USER INTERESTED OPTIMIZED RULE MINING WITH ARTIFICIAL BEE COLONY IN DOCUMENT CLASSIFICATION

Drashti Panchal[1],Mohit Patel[2]

[1] *Student, Department of Computer Engineering , SCET, Ahmedabad, INDIA*

[2]*A/Prof., Department of Computer Engineering, SCET, Ahmedabad, INDIA*

ABSTRACT

Document classification is problem in library science, information science and computer science. The task is to assign a document to one or more classes or categories. This may be done "manually" (or "intellectually") or algorithmically. In my base paper they have used "Association Rule Mining" for Document Categorization. So I am going to used Optimized Rule Mining with Artificial Bee Colony (ABC) algorithm recently introduced optimization algorithms to simulate intelligent foraging behavior of honey bee swarm was proposed by Karboga and Ozturk. We will customize the ABC Algorithm to improve accuracy of Document Classification.

Keyword : - ABC Algorithm, Document classification, Rule Mining

1. INTRODUCATION

Data mining is to extract or mine knowledge from a lot of data called Knowledge Discovery in Databases (KDD),which is the result of information technology natural which is the result of information technology natural evolution. Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to efficiency conclude the target class for each case in the data. For example a classification model could be used to identify loan applicants as low, medium or high credit risks.[1]

Traditional single-label classification is concerned with learning from a set of examples that are associated with a single label l from a set of disjoint labels L, |L| > 1. If |L| = 2, then the learning problem is called a binary classification problem (or filtering in the case of textual and web data), while if |L| > 2, then it is called a multi-class classification problem.[1]

In multi-label classification, the examples are associated with a set of labels Y ⊆ L. In the past, multi-label classification was mainly motivated by the tasks of text categorization and medical diagnosis. Text documents usually belong to more than one conceptual class. For example, a newspaper article concerning the reactions of the Christian church to the release of the Da Vinci Code film can be classified into both of the categories Society\Religion and Arts\Movies. Similarly in medical diagnosis, a patient may be suffering for example from diabetes and prostate cancer at the same time.[1]

TYPES OF DOCUMENT CLASSIFICATION

- **Content-based Classification:**
  is classification in which the weight given to particular subjects in a document determines the class to which the document is assigned. It is, for example, a common rule for classification in libraries that at least 20% of the

content of a book should be about the class to which the book is assigned. In automatic classification it could be the number of times given words appears in a document.[2]

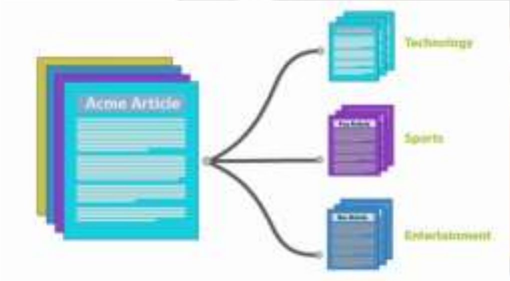
- **Request-oriented classification:**
  (or-indexing) is classification in which the anticipated request from users is affect how documents are being classified. The classifier asks himself Under which description should this entity be found and think of all the possible queries and choose for which particular the entity at hand is applicable. [2]

The Rest of this Paper is structured in the following way: We divide our research paper into five sections. First section delivers about our background and objective in this research. Our related works can be seen in second section. For third section, we define about our proposed method. Fourth section is used to define our experiment dataset and experiment scenario. Last section is used for providing our research conclusion and our further research.

2. RELATED WORK

Document classification can be applied as an information filtering tool and can be used to improve the retrieval results from a query process and to make good decisions.

The documents to be classified maybe texts, images, music etc.



Term Frequency & Inverse Document Frequency (Tf–Idf):

“Tf–idf is a numerical statistic that is shown that how important a word is to a document in a collection. It is often used as a weighting factor in information retrieval and text mining. The tf-idf value increases with respect to the number of times a word looks in the document.”

It is often used by search engines as a hub tool in scoring and ranking a documents relevance given a user query. Tf–idf can be generally used for stop- words filtering in various subject areas including text summarization and classification.

