is appropriate for features that have normal distribution in each class. For each feature, it separately

computes mean and standard deviation of training dataset in that class. MultiNominal (mn) distribution is fit for bag-of-words representation. MultiVariate MultiNominal (mvmn) distribution is fit for Boolean representation which is used for categorical features. In mvmn distribution, Naive Bayes classifier separately computes probabilities for the set of features for each class.

3.3 SVM

SVM is one of the most popular algorithms. SVM uses supervised learning technique and is used for both classification and regression. In general, linear SVMs are used for 2 class classification. For more than 2 classes, svm network is used for classification. To build a classifier, SVM finds a linear function of the form $f(x) = w.x + b$ so that an input vector xi is assigned to the positive class if $f(xi) \geq 0$, and to the negative class otherwise. In essence, SVM finds a hyper plane w.x+b=0 that separates positive and negative training examples. This hyper plane is called the decision boundary or decision surface [12]. The main objective function here is to maximize hyper plane's margin between positive and negative data points.

If the dataset is noisy, linear SVM will not be able to find a solution. In this case, soft margin SVMs are used. Also, if the dataset cannot be separated linearly, kernel functions are used. The kernel function transforms the input data from its original space into another space (usually of a much higher dimensional space) so that a linear decision boundary can separate positive and negative examples in the transformed space, which is called the feature space. Kernel functions can be polynomial functions, linear kernels etc.

There are various methods to find the separating hyper plane. The "Least Square (LS)" method finds solution by solving a set of linear equations. The "Sequential Minimal Optimization (SMO)" method breaks a problem into 2D sub-problems that may be solved analytically, eliminating the need of a numerical optimization algorithm.

3.4 Classification and Regression Tree

CART was developed by Breiman et. al [13]. CART is a non-parametric decision tree learning technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively. In CART, leaves represent class labels while branches represent conditions that will lead to any of the class labels i.e. leaves. The decision tree consists of linear combination of features that help in determining a class for test dataset. CART uses historical data to construct decision trees which thereafter classify new dataset. In order to use CART it is necessary to know number of classes a priori. Classification trees and regression trees predict responses to data. To predict a response, follow the decisions in the

tree from the root node down to a leaf node. The leaf node contains the response. Classification trees give nominal responses such as 'true' or 'false'. Regression trees give numeric responses.

3.5 Random Forest

Random Forest consists of an ensemble of decision trees that may be used for either classification or regression. To train each tree, different subsets of the training dataset (probably 2/3rd) are selected. To predict class labels of an ensemble of trees for testing dataset, Random Forest takes an average of predictions from individual trees. For estimating the prediction error, predictions are computed for each tree on its out-of-bag observations (those observations that were not used to train the trees). Thereafter these predictions are averaged over the entire ensemble for each observation and then compared with the true value of this observation.

3.6 PSO

The PSO is a population-based stochastic optimization method first proposed by Kennedy and Eberhart [3] in 1995. It is simple as well as efficient in global search. In PSO, each particle represents a possible solution. PSO finds optimal solution using this swarm of particles. PSO Algorithm is of two types: Global Best (gbest) PSO and Local Best (lbest) PSO. In gbest PSO, the neighbourhood of the particle is the entire swarm while in lbest PSO, a particle may have social or geographical neighbourhood. The PSO algorithm starts with initializing the position and velocity of each particle. The function that is to be optimized for the PSO algorithm is called the fitness function. For each iteration, the velocity of the particles is updated by considering the previous velocity along with the personal best and global best position.

$$V_{ij}(t+1) = V_{ij}(t) + C_1 * R_1(P_{ib}(t) - X_{ij}(t)) + C_2 * R_2(P_{igb}(t) - X_{ij}(t))$$

where Vij(t) is the velocity at iteration t, C1 and C2 are acceleration constants, R1 and R2 are random values in the range [0,1], Pib(t) is the personal best position of particle for iteration t, Xij(t) is the position of particle for iteration t and Pigb(t) is the global best position of particle. The personal best position is calculated by comparing the fitness of all the previous positions of the particle and selecting the position with the best fitness value. The global best position of the particle is obtained by selecting the personal best position of particle having the best fitness value. The position of the particle is updated using the new velocity and older position of the particle.

$$X_{ij}(t+1) = X_{ij}(t) + V_{ij}(t)$$

These iterations are repeated until the algorithm satisfies the stopping criteria. The stopping criteria may be no of iterations or when the motion of the particles ceases. The algorithm renders the position of the particle having the best fitness value.

Selection of appropriate parameters is essential for the algorithm to render best results. For distinguishing web pages, hyperplane is used. So initially 50 particles are used for PSO which are the hyperplanes obtained from SVM. The initial velocity for all particles is zero and the value for C1, C2 is 2 and 0.8 respectively. The algorithm is iterated 10 times.

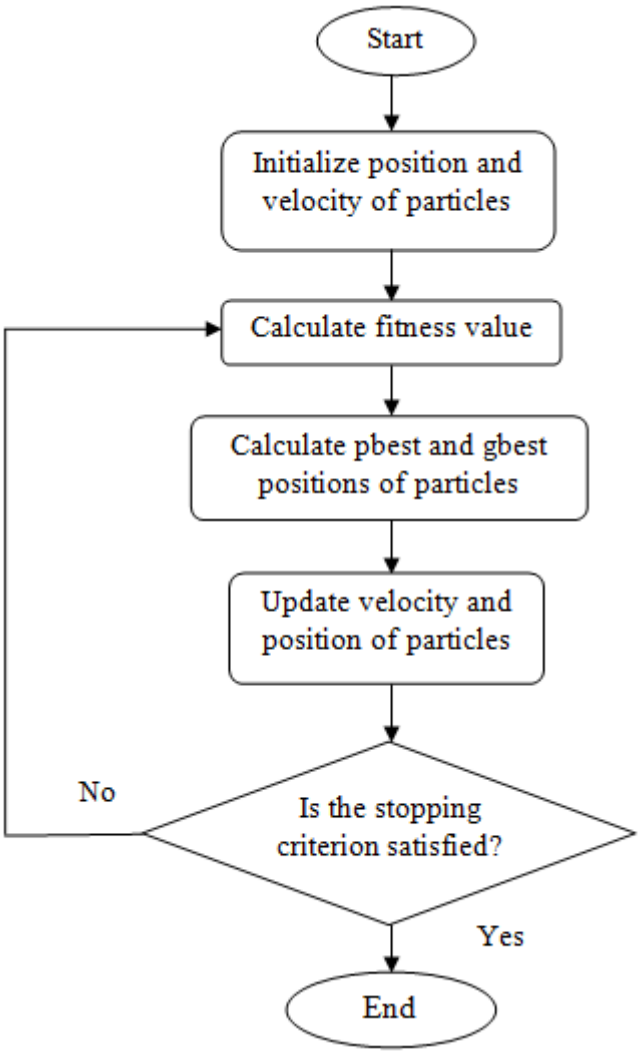


Figure 3.1: Particle Swarm Optimization

Chapter 4

Implementation Methodology

4.1 Dataset used

- WebKB dataset
  It consists of 4 classes respectively:- Project(Training 335, Testing 166 web pages), Course(Training 620, Testing 306 web pages), Faculty(Training 745 Testing 371 web pages) and Student(Training 1085 Testing 540 web pages).

