Asian Journal of Information Technology 15 (14): 2355-2366, 2016
ISSN: 1682-3915

Experimental Investigation for Text Categorization Based on Hybrid Approach Using Feature Selection and Classification Techniques

[1]K. Sridharan and [2]M. Chitra
[1]Department of IT, Panimalar Engineering College, Chennai, Tamil Nadu, India
[2]Department of IT, Sona College of Engineering, Salem, Tamil Nadu, India

Abstract: In the midst of the mounting accessibility of electronic documents and the fast development of the World Wide Web turned out to be the enormous task of automatic categorization of documents. It has been developed into the key method for systematizing the discovery of information and knowledge. The appropriate categorization of text mining is required for the digital libraries, e-documents, blogs, emails and online news, machine knowledge and usual language processing methods to get a significant knowledge. The primary objective of this paper is to accentuate the significant methods and procedures that are engaged in manuscript documents categorization, at the same time creating awareness about some of the intriguing challenges that continue to be solved, chiefly focused on text representation and machine learning techniques In this study, we initiate a new-fangled technique which associates the characteristic selection and categorization techniques to pace up the text classification process and then about the low time consumption. In this paper, we propose a new method of IG-ANN which is the combination of feature selection and classification technique that increases and improves the classification accuracy, feature selection rate. We demonstrate the effective of our process by means of a systematic assessment and similarity over 13 datasets. The performance can be improved thus achieved makes ANN comparable or higher to supplementary classifiers. The projected algorithm is revealed to do better than the further conventional techniques like Best First Search wrapper method and filtered attribute method.

Keyword: Information gain, PCA, artificial neural network, Latent Semantic Analysis (LSA), genetic algorithm, SVM, naive bayes classification

INTRODUCTION

In the modern days of technology text mining studies are advancing into next level due to mounting number of the electronic documents from a mixture of resources. The resources of unstructured and semi structured data includes the world wide web, governmental electronic repositories, news editorials, genetic directory, depositories of blog, online forums, digital libraries, electronic mail and chat rooms. Consequently, the appropriate categorization and knowledge detection from these sources and it marks a major role in the field for investigation.

Natural Language Processing (NLP), Information Mining and Machine Learning methods work reciprocally in categorizing the determine patterns instinctively from the electronic documents. The primary objective of the text mining is to facilitate clients to extort information from textual resources and compacts with the maneuvers like, repossession, categorization (supervised, unsupervised and semi supervised) and recapitulation. In contrast, how these documents can be aptly interpreted, presented and classified. In view of that, it consists of numerous challenges, like proper explanation to the documents, with appropriate file demonstration, dimensionality diminution to grip algorithmic concerns (Soon *et al*., 2001). Moreover, a suitable classifier jobs are occupied to accomplish good overview and evade over-fitting. Mining, incorporation and categorization of electronic datas from miscellaneous sources and knowledge discovery have been composed from these documents to channel it for the research societies.

At present, the web is the chief source for the text documents, the quantity of textual data existing to us is constantly mounting and approximately 80% of the data of an organization is piled up in unstructured textual format (Caldas and Soibelman, 2003) like reports, email, views and news, etc. The exhibits that just about 90% of the global data is apprehended in unstructured formats, as a result Information intensive business processes stipulate that we surpass since simple document retrieval to knowledge discovery (Matsuo and Ishizuka, 2004). The

Corresponding Author: K. Sridharan, Department of IT, Panimalar Engineering College, Chennai, Tamil Nadu, India

requirement of automatically retrieval of useful knowledge with the colossal amount of textual data is used to facilitate the support of human investigation is fully perceptible (Rahm and Do, 2000).

The movement of the market is depended on the information of the online news articles, reactions and proceedings is turned out to be an budding theme for investigation in the field data mining and text mining (Peng *et al*., 2008). To establish the outcome of modern methods for text classifications are elucidated in (Berry *et al*., 1995) which three problems were highlighted: documents demonstration, classifier erection and classifier assessment. Therefore, generating an information structure that can signify the datas and creating a classifier that can be utilized to visualize the class label of a document with high accuracy, develops into the major issues in text categorization.

The primary mission of research is to assess the existing known work, therefore it create**s** an attempt to accumulate what's identified about the data categorization and illustration. This research gives an opportunity to enfold the review of syntactic and semantic themes, domain ontology, tokenization alarm where it is pay more attention on the different machine in education techniques for text classification using the subsisting literature. The aggravated perception of the allied investigation areas of text mining are: Information Extraction Method; Information Retrieval Method and Natural Language Procession method.

Information Extraction (IE) methods intend is to dig out explicit information from text documents. This is the initial approach assumes that text mining fundamentally keep in touch to information extraction. Information Retrieval (IR) is used to discover the documents which contain answers to questions. It is to facilitate to achieve these target arithmetical procedures and techniques are employed for mechanical processing of text information and in contrast to the specified problem. Data recovery in wider sense compacts the means of the complete variety of data processing, from data retrieval to information retrieval (Dhillon and Modha, 2001).

Natural Language Processing (NLP) is used to achieve an enhanced perceptive of normal language through the utilization of computers and imply the documents semantically to advance the categorization and data retrieval methods. Semantic investigation is the technique of structurally based sentences and parts into major discernments, mostly verbs and proper nouns. By the means of statistics-backed technology, these terminologies are then compared to the classification. Ontology is the explicit and conceptual model representation of previously distinct finite sets of terms and concepts, implicated in information management, information engineering and smart data incorporation (Zhang *et al*., 2008). In order to determine the evidence and to confirm the quality of papers we carried out a comprehensively examination of the results provided from the study. In our forthcoming research, we will endeavor to make this step more strong and efficient. We have endeavored to get various reports by using tables and graphs on the base of existing investigations.

Representation of the documents: The document image is one of the preprocessing method that is utilized to decrease the complexity of the documents and formulate them to be more affordable to handle, the document have to be imprecise from the complete text edition to a manuscript vector. Transcript image is the one of most essential aspect in documents categorization, where it signifies the mapping of documents into a compact form from its contents. A text document is typically symbolized as a vector of expression weights (word features) from a set of stipulations (dictionary),wherever each expression takes place minimum one time in an undoubtedly in least number of document. The major trait of the text categorization problem is exceptionally resulted in elevated height of text data. The quantity of budding features frequently surpasses the number of guiding documents. A classification of a text is to completed of a mutual membership of terms which have an assortment of patterns of occurrence. The Text classification is a key component in many informational management tasks, still with the volatile augmentation of the web data, algorithms that canprogress the classification efficiency while maintaining accuracy are extremely preferred (Bizer *et al*., 2009).