Term frequency:

Assume that we have a set of English text documents and want to determine/find which document is most relevant to the query "the brown cow". A simple way to start out is by eliminating/removing documents that do not contain all three words "the", "brown", and "cow", but this still leaves many documents. To further illustrate them, we might count the number of times each term occurs in each document and sum them all together; the number of times a term occurs in a document is called its Term Frequency.

The weight of a term/definition that occurs in a document is proportional to the term frequency.

Inverse document frequency:

Given term "the" is common, term frequency will show to incorrectly emphasize documents which happen to use the word "the" more frequently/alternatively, without giving enough weight to the more meaningful terms "brown" and "cow". The term "the" is not a good keyword to illustrate relevant and non-relevant documents and terms, unlike the less common words "brown" and "cow". It supports the weight of terms that makes often in the document set and increases the weight of terms that occur rarely.

The specifically a term can be quantified as an inverse function of the number of documents in which it occurs.

ABC ALGORITHM

The ABC algorithm is based on bee's behavior in finding the food source positions without the benefit of visual information (Karaboga and Ozturk, 2011). The information exchange from bees is assimilation knowledge about which path to follow and quality of food through a waggle dance. Bees calculate their food source using probabilistic selection and abounding source by sharing their information through waggle dance and food source with less anticipation of producing new food source in neighborhood of old source in relation to their useful. The ABC has three necessary components: food source, employed bee, scout bee and onlooker bee, and the behaviors are: selection and rejection of the food source.[4]

- Employed Bee: The employed bees store the food source information which includes the gap the guidance and share with others according to a certain probability and shares with other bees waiting in the colony abundance and extraction of energy nectar taste and strength of the solution.

- Onlooker Bee: It takes the information from selected numbers of employed bee and decides the probability of higher nectar amount information of the food source is selected according to beneficial of food source.

- Scout Bee: If the position of food source is not better through best number of cycles food source will be removed from the community employed bee becomes a scout bee and names a new random food source. Based on the performance of strength value if the named new food source is better than dismiss one then scout bee becomes employee bee. This process is repeated until the maximum number of cycles to determine the optimal solution of food source.[5]

Classification Using ABC Algorithm

ABC algorithm is a new swarm intelligent algorithm and consists of three essential components:

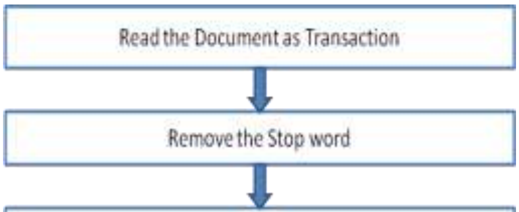
1. Food Sources: It represents a position of solution of the problem.

2. Employed Foragers: The number of employed bees is equal to the number of food sources. The employed bees store the food source information and share with others according to certain probability.

3. Unemployed Foragers: Their main task is exploring and exploiting food source. There are two choices for the unemployed foragers (i) It becomes an onlooker and determines the nectar amount of food source after watching the waggle dances of employed bee and collection food source according to beneficial (ii) It becomes a scout and randomly searches new food sources around the nest.[5]

3. PROPOSED METHOD

As referred base paper conclusion, The Associative Classification algorithm has proposed in this paper which has outperformed the traditional classification algorithms and other AC algorithm.