  Source: CSMINING Group

  WebKB dataset consists of 7771 features.

- Reuters-21578 R8 dataset
  It consists of 8 classes respectively:- acq(Training 1596, Testing 696 web pages), crude(Training 253, Testing 121 web pages), earn(Training 2840, Testing 1083 web pages), grain(Training 41, Testing 10 web pages), interest(Training 190, Testing 81, web pages), money-fx(Training 206, Testing 87 web pages), ship(Training 108, Testing 36 web pages) and trade(Training 251, Testing 75 web pages).

  Source: CSMINING Group

  R8 dataset consists of 17386 features.

4.2 Tools

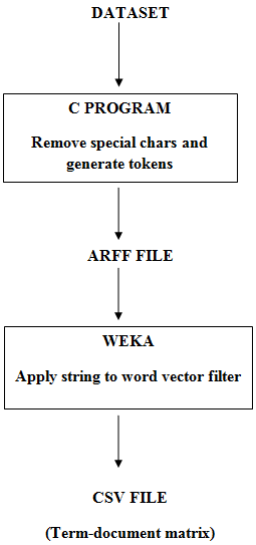
1. Matlab R2011a
   MATLAB (matrix laboratory) is a high level fourth generation programming language and interactive environment for numerical computing. Developed by MathWorks, the language, tools, and built-in math functions of MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written

in other languages, including C, C++, Java, and Fortran.
Usage: In Web Page Classification, Matlab was used for implementing all the algorithms on the dataset.

2. Weka 3.6
Weka (Waikato Environment for Knowledge Analysis) is a collection of machine learning algorithms for data mining tasks. The algorithms can either be used on a dataset or called from the Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It can also be used to develop new machine learning schemes.
Usage: In Web Page Classification, Weka was used especially for pre-processing the dataset.

4.3 Preprocessing

- Pre-processing of web pages is necessary to improve the quality of data thereby helping to improve subsequent classification processes.
- Pre-processing firstly involves converting the contents into lower case.
- Each word in the document is extracted and the stop words are removed from the dataset using weka.
- The Boolean, TFIDF and Bag-of-Words representations are obtained from the dataset using weka.
- Boolean representation consists only of zeros and ones- zero indicating the absence of the word in the web page while one indicating the presence of the word in the web page.
- Bag-of-Words representation counts the occurrence of words in the web pages.
- TFIDF representation computes the weight of each term (word) of the web page.



4.4 Implementation of Algorithms

The following algorithms are implemented in matlab for classification of web pages.

- KNN (K-Nearest Neighbor)
- Naive Bayes
- SVM (Support Vector Machine) (LS and SMO)
- Classification and Regression Tree
- Random Forest
- PSO (Particle Swarm Optimization)

4.5 Performance Measurement

Four parameters are used to measure the performance of the algorithms.

1. Precision (also called positive predictive value) is the ratio of true positives elements and the total no of elements that are predicted as positives (regardless of whether they are positive or not).

$$Precision = \frac{TruePositive}{TruePositive+FalsePositive}$$

For classifying multiple classes,

$$Precision_i = \frac{M_{ii}}{\sum_j M_{ji}}$$

Where i = no of rows and j = no of columns

2. Recall (also known as sensitivity) is the ratio of true positives elements and the total no of elements that are actually positive.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

For classification of multiple classes,

$$Recall_i = \frac{M_{ii}}{\sum_j M_{ij}}$$

Where i = no of rows and j = no of columns

3. F-measure is a measure that combines precision and recall. It is the harmonic mean of precision and recall. It is also known as F1 measure as precision and recall are evenly weighted. F-measure is used for better visualization.

$$F = \frac{2 * Precision * Recall}{Precision + Recall}$$

4. Accuracy is a measure of how well a given classifier can correctly classify new or previously unseen dataset.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + TrueNegative + FalsePositive + FalseNegative}$$

4.6 Feature Selection

- Feature selection also known as attribute selection is used to reduce the size of the dataset by removing redundant or irrelevant attributes. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context.
- Feature selection will reduce the set of terms to be used in classification, thus improving both efficiency and accuracy.

- Here feature selection is done using information gain. Information gain helps us determine which attributes in a given training set are most useful for discriminating between classes. It tells us how important a given attribute of the feature is.

$$InfoGain(Class, Attribute) = H(Class) - H(Class|Attribute)$$

where H stands for Entropy.
Entropy measures the level of impurity in a group.

$$Entropy = \sum_i -pilog_2pi$$

- Feature selection is applied on all the datasets and classification methods are applied on 5, 10, 15, 20, 30, 60, 100, 200, 300, 500, 1500, 2000, 3000, 4000, 5000... and likewise on all the features of the dataset.

- Figures 4.1 to 4.12 show the impact of feature selection on F-measure and Accuracy for Boolean, TFIDF and Bag-of-Words representation for WebKB and R8 dataset. Figures 4.13 to 4.18 compare the results of all the classification methods for WebKB and R8 datasets. The columns depict the no of features at which maximum f-measure is attained.

