Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/110522

A Hybrid Approach to Optimize Feature Selection Process Using iBPSO- BFPA for Review Spam Detection

*SP. Rajamohana and K. Umamaheswari**

Department of Information Technology, PSG College of Technology, Coimbatore-641004, India.

Received: 3 Apr. 2017, Revised: 20 May 2017, Accepted: 25 May 2017
Published online: 1 Sep. 2017

Abstract: With the increase in the customer reviews, feedbacks, suggestions posted in the web forum, blogs led to the emergence of spam. Spam detection is important for both the customer and service providers to arrive at a proper decision while purchasing as well as marketing the product. Most of the research works has been developed only for sentiment classification for the past few decades which favors the spammers to write fake reviews. Hence it is important to detect the spam reviews but the major issues in spam review detection are the high dimensionality of feature space which contains redundant, noisy and irrelevant features. To resolve this, optimization method for selecting subset of features is necessary. Hence, this paper proposes Hybridization of Improved Binary Particle Swarm Optimization (iBPSO) and Binary Flower Pollination Algorithm (BFPA) utilized with Naive Bayes and k-NN for optimization process to improve the classification performance. Experimentation result proves that hybrid iBPSO_BFPA outperformed the existing approach by obtaining the maximum accuracy of 94.43% for review spam dataset when compared with existing Cuckoo Search_NB(CS) and Shuffled Frog Leaping Algorithm_NB (SFLA) which achieved only 81.87% and 88.23%. The experimental result proves that the proposed hybrid method increases the classification accuracy.

Keywords: Feature selection, Improved Binary Particle Swarm Optimization (iBPSO), Binary Flower Pollination Algorithm (BFPA), Classification, Naive Bayes (NB) and k-Nearest Neighbor (k-NN).

1 Introduction

Online reviews have been popularly used in lots of applications. Because they can either encourage or harm the status of a product or a service, buying and selling. Fake reviews becomes a beneficial business and a big threat. Online reviews are generally used in many websites like Amazon, Yelp, and TripAdvisor for allowing users to share their personal experiences. Positive reviews can increase reputations and bring significant financial gains, while negative ones often cause dramatic sales loss. This information regrettably results in strong motivation for review spam i.e. writing fake reviews to mislead readers. Recently, many research works have been developed for review spam detection with help of different techniques and algorithm. Rule induction algorithms in order to resolve the tie that appear in special cases during the rule generation procedure for various data sets [1]. To find the greenery and used lands of the study area by classifying the satellite imagery they have used fuzzy-based, K-nearest neighbourhood, support vector machine classification methods [2]. For example, review graph model was designed in [3] to identify the relationships between reviewers, reviews and stores for detecting spammers. The problem of detecting spam online reviews from imbalanced data distributions was addressed in [4] where classifier technique was used for predicting review spam. E-Mail Abstraction Scheme was presented in [5] for effectively detain the near-duplicate phenomenon of spams. Hybrid of the Naive Bayes and K-means clustering was introduced in [6] for improved classification accuracy of spam detection. Spam Review Detection using a Hybrid Classification Method [7] was developed to detect whether a review is spam or not. Search Engine Spam Detection was presented in [8] for detecting search engine spam by using the integrated hybrid genetic based feature selection and decision tree classification. The survey of different spam detection methods was analyzed in [9]. Detecting Spam Review through Sentiment Analysis was designed in [10] to detect the spam store and spam review efficiently. In [11], Review Spam Detection was developed for identifying untruthful review spam detection using n-gram model and

* Corresponding author e-mail: monamohanasp@gmail.com

features selection techniques. In [12], the document clustering method was introduced for clustering and analyses of spam messages. A novel Hybrid Optimization Algorithm is implemented [13] using Improved Binary Particle Swarm Optimization (iBPSO) and Cuckoo Search in the Review Spam Detection. A Survey on Review Spam Detection techniques was presented in [14] to correctly identify the review as a spam or not. A novel review spam detection method was introduced in [15] using High-Order Concept Associations mining and Inferential Language modeling to improve the performance of untruthful review detection. An effective hybrid approach having Cuckoo with Harmony Search was proposed [16] for feature extraction in the review Spam Detection problem.

