Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/110422

Performance Analysis of iBPSO and BFPA Based Feature Selection Techniques for Improving Classification Accuracy in Review Spam Detection

SP. Rajamohana[*], *K. Umamaheswari and B. Abirami*

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 Jul. 2017

Abstract:
Objectives: Feature Selection is a important technique to reduce the quantity of features in various application domains where the data set includes thousands of features in it. The main objective of this study is to choose BFPA for feature selection to obtain better fitness function.
Methods/Statistical Analysis: Improved Binary Particle Swarm Optimization (iBPSO) approach is used for selecting the subset from the dataset and providing the better fitness values. Here, iBPSO used to obtain the feature subset and its performance is compared with BFPA. Naïve Bayes classifier is used to improve the classification accuracy.
Findings: The experimental results shows that BFPA overall accuracy is improved 0.7% in using NB and 0.4% using k-NN as compared to iBPSO.
Application/Improvements: To reduce the complexity and increase the accuracy the BFPA is used. Performance analysis shows that BFPA outperforms the iBPSO.

Keywords: Review Spam Detection, Feature Selection, BFPA, iBPSO, Classification.

1 Introduction

We are in an Internet era where people tends to spend more time online to buy as well as the vendors to sell their products and offer services. They often look into the reviews available on forums online to make their decisions. But the problem is the reviews posted online are not completely true. It may have fake reviews which can be a marketing one for own products as well as a demoting one for their competitors products and services. Detection of fake reviews from the huge data available online is complex. So, techniques to detect review spam are acquiring importance nowadays. Many researchers have been working on them. Feature selection are of two kinds namely Filter and wrapper where filter applies a statistical measure to assign a score to each feature on which they are ranked and either chosen to be kept or removed from the dataset. The methods are often univariate and consider the features independently, whereas wrapper considers selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared. The search process involves a best-first search, it may be stochastic such as a random hill-climbing algorithm, or it may use heuristics evaluations, like forward and backward passes to add and remove features. Umamaheswari et al. [1] proposed Consistency based Feature Selection Approach for Improving Text Classification Performance. Rajamohana et al. [2] proposed integrated evolutionary algorithms for review spam detection. Karaboga et al. [3] introduced a Artificial Bee Colony algorithm to solve multi-dimensional optimization problem. Yang et al. [4] presented the Flower pollination algorithm for global optimization. Rajamohana et al. [5] proposed a hybrid Cuckoo with Harmony Search for feature extraction.

The structure of the paper is organized as follows. Section 2 includes the Methodology used. Section 3 presents the classification techniques. Section 4 describes the performance measures. Section 5 details the experimental results of the system. Section 6 presents the

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

discussions about the results observed. Section 7 concludes the main contributions of this paper.

2 Methodology

In this paper the iBPSO and BFPA for Feature selection is implemented and their performance is analyzed. Figure 1 shows the steps of the methodology for feature selection.