Data pre-processing or dimensionality Diminution (DR) consents to a competent data management and demonstration. There are several considerations on the pre-processing and DR in the current literature and many representations and performances have been proposed. DR is an essential step in text classification, since the irrelevant and superfluous features often humiliate the performance of classification algorithms both in pace and categorization accuracy and also it has propensity to decrease over fitting.

DR techniques can be alienated into two approaches they are Feature Extraction (FE) (Cohen and Hersh, 2005) and Feature Selection (FS) methods as it is discussed below.

Feature extraction: The method of pre-processing is to make it simple to the border of every language structure and to eradicate to the extent of promising the language

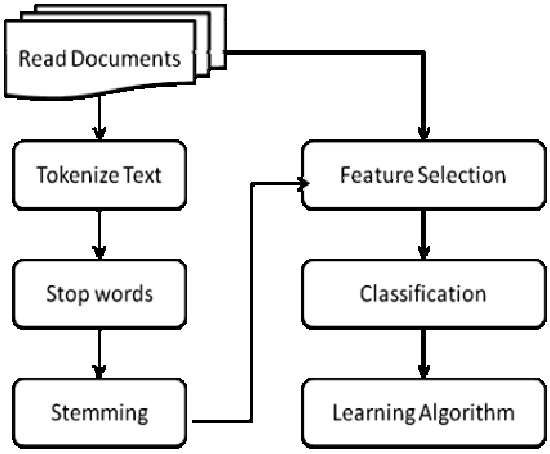

Fig. 1: Process for classification of text documents

contingent aspects, tokenization, prevent words removal and stalking (Krallinger *et al*., 2008). Feature Extraction is the primary procedure of pre processing which is deployed to depict the text documents in a comprehensible word format. Therefore, confiscating to prevent words and stalking words is the pre-processing responsibilities (Krallinger *et al*., 2008; Mahgoub *et al*., 2008). In text classification the documents are embodied by a great amount of features and the majority of documents might be immaterial or noisy. DR is the segregation of a large number of keywords, base rather on a statistical process, to produce a low dimension vector (Belkin and Niyogi, 2003). Recently, DR methods have innermost much attention since successful dimension reduction originates the learning assignments more professional and hoard more storage space (Kohonen *et al*., 2000). On the whole, the steps taken for the FE (Fig. 1) are.

Tokenization: A manuscript is applied as a series, in addition to that it is alienated into a catalog of indications.

Eradicate stop words: Prevent words such as “the”, “a”, “and” etc. are commonly takes place so the inconsequential words need to be detached.

Stemming word: Employing the stemming algorithm that renovates the different word structure into comparable canonical shape. This method is the formula of conflating symbols to their origin structure, e.g., connection to connect, computing to compute, etc.

Feature selection: In subsequent to characteristic mining, the significant procedure in pre processing of manuscript classification is attribute selection. It is employed to create vector space which progress the scalability, competence and exactness of a text classifier. In most cases, a good feature selection method should be considered the domain and algorithm features (Liu *et al*., 2005). The central design of FS is to opt for compartment of features from the novel documents. Moreover, FS is acted upon by keeping the words with highest score in according to prearranged measure of the significance of the word. The selected features conserves original substantial meaning and afford a better understanding for the information and learning process (Kalousis *et al*., 2007). For text classification, the key issues are the elevated facets of the feature space. Nearly each text domain has large quantity of features, the majority of these features are not significant and complimentary for text categorization task and still the various noise features may stridently diminish the classification accuracy (Sokolova and Lapalme, 2009). Therefore, FS is generally deployed in text categorization to shrink the dimensionality of feature space and progress the competence and exactness of classifiers.

There are generally two kinds of feature selection techniques in machine learning; wrappers and filters. Wrappers utilized in the classification precision of some learning algorithms as their assessment function. Since, wrappers have to instruct a classifier for every feature subset to be assessed, they are generally much more time consuming. In particularly, when the number of features is high. Accordingly, wrappers are by and large not suitable for text categorization. At the same time, as opposed to wrappers, filters act upon FS autonomously of the learning algorithm that determine to make use of the selected features. With the intention to assess a feature, filters exploit an assessment metric that calculates the capacity of the feature to distinguish each class (Mittermayer, 2004). In text categorization, a text manuscript may moderately match many categories. Therefore, we necessitate discovering the greatest identical group for the text document. The expression (word) regularity/converse Document Frequency (TF-IDF) technique is generally used to weight every word in the manuscript document according to how characteristic it is. In further expressions, the TF-IDF technique precincts the pertinent amongst words, text datas and fussy classifications.

Principle component analysis: PCA is a renowned technique that can decrease the dimensionality of information by converting the unique attribute space into smaller space. In other word, the intention of principle components investigation is to originate new variables that are mixture of the original variables and are uncorrelated. In further, this can be attained by converting the novel variables Y = [y1, y2, ... ,yp] (where

Table 1: Document classification semantic issues

Classification	Description
Stemming	If we reduce the terminology to their branches, how it will influence the connotation of the credentials
Tokenization	How the documents are tokenized and tokens are recordedor annotated, by word or phrase. This is importantbecause many downstream components need thetokens to be clearly identified for analysis
Sentence Splitting	How we Identifying sentence boundaries in a document
Stop wordlist	In which domain, the stop words should be considered and How stop expression directory will be assumed
Collocations	It is such about the technical terms and compounds
Domain and data understanding For Ontology	The construction of ontology using its relations, area and availability of data is defined
Syntax	How should formulate a syntactic or grammatical observation. What in relation to information craving, anaphoric harms
Word logic	How we clarify the connotation of the expression in the manuscript, uncertainty crisis
Text demonstration	Adjective or noun, concept or word, and phrases are more significant for the document representation
Noisy data	From the noisy data, the steps are necessitated to be authorizing for the document
Tagging of POS (Part-of-Speech)	It is such about the annotation of data and the parts of the characteristics of speech. POS information is token by tagging a POS assigned by such components

p is number of original variable) to an original set of variables, T = [t1, t2,..., tq] (where q is figure of innovative variables) which are permutations of the original variables. The distorted attributes are framed initially by computing the mean (μ) of the dataset and then covariance matrix of the novel elements is calculated (Kambhatla and Leen, 1997). And the next step is extorting its eigenvectors. The eigenvectors (Belkin and Niyogi, 2003) (principal components) establish as a linear revolution since the unique characteristic space towards a fresh space in which characteristics are uncorrelated. Eigenvectors can be arranged according to the quantity of variation in the original data. The preeminent eigenvectors (those one with highest Eigen values) are chosen as innovative characteristics while the remaining things are redundant.

