International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 12 | Dec -2016 www.irjet.net p-ISSN: 2395-0072

FEATURE SELECTION USING BINARY ARTIFICIAL BEE COLONY FOR SENTIMENT CLASSIFICATION

[1]SP.Rajamohana [1], [2]Dr.K.Umamaheswari [2]

[1]Assistant Professor (Sr.Gr), Department of Information Technology, PSG College of Technology, Coimbatore-4, Tamil Nadu, India.

[2] Professor, Department of Information Technology, PSG College of Technology, Coimbatore-4, Tamil Nadu, India.

Abstract - *Nowadays people are more interested to express and share their views, feedbacks, suggestions, and opinions about a particular topic on the web. People and company rely more on online opinions about products and services for their decision making. A major problem in identifying the opinion classification is high dimensionality of the feature space. Most of these features are irrelevant, redundant, and noisy which affects the performance of the classifier. Therefore, feature selection is an essential step in the fake review detection to reduce the dimensionality of the feature space and to improve accuracy. In this paper, binary artificial bee colony (BABC) with KNN is proposed to solve feature selection problem for sentiment classification. The experimental results demonstrate that the proposed method selects more informative features set compared to the competitive methods as it attains higher classification accuracy.*

Key Words: Feature selection, BABC, K-NN, Classification accuracy.

1. INTRODUCTION

Feature selection plays a vital role to remove noisy, irrelevant or redundant features from the dataset. This task improves the accuracy and reduces the computational cost. Many evolutionary algorithms have been used for feature selection, which includes genetic algorithms and swarm algorithms [4]. Swarm Intelligence algorithms such as, Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Bat Algorithm (BAT), and Artificial Bee Colony [7] [8]. Feature selection process can be categorized into two steps namely filter and wrapper approach. The filter model analyzes the intrinsic properties of data independent any learning algorithms. The filter method can perform both subset selection and ranking. Ranking involves identifying the importance of all the features. But this method is precisely used as a pre-process method since it selects redundant features. The wrapper model unlike other filter approaches considers the relationship between features. This method initially uses an optimizing algorithm to generate various subsets of features and then uses a classification algorithm to analyze the subsets generated. Due to the high-dimensional feature space, selecting the optimal feature subset is an NP-complete problem; In order to solve large scale feature selection problems, traditional optimization algorithm is not efficient. Therefore, meta-heuristic algorithms have been extensively applied to solve the feature selection problem. [8]

In 2005, Karaboga developed Artificial Bee Colony (ABC) algorithm which is a new population-based metaheuristic swarm intelligent algorithm. It is based on their foraging behavior of bees. Artificial bee colony algorithm consists of three types of bees and they are employee bee, onlooker bee and scout bees. Employee bee is one that recognizes the food sources depending upon the profitability. It carries all the information about the food source to the hive where the onlooker bees are waiting. The employee bee performs its dancing rituals before the onlooker bees. Each different move of the employee bee represents different direction of the food source from the hive. With the information obtained from the employee bees the onlooker bees select the food source with maximum profitability. Now the employee bee starts to collect food from the source as long as the food source becomes exhausted. Once the food source is exhausted the employee bee forgets the previous information and becomes scout bee and set out in search for new food sources randomly.

In this paper we proposed a binary artificial bee colony based feature selection (BABC) method to select optimal feature subset from the dataset. The remaining of the paper is organized as follows: Section I describes methodology of the proposed work. Section III describes the Experimental results and discussions about of the proposed method. Section IV concludes the proposed work.

2. METHODOLOGY

Fig 1 shows the proposed methodology framework. First the documents are preprocessed into tokens. After tokenization, stemming process is applied to obtain the root words. Then, Stopwords such as articles prepositions are removed from the dataset. Features are extracted using Sentiwordnet. Then the selected features are transformed into vectors by using TFIDF.

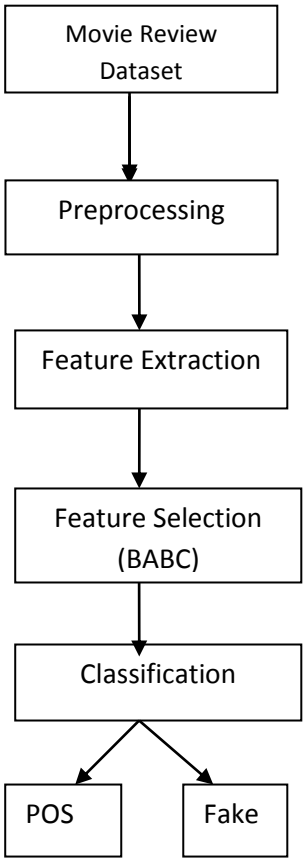

Fig1 Framework for proposed Methodology

Proposed feature selection method using Binary Artificial Bee Colony (BABC)

Binary Artificial Bee Colony (BABC) consists of three bees namely employed bees, onlooker bees and scout bees. The bees move around the search space to find the optimal solutions. Here features are represented as the food sources and nectar amount represent the fitness value of each food source.

