International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 4 | ISSN : 2456-3307

Survey on Feature Selection for Text Categorization

Sonali Suskar, Dr. S. D. Babar
Department of Computer Engineering, SIT College of Engineering, Lonavala, Maharashtra, India

ABSTRACT

In this massive amount of data, data is too vast so that text categorization is important issue. With the help of previously organize set of documents and classes we can automatically classify data. The filter approach is predominantly used in text categorization because of its simplicity and efficiency. However, the filter approach evaluates the goodness of a feature by only exploiting the intrinsic characteristics of the training data without considering the learning algorithm for discrimination, which may lead to an undesired classification performance. Given a specific learning algorithm, it is hard to determine which filter feature selection approach is the best for discrimination. This survey mainly focuses on the techniques used for feature selection method used for text categorization. This survey also presents the comparative analysis of such recent techniques along with their limitations.

Keywords : Classification, text categorization, feature selection, training data.

I. INTRODUCTION

In text classification, one commonly utilizes a 'sack of words' model: every position in the info include vector compares to a given word or expression. For instance, the event of the word "free" might be a helpful feature in separating spam email. The quantity of potential words frequently surpasses the quantity of preparing archives by more than a request of size. Include choice is important to make substantial issues computationally effective conserving computation, stockpiling and system assets for the preparation stage and for each future utilization of the classifier. Promote, well-picked features can enhance grouping precision significantly, or comparably, diminish the measure of preparing information expected to acquire a craved level of execution.

Feature selection is generally received to lessen dimensionality of information. As we said some time recently, the channel and the wrapper are the two sorts of highlight choice approach. The high computational cost makes the wrapper approach unfeasible, and we focuses on the channel approach in this work. Numerous channel approaches have been proposed in TC, including document frequency (DF), mutual information (MI),information gain (IG), Chi-square statistic, relevance score (RS), GSS coefficient, among others. The normal thought of these component choice criteria is to quantify reliance or importance between the twofold feature and the class as score of highlight significance.

The general feature selection method is to score every potential component as per a specific highlight determination metric, and after that take the best k highlights. Scoring includes tallying the events of a component in

preparing positive-and negative-class preparing illustrations independently, and at that point figuring a component of these.

There are number of ways to represent the document to calculate the form of term and the calculation of weight of term. But most widely used document representation for text categorization is by using one of the basic model called as "bag-of-words", where occurrence of each word is considered as a feature for training a classifier. This method of document representation is called as a Vector Space Model [3], where each feature in a feature space corresponds to term or a phrase in a vocabulary collected from a particular dataset. The value of each feature represents the importance of the term in the document, according to a specific feature measurement.

A major challenge in text categorization is the learning from high dimensional data. Firstly, a document may consist of large number of words that is hundreds, thousands of words. Applying original document directly which contains this large number of words, may result into a high computational load for the learning process. And second problem is that, many of the words from original document may be irrelevant to the topic or redundant, so use of those words will reduce the performance of classifier. To avoid the issue of the learning from high dimensional data and to make the learning process faster, it is important to reduce original feature space to few important words/feature preserving semantic meaning of original document. The process of reducing the original size of feature space is called as dimensionality reduction. A most common dimensionality reduction approach used in text categorization is feature selection. Feature selection aims at selecting only a subset of relevant features from the set of original large set of features. Reducing irrelevant and redundant features, feature selection approach improves the overall performance of classification.

II. LITERATURE REVIEW

In this paper [1], present a novel and efficient feature selection framework based on the Information Theory, which aims to rank the features with their discriminative capacity for classification. We first revisit two information measures: Kullback-Leibler divergence and Jeffreys divergence for binary hypothesis testing, and analyze their asymptotic properties relating to type I and type II errors of a Bayesian classifier. We then introduce a new divergence measure, called Jeffreys-Multi-Hypothesis (JMH) divergence, to measure multi-distribution divergence for multi-class classification.

