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 show 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.

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

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.
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

$$Strch\_space = \sum_{a=1}^{a} b_n$$

(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

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

(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

$$P_{pj}(t, pter) = \frac{\tau_{pj}^{u}(iter) \eta_{pj}^{\beta}(s)}{\sum_{j=1}^{Total\_terms} x_{p}^{r}(\tau_{pj}^{c}(iter) \eta_{pj}^{\beta}(s))}$$

(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

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

where $\rho$ 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.

Table 1.1: ACM Document Collection Dataset Description

<table>
<thead>
<tr>
<th>Class</th>
<th>Class</th>
<th>Docs.</th>
<th>Class</th>
<th>Class</th>
<th>Docs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACM-1</td>
<td>3D technologies</td>
<td>91</td>
<td>ACM-2</td>
<td>Usability and knowledge</td>
<td>86</td>
</tr>
<tr>
<td>ACM-1</td>
<td>Verification</td>
<td>72</td>
<td>ACM-2</td>
<td>Simulation</td>
<td>84</td>
</tr>
<tr>
<td>ACM-1</td>
<td>Wireless mobile multimedia</td>
<td>82</td>
<td>ACM-2</td>
<td>Software reusability</td>
<td>72</td>
</tr>
<tr>
<td>ACM-1</td>
<td>Solid and physical modeling</td>
<td>74</td>
<td>ACM-2</td>
<td>Virtual reality</td>
<td>83</td>
</tr>
<tr>
<td>ACM-1</td>
<td>Software engineering</td>
<td>82</td>
<td>ACM-2</td>
<td>Web intelligence</td>
<td>86</td>
</tr>
<tr>
<td>ACM-3</td>
<td>Computer architecture education</td>
<td>78</td>
<td>ACM-3</td>
<td>Networking and communications systems</td>
<td>75</td>
</tr>
<tr>
<td>ACM-3</td>
<td>Networking and communications systems</td>
<td>75</td>
<td>ACM-3</td>
<td>Privacy in the electronic society</td>
<td>98</td>
</tr>
<tr>
<td>ACM-3</td>
<td>Software and performance</td>
<td>83</td>
<td>ACM-3</td>
<td>Web information and data management</td>
<td>92</td>
</tr>
<tr>
<td>ACM-4</td>
<td>Embedded computer systems</td>
<td>90</td>
<td>ACM-4</td>
<td>Embedded systems</td>
<td>98</td>
</tr>
<tr>
<td>ACM-4</td>
<td>Information retrieval</td>
<td>90</td>
<td>ACM-4</td>
<td>Parallel algorithms and data structures</td>
<td>98</td>
</tr>
<tr>
<td>ACM-4</td>
<td>Virtual simulation</td>
<td>94</td>
<td>ACM-4</td>
<td>Software architecture</td>
<td>83</td>
</tr>
<tr>
<td>ACM-4</td>
<td>Web accessibility</td>
<td>71</td>
<td>ACM-4</td>
<td>Knowledge and data mining</td>
<td>105</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Label of the Class</th>
<th>Training Samples count</th>
<th>Testing Samples count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acq</td>
<td>1650</td>
<td>0719</td>
</tr>
<tr>
<td>Com</td>
<td>0181</td>
<td>0056</td>
</tr>
<tr>
<td>Crime</td>
<td>0389</td>
<td>0189</td>
</tr>
<tr>
<td>Earn</td>
<td>2877</td>
<td>1087</td>
</tr>
<tr>
<td>Grain</td>
<td>0433</td>
<td>0149</td>
</tr>
<tr>