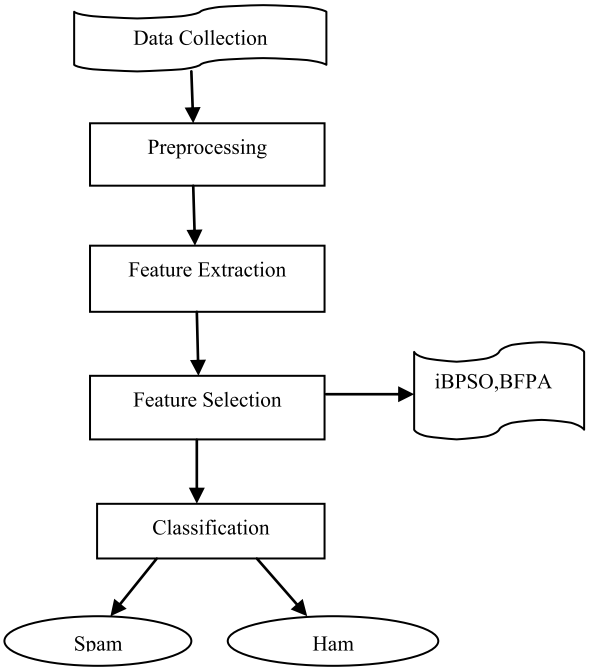

Fig. 1: Workflow of the methodology

2.1 Feature Selection Based on Ibpso

Kennedy and Eberhart introduced an optimization algorithm, Particle Swarm Optimization (PSO) [6],[10] which operates on a social flocking behavior of birds. PSO is based upon the movement and intelligence of swarms. The swarm size or population size is used as a possible solution to reach promising regions present in the search space.Particle Swarm Optimization is a computation tool for optimizing the specific problem and brings out informative candidates. The basic concept of PSO is accelerating each particle toward its particle best and the global best positions. It deals with the candidate solution and can obtain the quality of optimization. Assume that the search space is m-dimensional and the i-th particle of the swarm can be represented as a m-dimensional position Vector X$_i$=$(x_{i1}, x_{i2}, \ldots, x_{im})$. The velocity of the particle can be represented as $V_i = (v_{i1}, v_{i2}, \ldots, v_{im})$. Also consider the best visited position of the particle is $P_{ibest} = (p_{i1}, p_{i2}, \ldots, p_{im})$. And also the best position explored so far can be as $P_{gbest} = (p_{g1}, p_{g2}, \ldots, p_{gm})$. So the position of the particle and its velocity is being updated until the value of the fitness task converges. The convergence factor is combined with Linearly Decreasing inertia weight (W') to improve the performance of BPSO.

The updation rules in the iBPSO algorithm are given as follows.

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

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

$$W' = \frac{(w'_{start} - w'_{end})}{T_{max} + w'_{end}} * T_{max} - T \quad (3)$$

2.2 Feature Selection Based on BFPA

Flower Pollination Algorithm (FPA) is a global optimization algorithm, which was designed by Xin-She Yang in 2012 [4], inspired by the natural pollination process of flowering plants. It includes two key processes in FPA. One is global pollination otherwise known as cross or biotic pollination and the other is local or abiotic pollination. In the global pollination step, insects fly and moves for a longer distance and the fittest is represented by the g*. The flower pollination process with a longer distance covered is carried out with levy flights distribution. Mathematically, the global pollination process is represented as

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

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'* is the step size, *L'*(λ') is the Levy's flight step size. The step size L is drawn from Levy flight distribution,

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

where, (Γ-Standard gamma function and $\lambda' = 1.5$).

Local Pollination can be illustrated as follows,

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

x_i^t-solution vector at iteration $t, x_i^{(t+1)}$-solution vector at iteration $t+1, \varepsilon'$ – a random number ranges from [0, 1] and j,k - randomly 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 $j = 1, 2, \ldots, d$, at the iteration t and ε' belongs to (0, 1).

3 Classification

3.1 Naive Bayes

Naive Bayesian classifier is statistical classification system based on the Bayes theorem. It predicts membership probabilities, such as which class a specific tuple belongs to. It assumes it as a class conditional independence where it accepts the effect of an attribute value of a given class is independent of the value of the other attributes. This classifier is linear which is very simple to implement, fast to train,of total probability, given the vector $\overrightarrow{X} = X_1, X_2, \ldots, X_n$, of a text d, the probability that d belongs to category C is:

$$P\left(c = C \middle| \overrightarrow{X} = \overrightarrow{x}\right) = \frac{P\left(\overrightarrow{X} = \overrightarrow{x} \,|c = C\right)}{\sum_k P\left(\overrightarrow{X} = \overrightarrow{x} \,|c = k\right)} \tag{7}$$

It is well suited especially in the areas of document categorization and disease prediction. Some examples include the diagnosis of diseases and decisions making about treatment processes the classification of RNA sequences in taxonomic studies and spam filtering e-mail clients. In a text classification, the words (or terms/tokens) of the document are used in order to classify it on the appropriate class. We applied Naïve Bayes classifier on our hotel review dataset for the problem of review spam detection. This classifier gives around 92% accuracy for BFPA and 88 % accuracy for iBPSO.

3.2 k-Nearest Neighbour

k-Nearest Neighbour (k-NN) algorithm is one among the supervised classification algorithms that is being used in many domain areas like data mining, statistical pattern recognition etc. It follows a technique of classifying the objects with respect to the closest training examples present in the feature space. The *k*-NN classifier evaluates the closeness between the features and each training feature instance and uses the class labels of the *k* most similar neighbors to predict the class of the feature. “Closeness” is characterized as a separation metric, for example, Euclidean separation.

$$\begin{aligned} X_1 &= (x_{11}, x_{12}, \ldots x_{1n}), \\ X_2 &= (x_{21}, x_{22} \ldots x_{2n}), \\ Euc.D(X_1, X_2) &= \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \end{aligned} \tag{8}$$

The Euclidean distance (Euc.D) between two features selected from the dataset, is represented in equation 8. The underlying assumption is that features belonging to the same class will cluster together in the vector space. The value K must always be a positive integer. The neighbours are selected from a set of objects for which the correct classification is known. K-NN works well even in the presence of some missing data. K-NN classifier gives around 86% accuracy for BFPA and 83 % for iBPSO accuracy.