Latent semantic analysis: LSA method is a novel method in text classification. Generally, LSA explores the relationships between a term and concepts enclosed in a shapeless compilation of data. It is known as Latent Semantic Analysis, owing to its ability it is to associate semantically through associated expressions that are hidden in a text. LSA (Yeh *et al*., 2005) fabricates a group of perceptions which is smaller in dimension than the unique set, associated to datas and expressions. It deploys Singular Value Decomposing (SVD) to recognize pattern connecting the expressions and perceptions enclosed in the data and determine the relationships between the documents. Generally, this technique referred as concept searches. And, also it has capability to remove the abstract content of a body of text by ascertaining associations between those terms that occur in parallel contexts. LSA is typically used for page recovery systems and text clustering purposes. Among that LSA trounces two of the most challenging keyword queries: numerous words that have parallel connotations and expressions that encompass in excess of single meaning (Table 1).

Chi-square: The Chi-square (χ^2) (Liu *et al*., 2009) is a prominent feature selection technique that assesses the uniqueness individually by computing Chi-square data with reverence to the classes. It elucidates that the Chi-squared score, study the dependency flanked by the expressions and the class. If the expression is autonomous from the group, then it attain is equal to 0, other wise 1. An expression with a advanced Chi-squared score is more informative:

$$\chi^2(it,d) = DN\ (q(it,d)\ (q(\overline{it},\overline{d}) - q(it,\overline{d})\ q(\overline{it},d))^2$$

$$\chi^2(it) = avg_{j=1}^{n}[\chi^2(it,dj)]$$

Information gain: Information gain is a feature selection method that can shrink the amount of features by scheme the significance of every characteristic and grade the characteristics accordingly (Zhang *et al*., 2011). Next, we minimally make a decision to a threshold in the metric and remain the qualities with a measurement over it. It just remains those top ranking ones. In general, Information Gain opts for the features via scores. However, this method can be simpler than the earlier one. The fundamental initiative is that we simply have to calculate the score for each feature that it replicates in discrimination between classes and subsequently the characteristics are categorized according to this score and then just remain those top grading ones.

$$IG\ (\text{Information Gain})\ IG\ (x) =$$
$$-\sum_{i=1}^{j} Q(D_i)\log Q(D_i) +$$
$$Q(x)\sum_{i=1}^{j} Q(D_i|x)\log Q(D_i|x) +$$
$$Q(\overline{x})\sum_{i=1}^{j} Q(D_i|\overline{x})\log Q(D_i|\overline{x})$$
$$= K(\text{Sample}) - K(\text{Samples}|x)$$

Representation of ontology base documents and semantic web documents: In this segment, we concentrated on the

semantic, ontology techniques, language and the allied issues for documents classification. According to (Zhang *et al*., 2009) the arithmetical techniques, it is adequate for the data mining. Better categorization will be acted upon in allowing the semantic under observation. Ontology is a statistics model that embodies a set of concepts surrounded by a domain and the relationships between the concepts. It is used to locate the cause with reference to the substances surrounded by that domain. Ontology is the clear and conceptual model illustration of already elucidated the finite sets of terms and concepts, concerned in information management, information engineering and intelligent information absorption. The uniqueness of objects and creatures (individuals, instances) is a practical thing and linked (relations) with attribute which is used for the headings of the two thoughts or creatures. Ontology is alienated into three groups i.e., Natural Language Ontology (NLO) (Damljanovic *et al*., 2010), Domain Ontology (DO) (Jones *et al*., 2011) and Ontology Instance (OI) (Fensel *et al*., 2001). NLO subsists the connection linking the common lexical signs of reports based on normal language, DO exists the information in exacting domain and OI is mechanically produced web page performs similar to a point. Web Ontology Language (OWL) (Horrocks *et al*., 2003) is the ontology hold up language acquired from America DAPRA Agent Markup Language (DAML) based on ontology and implication of European Ontology Interchange Language (OIL) (Pulido *et al*., 2006). OWL alleges to be an addition in Resource Description Framework (RDF) (Ensel and Keller, 2012). In coherent to the logical statements it is not only depicting about the modules and properties but also affords the concepts of the namespace, import, cardinality association involving the modules and enumerated modules. Ontology encompasses its projection for handling semantically heterogeneity when extracting information from an assortment of text sources such as internet.

The mechanized Machine learning algorithms fabricates a classifier by educating the distinctiveness of the classifications from a group of categorized documents and consequently it used the classifier to categorize the documents into predefined categories. On the other hand, these machine learning methods have some drawbacks: with the intention of train classifier, human be required to accumulate large number of training text terms and the process is very conscientious. If the predefined categories distorted, these techniques must accumulate a new set of training text terms. The majority of these conventional techniques haven't measured the semantic relations flanked by words; therefore it is risky to growth the exactness of these classification methods (Ahlqvist, 2008). The apprehension of translatability, stuck between one natural language into a different natural language has dissimilar issues and this can be employed to categorize that machine understanding systems are facing problems. Such problems are discussed in the literature and some of these may be concentrated on if we have machine decipherable ontology and that's the reason to make this as a chief potential area for research. Through, the text mining process, ontology is able to be used to afford the expert, background knowledge about a domain. Recently, some of the investigation depicts the significance of the domain ontology in the text categorization method that open the mechanical categorization of inward news using hierarchical news ontology. It is based on the categorization on one hand and the client's profiles in contrast, the personalization engine of the system is capable to offer a modified paper to each user on to mobile reading device. A novel ontology-based mechanical categorization and grading technique is signified by Gauch *et al*. (2003) where Web documents are epitomized through a group of subjective expressions, categories are portrayed by ontology. In (Simon *et al*., 2006) that, the researchers have illustrated a new technique in the direction of mining ontology from the usual language. In which, the researchers measured a domain-special glossary for the telecommunication datas.