APPLY ABC ALGORITHM

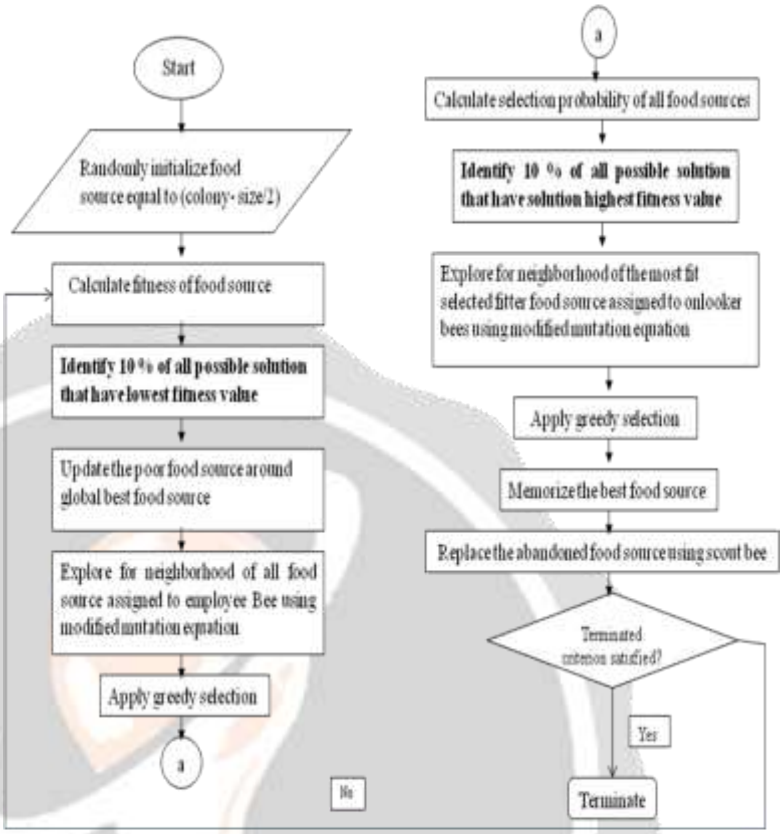

Start
Randomly initialize food source equal to (colony- size/2)
Calculate fitness of food source
Identify 10 % of all possible solution that have lowest fitness value
Update the poor food source around global best food source
Explore for neighborhood of all food source assigned to employee Bee using modified mutation equation
Apply greedy selection
a
a
Calculate selection probability of all food sources
Identify 10 % of all possible solution that have solution highest fitness value
Explore for neighborhood of the most fit selected fitter food source assigned to onlooker bees using modified mutation equation
Apply greedy selection
Memorize the best food source
Replace the abandoned food source using scout bee
Terminated criterion satisfied?
Yes
No
Terminate

Fig 1: Architecture of Proposed Work

PROPOSED ALGORITHM:

Step 1: Read the document as Transaction

Step 2: Remove stop word.

Step 3: Calculate Term Frequency and Inverse Document Frequency(TF/IDF)

Step 4: Filter terms based on support and confidence.

Step 5: Generate Rules.(Pair/Group wise terms)

Step 6: Get Predefined category documents.

Step 7: Get the abstract as training data.

Step 8: Load training samples. (Predefined rules)

[Apply ABC Algorithm For Best Category]

Step 9: Best solution updation.

Step 10: Assign category to document.

- **Fitness Function**

(nk + 1)/(n + vocabulary)

where,

n = “Total no of word set position in all training examples for Category(j)

nk = No. of times the word set found among all the training examples Category(j)

vocabulary =The total number of distinct word set found within all the training data.”

- **Precision:** Precision (P) is the ratio of number of categories correctly assigned to the total number of categories assigned.

In other word precision is the fraction of retrieved documents that are relevant = P (relevant retrieved)

$$Precision(P) = \frac{tp}{tp + fp}$$

tp = true positive fp = false positive

- **Recall:** “Recall (R) is the ratio of number of categories correctly assigned to the total number of correct categories that should be assigned.”

In other words recall is fraction of relevant docs that are retrieved = P (retrieved |relevant)

$$Recall\ (R) = \frac{tp}{tp + fn}$$

fn = false negative

To summarize the result we have also used **f1 measure.**

$$f1\,measure = \frac{2pr}{p + r}$$

pr = precision recall

- **Term frequency:** Term frequency is the ratio of the count of a term in a document to the entire number of terms in that document.

$$TF(t,d) = \frac{term\ count\ in\ d}{total\ terms\ in\ d}$$

Where d is particular document.