The paper is organized as follows: Section II, III describes the various feature selection algorithms. Section IV, explains the proposed methodology. Section V defines the Classification. Section VI states the experimental setting and results are compared with some other state of art methods. Conclusion and future work are discussed in section VII.

2 Improved Binary Particle Swarm Optimization (iBPSO)

Kennedy and Eberhart [17], considering the swarms behavior in nature, such as flock of birds, school of fish, etc. introduced the PSO algorithm which has particles driven from natural swarms with communications based on its evolutions. Here, candidate solution is a particle. It utilizes a collection of moving particles (changing solutions) in a search area (current and possible solutions) as well as the movement towards a promising region to reach a global optimum. The primitive concept of PSO is accelerating each particle towards its particle best and the global best positions. It deals with the individual candidate solution and can obtain the efficient optimization in a given problem. Assume that the search space is n-dimensional and the i-th particle of the swarm can be represented as a n-dimensional position Vector $Y_i = (y_{i1}, y_{i2}, \ldots, y_{in})$. The velocity of the particle can be represented as $V'_i = (v_{i1}, v_{i2}, \ldots, v_{in})$. The best visited position of a particle is $P_{iBest} = (p_{i1}, p_{i2}, \ldots, p_{in})$. And the best position explored in the entire swarm can be presented as $P_{gBest} = (p_{g1}, p_{g2}, \ldots, p_{gn})$. So the position of the particle and its velocity is being updated until the fitness value converges. The problem in PSO is it tends to fast and premature convergence in the mid optimum points. To solve this problem Improved Binary PSO is proposed. The convergence factor is combined with Linearly Decreasing Inertia Weight (LDIW) to improve the performance of BPSO [12]. For instance, in the best value for W' in (1) is set to 0.9, which linearly decreases to 0.4. The improvisation rules in the iBPSO algorithm are given as follows.

$$V'_{(t+1)} = \lambda' \left(W' \times \left(V'_{(t)} + C_1 \times random(0,1) \right) \times pBest'_{(t)} \right) - (currvalue_t + C_2 \times random(0,1)) \quad (1)$$

$$currvalue_{(t+1)} = currvalue_t + V'_{(t+1)} \quad (2)$$

$$\text{LDIW} = \left(w'_{start} - w'_{end} \right) \left(\frac{T_{max} - t}{T_{max}} \right) + w'_{end} \quad (3)$$

Where c_1, c_2 the acceleration parameters, are set to 2, *random()* = A random number between 0 and 1, *W':* Inertia weight.

Pseudo code for Improved Binary Particle Swarm Optimization:

Define Objective function max f '(x), where x= $(x_1, x_2, \ldots\ldots x_d)$
Initiate the PSO parameter coefficients-population size, no of iterations
Assign random positions, velocities for the particles
Initiate Personal best position
Calculate the fitness value of the particles
Determine particle best position of the swarm
Loop
Update the velocities as in equation (1)
Update the positions as in equation (2)
Determine personal best position
Determine global best particle of the swarm
Until stop criterion(Max no.of iterations)

3 Binary Flower Pollination Algorithm (BFPA)

Flower pollination algorithm is designed based on the idea behind the pollination process of flowering plants, the exchange of pollens. There exists Biotic pollination process where pollen is carried by Pollinators like insects, birds, bats, flies etc and an Abiotic pollination which occurs naturally via wind or diffusion in water. Pollination can be carried out by two ways such as self-pollination and cross-pollination. Abiotic/Self-pollination happens in one flower between the pollens of same flower or different flowers of the same plant. Biotic/Cross-pollination process happens between flowers of different plants with the help of pollinators. Pollinators are responsible for Biotic/Cross-pollination which covers longer distances in the transfer of pollens. So, these pollinators carryout the global pollination process. Xin-She Yang describes this flower pollination process in the following four rules [18]:

1. Biotic and cross-pollination where external pollinators performing Levy flights distribution to transfer pollens in to different flowers is considered as global pollination process.