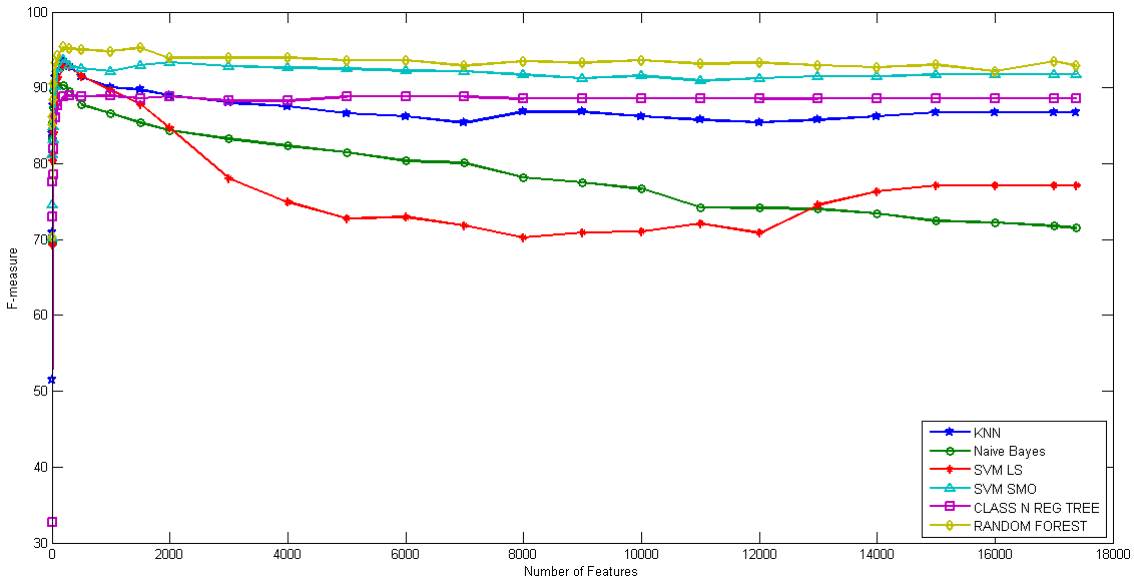


Figure 4.1: Impact of feature selection on F-measure (R8 Dataset, Boolean Representation)

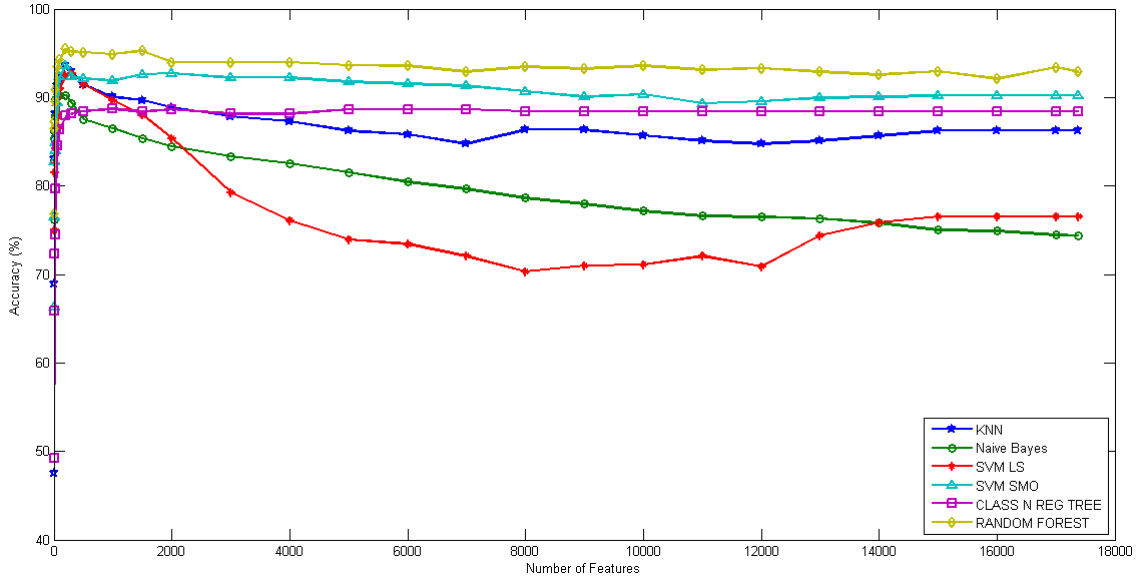


Figure 4.2: Impact of feature selection on Accuracy (R8 Dataset, Boolean Representation)

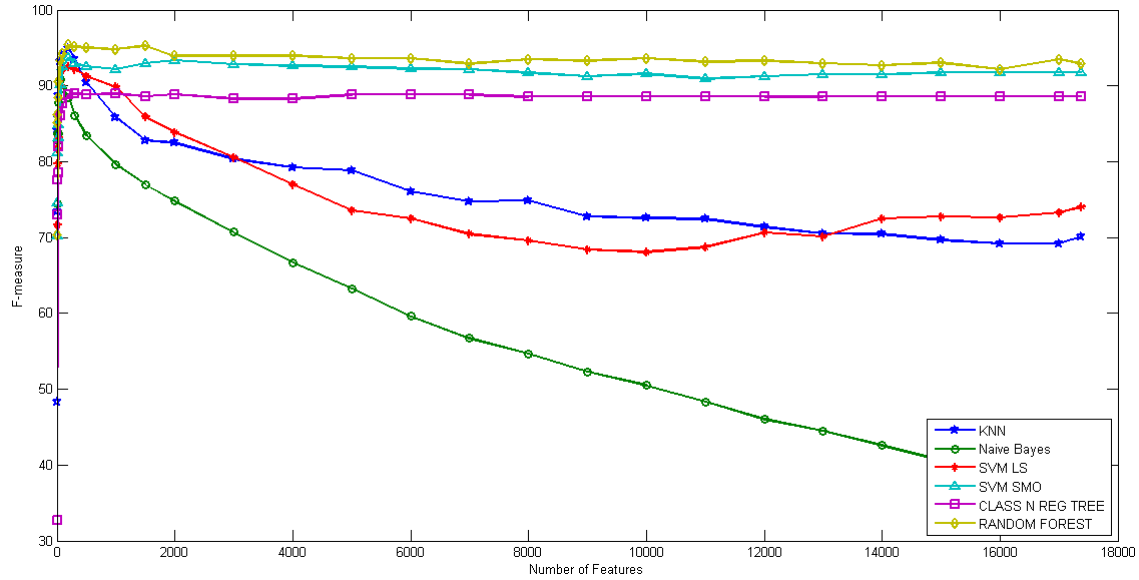


Figure 4.3: Impact of feature selection on F-measure (R8 Dataset, TFIDF Representation)

