Volume 1, Issue 8, November 2016 ISSN: 2456-0006
International Journal of Science Technology Management and Research
Available online at: www.ijstmr.com

Weighting Individual Optimal Feature Selection in Naïve Bayes for Text Classification

Bhusare Madhuri B.
Department of Computer Engg.
Gokhale Education Society 's R.H. Sapat College of Engg.
Nasik, MH, India

Mr.Barde Chandrakant R.
Department of Computer Engg.
Gokhale Education Society 's R.H. Sapat College of Engg.
Nasik, MH, India

Abstract: ***Over 100 variants of five major feature selection criteria were examined using four well-called classification algorithms. Automated feature selection is important for text categorization to reduce feature size and to speed up learning process of classifiers. In this paper, we 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. This paper reports a controlled study on a large number of filter feature selection methods for text classification. We first revisit two information measures: Kullback-Leiblerdivergence and Jeffreys divergence for binary hypothesis testing, and analyze their asymptotic properties relating to type I and type IIerrors 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. Based on the JMH-divergence, we develop two efficient feature selection methods, termed maximum discrimination (MD) and MD_ x2 methods, for text categorization. We will also apply algorithm by weighting each individual features for better accuracy of the system***.

Keywords: Indian scripts, OCR, Compound character.

I. INTRODUCTION

We present a feature selection method which ranks the original features, aiming to maximize the discriminative performance for text categorization, when naive Bayes classifiers are used as learning algorithms. Unlike the existing filter approaches, our method evaluates the goodness of a feature without training a classifier explicitly, and selects these features that offer maximum discrimination in terms of a new divergence measure. Specifically, the contributions of this paper are as follows:

1. We bring a new divergence measure for multi class classification by extending the J-divergencemeasure, termed Jeffreys-Multi-Hypothesis divergence (JMH-divergence).
2. We propose an efficient approach to rank the order of features to approximately produce the maximum JMH divergence. The on paper analysis shows that the JMH divergence is monotonically increasing when more features are selected.
3. We analyze the asymptotic distribution of the proposed test statistic, which leads to the x2 distribution. By doing so, we submit another simple and effective feature ranking approach by increasing the non centrality measurement of the noncentral x2 distribution.With the increasing availability of text documents in electronic form, it is of great importance to label the contents with a predefined set of thematic categories in an automatic way, what is also called as automated Text Categorization. In last decades, a growing number of advanced machine learning algorithms have been developed to address this challenging task by formulating it as a classification problem. Commonly, an automatic text classifier is built with a learning process from a set of relabelled documents. Documents need to be represented in a way that is appropriate for a general learning process. The most extensively used representation is "the bag of words": a document is represented by a vector of features, each of which corresponds to a term or a phrase in a vocabulary collected from a particular dataset.

II. LITERATURE SURVEY

Wai Lam @,Automatic Text Categorization and Its Application to Text Retrieving, They develop an automatic text categorization approach and investigate its application to text retrieval. The categorization approach is derived from a concatenation *01* a learning paradigm called as instance-based learning and an advanced document retrieval technique

called as retrieval feedback. We clearly show the effectiveness of our categorization approach using two real world document collections from the MEDLINE database [11].

HishamAl-Mubaid @ A New Text Categorization Technique Using Distributional Clustering and Learning Logic ,Text categorization is continuing to be one of the most researched NLP problems due to the ever-increasing amounts of electronic documents and digital libraries. In this paper, we present a new text categorization method that combines the distributional clustering of words and a learning logic technique, called Lsquare, for constructing text classifiers. The high dimensionality of text in a document has not been fruitful for the task of categorization, for which reason, feature clustering has been proven to be an ideal alternative to feature selection for reducing the dimensionality.[12]

Huan Liu @,Toward Integrating Feature Selection Algorithms for Classification and Clustering. This paper introduces concepts and algorithms of feature selection, surveys existing feature selection algorithms for classification and clustering, groups and compares different algorithms with categorizing framework based on search strategies, evaluation criteria, and data mining tasks, reveals unattempted combinations, and provides guidelines in selecting feature selection algorithms. With the categorizing framework, we continue our efforts toward building an integrated system for intelligent feature selection.[13]

