IJRSET August 2015 Volume 2, Issue 6 Pages: 1-8

ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

FEATURE SELECTION APPROACHES WITH TEXT MINING FOR CATEGORICAL VARIABLE SELECTION

[1] C. Kanakalakshmi, [2] Dr. R. Manicka chezian
[1] Research Scholar, [2] Associate Professor,
[1, 2] Department of Computer Science,
[1, 2] Nallamuthu Gounder Mahalingam College,
[1, 2] Pollachi, India.

Abstract:-

Feature Selection is the process of selecting a subset of relevant features for use in model construction. The Feature Selection methods are used to increase the overall efficiency of the classification model. The amount of text data is increasing rapidly in recent years, the feature selection approaches are important for the preprocessing textual documents for data mining. The feature selection method focuses on identifying relevant data that help to reduce the preprocessing of huge amount of data and reduce the data size by removing irrelevant or redundant attributes. The feature selection algorithm conducts a search for best subset using valuation algorithm. The valuation algorithm is run on the dataset with different set of features removed from the data. The main objective of this paper is to improve the accuracy of classification of text documents by removing the irrelevant, noisy features and compare the precision and recall of various Feature Selection methods. The performance of Feature Selection methods with various classifiers are compared and tabulated.

Keywords: - Text Classification, Feature Selection, Filter Approach, Precision, Recall, F-Measure.

1. INTRODUCTION

Text Mining is the process of extracting interesting information or knowledge or patterns from the unstructured text that are from different sources. Text classification [1] is the process of classifying documents into predefined categories based on their content. The text feature space is sparse and high dimensional. The high dimensional feature space will increase the training time and affect the accuracy of the classifiers. Text mining applications need to deal with large and complex datasets of textual documents that contain much relevant and noisy information. Feature selection [3] aims to remove that irrelevant and noisy information by focusing only on relevant and informative data for use in text mining. By focusing on the selected subset of features, simple and fast models can build by using only the subset and gain better understanding of the processes described by the data. Many techniques are developed for selecting an optimal subset of features from a larger set of possible features. The Feature selection methods are used to increase the overall efficiency of the classification model.

Feature selection techniques [4] [14] can be divided in to three types depending on how they interact with the classifier namely Filter method,

Wrapper method and embedded method. Filter method directly operate on the dataset, and provide a feature weighing, ranking or subset as output. Wrapper method performs a search in the space of feature subsets, guided by the outcome of the model. Embedded methods use internal information of the classification model to perform feature selection.

2. RELATED WORK

Ramaswami et al [3] presented the work on Feature Selection methods used in the educational field for the purpose of extracting useful information on the behaviors of students in the learning process. Divya et al [4] had given details about the steps in feature selection and the feature evaluation techniques filter and wrapper methods that are used for subset selection for text classification and text clustering. Bozhao Li et al [7] proposed the use of text categorization method to predict the trend of the stock. The several categorization methods including the feature selection methods are compared. The result show that the SVM method with information gain give the better performance for the predict of the stock with the news. Girish Chandrashekar et al [9] presented the feature selection methods to find a subset of variables which improves the overall prediction performance. The various filter and wrapper approaches are discussed for un-supervised and semi-supervised learning and the stability of the feature selection algorithms. George Forman et al [11] had presented an extensive comparative study of feature selection metrics for the high-dimensional domain of text classification. Sherin Mary Varghese et al [12] proposed a algorithm for feature subset selection. The proposed Pearson correlation measure focused on minimized redundant data, and a small number of discriminative features are selected. Jaga Priya Vathana et al [13] had proposed a FAST algorithm to identify and remove the irrelevant data set. Feature subset selection research is focused on searching for relevant features. The proposed fuzzy logic has focused on minimized redundant dataset and improves the feature subset accuracy. Sagar Imambi et al [14] introduced a novel feature weighting scheme GRW which improves the classification accuracy, and well in high dimension and unevenly distributed document classification. Nirmala Devi et al [16] had presented an overview of the different feature selection techniques for classification by reviewing the most important application fields in the bioinformatics domain.

3. FEATURE SELECTION OF METHODS

Feature selection is most important and frequently used techniques in data preprocessing for data and text mining [5] [8] [9]. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Based on the feature, the vector of unclassified document is compared with the vector of training set document. Feature selection can be divided in filter methods and wrapper methods [4] [5] [14]. The general Feature Selection methods are given in Table 1.