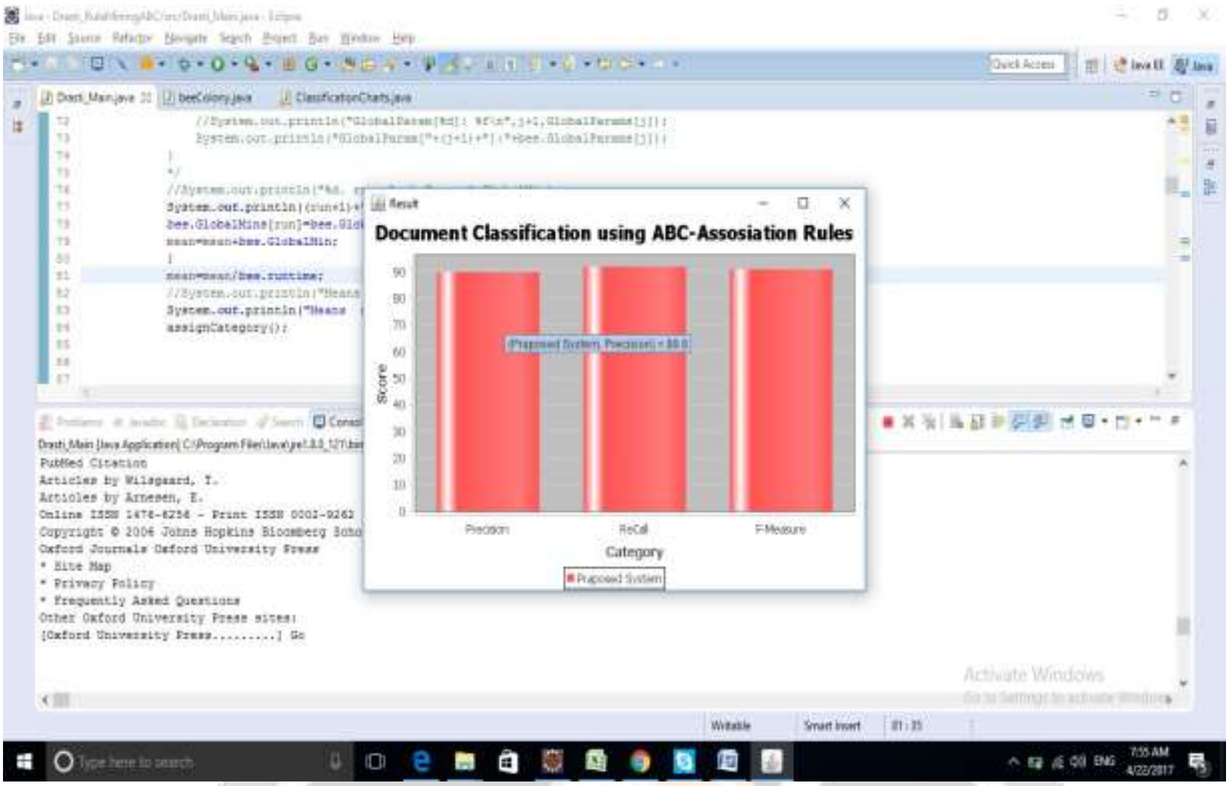
- **Inverse document frequency:** Inverse document frequency can be described as the ratio of log of whole number of documents to the number of documents in which that term appears.

$$IDF(t,d) = \frac{\log N}{number\ of\ documents\ in\ which\ t\ appears}$$

Where, N indicates the entire number of documents in the training documents set.

4. EXPERIMENTS

As We have applied Artificial Bee Colony Algorithm for Optimization of Associations Rules to categorize the documents. We have used NLM Dataset with 500 documents for this purpose. We have tried to get better precision for documents. In future this algorithm can be extended for Large Datasets and also we can use better Fitness functions or other variant of Swarm Intelligence Technique.


