RESEARCH ARTICLE OPEN ACCESS

Text Categorization Optimization By A Hybrid Approach Using Multiple Feature Selection And Feature Extraction Methods

K. Rajeswari*, Sneha Nakil**, Neha Patil**, Sebatini Pereira**, Neha Ramdasi**

*Assistant Professor, Department of Computer Engineering, Pimpri-Chinchwad College of Engineering, Pune-411044

**Student-BE, Department of Computer Engineering, Pimpri-Chinchwad College of Engineering, Pune-411044

ABSTRACT

Text categorization is the task of automatically sorting a set of documents into categories from a pre-defined set. This means it assigns predefines categories to free-text documents. In this paper we are proposing a unique two stage feature selection method for text categorization by using information gain, principle component analysis and genetic algorithm. In the methodology, every term inside the document is ranked depending on their importance for classification using the information gain (IG) methodology. In the second stage, genetic algorithm (GA) and principal component analysis (PCA) feature selection and feature extraction methods are applied individually to the terms which are ranked in decreasing order of importance, and a dimension reduction is carried out. Therefore, throughout the text categorization terms of less importance are ignored, and feature selection and extraction methods are applied to the terms of highest importance; so the computational time and complexity of categorization is reduced. To analyze the dimension reduction in our proposed model, experiments area unit conducted using the k-nearest neighbor(KNN) and C4.5 decision tree algorithmic rule on selected data set.

***Keywords* -** Information Gain, Principal Component Analysis, Genetic Algorithm, K-Nearest Neighbor classifier, C4.5 tree classifier.

I. INTRODUCTION

Text categorization is one of the challenging research topics due to the necessity to organize and categorize growing number of electronic documents worldwide. So far, text classification has been successfully applied to various domains such as topic detection [15], spam e-mail filtering [16], SMS spam filtering [17], author identification[18], web page classification [19]and sentiment analysis. A conventional text categorization framework consists of preprocessing, feature extraction, feature selection, and classification stages. It solves the problem of assigning text content to predefined categories. Text categorization has vital importance in applications used in the real world. For example, news stories are typically organized by subject categories (*topics*) or geographical codes; academic papers are often classified by technical domains and sub-domains; patient reports in health-care organizations are often indexed from multiple aspects, using taxonomies of disease categories, types of surgical procedures, insurance reimbursement codes and so on.

II. RELATED WORK

Text categorization is defined as assigning pre-defined categories to a given set of documents based on v classification patterns. Although many information retrieval applications such as filtering and searching for relevant information can benefit from text categorization research, a major problem of text categorization is the high dimensionality of the feature space due to a large number of terms. Some feature extractions have been successfully used in text categorization such as principal component analysis, latent semantic indexing, clustering methods etc. Among the many methods that are used for feature extraction, PCA has attracted a lot of attention. PCA is a statistical technique for reduction of dimensionality that aims at minimizing loss in variance in original data. Text categorization is the task of classifying a document into predefined categories based on the contents of the document. In recent years, more and more methods have been applied to the text categorization tasks based on statistical theories and machine learning, such as KNN, Naive bayes, Rocchio, Decision tree Support vector machine.

There are several techniques have been proposed for the text categorization such as:

2.1 Multi-label text categorization based on a new linear classifier learning method and a category-sensitive refinement method [2].

A new approach for dealing with multi-label text categorization based on a new linear classifier learning methodology and a category-sensitive refinement methodology. Here used is a replacement weighted classification technique to construct a multi-label linear classifier. We tend to use the degrees of similarity between classes to regulate the connection voluminous classes with relevancy a testing document. The testing document are often properly classified into multiple classes by employing a predefined threshold worth.

2.2 Text feature selection using ant colony optimization [1].

Feature selection and feature extraction are the most vital steps in classification systems. Feature selection is usually used to reduce spatiality of datasets with tens or many thousands of features which might be not possible to process more. One of the issues in which feature selection is crucial is text categorization. A significant drawback of text categorization is that the high spatiality of the feature space; thus, feature selection is the most vital step in text categorization. At the present there are several ways to affect text feature choice. To boost the performance of text categorization, we tend to gift a unique feature choice algorithmic program that's supported ant colony optimization. Ant colony optimization algorithm is impressed by observation on real ants in their search for the shortest paths to food sources. Planned algorithmic program is well enforced and since of use of a straightforward classifier there in, its process quality is incredibly low.