Approaches	Single Feature Evaluation
Filter	Mutual Information Chi square statistic Entropy Information Gain
Wrapper	Ranking accuracy using single feature

Table 1: Feature Selection Categories

FILTER METHOD:

Filter methods [6] [11] [12] is defined as using some actual property of the data in order to select feature using the classification algorithm. Features selected using the filter approach is the input variables to the different

classifiers. The various Filter methods [7] [13] [15] are Correlation Coefficient method, Chi-Squared, Information Gain, Gain Ratio. The filter techniques assess the relevance of features by using the intrinsic properties of the data. The feature relevance score is calculated, and the low features are removed. The subset of features is given as input to the classification algorithm. The Feature selection is performed only once and then different classifies can be evaluated.

WRAPPER METHOD:

Wrapper method [9] [10] [16] is a simple and effective way for variable selection. A wide range of search strategies can be used, including branch-and-bound, best-first and genetic algorithm simulated annealing to find a subset of variables which maximizes the classification performance. The general framework for feature selection for classification framework is shown in Figure 1.

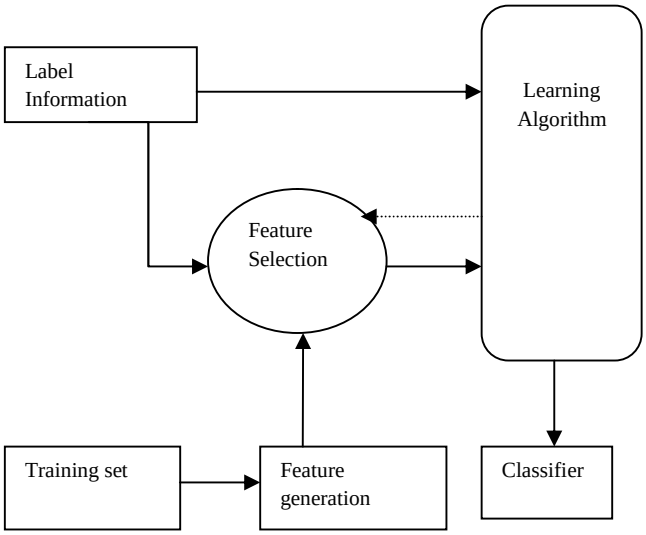

Figure 1: A General Framework of Feature Selection for Classification

4. CLASSIFIERS

Many algorithms [9] [15] [17] have been developed to deal with automatic text classification. The most common techniques used for this purpose include Association Rule Mining, Implementation of Naïve Bayes Classifier, Nearest Neighbors', and Decision Tree and so on. The attributes selected by using the feature selection method is given as input for the classifiers to evaluate the performance of the classification process.

5. PERFORMANCE MEASURES

There are various methods to determine effectiveness or the performance of the algorithms. The metrics Precision, Recall, and F-measure are most often used.

Precision [2] [18] is determined as the conditional probability that a random document d is classified under ci, or what would be deemed the correct category.

$$Precision = \frac{TP}{TP+FP}$$

Recall is defined as the probability that, if a random document (dx) should be classified under category (ci), this decision is taken.

$$Recall = \frac{TP}{TP+FN}$$

Precision and recall are often combined in order to get a better picture of the performance of the classifier given as F-Measure [15]

$$F - Measure = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

The Receiver Operating Characteristics (ROC) [3] curve is a graphical representation of the tradeoff between the false negative and false positive rates for every possible cut off. Equivalently, the ROC value is the representation of the tradeoffs between Sensitivity and Specificity.

6. EXPERIMENTS AND EVALUATION

The Performance metrics of the text classifiers such as Naïve Bayes, Support Vector and Decision Table are compared with the Filter methods. The Hepatitis dataset is used for the evaluation which contains 255 instances and 20 attributes. The dataset is obtained from the Universal Client Identification (UCI) repository. The Weka tool is used for the evaluation of the

classifiers. The attributes are selected using the different filter methods such as Correlation Coefficient, Chi Squared, Info Gain, Gain Ratio and Filtered Subset and the performance metrics value obtained for the corresponding attributes selected on applying different classifiers are tabulated.

Attribute Evaluator	No of Attributes Selected	Search Method
CfsSubset	10	Best First
ChiSquared	19	Ranker
ConsistencySubset	12	BestFirst
InfoGain	19	Ranker
GainRatio	19	Ranker
FilteredSubset	19	Greedy Stepwise

Table 2: No of Attributes Selected and the Search Method used for Feature Selection.