In paper [2] authors have developed a system which is automatically categorize text by making use of class specific features which is Bayesian classification. The proposed method allows selecting the vital features for every class. Authors have designed a navie Bayes rule by using Baggenstoss's PDF Projection Theorem for applying the class specific features for classification. The major advantage of derived technique is it can make use of present feature selection conditions.

In [3] present a novel exponentially embedded families (EEF) based classification method, in which the probability density function (PDF) on raw data is estimated from the PDF on features. With the PDF construction, we show that class-

specific features can be used in the proposed classification method, instead of a common feature subset for all classes as used in conventional approaches. We apply the proposed EEF classifier for text categorization as a case study and derive an optimal Bayesian classification rule with class-specific feature selection based on the Information Gain (IG) score.

In [4] extend the exponentially embedded family (EEF), a new approach to model order estimation and probability density function construction originally proposed by Kay in 2005, to multivariate pattern recognition. Specifically, a parametric classifier rule based on the EEF is developed, in which we construct a distribution for each class based on a reference distribution. The proposed method can address different types of classification problems in either a data-driven manner or a model-driven manner. In this paper, we demonstrate its effectiveness with examples of synthetic data classification and real-life data classification in a data-driven manner and the example of power quality disturbance classification in a model-driven manner.

In [5] proposed a new feature selection algorithm, named AD, which comprehensively measures the degree of relevance and distinction of terms occur in document set. We evaluated AD on three benchmark document collections, 20-Newsgroups, Reuters-21578 and WebKB, using two classification algorithms, Naive Bayes and Support Vector Machines. The experimental results, comparing AD with six classic feature-selection algorithms, show that the proposed method AD is significantly superior to Information Gain, Mutual Information, Odds Ratio, DIA association factor, Orthogonal Centroid Feature Selection and Ambiguity Measure when Naive Bayes classifier is used and significantly outperforms IG,MI,OR,DIA,OCFS and AM when Support Vector Machines is used.
In [6] propose here a supervised variant of the tf.idf scheme, based on computing the usual idf factor without considering documents of the category to be recognized, so that importance of terms frequently appearing only within it is not underestimated. A further proposed variant is additionally based on relevance frequency, considering occurrences of words within the category itself. In extensive experiments on two recurring text collections with several unsupervised and supervised weighting schemes, we show that the ones we propose generally perform better than or comparably to other ones in terms of accuracy, using two different learning methods.

In [7] proposes a novel feature selection method that first selects features in documents with discriminative power and then computes the semantic similarity between features and documents. The proposed feature selection method is tested using a support vector machine (SVM) classifier upon two published datasets, viz. Reuters-21578 and 20-Newsgroups. The experimental results show that the proposed feature selection method generally outperforms the traditional feature selection methods for text categorization for both published datasets.

In [8] propose a modified CHI feature selection approach which is called term frequency and distribution based CHI to overcome these weaknesses. We use sample variance to calculate the term distribution, and improve the classic CHI with maximum term frequency. Extensive and comparative experiments on three corpora

show that the proposed approach is comparable to the classic feature selection methods in terms of macro-F1 and micro-F1.

In [9] paper first presents the study of four well known frequency based feature selection methods, including Gini Index (GI), Document Frequency (DF), Class Discriminating Measure (CDM) and Accuracy Balanced (Acc2). Then we focus on calculating the importance of features through measuring the similarity of their contexts among the documents but the document frequency containing these features to incorporate context information. Hence we propose four new context similarity based feature selection methods, GI_{cs}, DF_{cs}, CDM_{cs} and $Acc2_{cs}$. They are evaluated on different data sets and compared against the four corresponding frequency based methods. Through experimental analysis, the results reveal that the context similarity based methods outperform the corresponding frequency based methods in terms of the micro and macro F1 measures both on binary and multi-classification problems. Benefit from the multi-words information surrounding features, the context similarity based feature selection methods are effective for article categorization.