Document Classification using ABC-Assosiation Rules
Score
90
80
70
60
50
40
30
20
10
0
Proposed System, Precision = 89.8
Precision
ReCall
F-Measure
Category
Proposed System

```
            //System.out.println("GlobalParam[%d]: %f\n",j+1,GlobalParams[j]);
            System.out.println("GlobalParam["+(j+1)+"]:"+bee.GlobalParams[j]);
        }
        */
        //System.out.println("%d. r
        System.out.println((run+1)+"
        bee.GlobalMins[run]=bee.Glo
        mean=mean+bee.GlobalMin;
        }
        mean=mean/bee.runtime;
        //System.out.println("Means
        System.out.println("Means
        assignCategory();
```

```
Drsti_Main [Java Application] C:\Program Files\Java\jre1.8.0_121\bin
PubMed Citation
Articles by Wilsgaard, T.
Articles by Arnesen, E.
Online ISSN 1476-6256 - Print ISSN 0002-9262
Copyright © 2006 Johns Hopkins Bloomberg Scho
Oxford Journals Oxford University Press
* Site Map
* Privacy Policy
* Frequently Asked Questions
Other Oxford University Press sites:
[Oxford University Press.........] Go
```

Fig-2 Result Analysis

Result

Precision 89.8

Recall 91.82004

F-measure 90.78879

5. CONCLUSIONS

Artificial Bee Colony has been used for different purposed in the field of Mining. In our research work we have used Artificial Bee Colony algorithm for document categorization. We have generated the Associations Rules from Training Dataset based on Terms and Category. We have applied ABC for better optimization of Rules and the prediction phase has been tested. We have got little bit better result for Prediction of the documents.

6. REFERENCES

(1) Grigorios Tsoumakas, Ioannis Katakis, "Multi-Label Classification: An Overview "

(2) C.Kavitha, Dr.G.Sudha Sadasivam, S.Kiruthika , "Semantic similarity based webdocument classification using Artificial Bee Colony (ABC) algorithm",Volume-13,2014.

(3) Mohammad Naved Qureshi,Hasan Faisal Hamood Aldheleai and Yahya Kord Tamandani "An Improved Documents Classification Technique Using Association Rules Mining", 2015 IEEE ICRCICN, pp. 460-465, 2015.

(4) Juan Luis Olmo Ortiz, Cristóbal Romero, Eva Gibaja, Sebastian Ventura, "Improving Meta-learning for Algorithm Selection by Using Multi-label Classification: A Case of Study with Educational Data Sets", 2015,

IJCIS

(5) J.Jayanth, Shivaprakash Koliwad, Ashok Kumar T, “Classification of remote sensed data using Artificial Bee Colony algorithm” , Feb-2015 , EJRSSS, pp. 1-8.

(6) V. Tam, A. Santoso, and R. Setiono, “A comparative study of centroid- based, neighborhood-based and statistical approaches for effective document categorization,” in Object recognition supported by user interaction for service robots, 2002, vol. 4, pp. 235–238.

(7) R. Irina, “An empirical study of the naive Bayes classifier,” IJCAI Work.Empir. methods Artif. Intell., vol. 3, no. 22, pp. 41–46, 2001

(8) U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Advances in Knowledge Discovery and Data Mining. The MIT Press, 1996.

(9) Naseer Ahmed Sajid, Muhammad Tanvir Afzal, Muhammad Abdul Qadir, “Multi-label Classification of computer science documents using fuzzy logic”,Jun-2016, JNFS, pp.155- 165.

(10) Fernando E. B. Otero, Alex A. Freitas, Colin G. Johnson, “A hierarchical multi-label classification ant colony algorithm for protein function prediction”, Jun-2010, Springer, pp. 165-181.

(11) Bo Li, Hong Li, Min Wu, PingLi, “Multi-label Classification based on Association Rules with Application to Scene Classification”, 2008,IEEE, pp. 36-41

(12) H Haripriya, Prathibhamol Cp, Yashwant RPai, M Sai Sandeep, “Multi Label Prediction Using Association Rule Generation and Simple k-Mea

(13) T.-Y. Wang and H.- M. Chiang, “Fuzzy support vector machine for multi- class text categorization,” Inf. Process. Manag., vol. 43, no. 4, pp. 914– 929, Jul. 2007.

(14) T.Sumathi, S.Karthik,M.Marikkannan, “Artificial bee colony optimization for feature selection in opinion mining” , August 2014, JATIT , pp. 368-379 .