The attribute evaluator selects the relevant attributes by using the corresponding search method. The Cfs subset selects 10 attributes from the given data set. The Chi Squared, Info Gain, Gain Ratio and Filtered subset methods selects same 19 attributes from the given dataset. By using the Classifier Subset Evaluator, the number of attributes selected by the corresponding classifiers is given in Table 3.

Classifier Subset Evaluator	No of Attributes Selected
BayesNet	5
NaiveBayes	2
SMO	1
IBK	4
DecisionTable	2
J48	4

Table 3: No of Attributes Selected using Classifier Subset Evaluator

The Precision, Recall, F-Measure and ROC values for different classifier using various feature selection methods are compared and tabulated. The chart representation for each table is also presented. The precision, Recall, F-Measure and ROC values for the classifiers for various feature selection methods is given in Table 4,5,6 and 7respectively.

Feature Selection method	Bayes Net	Naive Bayes	SMO	DecisionTable
Cfs Subset	0.73	0.744	0.683	0.618
Chis quared	0.704	0.69	0.71	0.627
InfoGain	0.704	0.69	0.71	0.627
GainRatio	0.704	0.69	0.71	0.627
Filtteredsubset	0.704	0.69	0.71	0.627
Consistency Subset	0.845	0.853	0.847	0.753

Table 4: Precision Values for various Feature Selection Methods

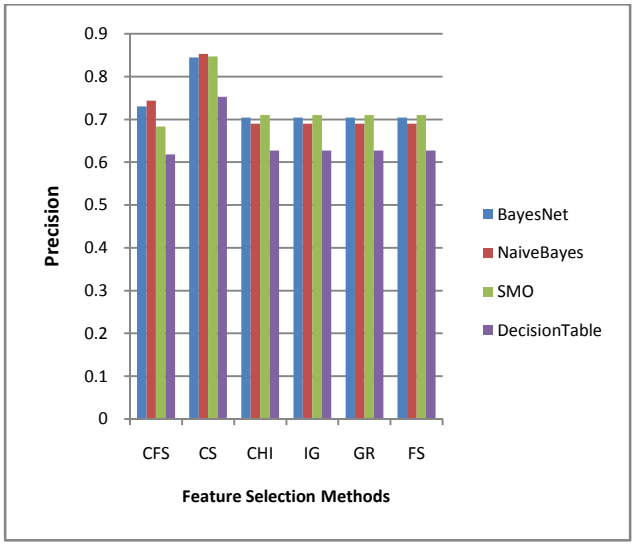

Figure 2: Precision Chart for different Feature Selection Methods

The precision values are high for Consistency subset feature selection method and the higher value is obtained for Naïve Bayes classification method.

Feature Selection method	Bayes Net	Naïve Bayes	SMO	Decision Table
CfsSubset	0.723	0.735	0.658	0.619
Chisquared	0.703	0.69	0.703	0.626
InfoGain	0.703	0.69	0.703	0.626
GainRatio	0.703	0.69	0.703	0.626
Filtteredsubset	0.703	0.69	0.703	0.626
Consistency subset	0.832	0.845	0.852	0.761

Table 5: Recall Values for Various Feature Selection Methods

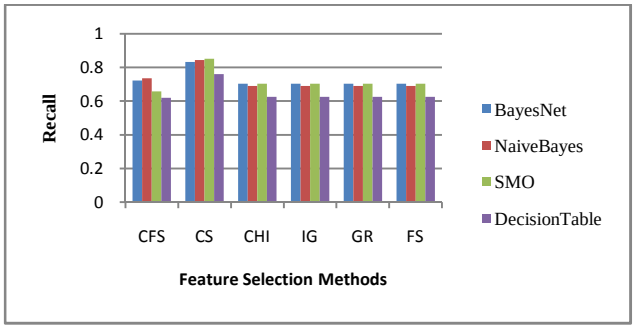

Figure 3: Recall Chart for different Feature Selection Methods

The Recall values are high for the Consistency subset evaluation method and the higher value is obtained for Support vector classifier.

