Conscientious Ant Colony Optimization Based Support Vector Machine for Text Document Classification

Deepa A, E. Chandra Blessie

Abstract: ***Document classification indicates the keyword extraction and it become a thrust research in text mining research. The main purpose of keyword extraction is to classify the documents in a more efficient manner. Misclassification of documents may lead the results to worst case. Hence, there exists a need for optimization to precede the document classification more efficiently. In this paper Conscientious Ant Colony Optimization based Support Vector Machine is proposed to classify the documents. Different keyword extraction methods are available for extracting the contents from documents. Proposed classifier is ensemble with selected keyword extraction methods to increase the classification accuracy. Results shows that the proposed classifier has got better accuracy when ensemble with different keyword extraction methods. The results show that the proposed classifier has better performance in terms of Classification Accuracy and F-Measure, than baseline classifiers.***

Key words: Classification,ACM,Mining,NBA,Reuters,Text

I. INTRODUCTION

The main task of document classification is grouping. Grouping is the process of categorizing the document based on its content. Classification of document is an important research problem which is at the main part of information management and retrieving tasks. Classification of document plays a crucial role in multiple applications that handles searching, organizing, and indicating the maximum amount of specific information. Classification is a prolong problem that exist in the domain of information retrieval. Classification of document can be segregate into three different categories, which are (i) unsupervised learning based classification of which are (i) unsupervised learning based classification of document, (ii) supervised learning based classification of document, and (iii) semi-supervised learning based classification of document. In unsupervised learning based classification of document, no external information is provided to the algorithm for classifying the document. In supervised learning based classification of document, external information related to documents are provided as input to the algorithm for classifying the document.

Revised Manuscript Received on January 05, 2020.

A. Deepa, Assistant Professor in the Department of MCA in Nehru College of Management, Coimbatore.

Dr. E.Chandra Blessie, Professor in the Department of MCA in Nehru College of Management, Coimbatore.

In semi-supervised learning based classification of document, partial inputs related to documents are fed as input to algorithm in the form of labels to the document.

Two important factors of classification of document are (i) extraction of features, (ii) ambiguity of topic. Extraction of features handles by picking the best features that correctly describe the document and assist in the developing of the better classification model. Ambiguity of topic is somewhat complicated when comparing with extraction of features, due to the difficulties faced during categorizing.

In everyday of life, problem of misclassification arises due to high dimensional feature-space. Due to the availability of increased set of words for extracting the feature for selection, the classification process becomes tedious and consumes more time. Hence, the need of optimization arises to classify more accurately with less time. This paper aims to propose ant colony optimization based support vector machine for classifying the documents.

II. LITERATURE REVIEW

Morphological Evaluation [1] proposed to analyze the sentiments in text by utilizing deep learning based classification. Preprocessing cum normalization used to enhance the results, but the results lead to low accuracy in classification. Efficient Text Classification [2] proposed to reduce the terms and weights assigned to text in classification. It mainly focused to identify frequency and concentrated in indices that occur in text. Over fitting problem got raised and degrades the rate of precision and recall. Cluster Classifier [3] proposed to classify the high-level dimensional text data that have multiple classes. It holds the set of clusters determined to identify the formation of new clusters. Subtrees generated to enhance the classification accuracy, but it has degraded the results with increased enhanced false positives.

Data Treatment Strategy [4] proposed to generate compound features for classifying the text. Compound features allowed co-occurring any number of times in the document to increase the classification accuracy. But the increased co-occurrence has decreased the accuracy of classification. Text Report Classification [5] proposed classify the radiology report for identifying the disease. Two deep learning methodologies were proposed for enhancing the classification efficiency, but it wasn't matched with identification of diseases and ended with low accuracy. Semi-Supervised Algorithm [6] proposed to classify text based on rough set cum ensemble learning. Dual classification used for classifying the text by labeling the data. Unlabelled data were used for learning the dataset. The theory of tolerance