2. Abiotic and self-pollination are considered as local pollination.
3. Flower constancy can be considered as the reproduction probability is proportional to the similarity of two flowers involved.
4. Local pollination and global pollination is controlled by a switching probability $p \in [0, 1]$.
The global pollination can be mathematically written as

$$x_i^{t+1} = x_i^t + \gamma' Le'(\lambda')(x_i^t - g*) \quad (4)$$

where x_i^t is the pollen i (solution vector) at the iteration level t and $g*$ is best solution so far, while γ' is the scaling parameter, s isthe step size, $Le'(\lambda)$ is the Levy's flight step size.

The step size Le' is derived from Levy flight distribution,

$$Le'(\lambda') = \frac{\lambda'.\Gamma(\lambda') \cdot \sin(\lambda')}{\pi} \cdot \frac{1}{s^{1+\lambda}}, \ S > 0 \quad (5)$$

where, (Γ- Standard gamma function and $\lambda' = 3/2$).

Local Pollination can be illustrated as follows,

$$x_i^{(t+1)} = x_i^t + \varepsilon'(x_j^t - x_k^t) \quad (6)$$

x_i^t - solution vector at iteration t, $x_i^{(t+1)}$- solution vector at iteration $t+1$, ε' – a random number ranges from [0, 1] and j,k - arbitrarily selected indices. In which $x_i^j(t)$ denotes the new pollen solution, i with the j th feature r, where $i = 1,2,\ldots m$ and $j = 1,2,\ldots,n$, at the iteration t and $\varepsilon'' U$ (0, 1)[21].

Pseudocode for Flower Pollination Algorithm:

Define Objective function max f '(x), where X= $(x_1, x_2, \ldots\ldots x_d)$
Initialize the population of'n' flowers/pollen gametes with random solutions
Find out the best solution B in the initial population
Define a switch probability p ∈ [0, 1]
Define a stopping criterion (number of iterations)
while (t <MaxGeneration)
for i = 1 to n(all n flowers in the population)
if random <p
Calculate Global pollination via equation 4
else
Calculate Local pollination via equation 6
end if
Evaluate the new solutions
if new solutions are better, update them in the population
end for
*Find the current best solution g**
end while
Output the best solution obtained

4 Hybrid IBPSO_BFPA For Review Spam Detection

Hybrid optimization Algorithm iBPSO_BFPA is proposed for feature selection. The merits of PSO are easier understandability, simpler process and faster searching. However, PSO converges in to a local optimum value in solving a large complicated problem. BFPA has advantages such as it includes very few control parameters and high efficiency, but it also possesses slow convergence rate. In BFPA, the Lévy distribution has the capability to generate new solutions which makes the search process quicker. It is more likely to not being trapped in the local optimal points. However, the randomness of the Levy flight leads to reduction in the convergence speed and search efficiency in obtaining the optimal solution. The proposed algorithm overcomes the drawbacks of PSO and utilizes the advantages of BFPA for optimization process. Hybrid iBPSO_BFPA algorithm, each particle positions in the iBPSO is updated using Levy Flight distribution in the Flower pollination to attain optimal solution convergence. PSO can successfully prevail over local optima convergence and the performance of searching for the optimal solution is improved. Hybrid Algorithm is defined as follows.

Step 1: Initialize a population of particles with random positions and velocities on search space D.

Step 2: For each particle, evaluate fitness using classification accuracy.

Step 3: Pbest and Gbest values are obtained for entire population.

Step 4: Compare fitness evaluation with the populations overall previous best. If current value is better than gbest, then update the current particle.

Step 5: Change the velocity and position of the particle according to equations (1) and (2) respectively.

Step 6: Loop to step 2 until a criterion (the maximum number of iterations) is met.

Step 7: The population/Flower size 'F' of possible solution is defined by a group of virtual flowers (n).

Step 8: Flowers are expressed in a vector X= $(x_1, x_2, \ldots x_d)$, where d represents number of features.

Step 9: The best particle position found by PSO is regarded as initial points for BFPA algorithm.