4 Performance Measures

To evaluate performance we calculated Precision (PR), Recall (RE), F-measure (F1’) and Accuracy (AC). Let TN: the number of legitimate reviews classified as legitimate (True negatives), TP: the number of spam reviews classified as spam (True positives), FP: the number of legitimate reviews classified as Spam (False Positives) and FN: the number of spam reviews classified as legitimate (false negatives), then we have:

$$PR = \frac{TP}{TP+FP} \tag{9}$$

$$RE = \frac{TP}{TP+FN} \tag{10}$$

$$F1' = \frac{2 \times PR \times RE}{PR+RE} \tag{11}$$

$$AC = \frac{TP+TN}{TP+FP+TN+FN}. \tag{12}$$

5 Experimental Results

The dataset used in the experimentation is Reviews about the 20 most popular Chicago hotels. It contains 1600 truthful and deceptive opinion [9]. Experiments are implemented in Windows 8 environment with core i4 processor and i4 Processor and 16 GB RAM. The optimization procedures developed in this work based on the two approaches iBPSO and BFPA. We used supervised learning to build review spam model based on the hotel features. In a hotel review, to find whether a review is spam or genuine, a model or classifier is constructed with class labels, such as “Spam” or “Ham”. Firstly Pre-processing techniques such as tokenization, stopwords removal and stemming were applied. To convert the text into numeric format TFIDF method is used. Features are extracted using iBPSO and BFPA. A supervised learning model builds the classifier by analyzing or “learning from” a training set made up of instances and their associated class labels. We used 10-fold cross validation to calculate performance of the system. Subsequently we applied Naive Bayes and k-NN as classifiers. We splitted the data set for training as well as test set and conducted 10-fold cross validation: the data set is randomly split into ten folds, where nine folds are selected for training the data and the tenth fold is selected for testing the accuracy of the data.

6 Discussions

To show the utility of proposed iBPSO based algorithm we compare the proposed algorithm with BFPA algorithm. Various values were tested for the parameters of proposed algorithm. The results show that the highest performance is achieved by setting the parameters to values as follow:

The parameters for BFPA includes The flower size which is chosen as 50,' $\lambda' = 1.5, \varepsilon' = 0.8$, the parameters for the iBPSO are say, population size is 50, the maximum number of iteration is 500, $C_1 = C_2 = 2$ and W is in the range of [0.5, 1.0].

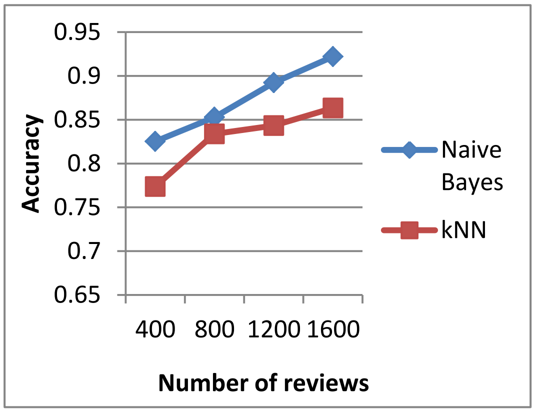

Fig. 2: Performance analysis of Classification Accuracy for BFPA

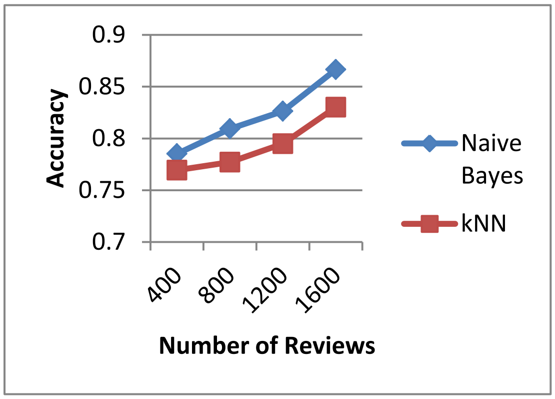

Fig. 3: Performance analysis of Classification Accuracy for iBPSO

From the figure 2 and 3 it is observed that the Accuracy for BFPA is better than iBPSO by 5.20% using Naïve Bayes and k-NN by 3.89 %. The proposed BFPA feature selection improves the precision by 3.99% when compared to iBPSO.

