Indian Journal of Science and Technology, Vol 9(39), DOI: 10.17485/ijst/2016/v9i39/90599, October 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Empirical Study of Feature Selection Methods for High Dimensional Data

S. DeepaLakshmi[1*] and T. Velmurugan[2]

[1]Bharathiar University, Coimbatore - 641046, Tamil Nadu, India; deepa.dgvc@gmail.com
[2]PG and Research Department of Computer Science, D. G. Vaishnav College, Chennai - 600106, Tamil Nadu, India; velmurugan_dgvc@yahoo.co.in

Abstract

Background/Objectives: Feature Selection is a process of selecting features that are relevant which is used in model construction by removing redundant, irrelevant and noisy data. A typical application of Text Mining is classification of messages and e-mails into spam and ham. **Methods/Statistical Analysis:** This article gives a comprehensive overview of the various Feature Selection methods for Text Mining. Various Filter methods like Pearson Correlation, Chi-square, Symmetrical Uncertainty and Mutual Information are applied to select the optimal set of features. **Findings:** Filter Feature Selection methods are used to classify Text data. Various Classification algorithms are applied using the optimal set of features obtained. The accuracy of classification algorithms are verified based on the chosen data set. **Novelty/ Improvements:** A comparative study of various filter methods for Feature Selection and classification algorithms for performance evaluation is conceded in this research work.

Keywords: Chi-Square, Feature Selection, Filter Method, Mutual Information, Pearson Correlation

1. Introduction

The quantity of high-dimensional data that exists has increased in the past few years. Feature Selection is an important technique in data pre-processing and is an important component of the machine learning process. The process of selecting the most relevant subset of attributes from large set of attributes according to some selection criteria is known as Feature Selection. High-dimensional data contains irrelevant or redundant features that results in decrease in the accuracy of data mining algorithms, Increase in the time taken for mining process, problems in retrieval and storage and Interpretation becoming hard. Text data contains a huge collection of documents. Text Mining processes unstructured data into meaningful numeric values which can be used by data mining algorithms. A common application for Text Mining is the classification of messages or e-mails as spam or ham to enable filtering of undesirable junk messages. Many research works has been carried out to find the related and relevant features to distinguish between spam and ham messages. Spam is junk email or unsolicited bulk e-mails that flood the Internet with many copies, mostly commercial advertisement for dubious products or quasi-legal services. Short Message Service (SMS) spam is annoying and expensive. There are two methods of detecting spam namely Collaborative based and Content based. Collaborative is based on feedback from the users and Content is based on analyzing the textual content of messages. Extensive research has been carried on Content based methods and some of the work is discussed below.

A Research work that combined both SMS Specific feature (SMSS) and Linguistic Inquiry and Word Count (LIWC) in the detection of SMS spam and a classification accuracy obtained was better than other methods[1]. A Feature Selection method which consists of filter and wrapper Feature Selection process was proposed[2]. The result has revealed that a combination of various methods was more effective than a single selection method. Filter methods using Mutual Information, Chi-square and Information Gain are used and Genetic Algorithm was

*Author for correspondence

applied for the Reuter's dataset and Newsgroup dataset and a high accuracy was obtained. Sin-Eon Kim proposed a FR (Frequency Ratio) measure which divides the spam class and ham class and evaluates the frequencies of words that appear in the SMS messages[3]. FR method was executed on SMS Spam Collection v.1 dataset and an accuracy of 94.7% was obtained using Naïve Bayes classifier and 94.82% using J-48 classifier. Yiping Zheng and F. Liu proposed Keyword selection by linear discrimination with Approximated Logistic Regression (KW-ALR) which uses linear discrimination analysis to extract features and uses logistic regression to train spam recognition model[4]. KW-ALR method was executed on SMS spam Collection v.1 and an accuracy of 89.4% was obtained.

