ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Efficient High Dimensional Text Data Classification Using Mutual Information Gain Based Feature Selection Technique

[I]Dr R.S. Vetrivel, [II]S. Maheswari, [III]Dr S.G.Shrinivas
[I]Professor, Computer Science, Subramanya College of Arts & Science, Palani
[II]Research Scholar, Computer Science, Subramanya College of Arts & Science, Palani
[III]Associate Professor, Computer Science, Subramanya College of Arts & Science, Palani

Abstract
Text data classification is the most important constraint of text retrieval systems, which recover texts to a user query. The process of classifying the text documents based on their content is referred as text classification that automatically classifies text documents. High dimensional text data classification is more considerable difficult in both supervised and unsupervised learning approach. Here, dimensionality classification is given with reduction method that classifies the data based on feature extraction and feature selection technique. Text mining is the process of retrieving the text information from textual data. The existing Biological based Genetic Algorithm (BGA) was offered that provides an effective and efficient text classification. The BGA fits a biological development into evolutionary process that complies with reasonable rules and process the resource allocation. It classifies the text data with least computational time and provides better classification accuracy. However, it does not consist of feature selection technique for selecting the text data during classification process. Therefore, Mutual Information Gain based Feature Selection technique is proposed to provide a mutual information gain for high dimensional text data classification. Initially, mutual information gain is introduced for measuring the occurrence or absence of features in text document. Then, the features are selection with the help of forward or backward approaches. Next, Discrete Function Learning algorithm is presented in high dimensional text data that identifies the relationship between the mutual information and the attributes of text data. Finally, Nearest Neighbour Classifier is used for classifying the high dimensional text data based on feature selection technique. The performance analysis of text data classification is performed using the metrics such as Feature Selection Time, Classification Accuracy and Information Gain.

Key words
Mutual Information Gain, Forward or Backward Approaches, Discrete Function Learning algorithm, Nearest Neighbour Classifier.

I. Introduction

A. Data Mining

Data mining is the method of data analysis from different user information and it is also known as knowledge discover. The extraction of hidden information from database and the classification is carried out with the help of data mining technique. The hidden information's are attained from large database to provide essential information. The unknown analytical information from large records is removed using data mining approach and helps predictive information to use the specialists with solution outside their expectations. It is described as a method of extraction and analysis of patterns, relationships and information from huge databases. Data Mining is prepared through various types of data mining software. Data mining field emerge highly efficient techniques and approaches like association rule learning. Data mining also executes interesting machine-learning algorithms like inductive-rule learning with the construction of decision trees to the development of large databases process

Data mining is generally denoted as mining of data to determine the knowledge and it is called as Knowledge Discovery in Database (KDD). It is described as the non trivial process of recognizing the reasonable patterns in data. Classification, clustering, regression and association rule learning is the major subtask used in data mining process. Data mining removes the interesting information or patterns from large information repositories like relational database, data warehouses and XML repository. The data mining based on KDD process generally consists of three phases namely, pre-processing, data mining and post-processing. Initially, pre-processing is used to obtain the operation before the data mining techniques are linked with an accurate data. It comprises of data cleaning, integration, selection and transformation. Next, data mining process in KDD produces the hidden knowledge with the use of many algorithms. Finally, post processing is computed with mining results along with user domain knowledge. Among all the process, data mining plays a significant role to KDD.

1. Knowledge Discovery in Database

Knowledge Discovery in Database is a routine process that examine and analysis the representation of data from data repositories. KDD process discovers the occupied data for improving the undefined pattern determination. The process of identifying the data or knowledge from different views of information is known as KDD process. Data mining is the method of determining the connections or patterns with dozens of fields in large relational records. Data Mining represents a method planned to examine large amount of data gathered. It is also a collection of tools employed to execute the process. Data collected from different areas like marketing, health, communication are employed in data mining.