Analyzing the accuracies, on average, the BFPA algorithm obtained a higher accuracy value than the iBPSO algorithm. To graphically depict the progress of the BFPA as it searches for optimal solutions, we take number of reviews as the horizontal coordinate and the Accuracy measure as the vertical coordinate. This should illustrate the process of improvement of the best flower as the number of reviews increases.

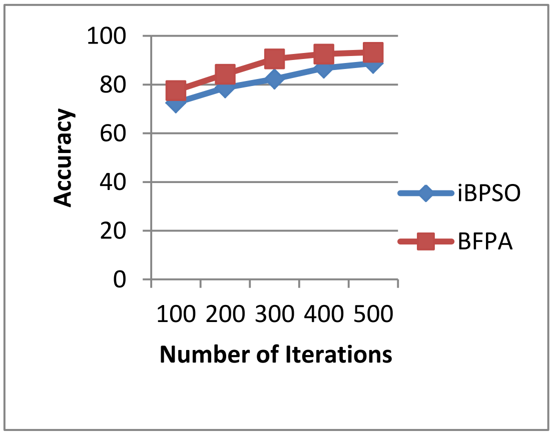

Fig. 4: Performance analysis of Feature Selection Techniques iBPSO and BFPA

From the figure 4, it is observed that the accuracy of the BFPA increases as the number of iteration increases when compared with the iBPSO algorithm.

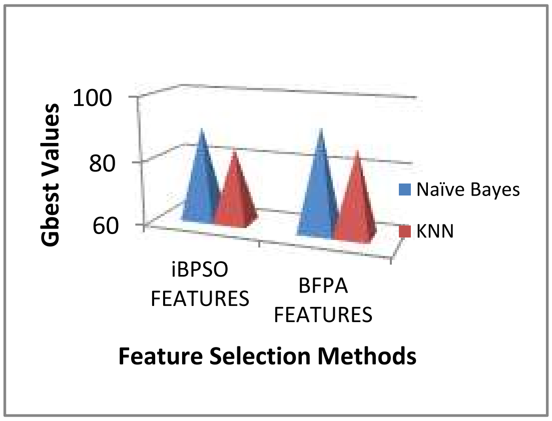

Fig. 5: Comparison of Feature Selection Techniques for iBPSO and BFPA

From the figure 5, it is observed that the gbest values for BFPA is high compared with iBPSO.

7 Conclusion

In this study, the paper presented that iBPSO and BFPA based algorithms can able to achieve an optimum selection of feature subset, that can be used to trained and classified by using Naive bayes and k-NN. The obtained results shows that BFPA gives the optimum fitness values for all classified subsets such as truthful and deceptive which is obtained by the Naive Bayes and k-NN. The proposed method shows that by using BFPA compared with iBPSO reduces the complexity of optimization. The experimental result shows that BFPA has increased overall accuracy than iBPSO.

References

[1] Umamaheswari. K. Sumathi, "Consistency based Feature Selection Approach for Improving Text Classification Performance", Proceedings of Third National Conference on Innovations in Information and Communication Technology, 2007.

[2] S. P. Rajamohana, Dr. K. Umamaheswari, "An Integrated Evolutionary Algorithm for Review Spam Detection on Online Reviews", Advances in Natural Applied Science (AENSI), 2016.

[3] Dervis Karaboga, Bahriye Basturk, "On the performance of artificial bee colony (ABC) algorithm, "Applied Soft Computing 8 (687–697), 2008.

[4] 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.

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

[6] J. Kennedy and R. Eberhart, "Particle swarm optimization," *Proc.Neural Networks, 1995. Proceedings.*, IEEE International Conference on, IEEE, 1995, pp. 1942–1948.

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

[8] Ajay Sharma, Anil Suryawanshi, A Novel Method for Detecting Spam Email using KNN Classification with Spearman Correlation as Distance Measure International Journal of Computer Applications (0975–8887) Vol. 136, No. 6, 2016

[9] MyleOtt, 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.

[10] Susana M. Vieira, Luís F. Mendonça, Gonçalo J. Farinha & João M. C. Sousa, "Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients", Applied Soft Computing, vol. 13, pp. 3494-3504, 2013.

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.

B. Abirami is received her B.E degree from the Department of CSE,Jansons Institute of Technology,Coimbatore. Currently she is pursuing her Masters from the Department of IT in PSG College of Technology, India. Her areas of interest include Data Mining and Evolutionary Computing.