In[5] used Genetic Algorithm (GAFS) to find the subset of features which is optimal for the Spam dataset was proposed by Mark Hopkins and with 4601 e-mails[5]. An accuracy of 91.8% and 89.1% was obtained using Bayesian and KNN classifiers. Subhajit Dey Sarkar proposed FS-CHICLUST which uses Chi-square and a feature clustering algorithm to select the important words[6]. SMS corpus dataset was one among the 13 datasets used. A Frequent Itemset and Ensemble Learning (FIEL) were used to find the item set which is frequent[7]. Naïve Bayes, LibSVM and Random Forest which use majority voting system were used. FIEL method was compared with other methods and the result showed that FIEL was more stable than other methods. Ashish Chandra used 32 low cost quality factors for classifying spam and ham messages[8]. The features were divided into URL features, Content features and Link features. Resilient Back-propagation algorithm was used as a classifier and an accuracy of 92% was obtained. A study of E-mail Classification was done which a new Feature Selection technique guided by F-selector package was used[9]. The filter methods enabled the classifiers to achieve maximum accuracy of 93.27%. Feature Selection using Principal Component Analysis and decision forest method was proposed and evaluated using social media data set[10]. A Feature Selection method using SVM-RFE and CV technique was proposed and genes were ranked based on Cumulative Ranking Score[11]. A Feature Selection algorithm to find the dependent attribute from a cluster using minimum variance method was proposed by Sivakumar[12]. A Research work in Haddop framework using random forest based on Rough-set Feature Selection was proposed by Thulasi Bikku[13]. Ha Van Sang integrated parallel Random Forest method and Feature Selection in credit scoring model[14].

The work presented in this paper is centered in Filter methods used for Feature Selection of High Dimensional data specifically Text Mining. The rest of the paper is structured as follows. In Section 2, the various Filter Feature Selection methods, the Classification algorithms and the performance metrics are elaborated. Section 3 discusses the results and the filter methods are compared. Finally, Section 4 concludes the research work.

2. Materials and Methods

Feature Selection has three general Approach[15] which are Filter Approach that selects the features which is not dependent on the classifier, the Wrapper Approach that selects the features using the classifier and Embedded Approach that is a combination of Filter and Wrapper approach. Filter methods are simple and Fast and independent of any mining algorithms. Filter approach uses Independent Criteria to evaluate the feature subset without using a learning algorithm. Filter methods are either univariate that considers one variable at a time or multivariate that considers more than one variable at a time[15]. Multivariate filter method shows Feature Dependencies and its computational complexity is better than wrapper methods. It is slower less scalable than univariate method.

Wrapper methods select the Features by using a specific mining algorithm as part of the Evaluation Function[16] but these methods are computationally expensive. Wrapper method selects an optimal subset which is best suited to learning algorithm. The wrapper approach is accurate but computationally expensive. Embedded methods combine both the Filter and Wrapper methods. It uses independent criteria to decide optimal subset and a learning algorithm to select the final optimal subset.

2.1 Filter Method

The most commonly used Filter method for Text Mining is the Correlation based feature selection. Correlation coefficient determines the statistical relation between features or attributes. The types of correlation coefficient are Pearson correlation coefficient, Rank correlation coefficient like spearman rank correlation, Kendall Tau and Goodman and Kruskal's gamma correlation coefficient.

2.1.1 Pearson Correlation

Pearson correlation coefficient ρ measures the strength of the relationship between two features, giving a value

between +1 and –1, where 1 indicates positive, 0 indicates no correlation and -1 is negative correlation[17]. Correlation coefficient based Feature Selection is used to find the best subset of features and is combined with search strategies. The formula for ρ can be written as:

$$\rho_{M,N} = \frac{E[(M - \mu_M)(N - \mu_N)]}{\sigma_M \sigma_N} \quad (1)$$

Where M and N are the features or attributes, μ_x is the mean of M and μ_y is the mean of N, σ_M is the standard deviation of M and σ_N is the standard deviation of N and E is the expectation. Rank Correlation coefficient measures the degree of similarity between two features that are ranked and can be used to assess the impact of the relation between them. The rank correlation statistics are Spearman correlation, Kendall correlation and Goodman and Kruskal coefficient. Spearman's rank correlation measures the relationship between two features using a monotonic function. Kendall correlation coefficient measures a portion of ranks between two data sets.