Techniques of machine learning: The documents can be classified into three methods, unverified, verified and semi verified methods. Many methods and algorithms are proposed freshly for clustering and classification of electronic documents. This fragment highlighted on the supervised classification techniques (Pal *et al*., 2002), new developments are highlighted some of the prospects and challenges using the current literature. The mechanical classification of documents is alienated into predefined categories has scrutinized as a dynamic consideration as the usage of internet rate has rapidly blown up. In the modern age, the assignment of mechanical text categorization have been widely investigated and speedy augmentation look as if in this field, together with the machine learning shift towards such as artificial neural network, genetic algorithm, support vector machine and naive bayes classification. Normally, for the automatic categorization of text, the techniques of supervised learning can be utilized wherever it is pre-defined classify tags are bestowed to the documents, a labeled documents training set can be proposed established on the likelihood.

Artificial neural network: Artificial neural networks are created from a great amount of elements with an input fan

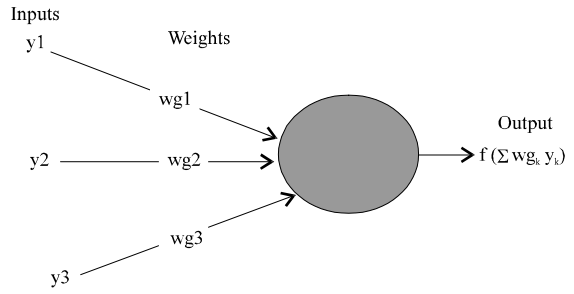

Fig. 2: Artificial neural network

into an order of magnitudes larger than in computational essentials of conventional architectures (Byvatov *et al*., 2003; Marinai *et al*., 2003). These elements, named as artificial neuron and these are interrelated into group using as a statistical model for information processing based on a association in approach to computation. The neural networks create their neuron sensitive to store items. It can be used for deformation open-minded piling up of a great amount of cases exemplified by high dimensional vectors.

There are various kinds of neural network methods have been accomplished for document classification tasks. Some of the researchers use the single-layer perception which restrains merely a key in layer and the productivity layer in consequence of its unfussiness of executing (Frean, 1990). However, inputs are supplied directly to the outputs via a sequence of weights. Thus it can be measured as the modest type of feed-forward network. The multi-layer observation which is more refined which consists of an key in layer, single or additional obscured layers and an output layer in its structure, moreover it is broadly executed for classification tasks (Marinai *et al*., 2005). Figure 2 represents the simple artificial neural network.

The ANN can get inputs y_k turned up through pre-synaptic connections, synaptic efficiency is reproduction of using the real weights wg_k and the reply of the neuron is a non linear function f of its subjective inputs. The productivity of neuron i for prototype q is P_{qi} where:

$$P_{qi}(net_i) = \frac{2}{1 + e^{-\alpha net_i}}$$

$$net_i = Wg_{bias} * bias + \Sigma_j p_{qj} Wg_{ij}$$

Neural network for document classification can fabricate positive results in complex domains and apposite for both separate and constant data. When the training is comparatively sluggish then there is fast in the testing and examining the results are complicated for users to understand than cultured systems (contrasting with Decision tree), Empirical Risk Minimization (ERM) composes ANN endeavor towards the reduce training error may lead to over fitting.

Support vector machine: Support Vector Machines (SVMs) are one of the main discriminative classification techniques which are generally documented to be further perfect. The SVM classification method (Zhang *et al*., 2008) is depended on the structural risk minimization code from computational learning theory. The proposal of this principle is to find a suggestion to assurance the lowest true error. In addition, the SVM is well-substantiated that unwraps for the theoretical understanding and investigation (Amari and Wu, 1999).

The SVM requires mutually optimistic and pessimistic guidance set which are exceptional for other classification methods. The positive and negative training sets are required for the SVM to hunt for the assessment surface that detaches the affirmative from the unconstructive data in the dimensional space, therefore it is known as hyper plane. The document authorities which are contiguous towards the assessment plane are termed as support vector. The presentation of the SVM categorization ruins the unaffected where the credentials that execute is not fit in to the sustain vectors are aloof from the set of training data.

Genetic algorithm: Genetic algorithm is aimed to find optimum feature parameters by means of the mechanisms of genetic advancement and endurance of the strongest in natural assortment (Ishibuchi *et al*., 1994). Biological algorithms construct it possible to eradicate the ambiguous judgments in the algorithms and advance the accurateness of document classification. And also, this is an adaptive prospect global optimization algorithm which replicated in a natural setting of organic and genetic advancement and it is mostly executed for their unfussiness and potency. At present, numerous researchers have deployed this technique for the enhancement of the text classification process. The different authors/(Tan *et al*., 2002) recognized the inherent algorithm to text classification and used to generate and optimize the user pattern and also conventionalized the replicated annealing to progress the shortcomings of genetic algorithm. In the experimental analysis, they depict that the developed method is realistic and efficient for text categorization.

Naive bayes classification: Naive Bayes classifier is an uncomplicated possible categorizer based on relating Bayes' Theorem with strong independence assumptions (Pop, 2006). The more eloquent term for the fundamental

possibility model would be termed as independent feature model. This independence hypothesis of features make the characteristics order which is inappropriate and in view of that portray about one feature does not influence new features in categorization tasks. These kind of assumptions formulate the computation of Bayesian classification approach more competent but this assumption brutally limits its applicability. Depending on the specific character of the possibility model, the naïve Bayes classifiers could be proficient very professionally by necessitating for a comparatively little amount of guiding data to fairly accurate the parameters needed for classification.

With account of its apparently over-simplified assumptions, the naïve Bayes classifiers frequently work much better than one may anticipate in many intricate real-world situations. The naïve Bayes classifiers have been reported to act upon astoundingly sound for many reliable world classification applications under some explicit circumstances (Pop, 2006).