III. METHODOLOGY

A two stage feature selection and feature extraction is used to reduce the high dimensionality of a feature space composed of a large range of terms, remove redundant and irrelevant features from the feature space and thereby decrease the computational complexity of the machine learning algorithms used in the text categorization and increase performances thereof. In the first stage, every term within the text is ranked depending on their importance for the classification in decreasing order using the Information gain method. Therefore, terms of high importance are assigned to the first ranks and terms of less importance are assigned to the subsequent ranks. Within the second stage, the PCA method selected for feature selection and the GA method selected for feature extraction are applied separately to the terms of highest importance, in accordance with IG strategies, and a dimension reduction is carried out. Thus, during text categorization, terms of less importance are ignored. The terms with highest importance undergo feature selection and feature extraction method. And therefore the computational time and complexity of the category are reduced. Datasets assortment that can be used for this could be Classic4 data set for analyzing the dimensionality reduction.

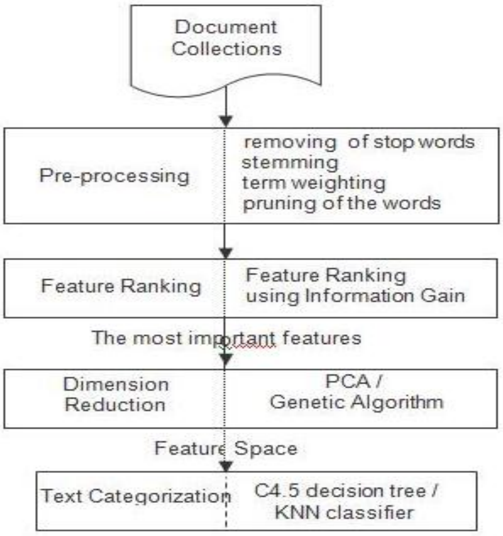

Fig 1 General architecture for two stage text categorization.

The proposed system is divided into following five modules:

3.1 Preprocessing:

Preprocessing is one of the key components in a typical text categorization framework [3]. Preprocessing impacts text categorization in various aspects such as categorization accuracy, text domain and dimension reduction and also promotes efficiency of text categorization [20].This process includes four parts:

3.1.1 Removal of stop-words: Words such as conjunctions and pronouns that are not related to the concept of the text are called stop-words. This process involves removing certain common words such as `a', `an', `the', etc., that occur commonly in all documents.

3.1.2 Stemming: The stemming process gives us the root form of all the terms. eg. For "*stemming*" the root word is "*stem*". Thus stemming process removes suffixes or prefixes like *ing, ed, ied, tion, ly , pre-* etc. Porters algorithm is the most commonly used algorithm for stemming purpose in textual data.

3.1.3 Term weighting: After the words are transformed into terms, each term is assigned a weight based on its frequency of appearance in the document. This process is called term weighting. The basic formula used for term weighting include term frequency ($tf_{i.}$) and inverse document frequency (idf_i) as follows :

$W_i = tf_{i.}idf_i$ [27]

Where, tf_i=term frequency

idf_i = Inverse Document Frequency

3.1.4 Pruning : The pruning process basically filters less frequent features in a document collection. For example words appearing just twice thrice in a document are not useful further hence are eliminated by setting a appropriate threshold value.

3.2 Information gain:

Information gain is one of the popular approaches employed as a term importance criterion in the text document data.

$$IG(t) = -\sum_{i=1}^{|C|} P(c_i)\log P(c_i) + P(t)\sum_{i=1}^{|C|} P(c_i|t)\log P(c_i|t) + P(\bar{t}) \times \sum_{i=1}^{|C|} P(c_i|\bar{t})\log P(c_i|\bar{t}),$$

where c_i represents the ith category, $P(c_i)$ is the probability of the ith category, P(t) and P(t) are the probabilities that the term t appears or not in the documents, respectively, $P(c_i|t)$ is the conditional probability of the ith category given that term t appeared, and $P(c_i|\bar{t})$ is the conditional probability of the ith category given that term t does not appeared[27].

