Global Journal of Advanced Engineering Technologies and Sciences

Mutual Information Gain Feature Selection Technique based on Bayes Classifier for high dimensional text data classification

M.Nivedha
PG Scholar
Department of Computer Science and Engineering, Kongu Engineering College, Erode,
Dr.VishnuRajaPalanisamy
Department of Computer Science and Engineering, Kongu Engineering College, Erode, Tamilnadu, India.
pvishnu@kongu.ac.in

Abstract

The text classification is based on constructing a model through learning from training examples to automatically classify text documents. With size of text document repositories grows rapidly storage requirement and computational cost of model learning become higher. The instance selection solve the above issues by reducing data size by filtering out noisy data from given training dataset. The existing work presented a biological based genetic algorithm (BGA) for effective and efficient text classification. The BGA fits a biological evolution into evolutionary process most streamlined process complies with reasonable rules. After long term evolution organisms find the most efficient way to allocate resources and evolve. It requires least computational time and provides better classification accuracy than GA. This method did not provide any feature selection criteria and marginal classification accuracy. The dimensionality of text datasets is very high. The propose work presents an efficient mutual information gain feature selection (MIGFS)technique based on naive bayes classifier for high dimensional text data classification.

INTRODUCTION

Knowledge Discovery in Databases (KDD) is a routine process, examining analysis and representation of huge amount of data repositories. KDD is classifying the lifecycle of useful, novel, and identifying valid and understandable patterns into huge and issues data sets. Data Mining (DM) is the fundamentals of the KDD process, occupying the deriving of method that discover the data, improve the design and determine preceding unidentified patterns. Classification is one of the data mining processes that allocate objects in a gathering to goal of groups or classes. The objective of classification is to correctly calculate the objective class for every case in the data. The rule of text classification is classifying documents from predefined class based on their content. In this method automated assignment of normal language text to predefined classes. The main requirement of text classification is text retrieval systems, which retrieve text in reply to a user query, and text perspective systems, which transmitting text in few via such as providing conclusions, response questions or gain data. Text classification is research concept of classification procedure of data mining where naive bayes classifier handled to classify text result viewed the different of the naive bayes classifier with associated procedures. However, it is ignores the unfavorable calculation for some individual class definition in few cases accuracy may fall.

II. RELATED WORK

In paper [1], the author clarifies the terminological confusion surrounding the view of mutual information from TC, and described MI techniques properly into information theory. Performance with the Reuter collection and OHSUMED collection view that the accurate MI techniques performance is parallel to that of IG, and it is considerably better than PMI.

In paper [2], the author discusses the problem of feature selection for the reason of classification and offers a result based on the notion of mutual information. In this function based on the information gain and obtains from reflection how features work together. Finally, the author detailed the metrics of this function combined to that of various measures which estimate features independently.

In paper [3], the author developed an ODE-based homotopy technique to go after this trajectory. The algorithm is capable to routinely discard unrelated features and too routinely to reverse and forth to remove restricted optima. The results on synthetic and real time datasets view that the technique improve low prediction error and is effective in individual relevant into irrelevant features.

In paper [4] variable and feature selection has becomes the view of more than research in areas of appliance for which datasets with tens or hundreds of thousands of variables are obtainable. In include the processing of internet documents, combinatorial chemistry and gene expression array analysis. The aim of variable selection is three-fold: developing the prediction presentation of the predictors, producing higher and most cost-effective predictors, and producing a well again understanding of the fundamental process that makes the data.

In paper [5], the author view that feature relevance only is inadequate for capable feature selection of high-dimensional data. It describes feature redundancy and suggests performing explicit redundancy analysis in feature selection. Frameworks initiated that decouples relevance analysis and redundancy

analysis. The author improves a correlation-based method for relevance and redundancy analysis, and accomplishes an empirical study of its competence and effectiveness comparing with representative methods.

In this paper [6], the author proposes a "filter" method for feature selection which is independent of any learning algorithm. In this method can be performed in also supervised or unsupervised fashion. The proposed method is based on the observation that, in many real world classification problems, data from the same class are often close to each other.

In paper [7], effectiveness of heuristic search is straightforwardly upon the worth of heuristic estimates of positions in a search space. Agreed the huge amount of research work dedicated to computer chess during the past half-century, inadequate attention has been paid to the problem of influential if proposed change to an evaluation function is beneficial.