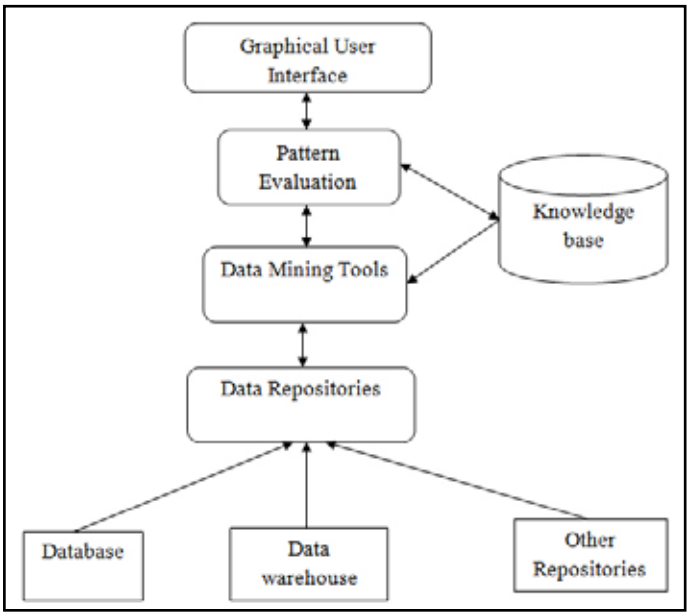

Fig. 1.1 : Knowledge Discovery in Database processes

Above figure 1.1 explains the knowledge discovery in database process along with different data mining techniques. Knowledge is estimated through definite rules like domain knowledge or ideas. The KDP model consists of a set of processing steps to be followed by practitioners when executing a knowledge discovery project. The essential KDD process occupies the derived data that discover and improves the prediction of unidentified patterns. Data classification is one of the data mining processes that allocate objects for collecting the data for every class.

Data cleaning
Data cleaning is also known as data purification for removing the noise data and irrelevant data from the collection.

Data integration
During data integration multiple data sources, often heterogeneous, are combined in a common source.

Data selection
The data selection is the process of analysis the relevant data and retrieved from the data collection.

Data transformation
It is also known as data consolidation. Here, the selected data is transformed into forms appropriate for the mining procedure.

Data mining
It is the essential step in which intellectual techniques are applied to extract potentially useful patterns.

Pattern evaluation
During the process of pattern evaluation, interesting patterns representing knowledge are identified based on given measures.

Knowledge representation
Finally, extracted knowledge is presented with the discovered knowledge to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.
Data mining is the method for determining the data connections or patterns with large relational records. It represents the methods that observes large amount of data collections. It is the procedure of identifying the data from various views and summarizing it into useful information. Data mining software is a variety of organized apparatus for analyzing the data. Data collected from different areas like marketing, health, communication are employed in data mining. . Data mining techniques are used in very large interesting organizations and data investigations. Most of the data mining approaches develop classification methods for identification of useful information from continuous data streams.

2. Classification of Data Mining
Data mining approach performs classification task to collect the data features to target categories or classes. There are different approaches for classifying the data mining that are according to the type of data source, database involved, kind of knowledge discover and used mining techniques. Spatial data, text data, multimedia data, world wide data are some of text data of data mining approaches that classified according to the data sources. Based on the database such as relational database, object oriented database, transactional database and so on performs the classification of data mining. The classification of data is performed based on the kind of knowledge discovered or data mining functionalities, such as categorization, discrimination, association, classification, clustering, etc. According to the data analysis approaches such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc performs the data mining classification.

3. Feature selection using Mutual information
Feature selection is an especially significant step during classification, because irrelevant and redundant features often degrade the performance of classification algorithms both in speed and prediction accuracy. Feature selection methods attempt to find reduced feature sets, which minimize the probability of error. The estimation functions determine a specific subset with discrimination between classes and can be divided into two main groups namely, filter and wrapper. Initially, Filters measure the significance of feature subsets that is separately given with classifier. Similarly, wrappers use the classifier's performance as the evaluation function. Filter is the most important process that is disturbed for feature selection than the wrapper process.
Feature selection has become the meeting point of much examine in areas of application for which datasets with tens or hundreds of thousands of features are available. These areas include text processing of internet documents, gene expression array analysis, and combinatorial chemistry. The prediction performance of the predictors are improved, providing faster and more cost-effective predictors and providing a better understanding of the underlying process that generated the data. The contributions of this special issue cover a wide range of aspects of such problems: providing a better definition of the objective function, feature construction, feature ranking, multivariate feature selection, efficient search methods, and feature validity assessment methods.

II. Literature Survey

1. Lexicon based Feature Extraction for Emotion Text Classification
Domain Specific Emotion Lexicon (DSEL) was developed in [1] for the generation of the feature extraction. Generated features are

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

expressed using Unigram Mixture Model (UMM) based DSEL learnt by harnessing labeled emotion. Here, the text can be used to remove effective features for emotion classification. Therefore, the features resulting using the proposed lexicon outperform those from state-of-the-art lexicons learnt using supervised Latent Dirichlet Allocation (sLDA) and Point-Wise Mutual Information (PMI).

The lexicons utilization extracts the new features of data with low dimensions for classification purposes and promising strategy. These results are predominantly impactful specified with efficient and effective need representations. Finally the hybrid features derived using the combination of n-grams and the proposed lexicon based features also result in consistent and significant improvements over n-gram features. However, the sentiment classification system was not developed for analyzing the expressive signatures imprinted by users in social. They do not calculate the prospective suggested emotions in the readers of creative text.