2.1.2 Chi-square

Chi-square Feature Selection χ^2 test is used to test whether the occurrence of a feature is independent of the class[18]. High values of χ^2 indicates that the feature and the class are independent.

$$\chi^2 = \sum \frac{(O-E)^2}{E} \quad (2)$$

Where O is the frequency that is Observed, E is the frequency that is Expected for each feature and class.

2.1.3 Mutual Information

Mutual Information measures mutual dependence between features in bits. It is a technique that determines how one variable is dependent on another variable.

$$I(M;N) = \sum_{n \in N} \sum_{m \in M} p(m,n) \log \left(\frac{p(m,n)}{p(m)p(n)} \right) \quad (3)$$

Where M and N are two features, p (m, n) is the joint probability distribution function of M and N, p (m) and p (n) are the probability distribution functions of M and N. I (M, N) is always greater than or equal to zero[18]. The dependency between M and N is stronger if mutual information value is higher.

2.1.4 Information Gain

Information Gain measures the amount of information in bits obtained for prediction of a class by determining the presence of a feature in a dataset. It determines the change in entropy when the feature is present vs. when the feature is absent. Entropy is a measure of uncertainty or unpredictability in a system. It is the basis for Information Gain attributes ranking methods. Entropy of a feature M, H (M), entropy of feature M after observing N, H (N/M) and Information Gain IG is given by:

$$H(M) = -\sum p(m) \log_2 (p(m)$$
$$H(N/M) = -\sum p(n) \sum p(m/n) \log_2 (p(m/n)$$
$$IG = H(M) - H(M/N) \quad (4)$$

2.1.5 Gini Index

Gini Index is a statistical measure of dispersion. It is a supervised multivariate method that measures the ability of a feature to distinguish between classes. Gini Index has smaller values for the most relevant features. Gini Index is determined for each attribute and the top attributes with smallest Gini Index values are selected. P_x is the relative frequency of class x in Data D and Gini Index G is given by:

$$GiniIndex(G) = 1 - \sum p_j^2 \quad (5)$$

2.2 Classification Methods

Classification refers to the task of predicting class labels or classification of data using the training data and the class attribute. Commonly used classifiers are Probabilistic classifier, Decision Tree classifier, Support Vector Machines and Memory based classifiers[19]. Probabilistic classification algorithms use statistical inference to find the best class for a given instance. The Naïve Bayes Classifier is based on Bayes' theorem that assumes that the features are not dependent on each other. An advantage is that it uses less number of training data to determine the parameters used for classification. Naïve Bayes is a popular method used for text categorization.

Multilayer perceptron is a feed forward Artificial Neural Network that uses a supervised learning technique called back propagation. It can be viewed as a logistic regression classifier which uses a non-linear transformation to transform the input. The intermediate layer is

the hidden layer and a single layer is sufficient to make MLP's a universal approximator. Decision tree learning is a supervised classification learning that assumes that all features have finite discrete domains. It is a tree in which input features are present in the internal node and each leaf is labeled with a class. ID3, C4.5, CART and MARS are some of the Decision Tree algorithms. J48 is a Java implementation of C4.5 algorithm that builds decision tree using information entropy. C4.5 selects the attribute that splits the data into subsets and information gain is the criteria used for splitting. C4.5 can handle continuous attributes, discrete attributes, attributes with missing values and attributes with differing cost.

Memory-based learning or Instance-based learning or exemplar-based learning is a classification technique based on k-nearest neighbor. It finds the appropriate class by learning from a set of examples. Memory-based learning is a part of the paradigm of Lazy Learning. Lazy learners store the data without making any modification. It is based on the assumption that instances that are similar belong to the same class. IB1, IBk, K star and LWL classifiers are examples of Memory-based learning.