In paper [8] Mini max Tree Optimization (MMTO) is study a heuristic estimate function of a realistic alpha-beta search program. The estimate function may be a linear or non- linear mixture of weighted features, and the weights are the metrics to be optimized. To organize the search effects so that the travel decisions concur with the game records of human authority, a well-modeled goal function to be minimized is designed.

In paper [9] naive Bayes classifier really make simples studying by high and mighty that features are self-governing given class. Although autonomy is normally a poor theory, in perform naive Bayes frequently competes well with more sophisticated classifiers. The main process of understand data characteristics which involve the presentation of naive Bayes. Monte Carlo simulations handled that permit a systematic learn of classification accuracy for different classes of arbitrarily makes problems.

In paper [10] the author describes heuristic solutions to a few issues with Naive Bayes classifiers, lecture to both systemic concerns with troubles that arise because text is not really produced according to a multinomial model. Find out the simple corrections result in rapid algorithm that is aggressive with state of-the-art text classification algorithms such as the Support Vector Machine.

In paper [11], the author presents a different alteration that is designed to contest potential issues from application of MNB to unbalanced datasets. In this method a suitable correction by correcting attributes priors. This modification can be realized as another data normalization step, and the author explain that it can considerably get better the area under the ROC curve.

In paper [12], the author adopts novel dimension reduction methods to reduce the dimension of the document vectors dramatically. This problem already initiates the decision functions for the centroid-based classification algorithm and support vector classifiers to utilize the classification issues where a document may belong to multiple classes. The important tentative results view that with some dimension decrease methods that are designed mainly for clustered data, higher efficiency for both training and testing can be attained without sacrificing calculation accuracy of text classification even when the dimension of the input space is significantly reduced.

In paper [13] presents mlogit is a package for R which facilitates the view of the multinomial logit models with being and/or alternative precise variables. The major expansions of the essential multinomial model (heteroscedastic, nested and random parameter models) are realized. The logit model is helpful when one tries to clarify discrete choices.

In paper [14] the author expand a robust estimator—the hyperbolic tangent (tanh) estimator—for over discrete multinomial regression models of count data. The tanh estimator offers correct estimations and dependable deductions even when the specified model is not good for as much as half of the data. Seriously ill-fitted counts— outliers—are identified as part of the estimation.

This study [15] proposes the assessment of Multinomial Probit models using Mendell-Elston's approximation to the collective multivariate normal for the computation of choice probabilities. The precision of this numerical approximation in computing probabilities is evaluated with other processes used in existing calibration programs. Finally, the proposed inference procedure is tested on simulated choice data.

III. PROBLEM DEFINITION

BGA did not provide any feature selection criteria and provides marginal classification accuracy. Information gained through BGA is less.

IV. PROPOSED WORK

Feature selection plays an important role in text categorization. Automatic feature selection methods such as document frequency thresholding, information gain, mutual information, and so on are commonly applied in text categorization. In information theory, the term "mutual information" refers to two random variables. It seems that (mostly in corpus-linguistic studies) that the term "mutual information" has been used for something which should correctly be termed "point wise mutual information" as it is applied not to two random variables, but rather to two particular events from the sample spaces on which the two random variables are defined. This is the version used in current studies, and really point wise mutual information (PMI). Bayesian classifiers assign the most likely class to a given example described by its feature vector.

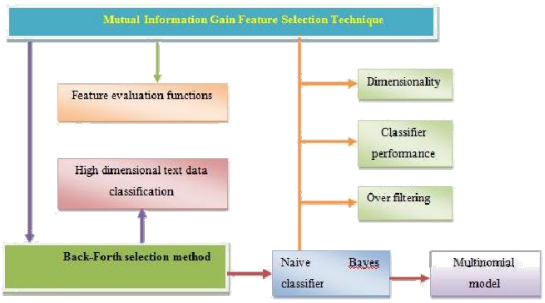

Fig: 3.1. Architecture Diagram of (MIGFS)

Thus the "mutual information" technique handle for feature selection in TC should properly be phrased "point wise mutual information". The achievement of naive Bayes in the incidence of feature dependencies can be explained as follows: optimality in terms of zero-one loss is not necessarily related to the quality of the fit to a probability distribution.