<td>Interest</td>
<td>0347</td>
<td>0131</td>
</tr>
<tr>
<td>Money-fx</td>
<td>0538</td>
<td>0179</td>
</tr>
<tr>
<td>ship</td>
<td>0197</td>
<td>0089</td>
</tr>
<tr>
<td>Trade</td>
<td>0569</td>
<td>0117</td>
</tr>
<tr>
<td>Wheat</td>
<td>0212</td>
<td>0071</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Class</th>
<th>Docs.</th>
<th>Class</th>
<th>Class</th>
<th>Docs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBA1_</td>
<td>Select_PoC</td>
<td>28</td>
<td>NBA5_</td>
<td>Library</td>
<td>30</td>
</tr>
<tr>
<td>NBA1_</td>
<td>Stud_Enroll_Capacity</td>
<td>30</td>
<td>NBA5_</td>
<td>Books</td>
<td>10</td>
</tr>
<tr>
<td>NBA1_</td>
<td>Earl_PoC</td>
<td>32</td>
<td>NBA5_</td>
<td>E-Sports</td>
<td>10</td>
</tr>
<tr>
<td>NBA1_</td>
<td>Admin_Process</td>
<td>28</td>
<td>NBA5_</td>
<td>Game_Databases</td>
<td>11</td>
</tr>
<tr>
<td>NBA1_</td>
<td>Admin_Registration</td>
<td>80</td>
<td>NBA5_</td>
<td>Films_videos</td>
<td>17</td>
</tr>
<tr>
<td>NBA1_</td>
<td>Final_Result</td>
<td>13</td>
<td>NBA5_</td>
<td>Lib_Mgr важая</td>
<td>20</td>
</tr>
<tr>
<td>NBA2_</td>
<td>F-Score</td>
<td>34</td>
<td>NBA6_</td>
<td>SS_Field_Walk</td>
<td>14</td>
</tr>
<tr>
<td>NBA2_</td>
<td>R-F</td>
<td>13</td>
<td>NBA6_</td>
<td>Working_Hours</td>
<td>25</td>
</tr>
<tr>
<td>NBA2_</td>
<td>R_F1</td>
<td>13</td>
<td>NBA6_</td>
<td>User_Feedback</td>
<td>13</td>
</tr>
<tr>
<td>NBA2_</td>
<td>R_F1</td>
<td>13</td>
<td>NBA6_</td>
<td>Intra_Lib_Network</td>
<td>17</td>
</tr>
<tr>
<td>NBA2_</td>
<td>R_F1</td>
<td>13</td>
<td>NBA6_</td>
<td>NWC_Collaborations</td>
<td>30</td>
</tr>
<tr>
<td>NBA2_</td>
<td>F-Score</td>
<td>34</td>
<td>NBA6_</td>
<td>NWC_Collaborations</td>
<td>30</td>
</tr>
<tr>
<td>NBA2_</td>
<td>R_F1</td>
<td>13</td>
<td>NBA6_</td>
<td>NWC_Collaborations</td>
<td>30</td>
</tr>
<tr>
<td>NBA2_</td>
<td>R_F1</td>
<td>13</td>
<td>NBA6_</td>
<td>NWC_Collaborations</td>
<td>30</td>
</tr>
<tr>
<td>NBA2_</td>
<td>R_F1</td>
<td>13</td>
<td>NBA6_</td>
<td>NWC_Collaborations</td>
<td>30</td>
</tr>
<tr>
<td>NBA2_</td>
<td>R_F1</td>
<td>13</td>
<td>NBA6_</td>
<td>NWC_Collaborations</td>
<td>30</td>
</tr>
</tbody>
</table>

C8062019320/2020©BEIESP
DOI: 10.35940/ijitee.C8062.019320
1058

Published By:
Blue Eyes Intelligence Engineering & Sciences Publication
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.

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

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>ACM</th>
<th>Reuters</th>
<th>NBA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>94</td>
<td>79</td>
<td>75</td>
</tr>
<tr>
<td>Bagging Random Forest</td>
<td>93</td>
<td>89.9</td>
<td>77</td>
</tr>
<tr>
<td>CACOSVM</td>
<td>98</td>
<td>95</td>
<td>82</td>
</tr>
</tbody>
</table>

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 adapt 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


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 Bharat University 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 Manonnanam 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.