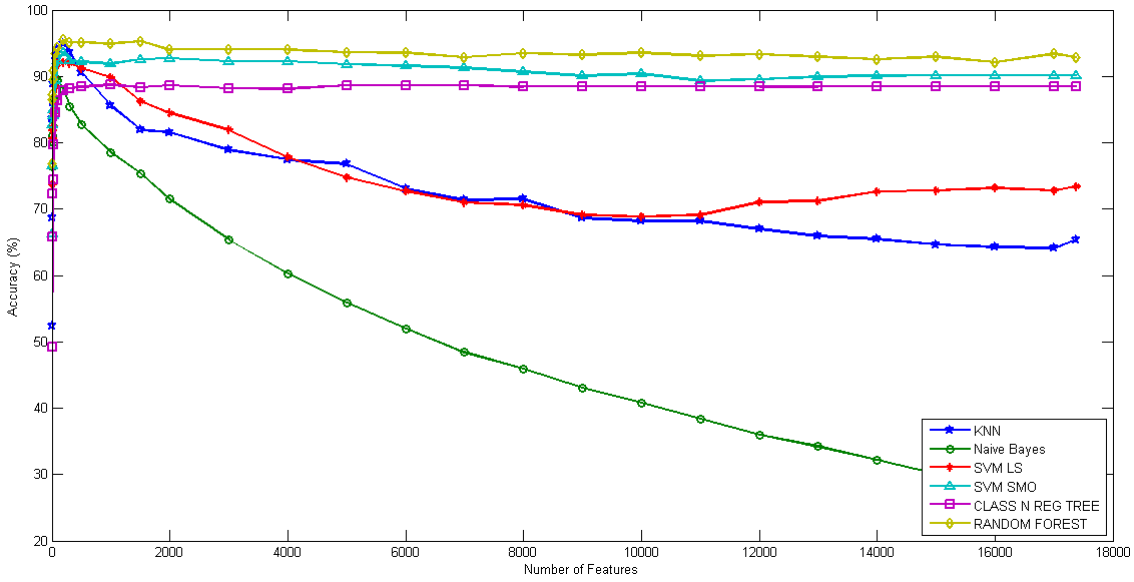


Figure 4.4: Impact of feature selection on Accuracy (R8 Dataset, TFIDF Representation)

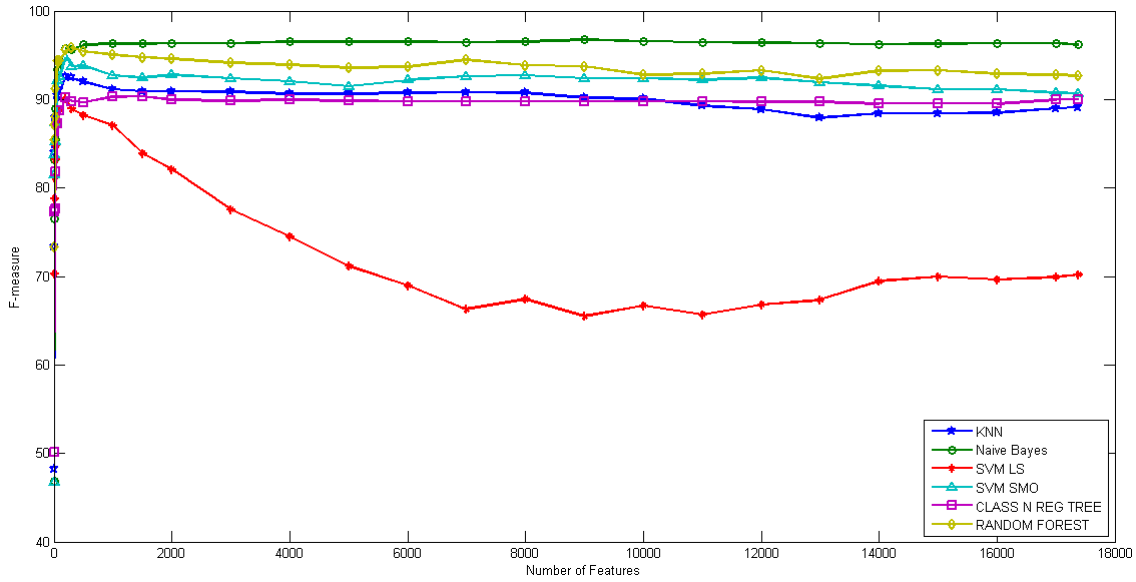


Figure 4.5: Impact of feature selection on F-measure (R8 Dataset, Bag-of-Words Representation)

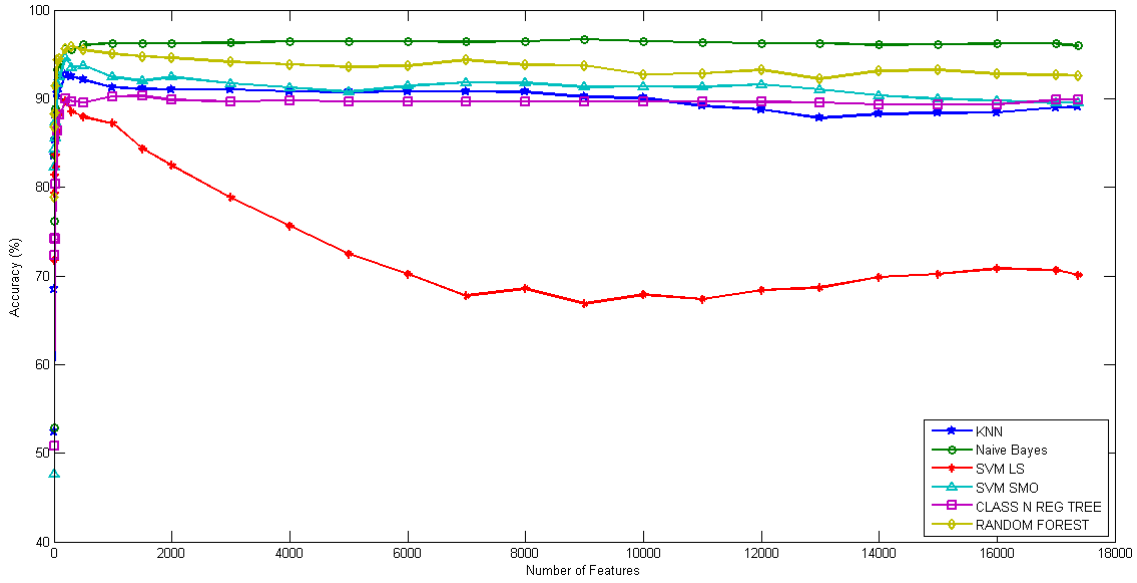


Figure 4.6: Impact of feature selection on Accuracy (R8 Dataset, Bag-of-Words Representation)