In the feature selection problem, the candidate solutions are represented by binary bit string N, where N represents total number of features. If the value at the corresponding position is 1 then the feature is selected as part of the subset to be evaluated or else if the value is 0 then the feature is not selected as part of the subset. The feature subset of each food source is given to the classifier to calculate the fitness value (nectar amount).

Steps for BABC Algorithm:

1. Assign the values for employee bees, onlooker bees and set maximum number of iterations, Pre-determined number of iterations (limit) for scout bees.

2. Initialize the population of food sources X=1, 2…..SN (possible set of features) randomly and allocate it to the employed bees.

$$X_{ij} = X_{minj} + rand(X_{maxj} - X_{minj}) \quad (1)$$

Where X_{maxj} , X_{minj} are the lower and upper bound of dimension j.

Evaluate the Fitness value using the accuracy:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (2)$$

Where TP=True Positive TN=True Negative, FP=False Positive, FN=False Negative.

3. Employed bees: Assign employee bee to its food source and calculate, a new solutions from its neighboring solution, The new feature subset is generated using eq(3)

$$V'_{i,j} = X'_{i,j} + \emptyset'_{i,j}(X'_{i,j} - X'_{k,j}) \quad (3)$$

4. Update Employed bees: For each employed bee, if its candidate solution (food source) is better than its present one, then replace it with the candidate solution.

5. Onlooker bees: For the onlooker bees, select one of the employer bee food sources probabilistically based on the following equation.

$$Probability(P_i) = \frac{f'_i}{\Sigma_{j=1} f'_i} \quad (4)$$

6. Onlooker bee produce new solution for each onlooker then swap any two randomly selected columns of their selected (food source) features. If the new (food source) features fitness value is

greater than the old food source (feature), then the new food source position will be updated in the corresponding employed bee's position Otherwise, do not change the employed bee food source.

7. Scout bees: If there is improvement in the current food source fitness until the pre defined number of iterations (Limit) and make the corresponding bee as a scout bee. Produce a new food source (solution) for scout bee.

8. Memorize the best food source solution obtained so far and repeat from step 3 until stopping criterion is met.

Pseudocode for Binary Artificial Bee Colony (BABC) Algorithm

1. Initialize ABC parameters
   a. population size SN
   b. max number of iterations N
   c. max number of trials T
   d. cycle=1
2. Initialize the food source positions Xi where i=1…….SN,
3. Evaluate the nectar amount (fitness) of food sources
4. Repeat Employed Bees Phase.
a. Construct solutions by the employed bees
   - Assign feature subset (binary bit string) to each employee bee
   - Generate new feature subset Vi
   - Evaluate the fitness value(fit i) of the new feature subset Vi
   - Calculate the probability pi of feature subset solution

b. Construct solutions by the onlookers
   - Select a feature based on the probability pi
   - Compute Vi using Xi and Xj
   - Apply the greedy selection process between Xi and Vi
   - Determine the scout bee and the abandoned solution
   - Calculate the best feature subset of the cycle
   - Memorize the best food source
   - Cycle = Cycle + 1

5. Repeat step 3 to 6 until pre-determined **number of iterations.**

3. K-NEAREST-NEIGHBOR CLASSIFIER

K-Nearest Neighbor (KNN) classification is one of the simplest classification methods and it should the first choices for a classification study when there is no prior knowledge about the distribution of the data. KNN classification was developed from the discriminant analysis which is used to perform when reliable parametric estimates the probability densities which are unknown or difficult to determine. It uses Euclidean distance as the distance measurement. For evaluating the accuracy of the selected subset, the KNN classifier acts as the fitness evaluator.

4. EXPERIMENTAL RESULTS

Dataset description

The evaluation of the proposed method was carried out using the movie review dataset is collected from the website cs.cornell.edu. It contains 2000 reviews consists of 1000 positive and 1000 negative reviews. The dataset is divided into training and testing dataset. From this review dataset, 80% (1480 instances) of the reviews are taken for the training dataset and the other 20% (520 instances) of reviews are taken for the testing dataset with the significant features.