2.3 Performance Measures

The research work measures the performance by using various parameters like Confusion matrix, Precision and Recall, Accuracy and Error Rate[20]. The Instances in a predicted class are represented by the column of the confusion matrix and the instances in the actual class are represented by the rows of the confusion matrix. Table 1 shows the confusion matrix, where True Positive rate is the proportion of positive cases and True Negative is the number of negative cases that were identified correctly, False Positive is the number of negative case and False Negative is the number of positive cases that were incorrectly classified.

$$\text{Specificity (\%)} = \frac{TN}{FP + TN} \times 100 \quad (6)$$

Table 1. Confusion matrix

Confusion Matrix		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Precision is the number of positive instances that were predicted accurately and Recall is the number of positive instances that were identified correctly.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

Accuracy (ACC) is the proportion of the total number of predictions that were correct and Error Rate is the proportion of instances that were misclassified. It is given by the Equation:

$$\text{Accuracy(\%)} = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \quad (9)$$

$$\text{Error rate} = 1 - \text{ACC} \quad (10)$$

F-measure finds the average of the precision and recall. Large F-measure value indicates a higher classification quality.

$$\text{F} - \text{measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Precision} + \text{Recall}} \quad (11)$$

3. Experimental Results

This research work proposes the usage of filter methods to remove irrelevant features from the set of features. T-Relevance, the relevance between the feature and the target class C is computed. If the T-Relevance is greater than a predetermined threshold, the feature is selected. Filter methods like Pearson Correlation, Chi-square, Symmetrical Uncertainty and Mutual Information are used to determine the T-Relevance of the features and the feature which has the T-Relevance value greater than the threshold is selected. Classification of the dataset using the selected features is done using the classifiers Naïve Bayes and J48.

3.1 Description of Data Set

Data set from UCI machine learning repository was used in this research work. The SMS Spam Collection v.1 is a set of SMS messages collected from SMS Spam research. It contains a set of SMS messages in English of 5,574 messages, tagged as ham or spam. The number of legitimate messages is 4,827 (86.6%) and the number of spam messages is 747 (13.4%). Each line in the file contains one message. Each line has label ham (legitimate) or spam and the raw text. The text data is converted into word which

is a set of attributes representing word occurrences. The number of attributes obtained is 1833. Table 2 gives the number of instances for legitimate and spam messages.

3.2 Results and Discussion

This research work used MATLAB software to find the optimal subset of attributes using the filter methods. Classification was done using WEKA software and classifiers like Naïve Bayes, IB1 and J48 were used. Table 3 illustrates the True Positive and Negative values obtained by Naïve Bayes Classifier using the various filter Feature Selection methods. Pearson Correlation has highest sensitivity and lowest specificity and Symmetric Uncertainty has the lowest sensitivity and highest specificity. Sensitivity and specificity of Chi-square and Mutual Information are also specified.

Figure 1 compares the sensitivity and specificity of Naïve Bayes classifier using Pearson Correlation, Symmetric Uncertainty and Chi-square Feature Selection methods. Accuracy obtained by Naïve Bayes Classifier using Features selected by Pearson Correlation is high compared to the accuracy obtained using other Filter Feature Selection methods. Table 4 gives the accuracy of the classifiers using various filter Feature Selection method. The accuracy of the classifiers Naïve Bayes, J48 and IB1 using the features selected by Pearson Correlation is 94.86, 94.68 and 93.38 respectively. The accuracy obtained using Symmetric Uncertainty Feature Selection method is the less compared to other Feature Selection methods. Figure 2 compares the Accuracy of classifiers Naïve Bayes, J48 and IB1 using the filter methods Pearson Correlation, Symmetric Uncertainty, Chi-square and Mutual Information.

Figure 3 shows the Error rate of Naïve Bayes classifier using the attributes selected by various filter methods. Pearson Correlation has the least Error rate compared to other filter Feature Selection methods. Symmetric Uncertainty has the highest Error rate.