Step 10: Switching probability p is chosen.

Step 11: If random value is less than p then global pollination via Levy flight is performed as equation 4.

Step 12: Else the local pollination according to equation 6 is performed.

Step 13: Evaluate new solutions when better solution is found, update them in the population.

Step 14: Find the current best solution.

Step 15: Output the best solution obtained.

5 Classification

5.1 Naive Bayes

Naive Bayes Classifier (NB) is a simple stochastic classifier derived from the Bayes theorem with strong (naive) independence assumptions [10]. It assumes each term is independent of each other. Depending on the precise nature of the probabilistic model, naive Bayes classifiers are trained efficiently as a supervised learning approach. By the assumption of an underlying probabilistic model allows us to identify the ambiguity about the model in a standard way by determining the probabilities of the outcomes. It can be helpful in diagnostic and predictive problems. This theorem determines the relationship between the updated probability P'(A|B), the conditional probability of A given the new knowledge B, and the probabilities of A and B,P'(A) and P'(B), and the conditional probability of B given A, P(B'|A'). Bayesian techniques uses an analytical formulae to analyze the content of the review text. Bayes theorem is represented as

$$P\left(A'|B'\right) = \frac{P\left(B'|A'\right) \cdot P(A')}{P(B')} \tag{7}$$

Bayesian classifier works on dependent events and the probability of an event occurrence in the future that can be predicted from the previous occurrence of the same event. Hence NB technique has become a very popular method in the text/spam categorization. The genuinity for a review R is calculated using the formula

$$S'[R] = \frac{T_{Spam(R)}}{T_{Spam(R)} + T_{Ham(R)}} \tag{8}$$

5.2 K-Nearest Neighbor

k-Nearest Neighbor (k-NN) is a supervised learning algorithm which classifies based on the majority of the nearest neighbors to k. It classify new objects based on the attributes. The classification is performed without the usage of a model but only based on a memory [2], [11]. k-NN works based on the minimum distance from the new data to the k's nearest neighbors. Once the k's nearest neighbors are obtained, predictions of a new class of data will be determined based on the majority of the k's nearest neighbors. To calculate the distance between the any two attributes say X and Y, Euclidean Distance formula is used. The algorithm for k-Nearest Neighbor classification data works as follows: First, Calculate the distance from all training data and test data using the following Euclidean distance formula [11]. After the data obtained, sort data from the smallest to largest after it took the majority of the data as much of the value of K to see the results of the diagnosis. Finally, the obtained the data value of K, is compared with each data value to find the smallest distance.

Table 1: Parameter Settings for iBPSO

Parameter	Values
PSO Swarm size	50
No.of Iterations	500
C1,C2	2
W(LDIW)	(0.4-0.9)

Table 2: Parameter Settings for BFPA

Parameter	Values
FPA Population size	50
No.of Iterations	500
ε'	(0,1)
λ	1.5
p	0.8

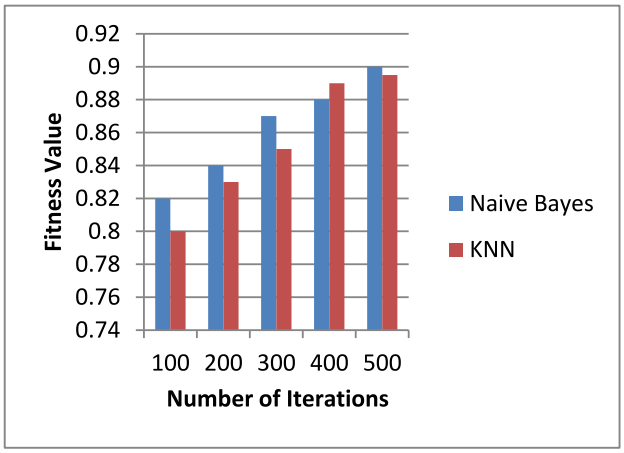

Fig. 1: Fitness value obtained for iBPSO.