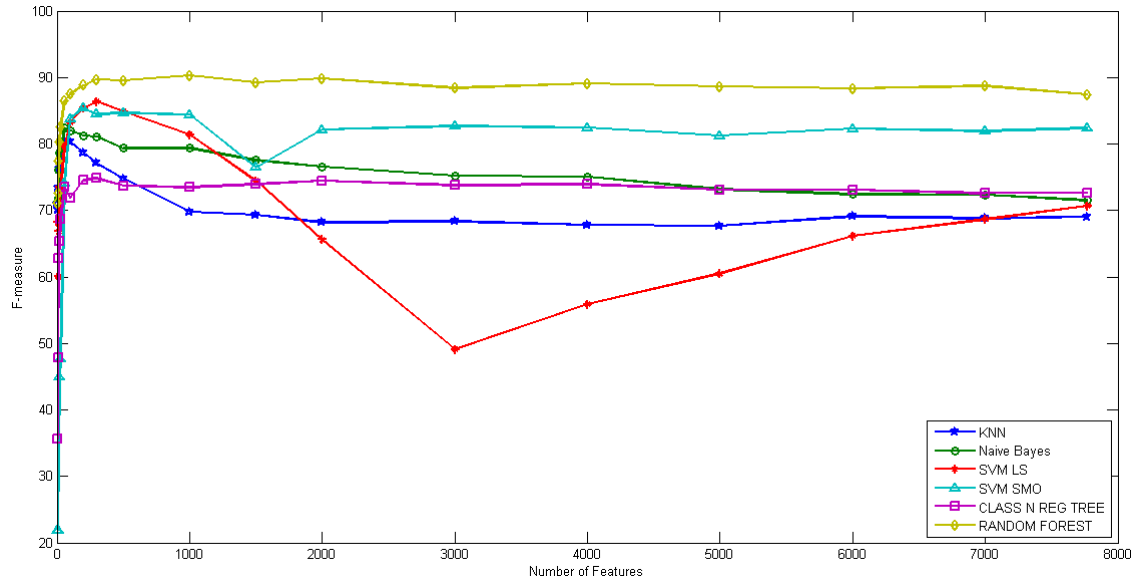


Figure 4.7: Impact of feature selection on F-measure (WebKB Dataset, Boolean Representation)

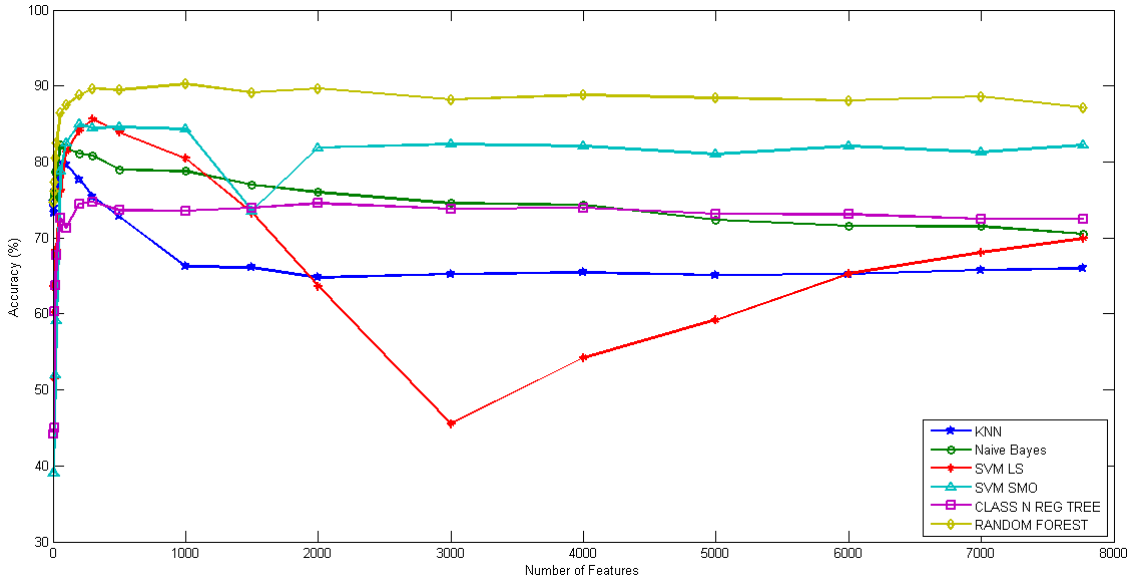


Figure 4.8: Impact of feature selection on Accuracy (WebKB Dataset, Boolean Representation)

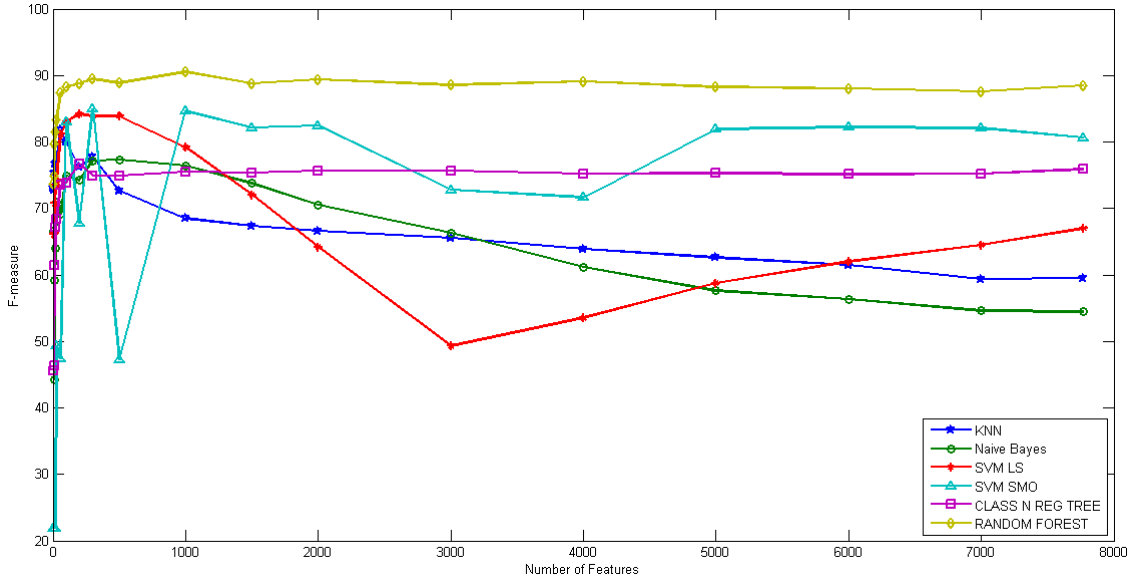


Figure 4.9: Impact of feature selection on F-measure (WebKB Dataset, TFIDF Representation)