Steven Kay @,Probability Density Function appreciation Using the EEF With Application to Subset/Feature Selection, The problem of multivariate probability density function (PDF) estimation is a critical one in almost all fields of scientific endeavour. We consider in particular PDF estimation for multiple hypotheses testing. In the engineering community this problem is called as classification, and in the statistical community it is called as discrimination. It requires the determination of the joint PDF of a set of important features or statistics.[14]

E. F. Combarro @,Introducing a family of linear measures for feature selection in text categorization

Text Categorization, which made up of automatically assigning documents to a set of categories, usually involves the management of a huge number of features. Most of them are irrelevant and others introduce noise which could cause the classifiers. Thus, feature reduction is often performed in order to increase the efficiency and effectiveness of the classification.[15]

S.-B. Kim @,Some effective techniques for naive Bayes text classification, Naive Bayes has been one of the popular machine learning methods for many years. Its simplicity makes the framework attractive in various assignment and reasonable performances are obtained in the tasks although this learning is based on an unrealistic independence assumption. For this reason, there also have been many interesting works of investigating naive Bayes. Especially, shows that naive Bayes can perform surprisingly well in the classification tasks where the probability itself calculated by the naive Bayes is not important. [16]

III. RELATED WORK

Learning of Classifiers Methods

- *Find Similar*

Our Find Similar method is a variant of Rocchio's method for relevance feedback (Rocchio,. 1971) which is a popular method for expanding user queries on the basis of relevance judgments. In Rocchio's formulation, the weight assigned to a term is a combination of its weight in an original query,and judged relevant and irrelevant documents.The parameters a, b, and g control the relative importance of the original query vector, the positive examples and the negative examples. In the context of text classification, there is no initial query, so a=0. We also set g=0 so we could easily use available code. Thus, for our Find Similar method the weight of each term is simply the average (or centroid) of its weights in positive instances of the category. There is no explicit error minimization involved in computing the Find Similar weights. Thus, there is no learning time so to speak, except for taking the sum of weights from positive examples of each category. Test instances are classified by comparing them to the category centroids using the Jacquard similarity measure. If the score exceeds a threshold, the item is classified as belonging to the category.

A. Decision Trees

A decision tree was constructed for each category using theapproach described by Chickering et al. (1997). The decision trees were grown by recursive greedy splitting, andsplits were chosen using the Bayesian posterior probabilityof model structure. We used a *structure prior* that penalized each additional parameter with probability 0.1, and derived parameter priors from a prior network as explained in Chickering et al. (1997) with an equivalent sample size of 10. A class probability rather than a binary decision is retained at each node.

B. Naive Bayes

A naïve-Bayes classifier is constructed by using the trainingdata to estimate the probability of each category given the document feature values of a new instance. We use Bayes theorem to estimate the probabilities:

$$P(C = c_k \mid \vec{x}) = \frac{P(\vec{x} \mid C = c_k)P(C = c_k)}{P(\vec{x})}$$

The quantity $P(x \mid C=c_k)$ is often impractical to compute without simplifying assumptions. For the Naïve Bayes classifier (Good, 1965), we assume that the features $X_1, \ldots X_n$areconditionally independent , given the category variable *C*. This simplifies the computations yielding:

$$P(\vec{x} \mid C = c_k) = \prod P(x_i \mid C = c_k)$$

Despite the fact the assumption of conditionalindependence is generally not true for word appearance in documents, the Naïve Bayes classifier is surprisingly effective.

C. Bayes Nets

Currently, there has been interest in learning more expressive Bayesian networks (Heckerman et al., 1995) as well as methods for learning networks specifically for classification (Sahami, 1996). Sahami, for example, allows for a restricted form of dependence between feature variables,thus relaxing the very restrictive assumptions of the NaïveBayes classifier. We used a 2-dependence Bayesianclassifier that allows the probability of each feature *xi* to be directly influenced by the appearance/non-appearance of atmost two other features.