6 Experimental Results

6.1 Data Set

The dataset used in the experimentation is opinion review hotel dataset. Opinions about the 20 most popular Chicago hotels. The opinion review hotel dataset includes examples of positive and negative deceptive opinion spam to conduct supervised learning and is liable evaluation of the task. This corpus includes a total of 1600 labeled examples of deceptive and truthful review. The corpora comprises of 400 truthful positive reviews, 400 truthful negative reviews, 400 deceptive positive reviews and 400 deceptive negative reviews. Deceptive opinions were generated using the Amazon Mechanical Turk, whereas (likely) truthful opinions were mined from reviews on TripAdvisor, Expedia, Hotels.com, Orbitz, Priceline, and Yelp [19].

The Table 1, 2 below specifies the parameter settings for iBPSO and BFPA algorithm.

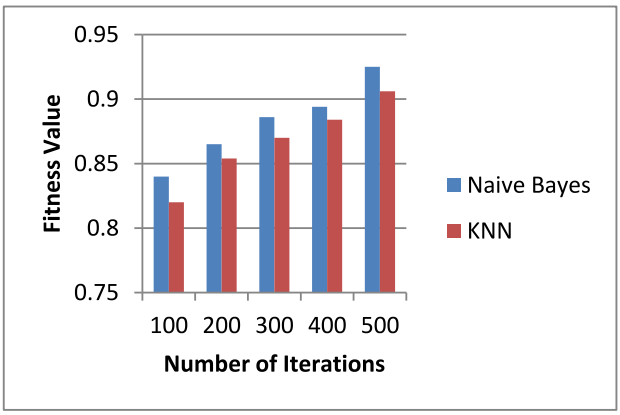

Fig. 2: Fitness value obtained for BFPA

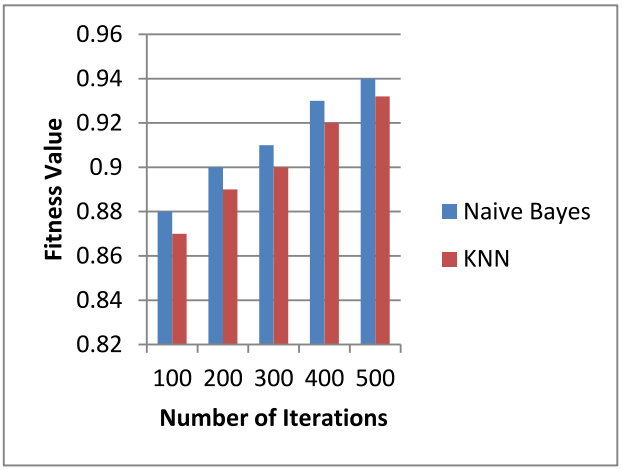

Fig. 3: Fitness value obtained for hybrid iBPSO_BFPA

From the Figure 1, 2 and 3 it is observed that the Fitness values of hybrid iBPSO_BFPA is higher than the standard iBPSO by 4.42 % and BFPA by 2.45%.

Figure 4 shows the reduction in the features count as the iteration increases in the hybrid iBPSO_BFPA compared with standard iBPSO and BFPA.

Table 3: Performance Metrics on various Feature Selection Algorithm

Algorithms	Accuracy	Precison	Recall
SFLA	0.8782	0.8628	0.8512
IBPSO	0.9047	0.8854	0.8795
Binary Cuckoo Search	0.9138	0.9063	0.8916
BFPA	0.9281	0.9191	0.9073
Hybrid iBPSO_BFPA	0.9443	0.9302	0.9274

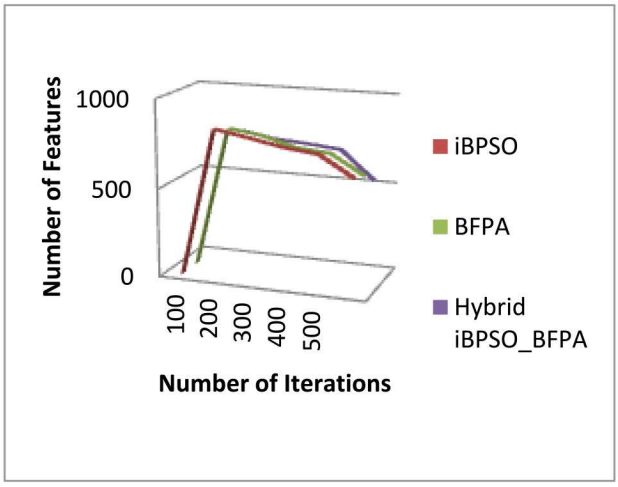

Fig. 4: Comparative analysis of Features count on Various feature Selection Algorithms