2. A Novel Feature Selection Technique for Text Classification Using Naive Bayes

A two-step feature selection method [2] was proposed based on a univariate feature selection and then features clustering. Here, the univariate feature selection method is used to reduce the search space and clustering process is applied to select comparatively independent feature sets. The naïve bayes classifier for text classification is improved with novel feature selection technique while comparing with other standard classifiers. Feature selection approach performs clustering process that provides minimum computational complexity.

Naive bayes text classification utilizes three steps for classifying the text. Initially, essential words are selected by using the chi-squared metric. Next, selected words are characterized by their occurrence in different documents and finally, clustering algorithm is applied for extracting the features. Therefore, it shows better efficiency on time and space complexity for both training and execution time. However, there is only limited number of clusters used during extraction process.

3. Simple-Random-Sampling-Based Multiclass Text Classification Algorithm

A simple random-sampling-based MTC (SRSMTC) algorithm was proposed in [3] to reconsider the power law. SRSMTC algorithm supports token level memory to store labeled documents and uses a text retrieval approach to solve text classification problems. The sharing of indication occurrence tracks the power law that is established with random phenomenon at least in web documents. According to the power law, many potential useless features are identified and succeed in SRS-based feature selection.

The index data structure is used in token level memory to store labeled documents. This structure has resident compressible possessions of raw texts. Each updating or retrieving of index has a constant time complexity, which can assure the space-limited and real-time requirements. The text retrieval approach with token level memory solves the text classification problems. Using linear combination, the collection of Bayesian conditional probabilities is calculated from token features. The straightforward incidence obtains promising classification performance during counting of token features thus it brings time reducing. However, is not applicable for multi label classification.

4. Terms-based Discriminative Information Space for Robust Text Classification

Discriminative information space for robust text classification (DIST) was proposed in [4]. It merges the structure of feature space and linear classifier learning for robust performance. It involves in applications with distribution shift between training and test data. Each term's discriminative power is measured by evaluating supervised term weighting and efficiently computed from the training data. The proposed method is the simplification of frequent generative, discriminative, and hybrid classification methods. They are classified with the help of naïve bayes classifier and support vector machines. DIST method uses six data sets for different applications and the data set consists of various training and testing distributions. But, only some of text features are classified and the classification of entire variants in not possible.

5. Improved global feature selection scheme for text classification

Feature selection scheme for high dimensional text data classification is one of the most common method used based on filter approach. An Improved Global Feature Selection Scheme (IGFSS) method [5] was developed for feature selection scheme to attain representative features. IGFSS method improves the classification performance of overall feature selection methods by creating a feature set representing all classes almost equally. According to their discriminative power on classes, a local feature selection method is used to label all the features.

IGFSS is a common explanation for all of the filter-based global feature selection methods that are different from the other approaches. Global feature selection method combines with class membership and non-membership for providing an efficient feature ranking. However, feature selection method with globalization produces unique features and it also extracts negative features.

6. Feature selection for high-dimensional imbalanced data

A new feature selection approach [6] was designed based on class decomposition that separates the classes. The proposed approach divides large classes into relatively smaller pretend sub class according to the class labels. Then, Hellinger distance method is established for selecting the features those measures with distribution divergence and produces effectiveness of the proposed approach by using Bayesian learning on synthetic data. Finally, the results illiterate both class decomposition and Hellinger distance method for selection high dimensional imbalanced data.

Decomposition-based feature selection approach is enforced by imbalanced data that indicates better measurement of real data. K-means is used in large classes for clustering the data points and maintain the computational cost of original classifier. An alternative feature selection method is designed based on Hellinger distance for imbalanced data. It fundamentally attains distribution of feature values on two different classes. However, Hellinger distance is insensitive to the class distributions, since the computation of this distance does not involve the class information.

7. Helmholtz principle based supervised and unsupervised feature selection methods

A new method for feature selection from textual data called Meaning Based Feature Selection (MBFS) [7] was developed for text classification. The proposed method is based on the Helmholtz

principle from the Gestalt theory of human sensitivity which is used in image processing. The performance of text data classification is effectively evaluated based on known classifiers on several datasets. In addition, feature selection algorithm is applied for text mining. Text mining based on meaning based feature selection measure is used for rapid change detection, keyword extraction and text summarization. But, Meaning score metric is not used in proposed feature selection approach for text classification.

8. Arabic Text Classification Using Polynomial Networks

Polynomial Neural Networks [8] was developed for Arabic statistical learning- based text classification system. Polynomial Networks consists of two layer such as input layer and output layer. The set of input features from input layer forms the set of monomial basis functions. Next, output layer combines all the output from networks thus produces better classification model. Identification and verification is the most important process used by Polynomial networks for text classification. The matching results of input features with given vector features are identified with the help of identification process. After feature identification, the features are accepted or rejected using verification phase. Text classification algorithm achieves better performance on dataset features with higher memory requirements. But, it does not perform direct comparisons between proposed PN classifier and a set of the state-of-the art Arabic TC algorithms.