rough-set used for approximation. Results ended with low classification accuracy. Ontology Guided Classification [7] proposed to use the taxonomy structure of unified-medical-system in order to improve the feature ranking. Similarity measures were used to improve the classification accuracy. It attempted to increase the classification accuracy, but the results with increased false positives decreased the classification accuracy. Rule based Classification [8] utilize the Naïve bayes algorithm method to reduce the rules generated for classification. It integrates the mining and ruling task to enhance the classification process. Due to reducing the rules used for classification, the accuracy level decreased. Cooperation framework [9] proposed to summarize and classify the enormous level of text data, where text summarizations were most commonly used in smart phones, radios, and television. In this weighting concept was used for features, later features were combined and processed for classification, but the F-Measure result was decreased. Polynomial Network Classification [10] proposed to classify the Arabic text. Initially advantages cum disadvantages of applying the polynomial network were studied, and then it was applied to classify the Arabic text. Results ended with higher level of false negatives leading to poor accuracy.

III. CONSCIENTIOUS ANT COLONY OPTIMIZATION BASED SUPPORT VECTOR MACHINE

A Conscientious Ant Colony Optimization

Natural characteristics of Ants tend to live in groups termed as colonies. Ants utilize chemical substances called pheromone to give a better communication system in a sophisticated manner. Ant move in a random manner to find the pheromone that are laid previously in order to find the better quantity of pheromone. This process is considered as a collection behavior and it is repeated until the best solution is found. The collection behavior of ants provided a way for inspiring the optimization in a metaheuristic method to solve many issues.

Optimization

Different levels are involved in proposed classifier for classifying the documents and it involves below steps.

Build of Search Space

The basic step of CACO begins with the concept to focus the search space. In CACO, the search spaces are made to isolate in making two equal vectors which are (i) rank based sub-graph, (ii) structure based sub-graph. These two sub graphs have same characteristics in handling the documents to segregate, and expressed as

$$Srch\_spce = \sum_{n=1}^{a} b_n \quad (1)$$

where b indicates the document count, a denotes the count of sub_graphs.

Pheromone Initialization

Probing quantity of pheromone that exist between source term and destination term (i.e., $term_p$ and $term_q$) is mathematically expressed as

$$\tau_{pq}(iter = 1) = \frac{1}{\sum_{n=1}^{a} x_n b_n} \quad (2)$$

In Eqn.(2), if p is falls in deliberated attribute, then the value of x_n is treated with 0, else it will be treated with the value 1.

Selection

With the objective of increasing the pheromone, the ants aim to maximize the document detection. It is processed by making a visit to sub-graphs. Ants aim to increase the protocol level to the attributes to detect the documents more accurately. With this protocol, selection process's likelihood is calculated using

$$P_{pj}(t, pter) = \frac{{\tau_{pj}}^{\alpha}(iter){\eta_{pj}}^{\beta}(s)}{\sum_{j=1}^{Total\_terms} x_p\left\{{\tau_{pj}}^{\alpha}(iter){\eta_{pj}}^{\beta}(s)\right\}} \quad (3)$$

Pheromone Level Updation

Persistence level of ant in next process leads in searching feasible search space and it is controlled by Eq.(3). The available pheromones are useful in providing the training CACO. The qualities of trail are treated as important in efficiently using the pheromone values. It is treated as road for the swarm to proceed with next move, and expressed as

$$\tau_{pq(t+1)} = (1+\rho)\tau_{pq}(t) - \left(1 + \left(\frac{1}{q} - Q\right)\right)\tau_{pj}(t) \quad ()$$

where ρ indicates pheromone evaporation probability and Q indicates quality of trail.

B. Support Vector Machine (SVM)

Supervised Machine Learning (SML) is considered as a special method that expects needed input and desired output from user. The user gives documents as input and it labeled in a clear manner for making better classification. It aims to provide better processing of data.