Feature Selection Method	Bayes Net	Naive Bayes	SMO	DecisionTable
CFS	0.715	0.728	0.63	0.607
CS	0.837	0.848	0.849	0.757
CHI	0.699	0.686	0.694	0.61
IG	0.699	0.686	0.694	0.61
GR	0.699	0.686	0.694	0.61
FS	0.699	0.686	0.694	0.61

Table 6: F-Measure Values for Various Feature Selection Methods

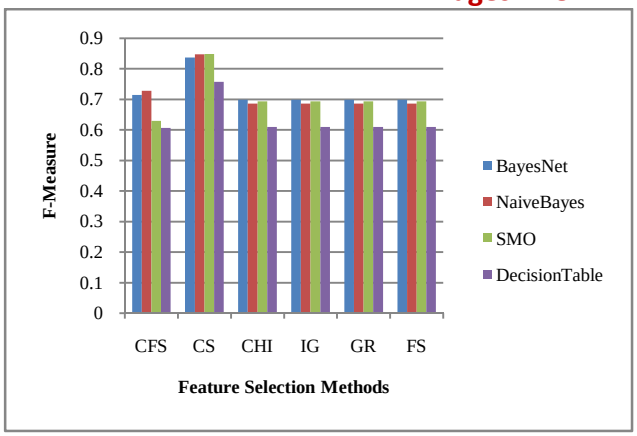

Figure 4: F-Measure Chart for different Feature Selection Methods

The F-Measure values are high for Consistency subset selection method and the value is higher for Naïve Bayes classifier .

Feature Selection Method	Bayes Net	NaiveBayes	SMO	DecisionTable
CFS	0.725	0.746	0.633	0.596
CHI	0.722	0.746	0.688	0.634
IG	0.722	0.746	0.688	0.634
GR	0.722	0.746	0.688	0.634
FS	0.722	0.746	0.688	0.634
CS	0.822	0.86	0.756	0.763

Table 7: ROC Values for Various Feature Selection Methods

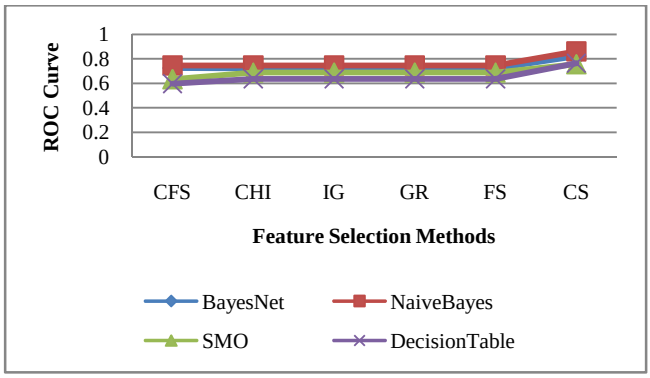

Figure 5: ROC Curve for different Feature Selection Methods

The ROC values are also high for Consistency subset method and the higher value is obtained for the Support vector classifier. The number of attributes selected by the Feature Selection Methods like Chi Squared, Info Gain, Gain Ratio and Filtered Subset is same. So the performance metric values for the corresponding feature selection methods using different classifiers is also same. The performance metrics is high for Consistency subset feature selection method which uses Best First Search method for selecting the relevant attributes.

CONCLUSION

In this paper the various Feature Selection methods are compared with each other on their performance by using different classifiers. Based on the evaluation the Consistency Subset evaluator has high metric values for the different classifiers used. It is observed that for specified Feature Selection method, the classification performance of the classifiers based on dataset, the corpuses is different. From the above discussion it is inferred that no single representation scheme and classifier can be mentioned as a general model for any application. Different Feature Selection methods perform differently for various classification algorithms depending on the data selection.

REFERENCES

[1] C.Kanakslakshmi, Dr.R.Manicka chezian, "An Analysis on Text Mining and Text Classification Techniques", Proceedings of the National Conference on Information and Image Processing, Volume 1, February 2015, pp: 132-135

[2] G.Angulakshmi, Dr.R.Manicka Chezian, " Three Level Feature Extraction For Sentiment Classification", International Journal of Innovative Research in Computer and Communication Engineering, Volume 2, Issue 8, August 2014, pp: 5501-5507.

[3] M. Ramaswami , R. Bhaskaran "A Study on Feature Selection Techniques in Educational Data Mining" Journal Of Computing, Volume 1, Issue 1, December 2009, pp: 7-11.

[4] Divya P, G.S. Nanda Kumar "Study on Feature Selection Methods for Text Mining" International Journal of Advanced Research Trends in Engineering and Technology , Volume 2, Issue 1, January 2015, pp: 11-19.

[5] Asha Gowda Karegowda, M.A.Jayaram , A.S. Manjunath, "Feature Subset Selection Problem using Wrapper Approach in Supervised Learning", International Journal of Computer Applications, Volume 1, Issue 7, 2010, pp. 0975–8887.