9. Term-Weighting Schemes for Text Classification

The learning approach named as Term-Weighting Schemes (TWSs) [9] was introduced in the context of text classification. The proposed scheme establishes the way of document representation with the help of classifier in vector space model. A genetic program was designed to improve the performances of text classification by combination of term-weighting schemes. As extensive result is proposed by combining the data set from thematic and non-thematic text classification, thus produces better results on image classification.

Text classification and image classification task produces efficient search process with different variations in vocabulary size of text. There are number of terms used in text classification during training and texting data set. Term-Weighting Schemes uses classifier for estimating the fitness function and more preferable for inter data set classification.

10. Global Information Gain

Feature selection is a essential preprocessing step for text classification task used to explain the dimensionality problem. Therefore, Global Information Gain (GIG) was proposed in [10] to avoid redundancy naturally. In addition, an efficient feature selection method called Maximizing Global Information Gain (MGIG) is also proposed. Initially, higher order feature selection metric is used for selecting features for text classification. Next, an efficient maximizing global information gain algorithm is developed that reduces the computational complexity from O (VK2) to O (VK). The proposed method performs testing on proposed approach to four other algorithms on six text collections.

Maximizing Global Information Gain algorithm tested on six datasets that are significantly different from all the other algorithms. The consequent matching evaluations demonstrate strong verification that MGIG is higher when compared to other algorithms. MGIG runs impressively faster than other higher order algorithms. Hence, MGIG method is proposed for selecting the features for text classification. It also be applied in other domains provided distributional clustering or information bottleneck technology can be used.

11. Multidimensional feature subset selection algorithm

The Multidimensional Feature Subset Selection (MFSS) algorithm [11] was proposed to yields a unique feature subset. It is applied to bench mark multidimensional datasets for reducing the number of features. MFSS is an efficient feature selection algorithm without affecting the classification accuracy even for the reduced number of features. MFSS algorithm is appropriate for solving both problem transformation and algorithm adaptation when they are based on applications generating multidimensional datasets.

The feature subset selection based on class-feature is an efficient and reliable algorithm for classifying the text data. Initially, each class feature and class correlation is considered to recognize the meaning of feature for each class. Next, weight is allocated to features based on the feature-class correlation for each class. Finally, feature weight is calculated based on the proposed weight method and develops the classifier for classifying the feature subsets. However, multidimensional classification such as different single-labeled classifiers and feature selection is not possible.

12. Probabilistic feature selection method for text classification

A novel filter based probabilistic feature selection method named Distinguishing Feature Selector (DFS) was proposed in [12] for text classification. The considered system is evaluated with recognized filter approaches including chi square, information gain, Gini index and deviation from Poisson distribution. DFS evaluates the class discrimination in a probabilistic approach and assigns certain importance scores to text approaches. Using different datasets, classification algorithms measures effectiveness of feature selector and compared against well known filter techniques. However, pattern classification on probabilistic feature selection is remaining unaddressed.

13. Class-indexing-based term weighting for automatic text classification

An automatic indexing method using the combination of document-based and class (category)-based approaches [13] was proposed named as Class-indexing-based term weighting. The proposed term weighting favors the rare terms and biased against frequent terms that gives a positive discrimination on both to rare and frequent terms. The proposed class-indexing-based term weighting approach is effective in high-dimensional and comparatively low-dimensional vector spaces than the state-of-the-methods. The proposed class-indexing method generates more informative terms based on a certain category through use of inverse class frequency and inverse class space density frequency functions. However, proposed system does not perform classification by combining class-indexing and semantic indexing.

14. Bilevel Feature Extraction-Based Text Mining for Fault Diagnosis

A bilevel feature extraction-based text mining [14] was proposed with improved classification performances. They perform classification by combining the features that are removed at both

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

syntax and semantic levels. Statistics-based feature selection method is improved at the syntax level to defeat the learning complexity produced by an imbalanced data set. A prior latent Dirichlet allocation-based feature selection is performed at the semantic level to reduce the data set into a low-dimensional topic space. Finally, fault features derived from both syntax and semantic levels are combined with serial fusion. The proposed method uses fault features at different levels and enhances the precision of fault diagnosis for all fault classes, particularly minority ones. However, parallel feature fusion is impossible for imbalances classifications.