Naive Bayes research well on numbered and textual information, easy to implement and computation in comparing with other algorithms, the data of real-world is violated by assuming conditional independence on the other side and perform very poorly when features are extremely interrelated and executed in not believing the amount of word rates (Ren *et al*., 2009; Zhang *et al*., 2009).

Proposed technique

Flowchart for the proposed technique: Figure 3 shows the flow chart for our proposed technique.

Step-by-step procedure for the proposed technique: We projected a new hybrid feature selection method by merging information gain assessment and artificial neural network. An information gain algorithm squeezes the number of attributes based on the SU measure, In information gain each attributes are contrasted with pair wise to find the resemblance and the attributes are compared to class attribute to find the quantity of contribution it provides to the class value, based on these the attributes are removed. The selected attributes from the IG algorithm is fed into artificial neural network for further reduction. Artificial neural network calculates the conditional probability for each attribute and the attribute which has the highest conditional probability is selected. Both the algorithms IG and ANN work on the conditional probability measure:

Input: V (W1; W2; :::; WD; M)//a training data set δ//a predefined threshold
Output: Vbest {Abest(highest IG)}//an optimal subset

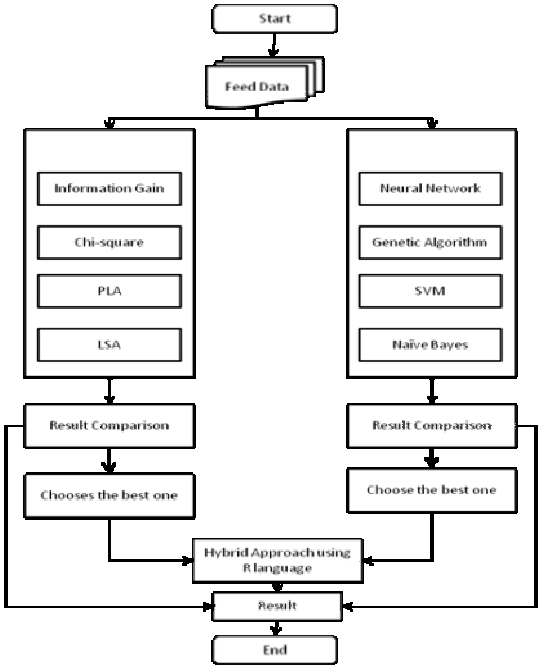

Fig. 3: Flowchart for the proposed technique

Step 1: start
Step 2: for t= 1 to D do begin
Step 3: calculate VIt;M for Wt
Step 4: if (Vit, M≥ δ)
Step 5: append Wt to V'list
Step 6: finish
Step 7: order V'list in descending Vit, M value
Step 8: We = getFirstElement(V'list)
Step 9: do start
Step 10: Wf = getNextElement(V'list, We)
Step 11: if (Wf<>NULL)
Step 12: do start
Step 13: W'f = Wf
Step 14: if (VIe,f , Vif, M)
Step 15: remove Wf from V'list
Step 16: Wf = getNextElement(V'list, W 'f)
Step 17: else Wf = getNextElement(V'list , Wf)
Step 18: finish until (Wf == NULL)
Step 19: We = getNextElement(V'list, We)
Step 20 finish until (We == NULL)
Step 21 Vbest = V'list
Step 22 Vbest={A1,A2,..AC}
Step 23 for j=1 to C begin
Step 24 for W=t+1 to C begin
Step 25 $E[M_D/(At,A_W)]=E[(At,A_W)/M_D]*E(M_D)$
Step 26 $E[M/(At,A_W)]=E[M1/(At,A_W)]+E[M2/(At,A_W)]+\ldots$
....+E[MC/(At,A_W)]
Step 27: If(E[M/At,A_W)]>Ù)
Step 28: {
Step 29: if ((E[M/At]>(E[M/A_W])
Step 30: Remove A_W from Vbest
Step 31: Vbest=IG(A)
Step 32: Else
Step 33: Remove At from Vbest
Step 34: Vbest=IG(A)
Step 35}
Step 36: finish

Table 2: Various datasets Characteristics

Datasets	No. of documents	No. of terms	No. of categories/classes
CNAE-9	1080	856	9
Books	50	3360	2
Computers	3232	100	2
Cook Ware	2000	1493	2
Flipkart	50	2370	2
Gender	194	4466	2
Hotel	279	3170	3
MyMail	50	3358	2
NB World	400	3043	2
NYdtm	5572	6631	2
Prosncons	50	3300	2
Reutiers	64	3723	2
SpamHam	3104	5587	27

Experiment: The experimental setup is explained in this section. It wraps up facts in relation to the datasets so as to deploy and dissimilar preprocessing methods that were practiced. The software device and enclosed that are utilized, Hardware and software particulars of the machine, on which the research was performed in.

Information of the datasets: The following Table 2 show the information about the datasets which are used for our experimental setup.

The software and hardware used are as follows: Processor: Intel Core i3 CPU M350 @ 2.27 GHz RAM: 3.00 GB Operating classification: Windows 7 Ultimate R: Version 2.15.3 (Zhao, 2012).

MATERIALS AND METHODS

For the reproducibility of the results, our experiment complies the following basic steps are described here.

Step 1: The Text credentials are exposed of space and punctuation.

Step 2: Numbers and stop words are removed.

Step 3: Stemming and lowercasing are applied.

Step 4: The expression document matrix is organized on the methodical document. The evaluating method that has been utilized is the tf-idf.

Step 5: The expression document matrix is divided into two subsets, 70% of the expression manuscript matrix is used for working out and the rest 30% is used for taxing categorization exactness.

Step 6: The results are obtained for the given feature selection techniques like PCA (Principle Component Analysis), Information Gain (IG), Chi-square and Latent Semantic Analysis (LSA) and compared.

Step 7: The results are obtained for the given classification techniques artificial Neural Network, Genetic algorithm, SVM and Naive Bayesian and it is compared.

Step 8: The results is obtained for our proposed algorithm IG-ANN.

Step 9: We also compare execution time taken by IG-ANN with other approaches of feature selection and classification.