SML algorithms tend to provide measuring capacity to find its future dimension. In short, SML have X input variables and Y output variables. The user use the algorithm to make a study about the classification with function $Y = f(X)$.

SVM is a treated as a special category of SML algorithms. It is fully used to in classifying and regression. Currently, different domains like business, education, medicine, etc., started using SVM algorithms for classification and prediction purpose.

SVM algorithm perform based on discovering the hyperplane which can divide the input or dataset (i.e., X) into two classes. Data points that close to the hyperplane are treated as the support vectors. If data points that are close to the hyperplane are removed then there exists a modification in the position of hyperplane. In short, the hyperplane is treated as a line which classifies the dataset in a linear manner

IV. KEYWORD EXTRACTION METHODS

A. Co-occurrence Statistical Information (CSI)

It target to give priority to important terms by validating the words that got repeated in same type of sentences. Firstly, repeated words were identified and it was used for finding the exact terms in text document [11].

B. Eccentricity Based (EB)

It uses the vertex centrality concept for resolving the issues identified in extracting the keywords. Documents are labeled for effective classification, and more relevant were identified by using the graphs. The document that occupies mid-position are identified as the most relevant document [12].

C. Most Frequent (MF)

It used to search the terms that are repeated in text documents. For searching process, it uses the keywords of the document in the matrix format. Count of the repeated words are used for classification [13]

D. Term Frequency Inverse Sentence Frequency (TFISF)

It works on the basis of statistics. It was considered as the enhance version of frequency based methods. It works by measuring the sentences in document. Each sentences used as a separate vector [14].

E. Text Ranking (TR)

It's a graph-based model utilized to manipulate the text processing. It was used in multiple natural language processing. Initially it seeks the vertices to find a value to process classification. Further, Syntax based filter was applied to generate graphs [15].

F. Grammar based Reduction of Term Frequency (GRBTF):

It seeks to find the grammar words and reduce those words from the whole document to identify keywords effectively. For making the identification process easy, grammar words are removed.. It involves reading and forming sentences, but in matrix format [16].

V. ABOUT DATASETS

A. ACM Document Collection Dataset

This dataset holds eight sub dataset and each have 5 different classes. The description of dataset is provided in Table 1.1. Deep experiments are carried with this dataset for analyzing the classifiers performance.

Table1.1: ACM Document Collection Dataset Description

Col.	Class #	Docs.	Col.	Class #	Docs.
ACM- 1	3D technologies	91	ACM- 5	Tangible and embedded interaction	81
	Visualization	72		Management of data	96
	Wireless mobile multimedia	82		User interface software and technology	104
	Solid and physical modeling	74		Information technology education	87
	Software engineering	82		Theory of computing	103
ACM- 2	Rationality and knowledge	86	ACM- 6	Computational geometry	89
	Simulation	84		Access control models and technologies	90
	Software reusability	72		Computational molecular biology	71
	Virtual reality	83		Parallel programming	96
	Web intelligence	86		Integrated circuits and system design	93
ACM- 3	Computer architecture education	78	ACM- 7	Database systems	104
	Networking and communications systems	75		Declarative programming	101
	Privacy in the electronic society	98		Parallel and distributed simulation	98
	Software and performance	81		Mobile systems, applications and services	95
	Web information and data management	92		Network and system support for games	73
ACM- 4	Embedded networked sensor systems	50	ACM- 8	Mobile ad hoc networking and computing	90
	Information retrieval	71		Knowledge discovery and data mining	105
	Parallel algorithms and architectures	98		Embedded systems	102
	Volume visualization	104		Hypertext and hypermedia	93
	Web accessibility	71		Microarchitecture	105

B. Reuters-21578 Document Collection Dataset

It holds the 10 classes of ModApte Split [18] that belongs to Reuters-21578. The information needed for concerning the training quantity and testing are given in Table 1.2.