A. High Dimensional Text Data Classification

The objective of text document classification is to consign naturally a new data from one or more predefined classes based on its contents. Statistical classification methods and machine learning algorithms build automatically a classifier by learning from previously labeled data. In text classification document representation using a bag-of-words approach is employed (each position in the feature vector representation corresponds to a given word) the representation scheme leads to very high dimensional feature space. Feature selection being crucial in text classification because irrelevant and redundant words degrade the performance of classification algorithms in speed and classification accuracy. Text document classification use an evaluation function applied to a single word. All words are independently evaluated and sorted according to the assigned criterion. The feature subset selection is Pre-defined number of best features is taken to form the best feature subset. The scoring of individual words is performed using: document frequency, term frequency, mutual information, information gain. The Information gain (IG) and frequency measures work well on text data classification.

B. Mutual Information Gain Feature Selection

Feature selection is used on learning text data. Text documents are characterized by high-dimensionality feature vector. Features selection is based on mutual Information between classes and word for both individual and sequential words. Information gain (IG) is evaluated for text classification. The best of individual features evaluate all the words individually according to a given criterion sort and select best words. Since the vocabulary has tens of thousands of words, better for text classification fast, efficient and simple. An evaluation of process is each word separately and completely ignores existence of other words and words relativity. The best pair of features need not contain best single features. The information gain is most effective in word selection. The evaluation function made on feature selection information gain criteria for high dimensional domain of text classification. Sequential feature selection selects best single word evaluated by given criterion. Add one word at a time until number of selected words reaches desired number of words. The feature selection is unable to optimal words subset but take dependencies between words. Not possible in single word selection used in text classification because of computation cost due to large vocabulary size.

Document Size	Classification Accuracy (%)	
	BGA (Existing)	EMI GS
10	52	58
20	64	72
30	78	84
40	90	96

C. Naive Bayesian Classifier

With bag-of-words representation, document can be represented by a feature vector consisting of one feature variable for each word in the given vocabulary. Classes are pre-defined document always belongs to at least one class. Given a new document probability of belongingness to a class is identified by Naive Bayesian rule. If the task is to classify a new document into a single class select the class with the highest posterior probability. In assigned a multinomial model and class-conditional independence of words yields. Naive Bayes classifier computes most probable class for document as number of occurrences of word in document. The classes priors are estimated by maximum-likelihood in turnestimates the fraction of documents in each class for evaluating multi- label classification accuracy, measures used are precision are measure the number of classes found and correct / total classes found. Recall measure the number of classes found and correct/ total classes correct.

VI RESULT

This work quantifies the performance of Efficient Mutual Information Gain Feature Selection (EMIGS) techniques based on naive bayes classifier for high dimensional text data classification. The performance of evaluation function is compared to normal information gain which evaluates features individually.

The performances of classification analysis help us to provide a better understanding of large data. Classification forecasts categorical and prediction models predict unbroken valued functions. Accuracy of classification refers to the ability of model to correctly predict the class label of new data.

Figure 6.1 demonstrates the classification accuracy performance. X axis measure the document size values whereas Y axis

measure classification accuracy both the concept of improve the classification method and the performance of gain feature selection method via accuracy of bayes classifier for high dimensional text data classification. When document increased, gain of information quality gets increases accordingly. The Efficient Mutual Information Gain Feature Selection (EMIGS) technique achieves the high performance of 15 to 20 % when compared to the existing system (BGA).

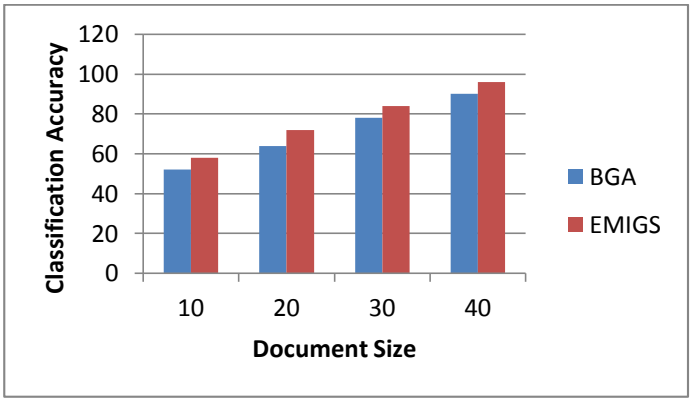

Figure 6.1. Document Size Vs Classification Accuracy

The rate of data reduction refers, as the size of text collections often high level of data which recent software and/or hardware gives correctly some shows on data reduction for text classification. The goal of instance selection is to lessen the data size by filtering out noisy data into given dataset, which could improve the probability of undignified the mining performance.