RESULTS AND DISCUSSION

We have employed the categorization accurateness which is an evaluation of how well a document is divided into its apposite category. It is simply the percentage of No. correctly classified documents/No. total documents. All the categorization correctness has been figured on taxing dataset. We submit the subsequent assessment and association, correspondingly.

Table 3 is represent the feature reduction using feature selection techniques like Information Gain (IG), Chi-square, PSA and LSA for various datasets. And, the Table 4 is used to display the feature reduction rate of the various techniques for given datasets. The classification accuracy of the given datasets using (a) artificial neural network, (b) Genetic algorithm, (c) SVM and (d) Naive Bayes are calculated and it is represented by Table 5. Table 6 gives the classification accuracy rate of the ANN, IG and IG-ANN. Figure 4 shows the chart which represents the classification accuracy rate of the ANN, IG and IG-ANN for the given various datasets. The feature space reduction using IG, Chi-Square, PCA, LSA and Proposed method IG-ANN is represented by Table 7.

Table 8 compares the classification accuracy of IG-ANN with ANN, Genetic algorithm, SVM and NB. The summarization of % improvement of the classification accuracy and % reduction of feature set over all the datasets using IG-ANN are given by the Table 9.

Table 10 and 11 are representing the execution time and classification accuracy of IG-ANN and it is compared with Filtered Attribute method and Best First search wrapper method. The outcomes of the classifiers can be balanced by utilizing Friedman's nonparametric investigation. Regarding the principal representation, there is no hypothesis by the specified inclination of Friedman test. $[sj_k]_{m*}$ is replaced for $[yj_k]$ for the given data where the rank is represented by sjk, an typical value is to be substituted for sjk in the case of a tie:

$$S_{jk} = \frac{1}{m}\sum_{j=i}^{m} s_{jk}$$

Table 3: Feature reduction using feature selection methods for various datasets

Datasets	No. of features	Using IG	Using χ^2	Using PCA	Using LSA
Books	3300	1223	1353	1425	1652
CNAE-9	856	425	586	654	632
Computers	3358	1185	1245	1321	1564
CookWare	2370	1024	1178	1324	1487
DBWorld	3723	1524	1687	1785	1879
Flipkart	3043	1825	1869	1898	2005
Gender	100	50	64	85	78
Hotel	3360	1324	1365	1452	1668
MyMail	4466	1852	1963	2014	2147
NYdtm	5587	2154	2653	2748	2898
Prosncons	1493	685	740	811	862
Reuters	3170	1439	1542	1687	1787
SpamHam	6631	3245	3541	3587	3698

Table 4: Feature selection rate for various datasets using feature selection techniques

	Feature selection rate (%)			
Database	IG	χ^2	PCA	LSA
Books	96	85	82	79
CNAE-9	89	83	80	78
Computers	95	90	92	89
CookWare	73	71	60	65
DBWorld	86	80	78	70
Flipkart	95	90	87	89
Gender	82	75	79	74
Hotel	90	82	83	79
MyMail	98	95	92	94
NYdtm	73	68	65	63
Prosncons	78	71	69	68
Reuters	94	85	84	82
SpamHam	83	87	85	80

Table 5: Classification accuracy for various datasets using classification techniques

	Classification accuracy (%)			
Dataset	ANN	GA	SVM	NB
Books	78	65	62	58
CNAE-9	58	50	45	38
Computers	85	82	79	80
CookWare	62	54	59	56
DBWorld	52	53	54	62
Flipkart	80	75	71	72
Gender	87	82	72	79
Hotel	83	80	75	79
MyMail	90	87	86	88
NYdtm	60	55	52	51
Prosncons	69	63	51	45
Reuters	88	82	86	81
SpamHam	89	78	82	79

Table 6: Accuracy in classification rate of ANN during the three phases of experiments

Datasets	ANN	IG	Using IG-ANN
Books	78	96	97
CNAE-9	58	89	92
Computers	85	95	96
CookWare	62	73	85
DBWorld	52	86	89
Flipkart	80	95	98
Gender	87	82	90
Hotel	83	90	91
MyMail	90	98	99
NYdtm	60	73	80
Prosncons	69	78	82
Reuters	88	94	96
SpamHam	89	83	93

Table 7: Feature reduction using IG-ANN with other techniques

	Feature reduction accuracy rate					
Datasets	Total feature	IG	χ^2	PCA	LSA	IG-ANN
Books	3300	45	47	52	54	23
CNAE-9	856	10	12	14	20	6
Computers	3358	52	54	59	60	21
CookWare	2370	115	126	124	135	18
DBWorld	3723	88	116	145	179	15
Flipkart	3043	53	57	63	67	19
Gender	100	10	16	20	21	4
Hotel	3360	105	120	121	132	28
MyMail	4466	120	128	136	149	33
NYdtm	5587	152	168	189	187	13
Prosncons	1493	36	45	47	50	11
Reuters	3170	40	48	54	55	5
SpamHam	6631	96	100	112	145	17

Table 8: Classification accuracy of IG-ANN with other classification methods

	Classification accuracy (%)				
Datasets	ANN	GA	SVM	NB	IG-ANN
Books	78	65	62	58	85
CNAE-9	58	50	45	38	65
Computers	85	82	79	80	88
CookWare	62	54	59	56	72
DBWorld	52	53	54	62	70
Flipkart	80	75	71	72	86
Gender	87	82	72	79	92
Hotel	83	80	75	79	87
MyMail	90	87	86	88	95
NYdtm	60	55	52	51	69
Prosncons	69	63	51	45	73
Reuters	88	82	86	81	93
SpamHam	89	78	82	79	95

Table 9: Summary of the classification accuracy and Feature reduction accuracy improvement

Datasets	Feature reduction improvement (%)	Classification accuracy improvement (%)
Books	98.00	99.00
CNAE-9	98.90	99.50
Computers	97.90	98.90
CookWare	98.20	98.70
DBWorld	99.10	99.60
Flipkart	99.90	99.90
Gender	95.50	96.30
Hotel	98.20	98.30
MyMail	97.80	97.90
NYdtm	99.00	99.00
Prosncons	97.00	98.00
Reuters	99.90	99.90
SpamHam	98.00	97.00