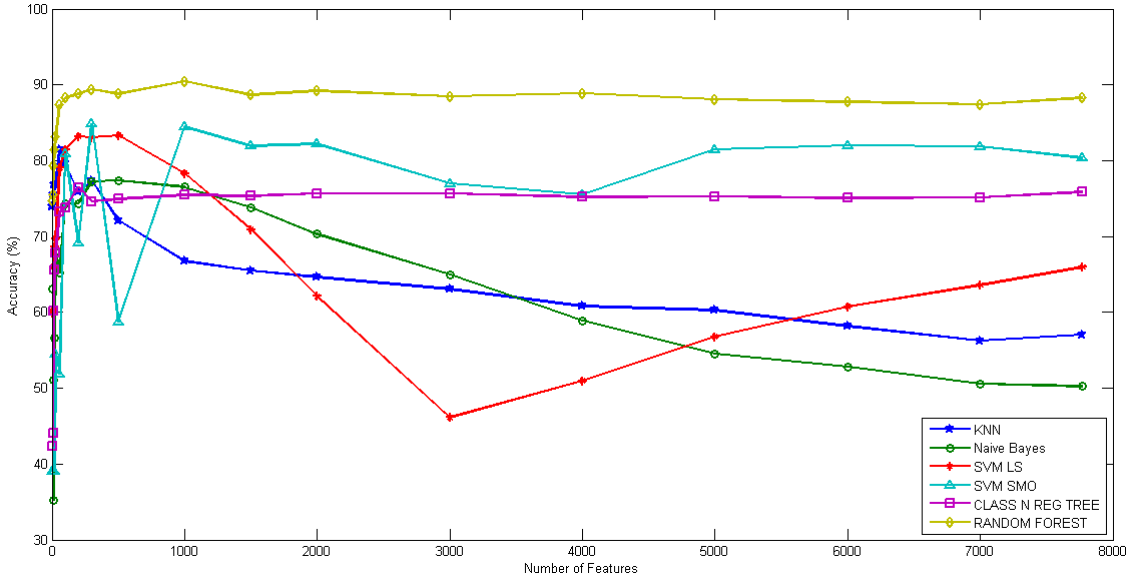


Figure 4.10: Impact of feature selection on Accuracy (WebKB Dataset, TFIDF Representation)

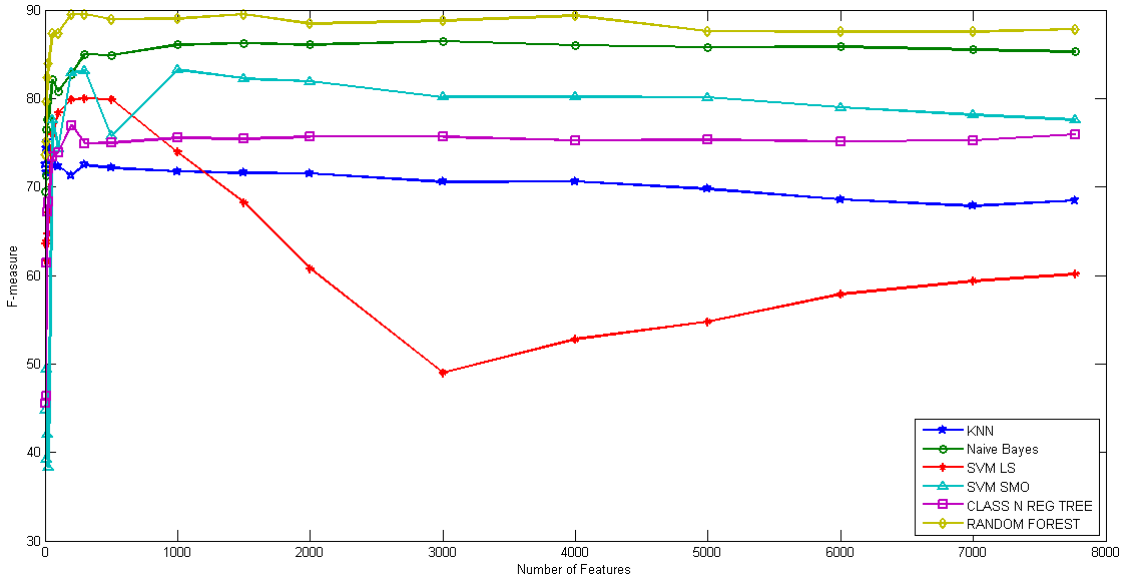


Figure 4.11: Impact of feature selection on F-measure (WebKB Dataset, Bag-of-Words Representation)

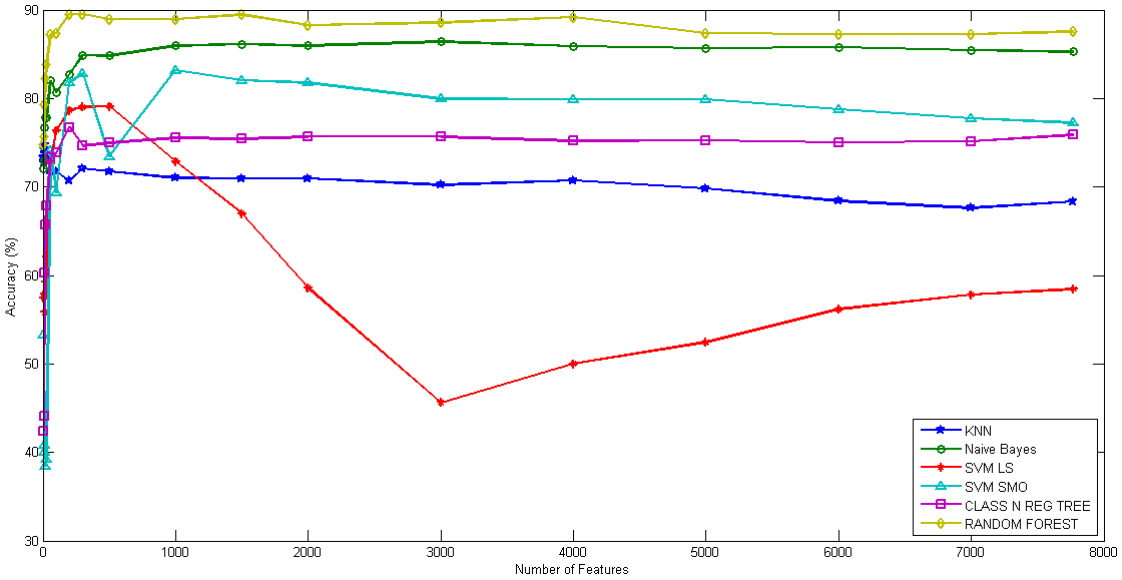


Figure 4.12: Impact of feature selection on F-measure (WebKB Dataset, Bag-of-Words Representation)