D. Support Vector Machines (SVMs)

Vapnik proposed Support Vector Machines (SVMs) in1979 (Vapnik, 1995), but they have only recently beengaining popularity in the learning community. In its simplest linear form, an SVM is a hyperplane that separates a set of positive examples from a set of negative.

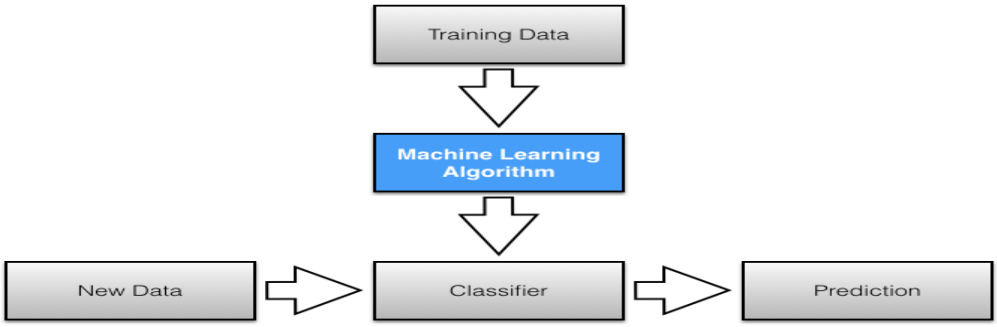

Fig1: Support Vector Machine

IV. RESULT OF EXISTING METHODOLOGY

Documents need to be represented in a way that is appropriate for a general learning process. The most widely used representation is "the bag of words": a document is representedby a vector of features, each of which corresponds toa term or a phrase in a vocabulary collected from a particulardata set. The value of each feature element representsthe importance of the term in the document, according to aspecific feature measurement.

A big challenge in text categorization is the learning fromhigh dimensional data. On one hand, tens and hundreds ofthousands terms in a document may lead to a high computationalburden for the learning process. On the other hand,some irrelevant and redundant features may hurt predictive performance of classifiers for text categorization. To avoidthe issue of the "curse of dimensionality" and to speed upthe learning process, it is necessary to perform featurereduction to reduce feature size.

A common feature reduction approach for text categorization is feature selection that this paper concentrates on, where only a subset of original features are selected as input to the learning algorithms. In last decades, a number of feature selection methods have been proposed, which canbe usually categorized into the following two types of approach: the filter approach and the wrapper approach [6].The filter approach selects feature subsets based on the generalcharacteristics of the data without involving the learning algorithms that will use the selected features. A scoreindicating the "importance" of the term is assigned to each individual feature based on an independent evaluation criterion, such as distance measure, entropy measure, dependency measure and consistency measure. Hence,the filter approach only selects a number of top ranked features and ignores the rest. Alternatively, the wrapper approach greedily searches for better features with an evaluation criterion based on the same learning algorithm.

Although it has been shown that the wrapper approach usually performs better than the filter approach,it has much more computational cost than the filter approach, which sometimes makes it impractical.Typically, the filter approach is predominantly used intext categorization because of its simplicity and efficiency.However, the filter approach evaluates the goodness of afeature 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, itis hard to determine which filter feature selection approachis the best for discrimination.

For a given data set, we first generate the vocabulary with a set of M unique terms from all documents. Then, for each document, a feature vector canbe formed by using various feature models. Typically, thevalue of a feature represents the information about this particular term in a document. Two feature models have been widely used. The first one is the binary feature model inwhich the feature takes value either 0 or 1 corresponding tothe presence or the absence of a particular term in the document.The distribution of such binary feature for each classcan be usually modeled by a Bernoulli distribution. Theother one is multi-value feature model in which the feature takes values in {0,1,..}corresponding to the number of occurrences of a particular term in the document, and thusit is also called term frequency (TF). The distribution of TFfor each class can be usually modeled by a multinomial distribution model. We note here that several other featuremodels also exist in literature, such as normalized term frequency and inverse document frequency (tf-idf) [7] and probabilistic structure representation [8]. Current work in learning vector representations of words using neural network shave shown superior performance in classification and clustering [9], [10], [11], [12], where both the ordering and semantics of the words are considered.