15. Optimal Feature Set in High-Dimensional Data by Swarm Search

A new feature selection scheme called Swarm Search [15] was proposed with an optimal feature set by using Meta heuristics. Different types of classifiers are integrated into its fitness function based on the flexibility of Swarm Search and plugging in any Meta heuristic algorithm to make possible heuristic search. Swarm Search over some high-dimensional datasets is carried out with different classification algorithms and various Meta heuristic algorithms. Swarm Search is proficient to achieve comparatively low error rates in classification without reduction in the size of feature subset. However, Swarm Search does not perform classification and produces the Naive Bayes as classification error on high dimensional data.

III. Efficient High Dimensional Text Data Classification Using Mutual Information Gain Based Feature Selection Technique

Feature selection technique is an important approach for high dimensional text classification. There are some of feature selection methods such as information gain, mutual information and filter method and so on. Data mining, machine learning, image processing are some of application that uses feature extraction technique. Feature selection utilizes joint mutual information for the identification of subset of features. They uses maximum of minimum approach with class labels that increases the mutual information. The effective relationship among relevancy and redundancy results in better feature selection and proposed joint mutual information maximization.

The feature selection technique frequently uses preprocessing step for classifying the high dimensional text data. With the help of induction algorithm, feature selection technique eliminates the irrelevant and a redundant feature thus produces better efficiency. Mutual Information (MI) metric represents the needs of features effectively. It measures the relevance of features with higher order statistical structures and evaluated based on different feature functions. Information gain is defined as the measures of information obtained for group calculation by significant presence or absence of a text in a document. Information gain also known as gain ratio is generally used for measuring the reduction in entropy required for category prediction by knowing the presence or the absence of a term or feature in the document. It is frequently used as a term goodness criterion in machine learning. Feature selection helps in reduction of instruct text data from document by selecting the features having non-zero value.

1. System Architecture

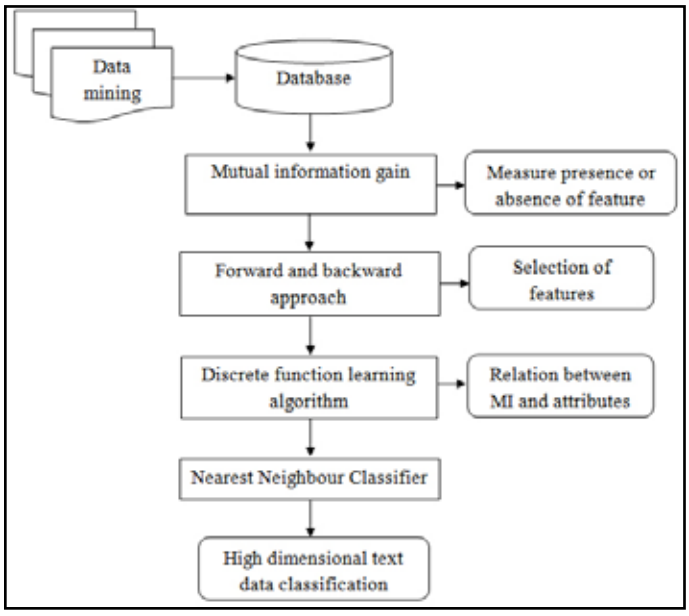

Fig. 3.1 Architecture diagram of Mutual Information Gain Based Feature Selection Technique for high dimensional text classification

Figure 3.1 shows the basic architecture diagram of mutual information gain based feature selection technique for high dimensional text classification. Initially, they identify the presence or absence of text feature by using mutual information gain technique. Then, based on forward or backward approach, feature is selected for classifying the high dimensional text data. Discrete function learning algorithm is introduced to detect the relationship between the mutual information and text data attributes. Finally, with the help of Nearest Neighbour Classifier along with feature selection technique obtains high dimensional text data classification. The high dimensional text classification using Mutual Information Gain based Feature Selection technique is divided into four type of process and it is given below.

a) Mutual Information Gain
b) Forward and Backward Feature Selection approaches
c) Discrete Function Learning Algorithm
d) Nearest Neighbour Classifier

a) Mutual Information Gain

Mutual information is used to properly predict the information based on feature selection when there is different mutual information. The term mutual information in information theory is referred with two variables. Information theory measures the uncertainty of random variables quantitatively and the amount of information shared by them effectively. Here, Information Gain measures the amount of information obtained for category prediction by knowing the presence or absence of a term in a document. Mutual information is a condition commonly used in numerical language modeling of word associations and related applications.