In this study [10], by using a new feature selection method based on IG (information gain) and PSO (particle swarm optimization) algorithms, text categorization process performed. Reuters 21.578 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. Proposed system performance was evaluated by using CA (Classification Accuracy), Precision, Recall and F-measure criteria.

Table 1. Survey Table

Sr. No.	Title	Technique Used	Advantages	Disadvantage
1.	Toward Optimal Feature Selection in Naive Bayes for Text Categorization	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	model involve the learning model in the feature filtering process	Doses not analyze feature dependence and develop feature selection algorithms by weighting each individual features
2.	A Bayesian Classification Approach Using Class-Specific Features for Text Categorization	presented a Bayesian classification approach for automatic text categorization using class-specific features	Many existing feature selection criteria can be easily incorporated	----------
3.	EEF: Exponentially Embedded Families with	Present a novel exponentially embedded	general framework for	the constructed distribution has

	Class-Specific Features for Classification	families (EEF) based classification method	building classifiers that deal with short and sparse text & Web segments	no closed form for a complex reference distribution
4.	A parametric classification rule based on the exponentially embedded family	This system extend the exponentially embedded family (EEF)	Improve performance	----------
5.	A term weighting scheme based on the measure of relevance and distinction for text categorization	Proposed a new feature selection algorithm, named AD, which comprehensively measures the degree of relevance and distinction of terms occur in document set.	System significantly outperforms IG,MI,OR,DIA,O CFS and AM when Support Vector Machines is used.	----------

III. PROBLEM STATEMENT

For given 20-NEWSGROUPS dataset, REUTER dataset, developed the system that will perform documents classification in several categories (topic) as an output using J48 classifier.

IV. PROPOSED SYSTEM

In proposed system we are working on a data set named as 20-NEWSGROUPS 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, J48 Classifier, Document Categories. First, in Preprocessing the Stemming, stop word 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. These features are given to J48 classifier for the purpose of classification and at the end the we well get the different document for each category.

V. CONCLUSION

This survey discusses the different technique for feature selection for text categorization. We have presented topic modeling for analyzing huge social media text data. We have also presented some recent approaches, their key concept and respected advantages. This survey also provide the limitations of some recent topic modeling techniques, which will helpful for further research. We conclude that the size of text data is the challenging problem of feature selection.

VI. REFERENCES

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

[2]. B. Tang, H. He, P. M Baggenstoss, and S. Kay, "A Bayesian Classification approach using class-specific features for text categorization," IEEE Trans. Knowl. Data Eng., vol. 28, no. 6, pp. 1602-1606, Jun. 2016.

[3]. B. Tang, S. Kay, H. He, and P. M. Baggenstoss, "EEF: Exponentially embedded families with class-specific features for classification," IEEE Signal Process. Lett., in press, 2016.

[4]. B. Tang, H. He, Q. Ding, and S. Kay, "A parametric classification rule based on the exponentially embedded family," IEEE Transactions on Neural Networks and Learning Systems, vol. 26, no. 2, pp. 367-377, 2015.

[5]. Yang, Jieming, et al. "A term weighting scheme based on the measure of relevance and distinction for text categorization." Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015 16th IEEE/ACIS International Conference on. IEEE, 2015.

[6]. Domeniconi, Giacomo, et al. "A study on term weighting for text categorization: a novel supervised variant of TF. IDF." Proceedings of the 4th international conference on data management technologies and applications (DATA). Candidate to the best conference paper award. 2015

[7]. Zong, Wei, et al. "A discriminative and semantic feature selection method for text categorization." International Journal of Production Economics 165 (2015): 215-222.

[8]. Jin, Chuanxin, et al. "Chi-square statistics feature selection based on term frequency and distribution for text categorization." IETE Journal of Research 61.4 (2015): 351-362.

[9]. Chen, Yifei, Bingqing Han, and Ping Hou. "New feature selection methods based on context similarity for text categorization." Fuzzy Systems and Knowledge Discovery (FSKD), 2014 11th International Conference on. IEEE, 2014.

[10]. F. Yi and O 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.