CONCLUSION AND FUTURE WORK

We have introduced new feature selection approaches 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. For future work, we will analyze feature dependence and develop feature selection algorithms by weighting each individual features aiming to maximize the discriminative capacity.

REFERENCES

[1] Apte, C., Damerau, F. and Weiss, S. Automatedlearning of decision rules for text categorization. *ACMTransactions on Information Systems, 12(3),* 233-251,1994.
[2] Apte, C., Damerau, F. and Weiss, S.. Text Miningwith decision rules and decision trees. *Proceedings ofthe Conference on Automated Learning and Discovery*,CMU, June, 1998.
[3] Boser, B. E., Guyon, I. M., and Vapnik, V., A TrainingAlgorithm for Optimal Margin Classifiers. *FifthAnnual Workshop on Computational Learning Theory*,ACM, 1992.
[4] Chickering D., Heckerman D., and Meek, C. ABayesian approach for learning Bayesian networkswith local structure. In *Proceedings of ThirteenthConference on Uncertainty in Artificial Intelligence*,1997.
[5] Cohen, W.W. and Singer, Y. Context-sensitivelearning methods for text categorization In *SIGIR '96:Proceedings of the 19th Annual International ACMSIGIR Conference on Research and Development inInformation Retrieval*, 307-315, 1996.
[6] Cortes, C., and Vapnik, V., Support vector networks.*Machine Learning, 20*, 273-297, 1995.
[7] Fuhr, N., Hartmanna, S., Lustig, G., Schwantner, M.,and Tzeras, K. Air/X – A rule-based multi-stageindexing system for lage subject fields. In *Proceedingsof RIAO'91*, 606-623, 1991.
[8] Good, I.J. *The Estimation of Probabilities: An Essayon Modern Bayesian Methods*. MIT Press, 1965.
[9] Hayes, P.J. and Weinstein. S.P. CONSTRUE/TIS: Asystem for content-based indexing of a database ofnews stories. In *Second Annual Conference onInnovative Applications of Artificial Intelligence*, 1990.
[10]Heckerman, D. Geiger, D. and Chickering, D.M.Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning, 20*,131-163, 1995.
[11]W. Lam, M. Ruiz, and P. Srinivasan, "Automatic text categorization and its application to text retrieval," IEEE Trans. Knowl. Data Eng., vol. 11, no. 6, pp. 865–879, Nov./Dec. 1999.
[12] H. Al-Mubaid and S. Umair, "A new text categorization technique using distributional clustering and learning logic," IEEE Trans. Knowl. Data Eng., vol. 18, no. 9, pp. 1156–1165, Sep. 2006.
[13] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," IEEE Trans. Knowl. Data Eng., vol. 17, no. 4, pp. 491–502, Apr. 2005.
[14] S. Kay, Q. Ding, B. Tang, and H. He, "Probability density function estimation using the EEF with application to subset/feature selection," IEEE Trans. Signal Process., vol. 64, no. 3, pp. 641–651, Feb. 2016.
[15] E. F. Combarro, E. Montanes, I. Diaz, J. Ranilla, and R. Mones, "Introducing a family of linear measures for feature selection in text categorization," IEEE Trans. Knowl. Data Eng., vol. 17, no. 9, pp. 1223–1232, Sep. 2005.
[16] S.-B. Kim, K.-S. Han, H.-C. Rim, and S. H. Myaeng, "Some effective techniques for naive Bayes text classification," IEEE Trans. Knowl. Data Eng., vol. 18, no. 11, pp. 1457–1466, Nov. 2006.