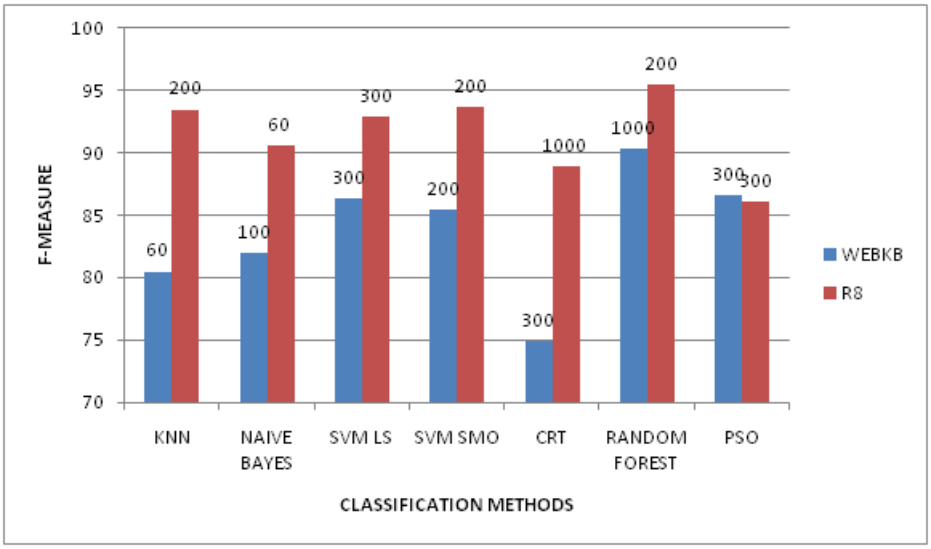


Figure 4.13: Best F-measure (Boolean Representation)

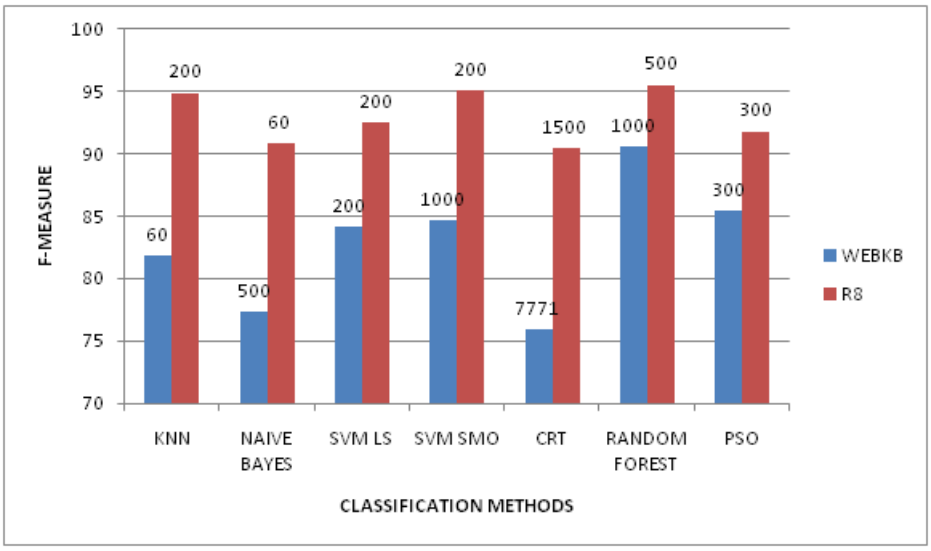


Figure 4.14: Best F-measure (TFIDF Representation)

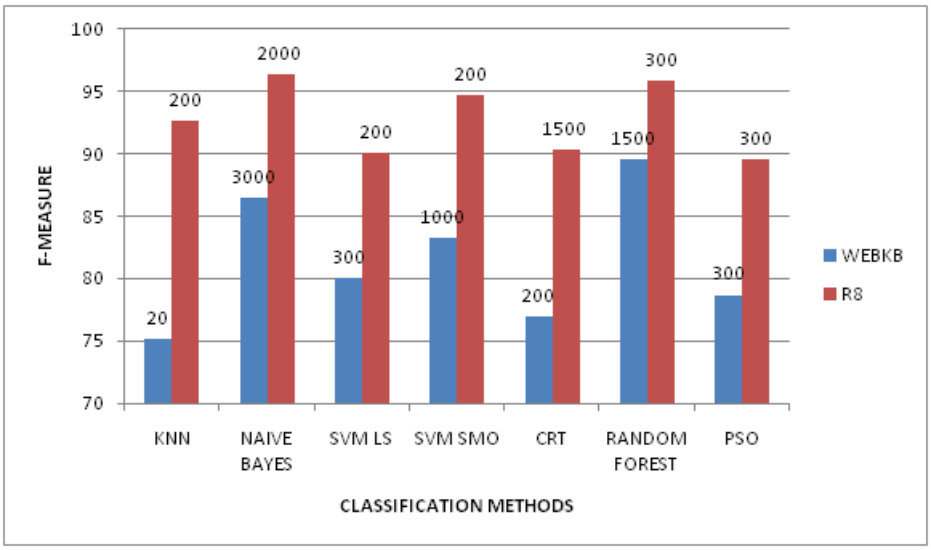


Figure 4.15: Best F-measure (Bag-of-Words Representation)

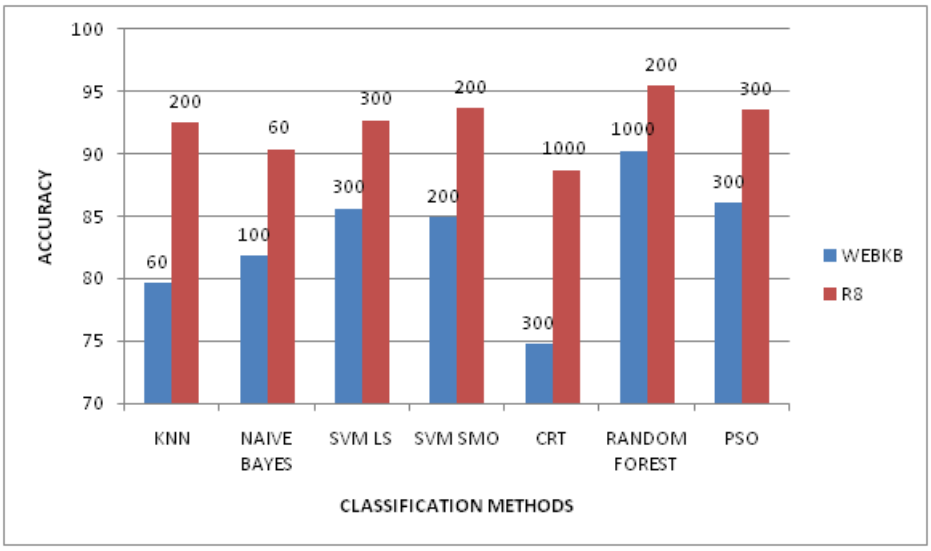


Figure 4.16: Best Accuracy (Boolean Representation)