Table 1.2 Description Of Reuuters 21578 Document Collection Dataset

Label of the Class	Training Samples count	Testing Samples count
Acq	1650	0719
Corn	0181	0056
Crude	0389	0189
Earn	2877	1087
Grain	0433	0149
Interest	0347	0131
Money-fx	0538	0179
Ship	0197	0089
Trade	0369	0117
Wheat	0212	0071

C. NBA Input Document Collection Dataset

It consist of 8 sub-division and each holds different number of classes. It's description is given in Table 1.3. This dataset is processed with different keywords for unique terms for extracting the keywords.

Table 1.3: NBA Input Document Collection Dataset

Col.	Class#	Docs	Col.	Class#	Docs
NBA1_ Student	Select_Proc	26	NBA5_ Library	Books	30
	Stud_Intake_Capacity	30		E-Journals	10
	Enrol_Proc	32		Online Databases	11
	Admn_Process	26		Films_videos	17
	Admn_Guidelines	30		Lib_Mgmt_S/W	20
	Final Result	13		SS_Field Work	14
NBA2_ Faculty	S-F Strength	24		Working_Hours	25
	F_S_R	13		Users_Feedback	13
	R_FT_PTF	18		Inter_Lib_Network	17
	F-Qual	16	NBA6_ Global_Input	N_IN_Collaborations	30
	F_Retention	20		NIN_AC_Partnerships	16
	Resrch_Proc_Fac	12		NIN_Strategic_Alliance	35
	Fac_Exposure	23		NIN_Exchg_prgms	27
	FDP_Observation	17		NIN_Corp_partners	26
	Out_Exp_CA	22		Resrch_Collaborations	23
	Resrch_Apt_Fac	12	NBA7_ Quality_ Assurance_Policy	Legacy_BSchool_QA	36
NBA3_ Physical_ Infrastructure	Nat_Geo_Access	40		IA_Process_EDU	26
	Dist_Loc	39		CC_Review_process	34
	Phys_Ambience	36		Emp_Orgs_Feedback	28
	Ava_resources	42		APP_Real_CC	33
NBA4_ IT_ Infrastructure	Operating_ICT	12	NBA8_ Finance	Fund_Effectiveness	18
	Use_Ins_Kits	30		Fin_Self_Suff	24
	H/W_S/W-State	26		Fin_Prf_3Yrs	22
	IT_Lab_Usage	18		IFCRS	22
	Wifi_Use	23		PI_Staff	23
	Video_conferencing	23		Scope_Range_FS	22
	Learn_Platforms	25		Ensr_Accountability	26

VI. EVALUATION MEASURES

The evaluations of this experiment are done in personal computer with configurations as Intel Core i7 processor having speed of 3.40 GHz, and random access memory of 8 gigabytes. The experiments are performed with MATLAB version R2013a. To measure the prediction performance of existing and proposed classification algorithms, this research work utilizes the traditional performance metrics classification accuracy and F-measure for the evaluation purpose.

- ✓ Classification Accuracy : Percentage of true values (positives and negatives) against the overall number of instances.
- ✓ Precision : Percentage of true positives over the total of false positives and true positives.
- ✓ Recall : Percentage of true positives over the total of false negatives and true positives.
- ✓ F-Measure : Percentage of precision and recalls harmonic mean.

VII. RESULTS AND DISCUSSION

A. Classification Accuracy Analysis

Figure 1, Figure 2 and Figure 3 discusses the performance of proposed classifier against RF [18] and Bagging RF [19]. Performances of the classifier are tested using three different datasets namely ACM Document Collection Dataset, Reuters-21578 Document Collection Dataset, and NBA dataset. Classifiers are ensemble with different keyword extraction methods for enhancing the results more. The proposed classifier has better performance with all keyword extraction methods and it gives more accuracy when ensemble with GRBTF.