Evaluation Metrics:

The evaluation of the proposed BABC_KNN was carried out using the following metrics.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (5)$$

$$\text{Precision} = \frac{TP}{TP+FN} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP+FP} \quad (7)$$

Where, TN=True Positive FP= False Positive FN= False Negative.TP= True Positive.

BABC_KNN parameter settings:

In order to reduce the computational time, the colony size is set to 40 and the maximum number of iterations is set to be as 100. After initializing the parameters, Features subsets are taken as the input for classifier. In order to evaluate the fitness (Accuracy) of both the employee phase and onlooker phase we uses two classifiers namely KNN and Naïve Bayes classifier to evaluate the selected subsets of features for each dataset. In each case, a 10 cross fold validation is used. The Table I, shows the parameters used for the algorithm and the Table II, shows the selected features in different runs. Table II shows the

accuracy for the BPSO and the modified BPSO using KNN and Naïve Bayes Classifier.

Table I BABC Parameters

Population size P	2*no.of features
Lowerbound	1
Upperbound	N
No.of runs	10
Max number of iterations	100

Table II Classification accuracy

	PSO Features	ABC Features	Proposed BABC Features
KNN	85.5	88.25	96.38
NB	82.5.5	85.46	88.84

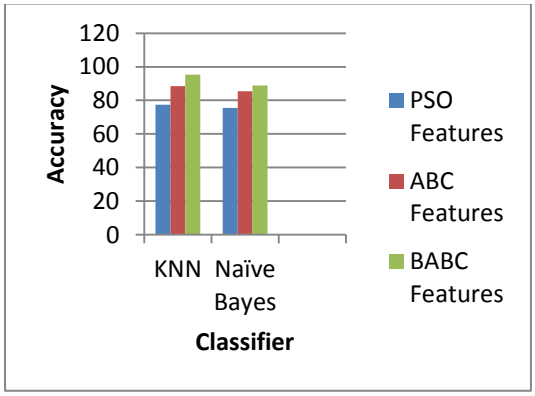

Figure 2 Accuracy

From Table II and fig.2, it is observed that Accuracy is improved for proposed BABC for feature selection when compared to PSO features and ABC features. On an average, Accuracy increases for Proposed BABC by 10.38% when compared to PSO, by 8.13% when compared to ABC.

Tables III and Table IV show the average precision, recall obtained for various feature extraction. The results are shown in Fig 3 and Fig 4.

Table III Precision

	PSO Features	ABC Features	Modified BABC Features
KNN	85.5	89.25	96.48
NB	82.5	86.46	89.82

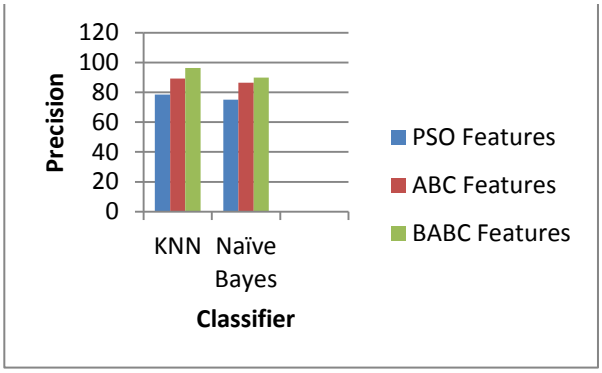

Figure 3 Precision

From Table III and Fig 2, it is observed that Precision is improved for Proposed BABC for feature selection when compared to PSO features and ABC features. On an average, Precision increases for Proposed BABC by 11.13% when compared to PSO, by 7.23% when compared to ABC.

Table IV Recall

	Features PSO	Features ABC	Features BABC
KNN	80.5	89.25	96.48
NB	75.58	86.46	89.88

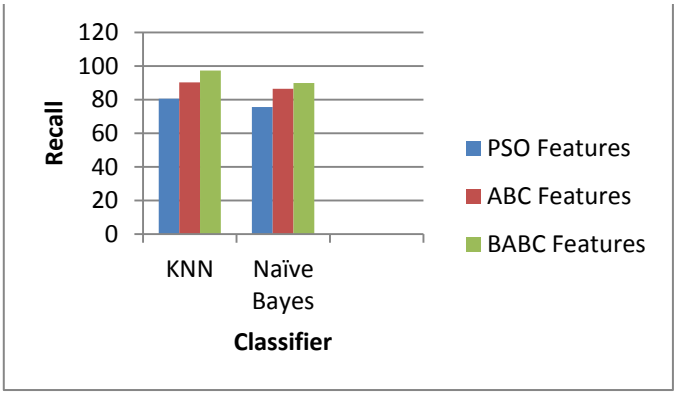

Figure 4 Recall

From Table IV and Fig 3, it is observed that Recall is improved for Proposed BABC for feature selection when compared to PSO features and ABC features. On an average, Recall increases for Proposed BABC by 15.48 % when compared to PSO, by 6.98% when compared to ABC

5. CONCLUSION

In this paper, BABC-KNN based feature selection has been used to classifying the reviews into positive and negative reviews. The proposed BABC based feature selection gives optimal feature subset. Evaluate the accuracy of the selected subset of features, the KNN algorithms is used as the classifier. Experimental results shows that the proposed BABC with k-NN selects the minimal number of features with the highest classification accuracy and reduces the computational complexity compared with other feature set.

