A Review of Feature Selection Methods for Classification Problem

[1]Nidhi B. Gadhia, [2]Gopi B. Sanghani
[1]PG Student, [2]Head of Department
[1] Department of computer engineering,
[1]Darshan institute of Engineering and Technology, Rajkot,Gujarat, India.

***Abstract*— The Classification are carried out using various feature selection technique. The feature selection methods allows the classification to be carried out more accurately and efficiently. Feature selection is one of the leading trends in the research work going on. There are various feature selection methods which are used along with the classification methods. According to the application the most appropriate feature selection method is selection for selection the feature. The selected feature is then supplied to the classifier to carry out the classification of data. Here we study 6 different Feature selection method which are Document Frequency (DF), Mutual Information (MI), Information Gain (IG), CHI Square Statistics, and Bi – Normal Separation. These methods are used separately for the text classification or a combination of methods are used.**

***Index Terms* - Feature Selection, Classification, Mutual Information, Information gain, CHI Square Statistics, Bi – Normal Separation, Text Classification.**

I. INTRODUCTION

With the increase in online information the need to classify data has become a major concern. Feature selection is an important method which provides input to the classifier and which improves the classification effectiveness, computational efficiency. The process through which the Feature dimensionality can be reduced is known as Feature Selection [1]. For some algorithm high dimensionality is not permitted. If there is noisy data along with the original data then it will hurt the precision of the classifier. So to avoid such problem we need to select few features from the original data to reduce the dimensionality and improve the efficiency of the classifier. The feature should be selected on the base to the label available that are used with that data.

Feature selection selects the features based on the following criteria. The criteria can be:

1. The classification accuracy does not significantly decrease [3].
2. The resulting class distribution, given only the values for the selected features, is as close as possible to the original class distribution, given all features [3].

The Feature Selection included few steps that are implemented to obtain the desired feature that are used for the classification. The steps are as follows [3].

1. Subset Generation.
2. Feature Subset Evaluation.
3. Stopping Criterion.

A. Subset Generation

The subset selection is a method through which a candidate subset is selected that would further be carried out to the next step of evaluation. The starting position of the search is decided. It generally starts with an empty set and adds features to it gradually as the search precedes. The search is started keeping a strategy in mind and according to that strategy the search of the data goes on [5]. There are basically three type of search that are carried out they are complete, sequential, and random search.

B. Feature Subset Evaluation

The newly generated subset needs to be evaluated and that is done with the help of Evaluation criterion. The criterion can be applied on the basis of the dependences of the algorithm. They are Independent criterion and the dependent criterion [5].

C. Stopping Criterion

The stopping criterion shows whether the feature selection process should be stopped or not. The process stops in the following condition, the search completes, maximum feature are selected, the same subset is generated every time, and a good subset is selected [5].

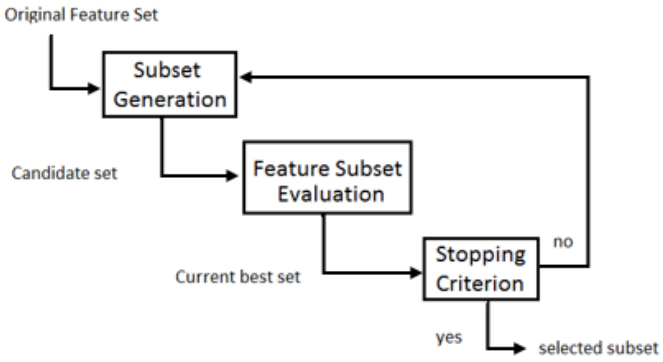

Figure 1: Feature Selection Process

II. LITERATURE REVIEW

Feature Selection is a fertile field for research and development. The unwanted data or the feature needs to be removed and the classifier needs to be made more efficient and provide an accurate result through the selected feature.

The feature selection has basically three methods Filter method, Wrapper method and Embedded Method [4]. The filter method selects the feature independently and it relies on the training data instead of the algorithm. Therefore it ignores the effect that is caused by the subset on the algorithm. The optimized subset is the one which depends on the specific biases and heuristics of the algorithm. But the Wrapper model requires a predictive algorithm and uses its performance to evaluate and determine which feature is selected. Based on the selected subset and the algorithm the utilizes a classifier and evaluates the quality of the selected feature [5].

In the early decades, there has been various research done on the Feature Selection for the classification problem. The various Feature selection methods have been considered and the accurate have been found in order to know whether which method is the most relevant for the given application of the classification. The most prominent methods that are studied are DF, IG, and CHI. A combination of methods was also considered and the most accurate result is considered [1].

In 1997, Yang and Pederson studied all the five methods in detail. And showed that, they improve can the classification accuracy. Their study was mainly focused on the Feature Selection Metrics and obtained the accurate among them [7].

Forman (2003) compared all the Feature Selection method on a classification problem and proposed a new method called the Bi-normal separation. And Forman also showed that the method worked very well in the evaluation metrics.

In 2006 Ng et al. found that the weighted log – likehood ratio had 87.1% accuracy when it was considered on the data that was available at the movies.

Shoushan Li, Rui Xia, Chengqing Zong, Chu- Ren Huang also proposed that the method Weighed Frequency and Odds which is an combination of methods provided an better result compared to the other methods. But the data that was taken into consideration was a static data and a balanced one [1].

III. FEATURE SELECTION METHOD

The Feature selection are done in order to obtain a specific set of feature which would be provided as input to the classifier which would result in the improvement of it accuracy and efficiency.

There are various feature selection methods that are present which are considered while bifurcating selected feature set. These parameter help to find out the most accurate of all the other set and the most accurate would be selected for the classification.

Before starting the study of methods we would keep the following data into consideration.