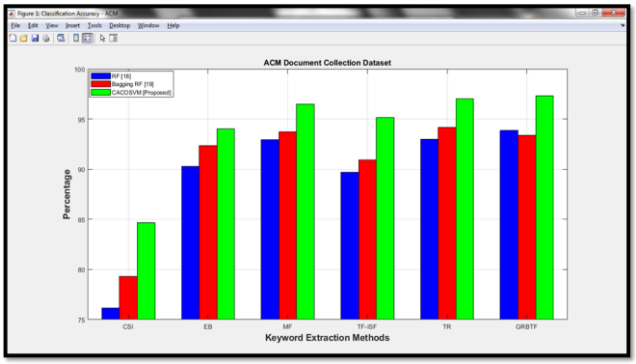
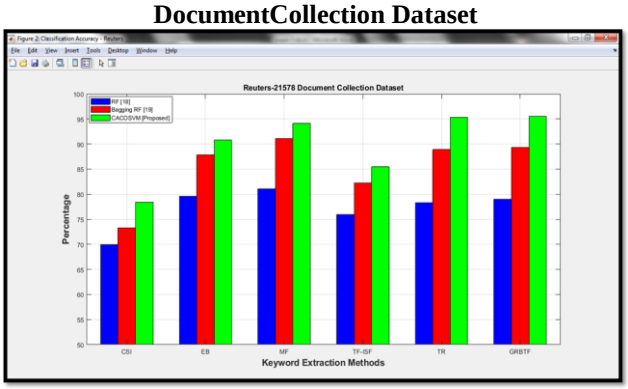


Figure. 1 Classification Accuracy vs ACM DocumentCollection Dataset



. Figure2:Classification Accuracy vs Reuters-21578 Document Collection Dataset

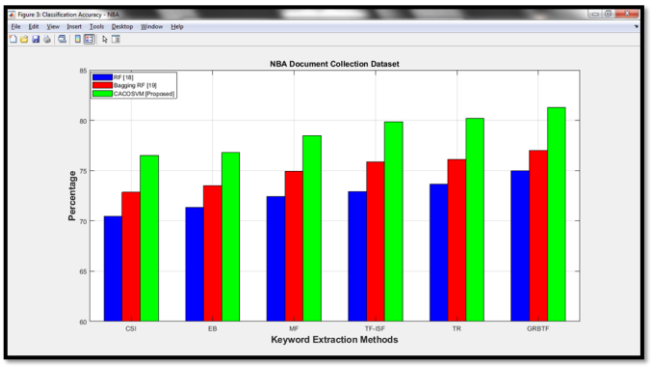


Figure. 3 Classification Accuracy vs NBA DocumentCollection Dataset

B. F-Measure Analysis

Figure 4, Figure 5 and Figure 6 discusses the performance of proposed classifier against RF [18] and Bagging RF [19]. The proposed classifier has better F-measure when it is ensemble with GRBTF, and it is due to performing the optimization and proceeding the classification.

