Survey on Feature Selection for Text Classification

Priyanka Mangalkar[1], **Sunil D Kale**[2]
[1, 2] Department of Computer Engineering
[1, 2] Rajashri Shahu College of Engineering,Savitribai Phule Pune University,Pune, India.

Abstract-*Machine learning for text classification is the foundation of document classification, news filtering, document routing, and personalization. Text classification is continuing to be a standout amongst the most researched NLP issues due to the ever-increasing amounts of electronic documents and digital libraries. In text domains, effective feature selection is essential to make the learning task effective and more precise. Automated feature selection is important for text classification to decrease the feature size and to speed the learning procedure of classifiers. Recent approaches to text classification have utilized two diverse first-order probabilistic models for classification, both of which make the naive Bayes assumption. This paper survey an approach to feature subset selection that considers issue specifics and learning algorithm attributes. Also focus on domains with many features that also have a highly unbalanced class distribution and asymmetric misclassification costs given only implicitly in the issue.*

Keywords-Feature selection, text categorization, Kullback-Leibler divergence, Jeffreys divergence, Jeffreys-Multi-Hypothesis divergence.

I. INTRODUCTION

The amount of huge data which is present as well as is publically available on the internet has greatly increased in the last few years. That's how, machine learning methods have difficulty in dealing with the large number of input features, which is posing an interesting challenge for programmer. For using machine learning techniques effectively, preprocessing of the information is must. Feature selection is one of the most frequent as well as valuable technique in data preprocessing, and has become an indispensable component of the machine learning process. This is called as variable selection, attribute selection, or variable subset selection in machine learning as well as statistics. It is the process of searching relevant features as well as deleting irrelevant, redundant, or noisy data. This process speeds up data mining algorithms, improves predictive accuracy, and increases comprehensibility. Irrelevant features are those that provide no useful information, and redundant features provide no more information than the currently selected features. In terms of supervised inductive learning, feature selection gives a set of candidate features using one of the three approaches:

- The specified size of the subset of features that optimizes an evaluation measure.
- The smaller size of the subset that satisfies a certain restriction on evaluation measures.
- In general, the subset with the best commitment among size and evaluation measure.

Therefore, the correct use of feature selection algorithms for choosing features improves inductive learning, either in term of generalization capacity, learning speed, or reducing the complexity of the induced model.

In the of feature selection, irrelevant and redundant features or noise in the data may be hinder in many situations, because they are not relevant and important with respect to the class idea like microarray data analysis. When the number of samples is much less than the features, then machine learning gets particularly difficult, because the search space will be sparsely populated. Therefore, the model will not able to differentiate accurately between noise and relevant data. There are two major approaches to feature selection. The first is Individual Evaluation, and the second is Subset Evaluation. Ranking of the features is known as Individual Evaluation. In Individual Evaluation, the weight of an individual feature is assigned according to its degree of relevance. In Subset Evaluation, candidate feature subsets are constructed using search strategy.

This paper presents a survey on feature selection techniques.

II. LITERATURE REVIEW

In paper [1] authors has introduced feature selection method based on the information measures for naive Bayes classifiers, aiming to select the features that offer the maximum discriminative capacity for text classification. We have also derived the asymptotic distributions of these measures, which leads to the other version of the Chi-square statistic approach for feature selection. The propose model involve the learning model in the feature filtering process, which provides us a theoretical way to analyze the optimality of the selected features.

In paper [2] authors developed a new feature selection technique based on Information Gain and Particle Swarm Optimization. Also stated the issues in text document categorization, which is, the numbers of extracted features are a lot of. In this study, by using a new feature selection method based on IG (information gain) and PSO (particle swarm optimization) algorithms, text categorization process performed. Reuters 2l.S78 and Classic3 corpus were used in the experiments. The roots of the words in the texts of corpus were taken as the features. Feature selection and categorization processes performed with k-Nearest Neighbors algorithm (K-NN) and Naive Bayes classifiers by using IG and PSO algorithms.

In paper [3] authors developed a Gaussian process dependant system for text categorization. The system has two major parts, that is LDA based feature learning and Gaussian process based classification. The first outcome shows that the GP classifier has the better performance than SVM and SRC. Both SVM as well as SRC are non-probabilistic classifiers. The main reason behind using GP classifier is that this model can provide the uncertainty of the predictions. In second test outcome, authors shows that the LDA-based feature can represent the document better in the small size of training dataset. The conventional TF-IDF feature considers only the frequency of words. Hence, IDF fails to represent the documents in a small dataset. Unlike the TF-IDF feature, LDA-based feature extracts the semantic meaning of the text documents.

In paper [4] authors have designed a novel text categorization system combining distribution clustering of words for document representation and linear LSTSVM for document classification. Unlike the conventional feature selection measures used in text categorization, distributional clustering of words uses all word features and generates a compact representation in low dimension space. LSTSVM benefits from this compact representation as training complexity depends on input dimension, yielding fast training with competitive results.