Mutual information is referred with two variables such as category and term. The results from the mutual information are always non-negative. The features from high dimensional text data is selected based on feature selection technique that involves searching process. The selection process selects the relevant features and rejects the irrelevant features according to various text data. Therefore, feature selection technique re-computes the

MI from unspecified occurrences of unselected features for each addition of a feature to a subset of the selected features. Hence, it minimizes the redundancy of the features by computing MI dynamically.

b) Forward and Backward Feature Selection approaches

There are two approaches on high dimensional text data classification namely, forward feature selection approach and backward feature selection approach. Initially, Forward Selection is an extremely efficient method for selection of relevant features from the data sets. Forward Selection refers to a search with no variables at the empty set of features. They add the feature variable at each step that decreases the error, until any further addition does not significantly decrease the error. Forward selection determines better solution on feature selection with two components from remaining input attributes. Sequential Forward Selection approach is the simplest greedy search algorithm.

Next, backward feature selection approach starts with all variables in high dimensional text data and removes the features one by one. While removing the features, they decrease the error at each stage until any additional removal increases the error significantly. In order to provide an appropriate reduction on feature error, the validation set is different from the training set.

c) Discrete Function Learning Algorithm

Discrete Function Learning algorithm as a filter feature selection method using high-dimensional mutual information to determine the correlation between the candidate feature subsets and the class attribute. The relationship is given between the mutual information of two variables and number of attributes. Information theory on discrete solution involves the quantification of information. It consists of continuous topics such as: analog signals, analog coding and analog encryption. Discrete probability theory arrangements with events are occurred in countable sample spaces. Categorization of machine learning tasks arises when one considers the desired output of a machine-learned system. They are classified as supervised learning, unsupervised learning and reinforcement learning.

d) Nearest Neighbour Classifier

The Nearest Neighbors Classifier Algorithm is used for text classification on high dimensional data. It is a non-parametric method used for classification and regression. The input consists of the k closest training examples in the feature space and the output depends on whether k-NN is used for classification or regression. Nearest Neighbors classification classifies the objects by a majority of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. Similarly k-NN regression gives the output value for the object. This value is the average of the values of its k nearest neighbors. Therefore, they perform the high dimensional text data classification with the help of Nearest Neighbors Classifier. The process of Mutual Information Gain based Feature Selection technique is given below:

Input : Number of text data in database
Output: High Dimensional Text Data Classification
Begin
Step 1: Mutual Information Gain
Step 2: Measures the presence or absence of features in text document
Step 3: Forward or Backward approaches
Step 4: Selection of features
Step 5: Discrete Function Learning algorithm
Step 6: Identifies relationship between the mutual information and the attributes
Step 7: Nearest Neighbour Classifier
Step 8: Classifying the high dimensional text data
End

IV. Results and Discussion

The performance analysis is carried out in this paper with the metrics of Feature Selection Time, Text Data Classification Accuracy and Information Gain Ratio. The performance metric of Mutual Information Gain based Feature Selection technique (MIG-FST) is evaluated and analyzes the values in java environment. Following metrics are used for experimental purposes.

- Feature Selection Time
- Text Data Classification Accuracy
- Information Gain Ratio

1. Feature Selection Time

The feature selection time is defined as the measure of time taken for selecting the features from high dimensional text data for the classification process. In order to provide minimum feature selection time, all the text data's are classified according to the stored datasets. It is measured in terms of milliseconds (ms). Lower data classification time ensures efficiency of the method.

Table 4.1 : Tabulation of Feature Selection Time (ms)

Number of high dimensional text data	Feature Selection Time (ms)	
	Existing BGA	Proposed MIG-FST
10	11.05	9.2
20	12.69	10.89
30	13.52	11.28
40	14.74	12.47
50	15.69	13.75

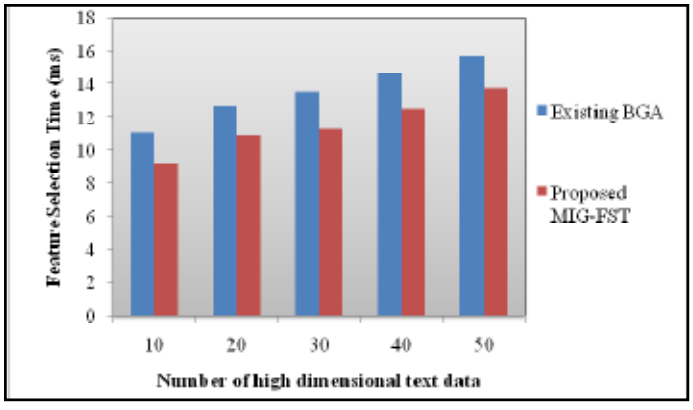

Fig. 4.1 Measure of Feature Selection Time (ms)

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Above figure 4.1 shows the analysis of feature selection time with respect to different number of high dimensional text data in text document. For experimental purpose, the high dimensional text data is considered in the ranges from 10 to 50. The figure shows the comparison made between Biological based Genetic Algorithm (BGA) and Mutual Information Gain based Feature Selection technique (MIG-FST) method. When the number of text data are increased, feature selection time is also get increased. Therefore, Mutual Information Gain based Feature Selection technique achieves minimum feature selection time that is used for classifying the text data. As a result, feature selection time is reduced by 15% when compared to the existing Biological based Genetic Algorithm (BGA).

