ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.3

Comparison of Feature Selection Techniques for Spam Email Classification

Prof. Priti Kulkarni[1], Dr. Haridas Acharya[2]
Research Scholar[2]
Symbiosis International University, Pune, Maharashtra, India[1]
Allana Institute of Management Sciences, Pune, Maharashtra, India [2]

Abstract:
Dimension reduction is problem pointing towards reducing high number of features in the dataset. Some of these features are irrelevant or redundant, may leads to inefficient model for classification. Dimension reduction using feature selection allows selecting minimal subset of features that are relevant to target concept for building efficient learning models. Email message is semi-structure, multidimensional in nature and represented by a large number of features. Directly applying classification makes learning intractable. This paper aims to target comparison of various feature selection techniques for email data. In our approach we have used only email header to find best subset of features. The resultant subsets of headers are used for email classification.

Keywords: Email, Classification, Dimension reduction, Feature Selection

I. INTRODUCTION

Email is preferred to use as quickest method of communication. Increasing number of emails in inbox causes inbox congestion. There are some emails in inbox which are very useful, at the same time some emails are useless. These unwanted emails are called as bulk or unsolicited commercial e-mails are curse of email communication. This result into consumption of bandwidth and inbox space. It is therefore required to separate emails into Spam and non-spam. When classifying email, this data contained in messages are very complex, multidimensional, or represented by a large number of features. Directly applying classification makes learning intractable. Feature selection is standard method use to reduce feature dimensionality. Feature selection allows selecting minimal number of features to build model for target concept. Feature selection helps to derive subset of features that are more efficient for classification. It is reported that feature selection can improve the efficiency and accuracy of text classification algorithms by removing redundant and irrelevant terms from the corpus [1]. Most of the research for feature selection on email data is addressed by using message content as features. In this technique words (text) are referred as feature. Email header consists of various attributes. In our approach we have used only header field to find potential useful subset of features for email data. Various feature selection techniques are applied on email data. The resultant subsets of features are used for email classification. The effectiveness of features is represented using Accuracy. This paper is organized as follows. First section describes concept and process of feature selection. Next section discusses the empirical result and comparison of various feature selection techniques.

II. FEATURE SELECTION

Feature selection refers to the problem of dimensionality reduction which allows selecting minimal number of features to build model for target concept. Feature selection helps to derive subset of features that are more efficient for classification. Feature selection reduces the number of features by eliminating features, weighting features or Normalizing features.[2] in his paper described the potential benefits of feature selection techniques as follows,

1. Feature selection facilitate data understanding
2. Reduced the measurement
3. Reduced storage requirement
4. Reduced computational processing time
5. Reduced dimensionality of data
6. Improves classification performance

We mean by the term feature selection in our approach is :
To retain only those features that are meaningful- "those help to build "good" classifier for email data classification.

III. FEATURE SELECTION PROCSS

There are four basic steps in a typical feature selection method [3]

1. A feature subset generation
2. An evaluation function to evaluate the subset under given condition.
3. A stopping criterion to decide when to stop and
4. A validation procedure to check whether the subset is valid.

Feature selection methods are broadly divided into filter and wrapper approaches. The filter model relies on some intrinsic characteristics of data to select features without involving classification learning; the wrapper model typically uses a classifier to evaluate feature quality. Wrapper model is computationally expensive than filter model and derived features are more biased towards classifier used.

IV. RELATED WORK

Email is a combination of structured and unstructured field. Structure part of email consists of email header whereas body of email is considering as unstructured part. Most of the researcher focuses email classification by using body of message as feature. The contents of message body are scan

and term or phrase based approach is used to derive feature from body text. Comparison of various feature selection techniques has been done on various dataset and it suggest that each method is best for some but not for all. [4] Proposed to reduce the feature sets using a simple univariate filter before applying clustering. chi-squared metric is used to select important words from text; [5] used Bayesian classifier using features like sender address, domain type, body and subject to classify spam and non spam and showed 100% accuracy in the result. [6] has classified email considering content based feature. For that they have considered subject, body, and from field to classify email into spam and non spam. They have compared different algorithm such as SVM, KNN, NN, AIS (***Artificial immune*** system) RS (***rough sets***) for performance evaluation on the Spam Assassin spam corpus. It is showed that Naïve bayes and rough sets methods has a very satisfying performance among the other methods. [7] has used header fields received, date, time, IP address for detection of spam emails and compared the performance of C4.5 Decision Tree (DT), Support Vector Machine (SVM), Multilayer Perception (MP), Nave Bays (NB), Bayesian Network (BN), and Random Forest (RF). RF classifier outperform among all classifiers. [8] used email message body and other features number of email sent-received, type and size, and compared AdaboostM1 and instance selection method (ISM), K mean cluster, with Random Forest.