Table 2. SMS spam collection v.1 dataset

No	Class Label	Number of Instances
1	Legitimate	4827
2	Spam	747
	Total	5574

Table 3. True Positive (Sensitivity) and False Negative (Specificity) values of Naïve Bayes Classifier

Feature Selection	Pearson correlation	Symmetric Uncertainty	Chi-square	Mutual Information
Sensitivity	0.949	0.937	0.945	0.944
Specificity	0.248	0.361	0.312	0.306

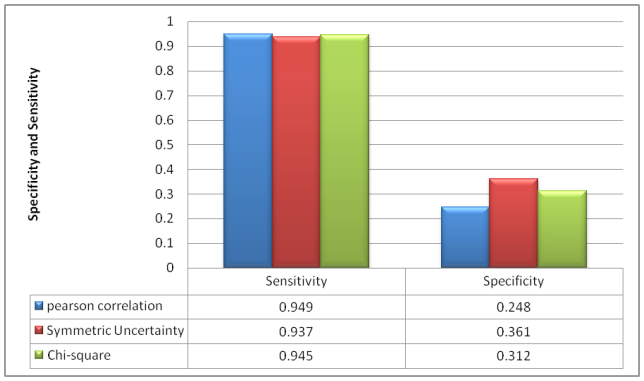

Figure 1. Specificity and sensitivity of Naïve Bayes.

Table 4. Accuracy of classifier for various Feature Selection methods

Feature Selection	Naïve Bayes	J48	IB1
Pearson correlation	94.869	94.689	93.38
Symmetric Uncertainty	93.702	93.702	85.378
Chi-square	94.474	94.887	93.541
Mutual Information	94.384	94.384	93.254

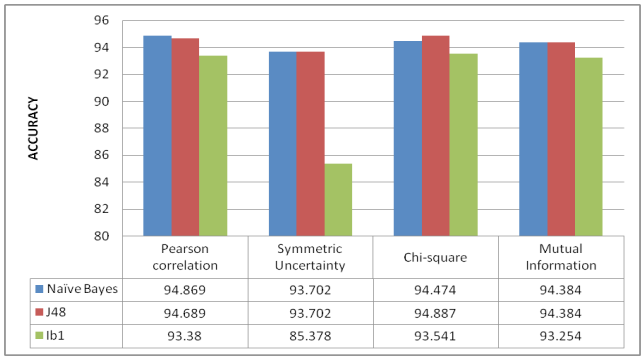

Figure 2. Accuracy of the classifiers.

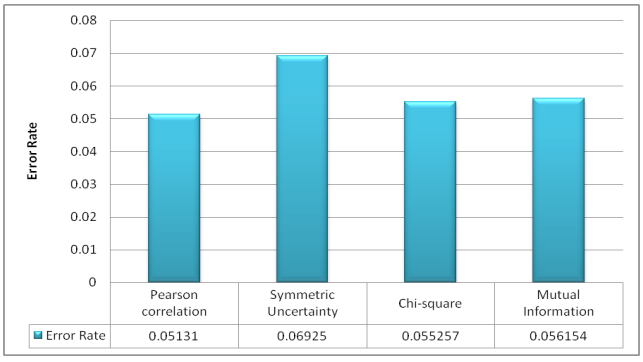

Figure 3. Error Rate of Naïve Bayes classifier.

Accuracy of Pearson Correlation is higher than other Feature Selection methods. The Error rate of Pearson Correlation is lower than other methods. Based on the Precision, F-measure, Accuracy and Error Rate, it is shown that Pearson Correlation has better performance in selecting the minimum number of features for Text Data.

4. Conclusion

This research work aims to compare the performance of the various filter methods used for Text Mining and specifically for Detection of spam and ham from a collection of SMS messages. Filter Feature Selection methods like Pearson Correlation, Chi-square, Mutual Information and Symmetrical Uncertainty have been applied on the SMS Spam Collection v.1 dataset in the UCI machine learning repository and relevant features have been selected. The main work of this research is to classify the SMS messages into spam or ham with minimum number of features. Classification algorithms namely Naïve Bayes and J48 have been used to classify the dataset with the features selected by the filter Feature Selection methods. Generally, filter Feature Selection methods are efficient for analyzing Text Data. This work identifies that the Pearson Correlation method perform well in order to classify the text data for the chosen data set in terms of selecting the minimum number of features. The future work will extract the use of other algorithms in the same context of the Feature Selection algorithms.