In paper [5] authors proposes a novel idea of segmenting the documents for calculating term weights. Tests with DynaPart-FiLa suggest that segmenting the documents before computing term weights helps in improving weighted average F-measure (based on macro-weighted averaging). For all the datasets, F-measure has improved for all classifiers with DynaPart-FiLa. From the improved F-measure, authors say that those relevant terms appear from the beginning of the document. Increasing their importance helps in improving the classification outcomes. At last they concluded that positional significance of terms are good indicators of context of the document.

In paper [6] authors developed an automatic text categorization system, HPHR, which is built on three text categorization models, PTC, a high-precision model, RTC, a high-recall model and, NFTC, a noise-filtered model. Authors claims a development of an efficient and effective clustering algorithm, Canopy-KM and a machine learning algorithm for automatic generation of three text categorization models, PTC, RTC and NFTC. The HPHR system was evaluated using two different collections of text documents, verbatim vehicle fault description, and Reuters text document corpus.

In paper [7] authors have implemented new context similarity based feature selection methods were introduced for text categorization research. They assign importance scores to features based on their similarity measure of context strings within certain text categories. Using two data sets from different application domain, the effectiveness of the proposed methods were investigated and compared against well known frequency based techniques. The proposed methods can achieve better performances on both binary and multi-classification problems.

In paper [8] authors have addresses a credit risk assessment issue from the perspective of machine learning. Authors formulate the decision making issue in credit risk assessment as a binary classification issue, which can be dealt with by learning a decision tree on a training data set. The developed technique uses an association rule-based feature selection technique to minimize the dimensionality of the data before training classifiers. The process of feature selection helps identify the key subset of the features that can provide simple but accurate representations of the original data. The association rule based feature selection includes three steps: mining association rule by Apriori algorithm, ranking the association rules with a certain strategy, and construct the subset of features by selecting them from the association rules.

III. PROPOSE SYSTEM

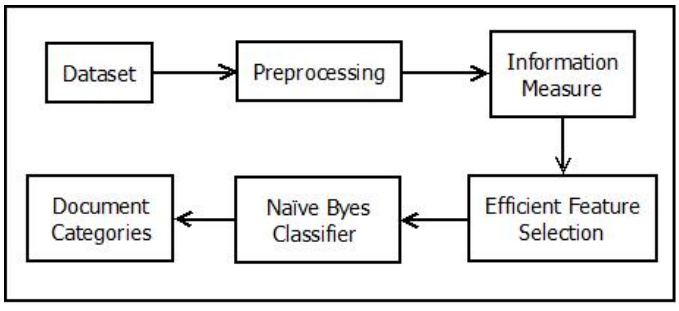

Fig. 1: System Flow

In proposed system we are working on a data set (20-NEWSGROUPS dataset, REUTER dataset) which contains the data related to different news category this data set is taken as input to the system and as a output we will get documents classification in several categories (topic) as an output.

This proposed system has several phases such as Preprocessing, Information measure, Efficient Feature Selection, Naïve Byes Classifier, Document Categories. First, in Preprocessing the Stemming , Stopword are removed from the input file, next on the out of the phase the Information measure is performed using this information measure for efficient feature selection. The this features are given to naïve byes classifier for the purpose of classification and at the end the we well get the different document for each category.

IV. CONCLUSION

This paper survey on different existing feature selection methods for text classification and discuss their pros and cons. However, the existing methodology assesses the goodness of a feature by only exploiting the inherent attributes of the training data without considering the learning algorithm for discrimination, which may prompt an undesired classification performance. From this survey we conclude that, there is need to presented new feature selection methodologies based on the information measures for naive Bayes classifiers, intending to choose the features that offer the maximum discriminative capacity for text classification.

REFERENCES

[1] B. Tang, S. Kay and H. He, "Toward Optimal Feature Selection in Naive Bayes for Text Categorization," in IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 9, pp. 2508-2521, Sept. 1 2016.

[2] F. Yiğit and Ö K. Baykan, "A new feature selection method for text categorization based on information gain and particle swarm optimization," 2014 IEEE 3rd International Conference on Cloud Computing and Intelligence Systems, Shenzhen, 2014, pp. 523-529.

[3] S. H. Chen, Y. S. Lee, T. C. Tai and J. C. Wang, "Gaussian process based text categorization for healthy information," 2015 International Conference on Orange Technologies (ICOT), Hong Kong, Hong Kong, 2015, pp. 30-33.

[4] M. A. Kumar and M. Gopal, "Least squares twin support vector machines for text categorization," 2015 39th National Systems Conference (NSC), Noida, 2015, pp. 1-5.

[5] A. Kulkarni, V. Tokekar and P. Kulkarni, "Term weighting using contextual information for categorization of unstructured text documents," 2015 Annual IEEE India Conference (INDICON), New Delhi, 2015, pp. 1-4.

[6] D. Li and Y. L. Murphey, "Automatic text categorization using a system of high-precision and high-recall models," Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on, Orlando, FL, 2014, pp. 373-380.

[7] Y. Chen, B. Han and P. Hou, "New feature selection methods based on context similarity for text categorization," Fuzzy Systems and Knowledge Discovery (FSKD), 2014 11th International Conference on, Xiamen, 2014, pp. 598-604.

[8] X. Mei and Y. Jiang, "Association rule-based feature selection for credit risk assessment," 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS), Chongqing, China, 2016, pp. 301-305.