[6] Sunita Beniwal, Jitender Arora "Classification and Feature Selection Techniques in Data Mining" International Journal of Engineering Research & Technology, Volume 1, Issue 6, August – 2012, pp: 1-6

[7] Bozhao Li, Na Chen, Jing Wen, Xuebo Jin , Yan Shi "Text Categorization System for Stock Prediction" International Journal of u- and e-Service, Science and Technology, Volume 8,Issue 2, pp: 35-44, 2015.

[8] Tien Dung Do, Siu Cheung Hui and Alvis C.M. Fong "Associative Feature Selection for Text Mining" International Journal of Information Technology, Volume 12, Issue 4, 2006, pp: 59-68.

[9] Girish Chandrasekhar, Ferat Sahin "A survey on feature selection methods", Computers and Electrical Engineering, Volume 40, 2014, pp: 16–28.

[10] Bangsuk Jantawan, Cheng-Fa Tsai "A Comparison of Filter and Wrapper Approaches with Data Mining Techniques for Categorical Variables Selection" International Journal of Innovative Research in Computer and Communication Engineering, Volume 2, Issue 6, June 2014, pp: 4501-4508.

[11] George Forman "An Extensive Empirical Study of Feature Selection Metrics for Text Classification" Journal of Machine Learning Research, Volume 3, 2003, pp: 1289-1305.

[12] Sherin Mary Varghese, M.N.Sushmitha "Efficient Feature Subset Selection Techniques for High Dimensional Data" International Journal of Innovative Research in Computer and

Communication Engineering, Volume 2, Issue 3, March 2014, pp: 3509-3515.
[13]T.Jaga Priya Vathana, C.Saravanabhavan, Dr.J.Vellingiri "A Survey On Feature Selection Algorithm For High Dimensional Data Using Fuzzy Logic", International Journal Of Engineering And Science, Volume 2, Issue 10, 2013, Pages 27-38.
[14] S.Sagar Imambi, T.Sudha "A Novel Feature Selection Method for Classification of Medical Documents from Pubmed" International Journal of Computer Applications, Volume 26, Issue 9, July 2011, pp: 29-33.
[15] Aansi A. Kothari, Warish D. Patel "A Contemporary Overview on Feature Selection and Classification Techniques in Opinion Mining" , International Journal of Computer Applications , Volume 110 , Issue 10, January 2015 , pp: 10-14.
[16] S.Nirmala Devi, Dr. S.P Rajagopalan "A study on Feature Selection Techniques in Bio-Informatics" International Journal of Advanced Computer Science and Applications, Volume 2, Issue 1, January 2011, pp: 138-144.
[17] S. Ramasundaram and S.P. Victor, "Algorithms for Text Categorization: A Comparative Study", World Applied Sciences Journal, Volume 22, Issue 9, 2013, pp: 1232-1240.
[18] Trilok Chand Sharma, Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm", International Journal of Advanced Research in Computer and Communication Engineering, Volume 2, Issue 4, April 2013, pp: 1925-1931.

BIOGRAPHIES

Dr. R.Manickachezian received his M.Sc., degree in Applied Science from P.S.G College of Technology, Coimbatore, India in 1987. He completed his M.S. degree in Software Systems from Birla Institute of Technology and Science, Pilani, Rajasthan, India and Ph D degree in Computer Science from School of Computer Science and Engineering, Bharathiar University, Coimbatore, India. He served as a Faculty of Maths and Computer Applications at P.S.G College of Technology, Coimbatore from 1987 to 1989. Presently, he has been working as an Associate Professor of Computer Science in N G M College (Autonomous), Pollachi under Bharathiar University, Coimbatore, India since 1989. He has published more than One Hundred and Twenty papers in international/national journal and conferences. He is a recipient of many awards like Desha Mithra Award and Best Paper Award. His research focuses on Network Databases, Data Mining, Distributed Computing, Data Compression, Mobile Computing, Real Time Systems and Bio-Informatics.

C.Kanakalakshmi is a Research Scholar in Department of Computer Science, Nallamuthu Gounder Mahalingam College, and Pollachi. She received her Master of Computer Applications (M.C.A) in 2011 from Nallamuthu Gounder Mahalingam College, Pollachi under Bharathiar University, Coimbatore. She has presented papers in International/National conferences and attended Workshop, Seminars and published paper in international journal. Her research focuses on Data Mining.