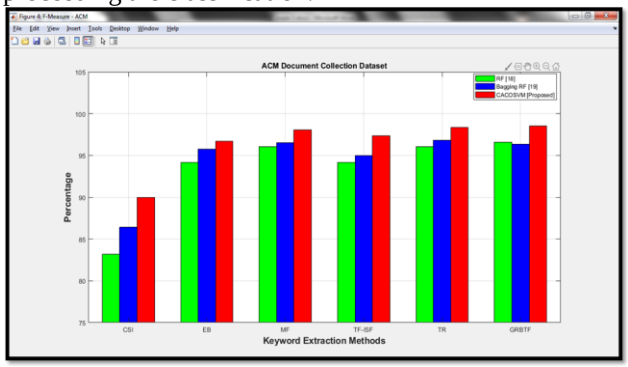


Figure. 4 F-Measure vs ACM Document Collection Dataset

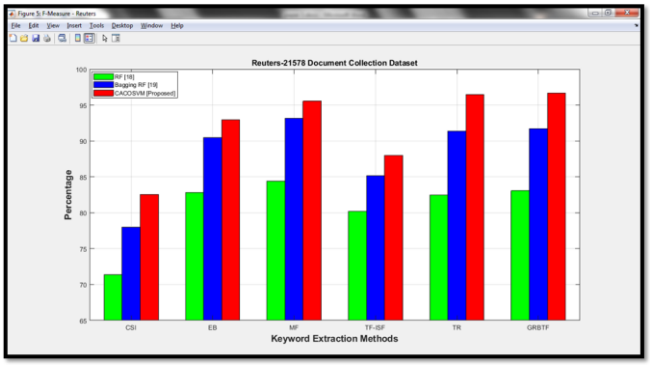


Figure. 5 F-Measure vs Reuters-21578 Document Collection Dataset

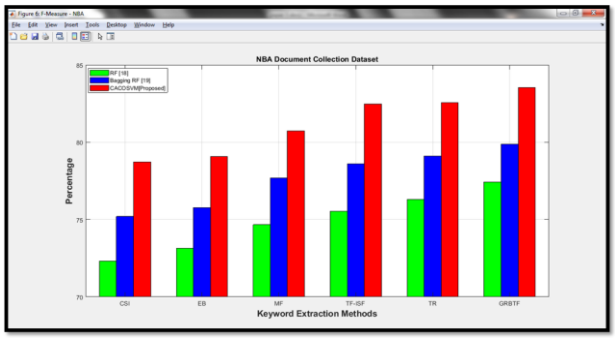


Figure. 6 F-Measure vs NBA Document Collection Dataset

Table1: Tabular Representation for Accuracy computation of classifiers on 3 datasets

Classifiers	ACM	Reuters	NBA
Random Forest	94	79	75
Bagging Random Forest	93	89.9	77
CACOSVM	98	95	82

VIII. CONCLUSION

This paper has proposed CACOSVM have classified the documents with more accurately based on the keywords extracted by different methods. Most available classifiers are suitable for more small or specific dataset. Those classifiers won't have better performance with huge dataset. The proposed classifier is designed to adopt dataset of any size and perform classification by segregating the dataset into multiple parts and perform classification in a random manner which results in an improved classification accuracy and f-measure. For evaluating the performance of the proposed classifier ACM document collection dataset, Reuters-21578 document collection dataset, and NBA Input Document Collection Dataset are used.