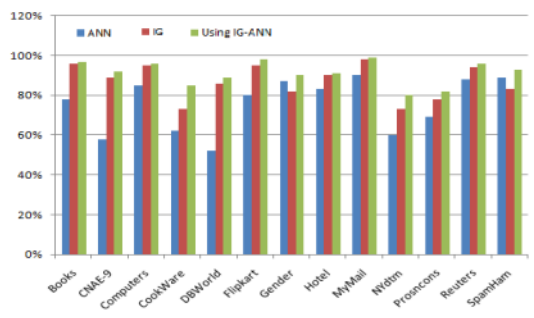

Fig. 4: Classification accuracy rate of ANN, IG and IG-ANN for various given database

Table 10: The classification accuracy and execution time of IG-ANN with best first search

	Best first search		Proposed IG-ANN	
Datasets	Execution time (min)	Classification accuracy (%)	Execution time (min)	Classification accuracy (%)
Books	56.80	78.00	4.50	97
CNAE-9	69.50	65.00	1.60	92
Computers	102.1	79.00	0.98	96
CookWare	72.30	70.00	2.36	85
DBWorld	71.50	62.50	3.54	89
Flipkart	120.6	80.50	1.85	98
Gender	45.00	59.00	0.65	90
Hotel	112.0	63.00	4.27	91
MyMail	132.0	68.00	2.87	99
NYdtm	66.00	72.00	4.89	80
Prosncons	58.00	63.30	1.47	82
Reuters	119.0	75.00	3.89	96
SpamHam	105.0	81.00	4.74	93

Table 11: Execution time and classification accuracy of IG-ANN with Filtered Attribute method

	Filtered attribute method		Proposed IG-ANN	
Datasets	Execution time (min)	Classification accuracy (%)	Execution time (min)	Classification accuracy (%)
Books	68.8	72	4.5	97
CNAE-9	75.5	69	1.6	92
Computers	152	85	0.98	96
CookWare	78.9	63	2.36	85
DBWorld	59.6	59	3.54	89
Flipkart	123	82	1.85	98
Gender	69	56	0.65	90
Hotel	83.65	45	4.27	91
MyMail	119	70	2.87	99
NYdtm	45	49	4.89	80
Prosncons	107	58	1.47	82
Reuters	145	66	3.89	96
SpamHam	89	71	4.74	93

Table 12: Comparing mean rank of proposed method with other classifiers

Algorithms	Mean rank
IG-ANN	1.56
ANN	2.25
GA	3.45
SVM	4.87
NB	4.65

Table 12 summarizes the mean ranks and it is compared with the other four methods to prove that our proposed method is better than others.

Analysis: From Table 2, we can conclude that the information gain feature selection methods and its accuracy rate gives the better result than the other methods like chi-square, Principle Component Analysis (PCA) and Latent Semantic Analysis (LSA). The classification technique of Artificial Neural Network produces the better classification accuracy than the other classification techniques like genetic algorithm, support vector machine and Naive Bayes in Table 3. The proposed method of IG-ANN gives the good classification accuracy rate when it is compared with IG and ANN and it is given by Table 5. From the Table 6 and 7, IG-ANN produces the least number of feature space and better classification accuracy than other methods. Using IG-ANN, the feature reduction and classification accuracy are improved than using previous methods. From Table 9, we can conclude than the proposed method of IG-ANN gives the better execution time and classification accuracy when it is compared with best first search wrapper method and filtered attribute method. In Table 10, the mean rank of IG-ANN is least than the other classifiers using Friedman test.

CONCLUSION

Our earlier researches and findings of other researchers illustrate ANN to be a substandard classifier particularly for text categorization. In this study, we have projected an innovative two-step characteristic assortment algorithm which can be implemented in conjunction with ANN to advance the presentation. In addition, we have assessed our algorithm IG-ANN over thirteen datasets and tested the wide-ranging assessments with other categories along with other characteristic assortment techniques like to chi-square, PCA and so forth. The detailed synopsis of our results are presented below.

IG-ANN improves the classification accuracy than the other classification techniques like artificial neural network, genetic algorithm, support vector machine and naïve bayes. IG-ANN reduces the features of the dataset than the other feature selection methods like information gain, Chi-square, PCA and LSA and it produces better classification accuracy than the others. The IG-ANN improves the performances of the feature reduction and classification accuracy. And it takes less execution time and more classification accuracy when it is compared with both filtered and wrapper methods.

So, from this study, we can conclude that the IG-ANN will improve the performance of text classification and it make this simple to implement intuitive classifiers suitable for the task of text classification.

REFERENCES

Ahlqvist, O., 2008. Extending post-classification change detection using semantic similarity metrics to overcome class heterogeneity: A study of 1992 and 2001 US national land cover database changes. Remote Sens. Environ., 112: 1226-1241.

Amari, S. and S. Wu, 1999. Improving support vector machine classifiers by modifying kernel functions. Neural Networks, 12: 783-792.

Belkin, M. and P. Niyogi, 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput., 15: 1373-1396.

Berry, M.W., S.T. Dumais and G.W. O'Brien, 1995. Using linear algebra for intelligent information retrieval. SIAM. Rev., 37: 573-595.

Bizer, C., J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak and S. Hellmann, 2009. DBpedia-A crystallization point for the web of data. Web Semant.: Sci. Serv. Agents World Wide Web, 7: 154-165.

Byvatov, E., U. Fechner, J. Sadowski and G. Schneider, 2003. Comparison of support vector machine and artificial neural network systems for drug nondrug classification. J. Chem. Inf. Comput. Sci., 43: 1882-1889.

Caldas, C.H. and L. Soibelman, 2003. Automating hierarchical document classification for construction management information systems. Autom. Constr., 12: 395-406.

Cohen, A.M. and W.R. Hersh, 2005. A survey of current work in biomedical text mining. Briefings Bioinf., 6: 57-71.

Damljanovic, D., M. Agatonovic and H. Cunningham, 2010. Combining Syntactic Analysis and Ontology-Based Lookup Through the User Interaction. In: The Semantic Web: Research and Applications. Lora, A., G. Antoniou, E. Hyvonen, A.T. Teije and S. Heiner *et al.* (Eds.). Springer Berlin Heidelberg, Berlin, Germany, ISBN: 978-3-642-13485-2, pp: 106-120.