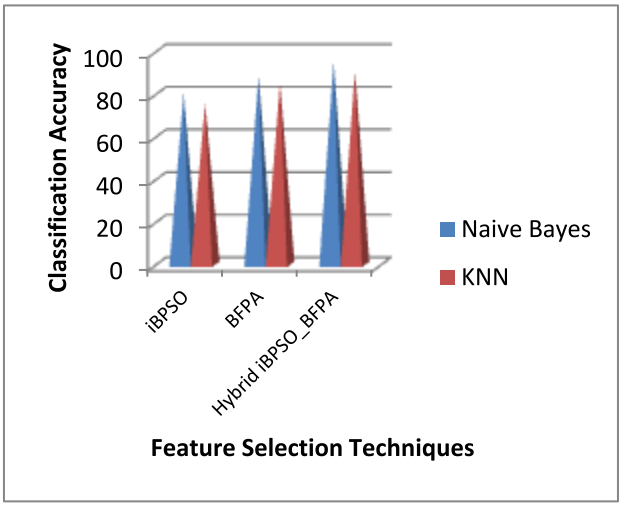

Fig. 5: Classification performance of iBPSO, BFPA and proposed Hybrid iBPSO_BFPA

Figure 5 explains the novel hybrid iBPSO_BFPA outperforms the other two standard iBPSO and BFPA algorithms in terms of classification and the accuracy is 94.42 %. The Table 3 shown below describes comparison of the performance metrics of different algorithms.

7 Conclusion

In this paper, the proposed hybrid iBPSO_BFPA algorithm for feature subset selection has been implemented and evaluated for the hotel review dataset. This algorithm combines the PSO algorithm's

strengths in which the ability to operate faster with good performance and the BFPA algorithm's faster convergence rate as well as global search. Therefore, it possesses a trade-off between neglecting premature convergence and exploring the whole search region. Hence, the proposed method effectively selects the feature subset which in turn increases the classification accuracy. Experimental results obtains better search performance and it outperforms well when it is compared with other well known evolutionary algorithms such as iBPSO, BFPA, Cuckoo Search and SFLA.

References

[1] S. Appavu alias Balamurugan, "Effective solution for unhandled exception in decision tree induction algorithms", ACM digital Library 2009.

[2] Suresh Kumar Nagarajan, Shanmugam Saravanan, "Content-based medical image annotation and retrieval using perceptual hashing algorithm", International Organisation of Scientific Research Journal, 2012.

[3] Guan Wang, Sihong Xie, Bing Liu, Philip S. Yu, "Review Graph based Online Store Review Spammer Detection", 2011 IEEE 11th International Conference on Data Mining (ICDM), Pages: 1242 – 1247, 2011.

[4] Hamzah Al Najada and Xingquan Zhu, "iSRD: Spam Review Detection with Imbalanced Data Distributions",2014 IEEE 15th International Conference on Information Reuse and Integration (IRI), Pages: 553 – 560, 2014.

[5] Chi-Yao Tseng, Pin-Chieh Sung, and Ming-Syan Chen, "Cosdes: A Collaborative Spam Detection System with a Novel E-Mail Abstraction Scheme" IEEE Transactions on Knowledge and Data Engineering, Volume 23, Issue 5, 2011.

[6] Nadir Omer FadlElssied, Othman Ibrahim, "K-Means Clustering Scheme for Enhanced Spam Detection", Research Journal of Applied Sciences, Engineering and Technology, Volume 7(10), Pages: 1940-1952, 2014.