Ai: the number of the documents that contain the term t and also belong to category ci

Bi : the number of the documents that contain the term t but do not belong to category ci

Ni : the total number of the documents that belong to category ci

Nall : the total number of all documents from the training data.

Ci : the number of the documents that do not contain the term t but belong to category ci,

$$Ni - Ai$$

Di : the number of the documents that neither contain the term t nor belong to category ci

$$Nall - Ni - Bi$$

A. Document Frequency

The Document frequency can be defined as the number of documents in which a term occurs.

$$DF = \sum_{i=1}^{m} (A_i)$$

We find out the unique terms from the document based on their frequency of occurrence. The feature that have low DF are discarded. The removal is done keeping in mind that the data is not relevant and are non-informative. The documents that have high DF are more informative for classification. This method performs very well for some topic based classification [1][7].

B. *Mutual Information*

Mutual Information are generally used in statistical language modeling of word associations and other application that are related to it. Using term t and class ci Mutual Information can defined as follows [1].

$$MI = \log \frac{A_i \times N_{all}}{(A_i + C_i)(A_i + B_i)}$$

The term that has higher ratio are more accurate for classification. The MI has the time complexity of O(Vm) [7]. There are basically two types of MI used they are with maximum value and one with average value. The main drawback of MI is that the value of MI is influenced by marginal Probability. So the score would be more for rare term compared to that of common term. Document Frequency and Mutual Information are the base of all the methods that are being used for feature selection.

C. *Information Gain*

Information Gain is used to measures the number of bits of information that are obtained for category prediction by recognizing the presence or absence of a term in document [7].

$$IG = \{-\sum_{i=1}^{m} \frac{N_i}{N_{all}} \log \frac{N_i}{N_{all}}\}$$
$$+(\sum_{i=1}^{m} A_i / N_{all})[\sum_{i=1}^{m} \frac{A_i}{A_i + B_i} \log \frac{A_i}{A_i + B_i}]$$
$$+(\sum_{i=1}^{m} C_i / N_{all})[\sum_{i=1}^{m} \frac{C_i}{C_i + D_i} \log \frac{C_i}{C_i + D_i}]$$

In the field of Machine Learning, Information gain is used as a term goodness criterion. The number of bits of information is obtained according to the presence or absence of. The IG considers a training set and finds out the IG of each unique term. And the one with less value are removed. The IG measures the decrease in entropy when the feature is not given. The Information Gain also has time complexity of O(Vm). The Information gain are compared better than the other two methods and also provides accurate result in various application.

D. *CHI Square Statistic*

The CHI Square Statistic can be given as follows

$$CHI = \frac{N_{all} \cdot (A_i D_i - C_i B_i)^2}{(A_i + C_i) \cdot (B_i + D_i) \cdot (A_i + B_i) \cdot (C_i + D_i)}$$

In this method the training set is taken into account to find out the value of CHI. It considers frequency measurement and ratio measurement to obtain the result. The value with high Document frequency has high value of CHI. The CHI measures the independences between the values. The value evaluated is dependent because the value of CHI is dependent on the Document Frequency. It has a natural value 0 is both the data are independent. The CHI square has normalized values so the values are comparable across term in different category. If the Normalization breaks down this method would not be relevant any more for low frequency terms.

E. *Bi – Normal Separation*

The Bi – Normal Separation was proposed by Forman (2003), and he defined it as

$$BNS = \left| F^{-1}(\frac{A_i}{N_i}) - F^{-1}(\frac{B_i}{N_{all} - N_i}) \right|$$

The Bi – Normal separation takes the difference between the values into consideration for the normal distributions. The score of the method increases as difference between the value increases. And also if the there is more data in the document then also the score of the method increases. Unlike CHI square the Bi – Normal Separation is not dependent to the Document Frequency and is biased towards the term with high category ratio.

IV. CONCLUSION

In this paper, from the study of various Feature Selection methods we could conclude which method is best suitable for a given application. Each method have their own particular advantage as well as disadvantages. According to the advantages the best feature selection method is taken to carry classification of data. To obtain the most accurate data for classification, combination of methods were also selected. The combination of methods proved to be more efficient.

V. FUTURE SCOPE

The Feature Selection methods that we reviewed contained only static data. For our future work we would consider dynamic and unbalanced data set. On the basis of the data set the most appropriate method would be taken. To find the accurate result combination of FS method can also be taken.

REFERENCES

[1] Shoushan Li, Rui Xia, Chengqing Zong, and Chu-Ren Huang. "A Frame Work of Feature Selection". Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 692–700, Suntec, Singapore, 2-7 August 2009.2009 ACL and AFNLP.

[2] Wenqian Shang, Houkuan Huang a, Haibin Zhu, Yongmin Lin, Youli Qu, Zhihai Wang. "A novel feature selection algorithm for text categorization". Expert Systems with Applications 33 (2007) 1–5.

[3] M. Dash, H. Liu. "Feature Selection for Classification". Department of Information Systems & Computer Science, National University of Singapore, Singapore 119260. Intelligent Data Analysis 1 (1997) 131–156.

[4] Jiliang Tang, Salem Alelyani and Huan Liu. "Feature Selection for Classification: A Review".

[5] Sunita Beniwal, Jitender Arora. "Classification and Feature Selection Techniques in Data mining". International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 6, August – 2012 ISSN: 2278-0181.

[6] Nicolette Nicolosi. "Feature Selection Methods for Text Classication". November 7, 2008.

[7] Yiming Yang, Jan O. Pedersen. "A Comparitive Study on Feature Selection in Text Categorization". School of Computer Science, Carnegie Mellon University.

[8] Huan Liu and Lei Yu. "Toward Integrating Feature Selection Algorithms for Classification and Clustering". Department of Computer Science and Engineering Arizona State University, Tempe, AZ 85287-8809.