Dhillon, I.S. and D.S. Modha, 2001. Concept decompositions for large sparse text data using clustering. Mach. Learn., 42: 143-175.

Ensel, C. and A. Keller, 2012. An approach for managing service dependencies with xml and the resource description framework. J. Netw. Syst. Manag., 10: 147-170.

Fensel, D., I. Horrocks, V.F. Harmelen, D. McGuinness and P.P.F. Schneider, 2001. OIL: Ontology infrastructure to enable the Semantic Web. IEEE. Intell. Syst., 16: 38-45.

Frean, M., 1990. The upstart algorithm: A method for constructing and training feedforward neural networks. Neural Comput., 2: 198-209.

Gauch, S., J. Chaffee and A. Pretschner, 2003. Ontology-based personalized search and browsing. Web Intel. Agent Syst. Int. J., 1: 219-234.

Horrocks, I., P.F. Patel-Schneider and F.V. Harmelen, 2003. From SHIQ and RDF to OWL: The making of a web ontology language. J. Web Semantics, 1: 7-26.

Ishibuchi, H., K. Nozaki, N. Yamamoto and H. Tanaka, 1994. Construction of fuzzy classification systems with rectangular fuzzy rules using genetic algorithms. Fuzzy Sets Syst., 65: 237-253.

Jones, M.V., N. Coviello and Y.K. Tang, 2011. International entrepreneurship research (1989-2009): A domain ontology and thematic analysis. J. Bus. Venturing, 26: 632-659.

Kalousis, A., J. Prados and M. Hilario, 2007. Stability of feature selection algorithms: A study on high-dimensional spaces. Knowl. Inf. Syst., 12: 95-116.

Kambhatla, N. and T.K. Leen, 1997. Dimension reduction by local principal component analysis. Neural Comput., 9: 1493-1516.

Kohonen, T., S. Kaski, K. Lagus, J. Salojarvi and J. Honkela *et al.*, 2000. Self organization of a massive document collection. IEEE. Trans. Neural Netw., 11: 574-585.

Krallinger, M., A. Valencia and L. Hirschman, 2008. Linking genes to literature: Text mining, information extraction and retrieval applications for biology. Genome Biol., 9: 1-14.

Liu, L., J. Kang, J. Yu and Z. Wang, 2005. A comparative study on unsupervised feature selection methods for text clustering. Proceedings of the 2005 International Conference on Natural Language Processing and Knowledge Engineering, October 30-November 1, 2005, IEEE, Beijing, China, ISBN: 0-7803-9361-9, pp: 597-601.

Liu, Y., H.T. Loh and A. Sun, 2009. Imbalanced text classification: A term weighting approach. Expert Syst. Appl., 36: 690-701.

Mahgoub, H., D. Rosner, N. Ismail and F. Torkey, 2008. A text mining technique using association rules extraction. Int. J. Comput. Ontell., 4: 21-28.

Marinai, S., M. Gori and G. Soda, 2005. Artificial neural networks for document analysis and recognition. IEEE Trans. Pattern Anal. Mach. Intell., 27: 23-35.

Matsuo, Y. and M. Ishizuka, 2004. Keyword extraction from a single document using word co-occurrence statistical information. Int. J. Artif. Intell. Tools, 13: 157-169.

Mittermayer, M.A., 2004. Forecasting intraday stock price trends with text mining techniques. Proceedings of the 37th Annual Hawaii International Conference on System Sciences, January 5-8, 2004, IEEE, Bern, Switzerland, ISBN: 0-7695-2056-1, pp: 1-10.

Pal, S.K., V. Talwar and P. Mitra, 2002. Web mining in soft computing framework: Relevance, state of the art and future directions. IEEE. Trans. Neural Netw., 13: 1163-1177.

Peng, Y., G. Kou, Y. Shi and Z. Chen, 2008. A descriptive framework for the field of data mining and knowledge discovery. Int. J. Inf. Technol. Decis. Making, 7: 639-682.

Pop, I., 2006. An approach of the Naive Bayes classifier for the document classification. Gen. Math., 14: 135-138.

Pulido, J.R.G., M.A.G. Ruiz, R. Herrera, E. Cabello and S. Legrand *et al*., 2006. Ontology languages for the semantic web: A never completely updated review. Knowl. Based Syst., 19: 489-497.

Rahm, E. and H.H. Do, 2000. Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23: 1-11.

Ren, J., S.D. Lee, X. Chen, B. Kao and R. Cheng *et al*., 2009. Naive bayes classification of uncertain data. Proceedings of the 2009 9th IEEE International Conference on Data Mining, December 6-9, 2009, IEEE, Hong Kong, China, ISBN: 978-1-4244-5242-2, pp: 944-949.

Simon, J., D.M. Santos, J. Fielding and B. Smith, 2006. Formal ontology for natural language processing and the integration of biomedical databases. Int. J. Med. Inf., 75: 224-231.

Sokolova, M. and G. Lapalme, 2009. A systematic analysis of performance measures for classification tasks. Inform. Process. Manage., 45: 427-437.

Soon, W.M., H.T. Ng and D.C.Y. Lim, 2001. A machine learning approach to coreference resolution of noun phrases. Comput. Ling., 27: 521-544.

Tan, C.M., Y.F. Wang and C.D. Lee, 2002. The use of bigrams to enhance text categorization. Inf. Process. Manage., 38: 529-546.

Yeh, J.Y., H.R. Ke, W.P. Yang and I.H. Meng, 2005. Text summarization using a trainable summarizer and latent semantic analysis. Inform. Process. Manage., 41: 75-95.

Zhang, M.L., J.M. Pena and V. Robles, 2009. Feature selection for multi-label naive Bayes classification. Inform. Sci., 179: 3218-3229.

Zhang, W., T. Yoshida and X. Tang, 2011. A comparative study of TF IDF, LSI and multi-words for text classification. Expert Syst. Appl., 38: 2758-2765.

Zhang, W., T. Yoshida, and X. Tang, 2008. Text classification based on multi-word with support vector machine. Knowledge-Based Syst., 21: 879-886.

Zhao, Y., 2012. R and Data Mining: Examples and Case Studies. Academic Press, USA., ISBN: 978-0-123-96963-7, Pages: 233.