[7] KishanThumar, DulariBosamiya, "Evaluate spam detection using hybrid technique of Support Vector Machine" International Journal of Engineering Development and Research, Volume 3, Issue 4, Pages: 807-811, 2015.

[8] D. Saraswathi, A. Vijaya, "Search Engine Spam Detection using an Integrated Hybrid Genetic Algorithm based Decision Tree", International Journal of Computer Applications, Volume 133, Issue 10, Pages: 20-27, 2016.

[9] Rekha, SandeepNegi, "A Review on Different Spam Detection Approaches", International Journal of Engineering Trends and Technology (IJETT), Volume 11, Number 6, Pages: 315-318, 2014.

[10] Qingxi Peng and Ming Zhong, "Detecting Spam Review through Sentiment Analysis", Journal of Software, Volume 9, Issue 8, Pages: 2065-2072, 2014.

[11] Kustiyo, Aziz, Classification Methods: Quantitative Methods Lecture, Department of Computer Science FMIPA IPB. 2010.

[12] SP. Rajamohana and K. Umamaheswari, "Hybrid Optimization algorithm of improved Binary Particle Swarm optimization (iBPSO) and Cuckoo Search for review spam detection", Proceedings of the 9th International Conference on Machine Learning and Computing, Singapore 2017.

[13] Manali S. Patil, A. M. Bagade, "Online Review Spam Detection using Language Model and Feature Selection" International Journal of Computer Applications, Volume 59, Issue 7, Pages: 33-37, December 2012.

[14] Rasim M. Alguliev, Ramiz M. Aliguliyev, and Saadat A. Nazirova, "Classification of Textual E-Mail Spam Using Data Mining Techniques", Hindawi Publishing Corporation, Applied Computational Intelligence and Soft Computing, Volume 2011, Article ID 416308, 8 pages.

[15] Vyas Krishna Maheshchandra, Ankit P. Vaishnav, "A Survey on Review Spam Detection techniques"International Journal of Engineering Research & Technology (IJERT),Volume 4, Issue 04, Pages: 368-371, April-2015

[16] SP. Rajamohana, Dr. K. Umamaheswari, "An Effective Hybrid Cuckoo Search with Harmony Search for Review Spam Detection", Proceedings of the 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB17), 2017.

[17] Kusum Kumari Bharti & Pramod Kumar Singh, "Opposition chaotic fitness mutation based adaptive inertia weight BPSO for feature selection in text clustering", Appl. Soft Computing, Vol. 43, pp. 20-34, 2016.

[18] Yang, X.S, "Flower pollination algorithm for global optimization", UCNC'12 Proceedings of the 11th international conference on Unconventional Computation and Natural Computation (240-249), 2012.

[19] Myle Ott, YejinChoi, Claire Cardie&Jeffrey T. Hancock, "Finding deceptive opinion spam by any stretch of imagination", in the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, pp. 309-319, ACM, 2011.

[20] Haiyi Zhang, Di Li, "Naïve Bayes Text Classifier" IEEE International Conference on Granular Computing, 2007.

[21] SP. Rajamohana, K. Umamaheswari and B. Abirami "Adaptive Binary Flower Pollination Algorithm for Feature Selection in Review Spam Detection", Proceedings of Innovations in Green Energy and Healthcare Technologies, 2017.

SP. Rajamohana is working as Assistant Professor (Sr.Gr) in the Department of IT, PSG College of Technology, Coimbatore, and TamilNadu, India. She obtained her Bachelor's degree from Thiagarajar College of Engineering in 2006. She received her Master degree from PSG College of Technology in 2008. She is currently doing research in the area of Review spam detection. Her research areas include Data mining, Evolutionary Computation, Software Engineering and Open source systems.

K. Umamaheswari is working as a Professor in the Department of IT, PSG College of Technology, Coimbatore, Tamilnadu, India. She received her PhD in the year of 2010. Her research areas include classification techniques in data mining and other areas of interest are information retrieval, software engineering, theory of computation and compiler design. She has published more than 50 papers in international, national journals and conferences. She is the editor for National Journal of Technology, PSG College of Technology and reviewer for many national and international journals.