5. References

1. Karami K, Amir A, Zhou L. Improving static SMS spam detection by using new content-based features. AISeL. 2014. p. 1–9.
2. Uysal AK, Gunal S. A novel probabilistic Feature Selection method for text classification. Knowledge-Based Systems. 2012; 36:226–35.
3. Kim K, Sin-Eon S, Jo J, Choi SH. SMS Spam filtering using Keyword Frequency Ratio. International Journal of Security and its Applications. 2015; 9(1):329–36.
4. Zheng Z, YipingY, Liu F. Filtering Network Spam Message using approximated logistic regression. Journal of Networks. 2014; 9(9):2462–7.
5. Kalaibar K, Sorayya Mirzapour S, Naser Razavi S. Spam filtering by using Genetic based Feature Selection. International Journal of Computer Applications Technology and Research. 2014; 3(12):839–843.
6. Sarkar D, Subhajit S, Goswami S, Agarwal A, Aktar J. A Novel Feature Selection Technique for Text Classification using Naïve Bayes. International Scholarly Research Notices; 2014. p. 10.
7. Ahmed A, Ishtiaq I, Ali R, Guan D, Lee YK, Lee S, Chung TC. Semi-supervised learning using frequent item set and ensemble learning for SMS classification. Expert Systems with Applications. 2015; 42(3):1065–73.
8. Chandra C, Ashish A, Suaib M, Beg B. Web spam classification using supervised Artificial Neural Network algorithms. 2015; 2(1):1–10.
9. Parimala R, Nallaswamy R. A Study of Spam E-mail classification using Feature Selection package. Global Journal of Computer Science and Technology. 2011; 11(7):1–11.
10. Jotheeswaran J, Koteeswaran S. Feature Selection using Random Forest Method for sentiment analysis. Indian Journal of Science and Technology. 2016 Jan; 9(3). DOI: 10.17485/ijst/2016/v9i3/86387.
11. George GVS, Raj VC. Accurate and stable Feature Selection powered by iterative backward selection and cumulative ranking score of features. Indian Journal of Science and Technology. 2015; 8(11). DOI: 10.17485/ijst/2015/v8i11/71766.
12. Sivakumar V, Sivakumar S, Selvaraj R. A novel clustering based Feature Subset Selection Framework for effective data classification. Indian Journal of Science and Technology. 9.4 2016; 9(4). DOI: 10.17485/ijst/2016/v9i4/87038.
13. Bikku T, Sambasiva Rao N, Akepogu AR. Hadoop based Feature Selection and Decision Making Models on Big Data. Indian Journal of Science and Technology. 2016; 9(10). DOI: 10.17485/ijst/2016/v9i10/88905.
14. Van Sang H, Ha Nam N, Duc Nhan N. A novel credit scoring prediction model based on Feature Selection approach and parallel random forest. Indian Journal of Science and Technology. 2016; 9(20). DOI: 10.17485/ijst/2016/v9i20/92299.
15. Liu L, Huan H, Motoda M. Feature Selection for knowledge discovery and Data Mining. Springer Science and Business Media. 2012; 454:214.
16. Kohavi K, Ron R, George H, John J. Wrappers for feature subset selection. Artificial Intelligence.1997; 97(1):273–324.
17. Hall H, Mark A. Correlation-based Feature Selection for machine learning. The University of Waikato; 1999. p. 1–198.
18. Yang Y, Yiming Y, Jan O, Pedersen P. A comparative study on Feature Selection in text categorization. ICML; 1997. p. 1–9.
19. Taneja T, Gaurav G, Ashwini Sethi A. Study of classifiers in Data Mining. International Journal of Computer Science and Mobile Computing. 2014; 3(9):263–9.
20. Sokolova S, Marina M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information Processing and Management. 2009; 45(4):427–37.