REFERENCES

1. J. Singh, G. Singh, R. Singh, P. Singh, "Morphological evaluation and sentiment analysis of Punjabi text using deep learning classification", *Journal of King Saud University - Computer and Information Sciences*,2018.
2. K. Shi, J. He, H. Liu, N. Zhang, W. Song, "Efficient text classification method based on improved term reduction and term weighting", *The Journal of China Universities of Posts and Telecommunications*, Volume 18, Supplement 1, 2011, Pages 131-135.
3. Y. Li, E. Hung, K. Chung, "A subspace decision cluster classifier for text classification", *Expert Systems with Applications*, Volume 38, Issue 10, 2011, Pages 12475-12482.
4. F. Figueiredo, L. Rocha, T. Couto, T. Salles, M. A. Gonçalves, W. M. Jr., "Word co-occurrence features for text classification", *Information Systems*, Volume 36, Issue 5, 2011, Pages 843-858.
5. I. Banerjee, Y. Ling, M. C. Chen, S. A. Hasan, C. P. Langlotz, N. Moradzadeh, B. Chapman, T. Amrhein, D. Mong, D. L. Rubin, O. Farri, M. P. Lungren. "Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification", *Artificial Intelligence in Medicine*, 2018.
6. L. Shi, X. Ma, L. Xi, Q. Duan, J. Zhao, "Rough set and ensemble learning based semi-supervised algorithm for text classification", *Expert Systems with Applications*, Volume 38, Issue 5, 2011, Pages 6300-6306.
7. V. N. Garla, C. Brandt, "Ontology-guided feature engineering for clinical text classification", *Journal of Biomedical Informatics*, Volume 45, Issue 5, 2012, Pages 992-998.
8. W. Hadi, Q. A. Al-Radaideh, S. Alhawari, "Integrating associative rule-based classification with Naïve Bayes for text classification", *Applied Soft Computing*, Volume 69, 2018, Pages 344-356.
9. H. Jeong, Y. Ko, J. Seo, "How to Improve Text Summarization and Classification by Mutual Cooperation on an Integrated Framework", *Expert Systems with Applications*, Volume 60, 2016, Pages 222-233.
10. M. M. Al-Tahrawi, S. N. Al-Khatib, "Arabic text classification using Polynomial Networks", *Journal of King Saud University - Computer and Information Sciences*, Volume 27, Issue 4, 2015, Pages 437-449.
11. Y. Matsuo, M. Ishizuka, "Keyword Extraction from a Single Document using Word Co-Occurrence Statistical Information", *International Journal on Artificial Intelligence Tools*, Vol. 13, Issue. 1, pages. 157–169, 2004.
12. G. K. Palshikar, "Keyword Extraction from a Single Document Using Centrality Measures", In: Proc. *Second International Conference on Pattern Recognition and Machine Intelligence*, India. *Lecture Notes in Computer Science*, Vol 4815, pages 503-510, 2007.
13. R. G. Rossi, R. M. Maracini, S. O. Rezende, "Analysis of Domain Independent Statistical Keyword Extraction Methods for Incremental Clustering", *Learning and Nonlinear Models*, Vol. 12, Issue. 1, pages 17–37, 2014.
14. L. J. Neto, A . D. Santos, C. A .Kaestner, A. A. Freitas, "Document Clustering and Text Summarization", In: Proc. *4th International Conference on Practical Applications of Knowledge Discovery and Data Mining*, United Kingdom, pp. 41–55, 2000.
15. R. Mihalcea, P. Tarau, "TextRank: Bringing Order into Text", In: Proc. *2004 Conference on Empirical Methods in Natural Language Processing*, Spain, pp. 404–411, 2004.
16. D. Roland, F. Dick, J.L. Elman, Frequency of Basic English Grammatical Structures: A Corpus Analysis, Journal of Memory and Language, Vol 57, Iss. 3, 2007, Pages 348-379.
17. A. K. Uysal, "An Improved Global Feature Selection Scheme for Text Classification", *Expert Systems with Applications*, Vol. 43, 2016, Pages 82-92.
18. L. Breiman, Random Forests, *Machine Learning*, Vol. 45, Issue. 1, 2001, Pages 5-32
19. A. Onan, S. Korukoglu, H. Bulut, "Ensemble of Keyword Extraction Methods and Classifiers in Text Classification", *Expert Systems with Applications*, Vol. 57, 2016, Pages 232-247.

AUTHORS PROFILE

***A.* Deepa,** is currently working as Assistant Professor in the Department of MCA in Nehru College of Management, Coimbatore. She has pursued her Bachelor's degree under Calicut University and Master's degree under Indiragandhi National Open University and M.Phil. in Bharatidasan University. She has registered for Ph.D. in Bharatiar Uniiversity in November 2016.She published papers in reputed international/national journals, Scopus indexed Journals in UGC Journals and in international conferences. Her main research work focuses on Data Mining, and, Big Data Analytics. She has 15 years of teaching experience.

Dr. E.Chandra Blessie, is currently working as Professor in the Department of MCA in Nehru College of Management, Coimbatore. She has pursued her Bachelor's and Master's degree under Manonmaniam Sudaranar University and M.Phil. in Alagappa University. She has completed her Ph.D. in Karunya University.She got Best Paper Awards in Conferences and published papers in reputed international/national journals. She is the Student Branch Counselors –HEAD , CSI, Entire Coimbatore Chapter, Management Committee member of CSI, Coimbatore Chapter. Member of Institute of Advanced Scientific Research, Member of IACSIT. Her main research work focuses on Data Mining, Preprocessing in DM, Big Data Analytics. She has 17 years of teaching experience