Document Size	Data Reduction Rate (%)	
	BGA	EMIGS
10	15	17
20	22	25
30	34	38
40	43	47

Figure 6.2 demonstrates the data reduction rate performance. X axis measure the document size values whereas Y axis measure data reduction rate both the concept of improve the classification method and the performance of rate of data reduction via accuracy of bayes classifier for high dimensional text data classification. When document increased, gain of information quality gets increases accordingly. The Efficient Mutual Information Gain Feature Selection (EMIGS) technique achieves the high performance of 5 to 10 % when compared to the existing system (BGA).

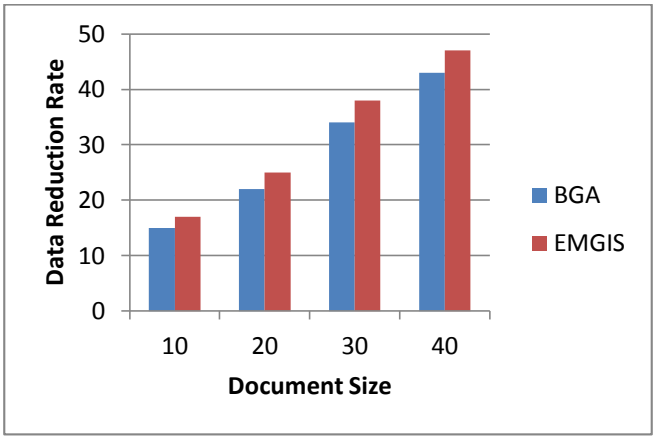

Figure 6.2. Document Size Vs Reduction Rate

VII CONCLUSION

In this paper described a mutual information gain feature selection technique based on bayes classifier for high dimensional text data classification. The Back-Forth selection method based on mutual information gain is adapted for feature selection of high dimensional text data. This work analyzed and implemented the metrics in java environment.

Performance of evaluation functions is done with compared to normal information gain which evaluates features individually. In future work, to extend the naive bayes utilize the word occurrence dependencies. This type of modifications better align naive bayes with the realities of bag-of-words textual data and view empirically, significantly develops its metrics on a number of data sets.

REFERENCES

[1] Yan Xu, Gareth Jones, JinTao Li, Bin Wang and ChunMing Sun, " A Study on Mutual Information-based Feature Selection for Text Categorization", Institute of Computing Technology, Chinese Academy of Sciences, Beijing,100080, China.

[2] Ahmed Al-Ani and Mohamed Deriche, "Feature Selection Using a Mutual Information Based Measure", Queensland University of Technology, GPO Box 2434, Brisbane, Q 4001, Australia

[3] Fuxin Li and Cristian Sminchisescu, "The Feature Selection Path in Kernel Methods", Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy.

[4] Isabelle Guyon and Andr´e Elisseeff, "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research 3 (2003) 1157-1182.

[5] Lei Yu and Huan Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy", Journal of Machine Learning Research 5 (2004) 1205–1224.

[6] Xiaofei He Deng Cai and Partha Niyogi, "Laplacian Score for Feature Selection", Department of Computer Science, University of Chicago.

[7] D. Gomboc, M. Buro and T. A. Marsland, "Tuning evaluation functions by maximizing concordance", Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada.

[8] Kunihito Hoki and Tomoyuki Kaneko, "Large-Scale Optimization for Evaluation Functions with Mini max Search", Journal of Artificial Intelligence Research 49 (2014) 527-568.

[9] I. Rish, "An empirical study of the naive Bayes classifier", T.J. Watson Research Center, 30 Saw Mill River Road, Hawthorne, NY.

[10] Jason D. M. Rennie, Lawrence Shih, Jaime Teevan And David R. Karger, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers", Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

[11] Eibe Frank and Remco R. Bouckaert, "Naive Bayes for Text Classification with Unbalanced Classes", Computer Science Department, University of Waikato, New Zealand.

[12] Hyunsoo Kim, Peg Howland and Haesun Park,"Dimension Reduction in Text Classification with Support Vector Machines", Journal of Machine Learning Research 6 (2005) 37–53.

[13] Yves Croissant, "Estimation of multinomial logit models in R: The mlogit Packages", University de la Reunion.

[14] Walter R. Mebane and Jasjeet S. Sekhon, "Robust Estimation and Outlier Detection for Over dispersed Multinomial Models of Count Data", American Journal of Political Science, Vol. 48, No. 2, April 2004, Pp. 392–411.

[15] Wangner A. Kamakura, "Estimation of Multinomial Probit Models: A New Calibration Algorithm", Vanderbilt University, Nashville, Tennessee 37203.