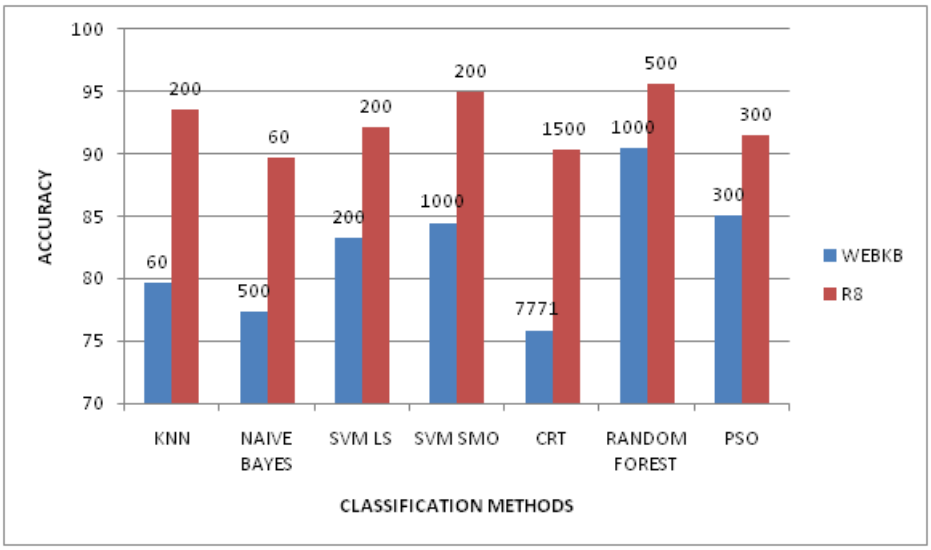


Figure 4.17: Best Accuracy (TFIDF Representation)

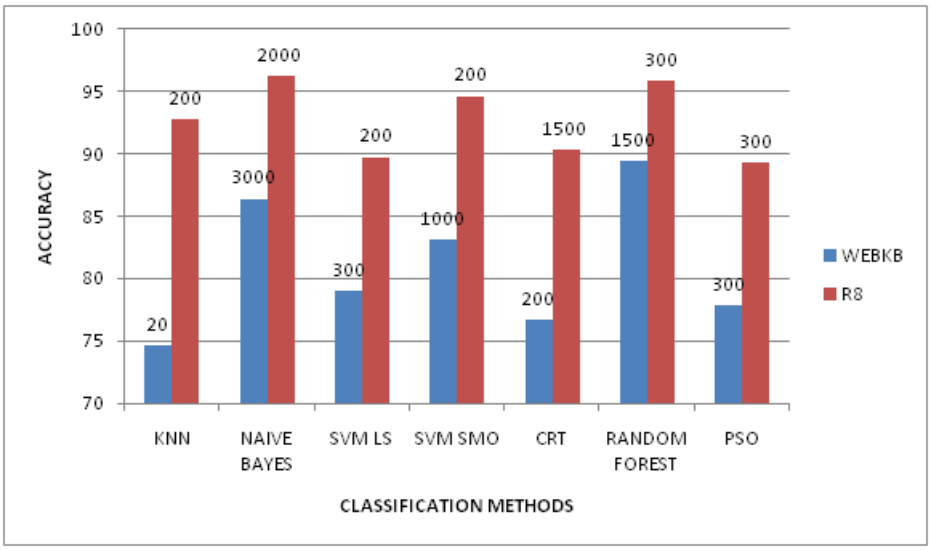


Figure 4.18: Best Accuracy (Bag-of-Words Representation)

Chapter 5

Conclusion

- This disertation work addresses the task of classifying web pages using various classification techniques.
- Performance of KNN, NB, SVM, CART, RF and PSO is compared for different possible representation of web pages.
- Among all the methods, random forest (RF) gives best overall result. Results also demonstrate that the performance of the classifier is affected by the representation used.
- It can be seen from the results that different classification techniques perform best for different representation of the web pages. This implies that there is no single representation which works best for all the classification techniques.
- One should select the representation based on the techniques to be used.
- Impact of feature selection is also studied here and results show that selecting right number of features definitely improves the performance of the classifier.

Bibliography

[1] T. N. Phyu, "Survey of classification tchniques in data mining," *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, 2009.

[2] M. A. Nayak, "A comparative study of web page classification techniques," *GIT-Journal of Engineering and Technology*, vol. 6, 2013.

[3] J. Materna, "Automaticweb page classification," 2008.

[4] W.-K. N. Aixin Sun, Ee-Peng Lim, "Web classification using support vector machine." *Proceedings of the fourth international workshop on Web information and data management - WIDM '02*, 2002.

[5] R.-C. Chen and C.-H. Hsieh, "Web page classification based on a support vector machine using a weighted vote schema," *Expert Systems with Applications*, vol. 31, 2006.

[6] L.-b. X. Yong Zhang, Bin Fan, "Web page classification based-on a least square support vector machine with latent semantic analysis," *Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, 2008.

[7] W. Xue, W. H. Hong Bao, Weimin Xue, and Y. Lu, "Web page classification based on svm," *6th World Congress on Intelligent Control and Automation*, 2006.

[8] D. Radha Damodaram, "Phishing website detection and optimization using particle swarm optimization technique," 2011.

[9] A. K. J. Sarita Mahapatra and B. Naik, "Performance evaluation of pso based classifier for classification of multidimensional data with variation of pso parameters in knowledge discovery database," vol. 34, 2011.

[10] D. Z. Ziqiang Wang, Qingzhou Zhang, "A pso-based web document classification algorithm," *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD 2007)*, 2007.

[11] E. T. De Falco, A. Della Cioppa, "Facing classification problems with particle swarm optimization," *Applied Soft Computing*, vol. 7, 2007.

[12] B. Liu, "Web data mining: Exploring hyperlinks, contents, and usage data (data-centric systems and applications)," *Springer-Verlag New York, Inc., Secaucus, NJ*, 2006.

[13] L. Breiman, J. H. Friedman, R. A. Olshen, and Stone, "Classification and regression trees," 1984.