V. OUR APPROACH FEATURE SELECTION

1. Input : Email corpus
2. Extract Email headers
3. Apply feature selection techniques
4. Generate feature subset and select top efficient (60%) features
5. Apply classification techniques
6. Classify email into spam and non spam.

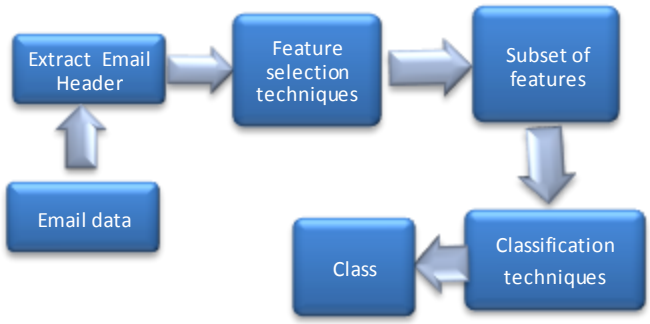

Figure.1. Feature selection

VI. EXPERIMENT AND RESULT

Email in the personal inbox is used as a sample data for extraction into a form which could be submitted as input to the various algorithms chosen. We extracted a total of 849 Emails from the period December 2016 to March 2017 with 41 features. Out of 41 features, 9 features carrying only 5-10% values are most likely not useful to carry any classification and such features are removed. Also we have ignored features that are not directly occurred in our collected sample. In order to obtain a training corpus for supervised learning algorithms, emails were classified as spam and non spam. The main purpose is to find out whether all features are required for email classification. Whether accuracy of classifiers are affected by feature subset selection.

TABLE.1. DATA DESCRIPTION TABLE

No. of features	No. of Instances	No. of Classes	Missing values
32	*849*	*2*	*yes*

Experiment is conducted using Weka tool with four different feature selection methods along with supporting search techniques. The resulting subset are used further with standard classifiers for classification

TABLE. 2. EXPERIMENT OUTPUT

Feature selection technique	Search Techniques	No of features selected	Classifiers Accuracy		
			J48	SMO	Adaboost
CfsSubsetEval	Best-First	3	89.63%	88.93%	83.8634
Chisquareattribute eval	Ranker	10	92.34%	94.82%	88.34%
GainRatioAttribute Eval	Ranker	10	91.99%	93.99%	84.33%
Wrapper subset Eval	Greedy search	7	93.05%	95.05%	88.34%

[9]Cfs Subset Val evaluates subsets of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. Chi Squared Attribute Eval evaluates chi squared statistics with respect to class. Gain Ratio Attribute Eval methods evaluate by measuring the gain ratio with respect to the class. Wrapper subset eval evaluates attribute sets by using a learning scheme. SMO classifier shows higher accuracy among all classifiers but it takes more time to build model. Comparatively all three classifiers shows decrease in accuracy with feature subsets used by Cfs Sub set Eval technique. All classifiers shows higher accuracy when used with Wrapper feature selection as it evaluates feature subset using classifier. Chi square attribute eval performs well than Gain Ratio Attribute Eval feature selection technique.

VII. CONCLUSION

Experiment result shows that use of various feature selection techniques with search method to reduce dimension of dataset. For email dataset SVM (SMO) shows higher accuracy but took more time to build. Adaboost performs poor with selected feature set. The time to build model and size of feature subsets vary with each method. Once best feature selection technique is detected for a data set same can be use along with classifier to increase the accuracy.

VIII. REFERENCES

[1]. Yanjun Li, Congnan Luo,(2008) ," text clustering with feature selection by using statistical data", IEEE Transactions on Knowledge and Data Engineering, May 2008 vol: 20 no:5

[2]. Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. Advances in bioinformatics, 2015.

[3]. Dash, M., & Liu, H. (1997). Feature selection for classification. Intelligent data analysis, 1(1-4), 131-156.

[4]. (Dey Sarkar, S., Goswami, S., Agarwal, A., & Aktar, J. (2014). A Novel Feature Selection Technique for Text Classification using Naïve Bayes. International Scholarly Research Notices, 2014.

[5]. Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998, July). A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization: Papers from the 1998 workshop (Vol. 62, pp. 98-105).

[6]. W.A. Awad1 and S.M. Elseuofi, "Machine Learning Methods For Spam E-Mail Classification", International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011

[7]. Al-Jarrah, O., Khater, I., & Al-Duwairi, B. (2012). Identifying potentially useful email header features for email spam filtering. In The Sixth International Conference on Digital Society (ICDS).

[8]. Rafiqul Islam and Yang Xiang, Email Classification Using Data Reduction Method, (IEEE) in CHINACOM 2010 : Proceedings of the 5th International ICST Conference on Communications and Networking in China, IEEE, Piscataway, N.J., pp. 1-5

[9]. Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2010). WEKA manual for version 3-7-3. The university of WAIKATO.