2. Text Data Classification Accuracy

The text data classification accuracy is characterized as the measure of number of high dimensional text data that are correctly classified from the text document. Text data classification is considered with number of high dimensional text data given by the text document. The text data classification accuracy is measured in terms of percentage (%).

Table 4.2 : Tabulation of Text Data Classification Accuracy (%)

Number of high dimensional text data	Text Data Classification Accuracy (%)	
	Existing BGA	Proposed MIG-FST
10	68.56	71.85
20	69.58	73.44
30	71.58	76.98
40	73.2	78.65
50	75.74	80.74

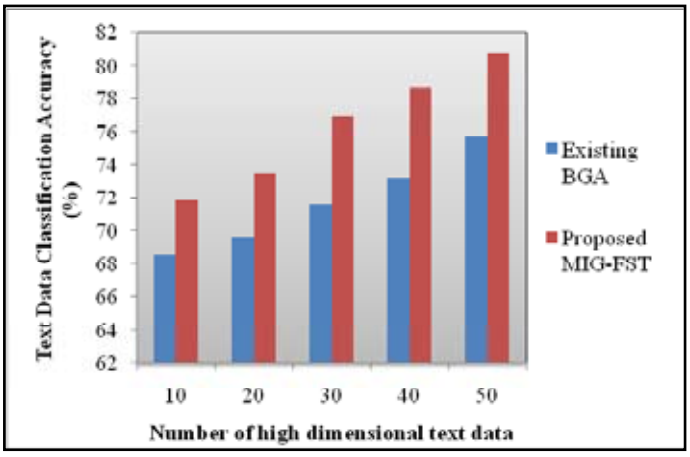

Fig. 4.2 : Measure of Text Data Classification Accuracy (%)

Above figure 4.2 demonstrate the analysis of text data classification accuracy with respect to different number of high dimensional text data in text document. For experimental purpose, the high dimensional text data is considered in the ranges from 10 to 50. The figure shows the comparison made between Biological based Genetic Algorithm (BGA) and Mutual Information Gain based Feature Selection technique (MIG-FST) method. When the number of text data are increased, text data classification accuracy is also get increased. Therefore, Mutual Information Gain based Feature Selection technique achieves higher text data classification accuracy by classifying the text data. As a result, text data classification accuracy is improved by 6% when compared to the existing Biological based Genetic Algorithm (BGA).

3. Information Gain Ratio

The large amount of data is increasing rapidly; text classification has become one of the techniques for managing large scale text repositories. Feature selection is an extremely important step in text classification, since irrelevant and redundant words frequently degrade the performance of information gain in high level of speed in proposed scheme.

Table 4.3 : Tabulation of Information Gain Ratio (%)

Number of high dimensional text data	Information Gain Ratio (%)	
	Existing BGA	Proposed MIG-FST
10	72.56	77.68
20	74.85	79.69
30	75.96	81.27
40	77.42	84.63
50	79.14	86.87

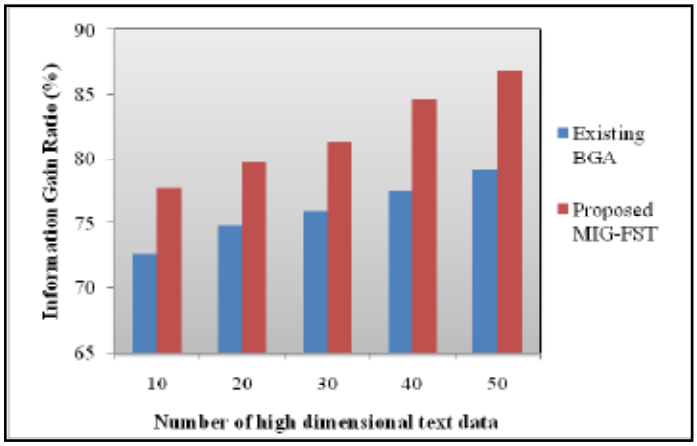

Fig. 4.3 : Measure of Information Gain Ratio (%)

Above figure 4.3 illustrate the study of information gain ratio with respect to different number of high dimensional text data in text document. For experimental purpose, the high dimensional text data is considered in the ranges from 10 to 50. The figure shows the comparison made between Biological based Genetic Algorithm (BGA) and Mutual Information Gain based Feature Selection technique (MIG-FST) method. When the number of text data are increased, information gain ratio is also get increased. Therefore, Mutual Information Gain based Feature Selection technique achieves higher information gain ratio during classifying the text data. As a result, information gain ratio is improved by 8% when compared to the existing Biological based Genetic Algorithm (BGA).