3.3 Principal component analysis:

PCA is a statistical technique used for extracting information from a multi-variety dataset. This process is performed via having principal components of original variables with linear combinations identified. While the original dataset with the maximum variability is represented with the first principal component, the dataset from the remaining with the maximum variability is represented with the second principal component [28,29]. The process goes on like this one after other as such, with the remaining dataset having the maximum variability being represented with the next principal component. Here m represents the number of all principal components, and p represents the number of the significant principal components among all principal components. Further p may be defined as the number of those principal components of the m dimensional dataset with the highest variance values. It is clear therein that $p \leq m$. Therefore, PCA may be defined as a data-reducing technique. In other words, PCA is a technique used for producing the lower-dimensional version of the original dataset.

3.4 Genetic algorithm:

Genetic Algorithms [23,24] are based on the principles of biological inheritance and evolution. Every solution which is potential is called an individual (i.e. a chromosome) in a population. GAs work iteratively, applying the genetic operations such as selection, crossover and mutation to a population of individuals, aimed at creating better adapted individuals in subsequent generations. For each category of problems solved by a GA, a fitness function must be provided. This choice is crucial to maximize the GA's performance. The fitness function assigns a fitness value for each individual called the individual score, and it is well-known to play an essential role in the genetic evolution [25]. This is because this score is used in the selection processes of parents for crossover and of survivals for each next generation. Thus, the highest probabilities to reproduce and survive must be given to the best adapted individuals, that is, the ones who provide the best features for an accurate query process. Due to its importance to GA, the fitness function must be tailored to the problem at hand. Thus, a fitness function should model the solution space in such a way that, when solving a maximization problem, better solutions receive higher scores. Associated with the characteristics of exploitation and exploration, GAs can efficiently deal with large search spaces, and hence are less prone to get stuck into a local optimum solution when compared to other algorithms. This derives from the GAs ability to handle multiple concurrent solutions (individuals) in the search space and apply probabilistic genetic operators [23,24]. Genetic algorithm has been successfully used as an optimization technique [4][5].The various stages in genetic algorithm [26] are as follows :

3.4.1 Individual Encoding *:* The candidate feature set can be represented by a binary string called chromosome. In the chromosome the ith bit represents the presence of the ith feature. If the value of the gene if coded in binary system as 1, it means that the corresponding feature is selected and vice versa. The length of the chromosome is equal the number of features.

3.4.2 Fitness Function : Fitness function is used to decide which individuals are to optimum solution. Every individual has its own fitness value. A higher value of fitness means that the individual is more appropriate as a problem solution, on the other hand, a lower value of fitness means that the individual is less appropriate as a problem solution.

3.4.3 Selection : The object of the selection process is to choose the individuals of the next generation according to the selected fitness function and selection method among the existing population. In the selection process, the transfer possibility of the fittest individuals chromosome to the next generation is higher than others. The decision of the individuals characteristic which will be transferred to the next generation is based on the values evaluated from the fitness function and shows the quality of individual . The roulette wheel selection method is generally used for selection purpose.

3.4.4 Crossover : In this phase an individual is selected for mating . Forming of new generation by mating is called crossover. The method is of forming new individuals from two chromosome. Different type of crossover procedures are there like single point, double point and multile point crossover.

3.4.5 Mutation : To increase the variety of chromosomes the which are applied on crossover, process mutation process can be applied. Mutation introduces local variations to the individuals for searching different solution spaces and keeps the diversity of the population. In our study, the number of chromosomes that will be mutated is determined according to the mutation rate and their values are changed from 1 to 0 or 0 to 1 respectively.

3.5 Text categorization methods*:*

For categorization of text documents effectively following two methods have been considered.