REFERENCES

1. Nitin Jindal, Bing Liu, "Analyzing and Detecting Review Spam", Seventh IEEE International Conferenceon Data Maing ,pp.547-552, 2007.

2. Banerjee, S., & Chua, A,Applauses in hotel reviews: Genuine or deceptive? In Science and information conference (SAI) (pp. 938–942). IEEE,2014.

3. Lin, Y., Zhu, T., Wang, X., Zhang, J., & Zhou, A, Towards online anti-opinion spam: Spotting fake reviews from the review sequence. In 2014 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM) (pp. 261–264). IEEE, 2014

4. Banerjee, S., & Chua, A,Applauses in hotel reviews: Genuine or deceptive? In Science and information conference (SAI) (pp. 938–942). IEEE,2014.

5. Aksu, Y, Miller, D. J and Kesidis, G, "Margin–maximizing feature elimination methods for linear and nonlinear kernel–based discriminant functions, IEEE Trans. Neural networks, vol. 21, pp.701–717, 2010.

6. Gutlein, M, Frank, E and Karwath, A, "Large–scale attribute selection using wrappers, In IEEE symposium computational intelligence and data mining, pp. 332–339, 2010.

7. Alper, U, Alper, M. and Ratna, B. C, "PSO:A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification, Information Science, vol. 181, pp. 4625-4641, 2011.

8. Vieira, S. M, Luis, F, Mendonca, L. F, Farinha, G. J and Sousa, M. C. J, "Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients, Applied soft computing vol. 13, pp. 3494–3504, 2013.

9. Thananan Prasartvit, Anan Banharnsakun, Boonserm Kaewkamnerdpong and Tiranee Achalakul, "Reducing bio– informatics data dimension with ABC–K-NN, Neurocomputing, vol. 116, pp. 367–381, 2013.

10.Farhadloo.M, Rolland.E, "Multi-Class Sentiment Analysis with Clustering and Score Representation" 13th International Conference on Data Mining Wokshop,IEEE,2013.

11. S.Njolstd.P, S.Hoysaeter.L, Wei.W and Atle Gull.J. "Evaluating Feature Sets and Classifiers for Semtiment Analysis of Financial News," IEEE/WIC/ACM International Joint Conferences on Web Intelligence (W1) and Intelligent Agent Technologies (IAT), IEEE, 2014.

12. Pampara G, Engelbrecht A P: Binary artificial bee colony optimization. In: IEEE Symposium on Swarm Intelligence, IEEE, Perth, pp. 1-8 (2011)

13. Kashan M H, Nahavandi N, A.H. Kashan: DisABC: A new artificial bee colony algorithm for binary optimization. Applied Soft Computing, 12(1), pp. 342-352 (2012).

14. C. Ozturk and D. Karaboga. Hybrid artificial bee colony algorithm for neural network training. In *Evolutionary Computation (CEC), 2011 IEEE Congress on,* pages 84-88. IEEE, 2011.

15. Fei Kang, Junjie Li, Haojin Li, Zhenyue Ma and Qing Xu, An Improved Artificial Bee Colony Algorithm,Proc. IEEE International Workshop on Intelligent Systems and Applications, 2010, pp 1 – 4.

16. Abdul-Rahman, A.A. Bakar, Z.A. Mohamed-Hussein, Optimizing big data in bioinformatics with swarm algorithms., in: In: 16th International Conferenceon Computational Science and Engineering (CSE), Sydney, NSW, IEEE, 2013, pp.1091–1095.

17. Umamaheswari K, Rajamohana SP, "Opinion Mining using Hybrid Methods", International Journal of Computer Application, pp 18-21, 2015.

18. Rajamohana SP &Umamaheswari K, "Sentiment Classification based on LDA using SMO Classifier", International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10, No.55 ,pp. 1045-1049, 2015.

19. Rajamohana SP & Umamaheswari K, "Sentiment Classification based on Latent Dirichlet Allocation", International Journal of Computer Application, pp 14-16.