V. Conclusion and Future Work

Therefore, a mutual information gain for high dimensional text data classification is provided using proposed Mutual Information Gain based Feature Selection technique. The presence or absence of features in text document is measured by introducing mutual information gain. They are measured based on forward or backward feature selection approaches. Next, Discrete Function Learning algorithm is presented in high dimensional text data that

identifies the relationship between the mutual information and the attributes of text data. Finally, Nearest Neighbour Classifier is used for classifying the high dimensional text data based on feature selection technique. Further, the work is extended naive bayes classifiers for classifying the text data of bag-of-words textual data. They are significantly developed with number of data sets.

References

[1] *Anil Bandhakavi, Nirmalie Wiratunga, Deepak P and Stewart Massiea, "Lexicon based Feature Extraction for Emotion Text Classification", Pattern Recognition Letters, Elsevier December 2016, Pages 1-12.*

[2] *Subhajit Dey Sarkar, Saptarsi Goswami, Aman Agarwal and Javed Aktar, "A Novel Feature Selection Technique for Text Classification Using Naive Bayes", Hindawi Publishing Corporation, International Scholarly Research Notices, Volume 2014. Pages 10.*

[3] *Wuying Liu, Lin Wang and Mianzhu Yi, "Simple-Random-Sampling-Based Multiclass Text Classification Algorithm", Hindawi Publishing Corporation, The Scientific World Journal, Volume 2014.*

[4] *Khurum Nazir Junejo, Asim Karim, Malik Tahir Hassan and Moongu Jeon, "Terms-based Discriminative Information Space for Robust Text Classification" Information Sciences, Elsevier, Volume 372, 1 December 2016, Pages 518–538.*

[5] *Alper Kursat Uysal, "An improved global feature selection scheme for text classification", Expert Systems with Applications, Elsevier, Volume 43, January 2016, Pages 82–92.*

[6] *Liuzhi Yin, YongGe, Keli Xiao, Xuehua Wang and Xiaojun Quan, "Feature selection for high-dimensional imbalanced data" Neuro computing, Elsevier, Volume 105, 1 April 2013, Pages 3–11.*

[7] *Melike Tutkan, Murat Can Ganiz and Selim Akyokus, "Helmholtz principle based supervised and unsupervised feature selection methods for text mining", Information Processing & Management, Elsevier, Volume 52, Issue 5, September 2016, Pages 885–910.*

[8] *Mayy M. Al-Tahrawia and Sumaya N. Al-Khatibb, "Arabic Text Classification Using Polynomial Networks" Journal of King Saud University - Computer and Information Sciences, Volume 27, Issue 4, October 2015, Pages 437–449.*

[9] *Hugo Jair Escalante, Mauricio A. García-Limón, Alicia Morales-Reyes, Mario Graff, Manuel Montes-y-Gomez, Eduardo F. Morales and Jose Martinez-Carranza, "Term-weighting learning via genetic programming for text classification", Knowledge-Based Systems, Elsevier, Volume 83, July 2015, Pages 176–189.*

[10] *Changxing Shang, Min Li, Shengzhong Feng, Qingshan Jiang and Jianping Fan, "Feature selection via maximizing global information gain for text classification", Knowledge-Based Systems, Elsevier, Volume 54, December 2013, Pages 298–309.*

[11] *Senthilkumar Devaraj and S. Paulraj, "An Efficient Feature Subset Selection Algorithm for Classification of Multidimensional Dataset", The Scientific World Journal, Hindawi, Volume 2015, Pages 9.*

[12] *Alper Kursat Uysal and Serkan Gunal, "A novel probabilistic feature selection method for text classification", Knowledge-Based Systems, Elsevier, Volume 36, December 2012, Pages 226–235.*

[13] *Fuji Ren and Mohammad Golam Sohrab, "Class-indexing-based term weighting for automatic text classification", Information Sciences, Elsevier, Volume 236, 1 July 2013, Pages 109–125.*

[14] *Feng Wang, Tianhua Xu, Tao Tang, Meng Chu Zhou and Haifeng Wang, "Bilevel Feature Extraction-Based Text Mining for Fault Diagnosis of Railway Systems", IEEE Transactions on Intelligent Transportation Systems, Volume 18, Issue 1, January 2017, Pages 49-58.*

[15] *Simon Fong, Yan Zhuang, Rui Tang, Xin-She Yang and Suash Deb, "Selecting Optimal Feature Set in High-Dimensional Data by Swarm Search", Journal of Applied Mathematics, Hindawi, Volume 2013, 18 pages.*