3.5.1 KNN classifier**:**

KNN is typical methods of example-based classifiers. They have been called lazy learners due to the fact that they do not learn before they start to classify a document. The KNN method classifies a testing document by the top k nearest training documents in the vector space. The difference between traditional linear classifiers and KNN is that KNN does not divide the document space linearly and does not have the drawback of Rocchio's method. A number of experiments have shown that the KNN method is quite effective [21][22].The KNN algorithm is a well-known instance-based approach that has been widely applied to text categorization due to its simplicity and accuracy. To categorize an unknown document, the KNN classifier ranks the document's neighbors among the training documents and uses the class labels of the k most similar neighbors. Similarity between two documents may be measured by the Euclidean distance, cosine measure, etc. The similarity score of each nearest neighbor document to the test document is used as the weight of the classes of the neighbor document. If a specific category is shared by more than one of the k-nearest neighbors, then the sum of the similarity scores of those neighbors is obtained from the weight of that particular shared category.

3.5.2 C4.5 decision tree classifier**:**

C4.5 algorithm is an improved version of ID3. This algorithm uses a value called Gain Ratio as a splitting criterion, unlike ID3 algorithm where gain is used for splitting criteria in tree growth phase. Hence C4.5 is an evolution of ID3 [8]. This algorithm handles both continuous and discrete attributes- In order to handle continuous attributes, in C4.5 a threshold is created and then the list gets split into those whose attribute value is above the threshold and those that are less than or equal to it[9]. Same as ID3 the data is sorted at every node of the tree in order to determine the splitting attribute that is best. The splitting stops when the number of instances to be split is below a certain threshold value. The main plus point of C4.5 is when building a decision tree, C4.5 can deal with datasets having unknown attribute valued patterns and datasets with attributes having continuous domains by discretization. This algorithm can deal with training data with attribute values by allowing an attribute value to be marked as missing. A missing attribute value is simply avoided in gain and entropy calculations. It includes an enhanced method of tree pruning that decreases the misclassification errors due to noise or too much detail in the training data set.

IV. CONCLUSION

In this paper, we have presented a unique theme called two stage feature selection method for text categorization. Every term inside the document is ranked depending on their importance for classification using the information gain (IG) methodology. Information Gain methodology is proved to be best among other feature selection techniques[27]. In the next stage genetic algorithm (GA) and principal component analysis (PCA) feature selection and feature extraction methods are applied individually to the terms which are ranked in decreasing order of importance, and after that a dimension reduction is carried out. This proves to be better than the other methodologies generally used which involve single stage feature selection and extraction. Two stage feature selection improves performance since filtering for only important relevant features is done twice ,hence , reducing noise and dimensionality of feature space. To analyze the dimension reduction we use the k-nearest neighbor (KNN)[22] and C4.5 decision tree algorithmic rule on selected data set. The evaluation results shows that, the computational time and complexity of categorization is reduced.

REFERENCES

[1] M.H. Aghdam, N. Ghasem-Aghaee, M.E. Basiri, Text feature selection using ant colony optimization, *Expert Systems with Applications 36 (2009) 6843–6853.*

[2] Yu-Chuan Chang, Shyi-Ming Chen, Churn-Jung Liau, Multi-label text categorization based on a new linear classifier learning method and a category-sensitive refinement

method, *Expert System with Application 34(2008) 1948-1953.*

[3] Jiawei Han, Jaian Pei, Micheline Kamberg, *Data mining concepts and techniques, third edition.*

[4] F.J. Damerau, T. Zhang, S.M. Weiss, N. Indurkhya, Text categorization for a comprehensive time-dependent benchmark, *Information Processing and Management 40 (2004) 209–221.*

[5] M.E. ElAlami, A filter model for feature subset selection based on genetic algorithm, *Knowledge-Based Systems 22 (2009) 356–362.*

[6] M. Gen, R. Cheng, Genetic Algorithms and Engineering Optimization, *vol. 68, Wiley Interscience Publication, 2000*

[7] Y. Li, D.F. Hsu, S.M. Chung, Combining multiple feature selection methods for text categorization by using rank-score characteristics, in*: 21st IEEE International Conference on Tools with Artificial Intelligence, 2009, pp. 508–517.*

[8] M atthew N. Anyanwu & Sajjan G. Shiva, " Comparative Analysis of Serial Decision Tree Classification Algorithms*", International Journal of Computer Science and Security, 2009, (IJCSS)Volume (3): Issue (3).*

[9] J. R. Quinlan, 1996, "Improved use of continuous attributes in C4.5" , *Journal of Artificial Intelligence Research, Vol. 4, pp. 77-90*

[10] V. Mitra, C.-J. Wang, S. Banerjee, Text classification: a least square support vector machine approach, *Applied Soft Computing 7 (2007) 908–914*

[11] D. E. R. O. Alonso and B. Stewart, Crowdsourcing for relevance evaluation,*" SIGIR Forum,November 2008.*

[12] F. Sebastiani, A tutorial on automated text categorisation, in: *Proceedings of the ASAI-99, in: 1st Argentinian Symposium on Artificial Intelligence, Buenos Aires, AR., 1999, pp. 17–35.*

[13] A. McCallum, K. Nigam, A comparison of event models for Naive Bayes text classification, in: *AAAI'98 Workshop on Learning for Text Categorization, 1998, pp. 41–48.*

[14] W. Lam, Y. Han, Automatic textual document categorization based on generalized instance sets and a metamodel, *Proceeding of the IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (5) (2003) 628–633.*

[15] Ghiassi, M., Olschimke, M., Moon, B., & Arnaudo, P.. Automated text classification using a dynamic artificial neural network model. *Expert Systems with Applications, (2012), 39, 10967–10976.*

[16] Ergin, S., Gunal, E. S., Yigit, H., & Aydin, R. Turkish anti-spam filtering using binary and probabilistic models. AWERProcedia *Information Technology and Computer Science, (2012),1, 1007–1012.*

[17] Uysal, A. K., Gunal, S., Ergin, S., & Gunal, E.S.. A novel framework for sms spam filtering. *In Proceedings of the IEEE international symposium on innovations in intelligent systems and applications. Trabzon, Turkiye. (2012)*

[18] Cheng, N., Chandramouli, R., & Subbalakshmi, K. P.. Author gender identification from text. *Digital Investigation, (2011) ,8, 78–88.*

[19] Ozel, S. A. , A web page classification system based on a genetic algorithm using tagged-terms as features. *Expert Systems with Applications, (2011),38,3407–3415.*

[20] Alper Kursat Uysal, Serken Gunal , The impact of preprocessing on text classification, *Information Processing and Management 50 (2014) 104–112*

[21] Yang, Y. , An evaluation of statistical approaches to text categorization.Information Retrieval, *(1999). 69–90.*

[22] Tan, S. (2005). Neighbor-weighted K-nearest neighbor for unbalanced text corpus. *Expert Systems with Applications, 28(4), 667–671.*

[23] D.E. Golberg, *Genetic algorithms in search, optimization and machine learning*,Addison Wesley,1989

[24] R.L. Haupt, S.E. Haupt, Practical Genetic Algorithms, *second edition,* John Wiley & Sons, *New Jersey, United States, 2004.*

[25] C. López-Pujalte, V.P. Guerrero-Bote, F. Moya-Anegón, Order- based fitness functions for genetic algorithms applied to relevace feedback, *Journal of the American Society for Information (2) Science 54(2003) 152–160.*

[26] Sergio Francisco da Silva, Marcela Xavier Ribeiro, João do E.S. Batista Neto , Caetano Traina-Jr., Agma J.M. Traina, Improving the ranking quality of medical image retrieval using a genetic feature selection method , *Decision Support Systems 51 (2011) 810–820.*

[27] Yang, Y., & Pedersen, J. O. *(1997).* A comparative study on feature selection in text categorization. *In 14th international conference on machine learning (pp. 412–420). Morgan Kaufmann Publishers Inc.*

[28] Harun Uguz, "A hybrid approach for text categorization by using x2 stastics, Principal Component Analysis and partical swam optimization*", Academic Journals, vol(8) 37, PP 1818-1828, 2013.*

[29] Kemal Polat, Salih Gunes, " Prediction of hepatitis disease based on principal component analysis and artificial immune recognition system", *Applied Mathematics and computation 189(2007) 1282-1291.*