International Journal of Engineering &Technology, 7 (1.7) (2018) 205-209

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



An intelligent recursive feature reduction methods for efficient classification of medical blogs

Neeba E A [1] and Koteeswaran S [2]

[1]Assistant Professor, Department of information Technology, Rajagiri School of Engineering & technology Rajagiri Valley P O, Kochi-682039, Kerala,India
[1]Research Scholar, Dept of CSE, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology
[2] Associate Professor , Dept of CSE, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology
**Corresponding author E-mail: neebaea@rajagiritech.edu.in*

Abstract

Nowadays people express themselvesin the form of weblogs attracting and benefiting users in their perspective. Blogs are maintained by individual users with specific software reporting activities that removing all possible barriers and developing learning blogs. Blogs together forms a linked community as blogosphere. Procuring information from these blogs has become very tough due to the intense growth of blogs. The proposed system classifies and compiles the blog ensuring appropriate chemo taxis analysis and reporting. These data preparation and analysis helps the healthcare organizations to solve many health related problems and identify new improvements. The movement of organism to chemical stimulus is recognized in chemo taxis. Chemo taxis is subverted in cancer phases. Weblogs not only manage and organize these data but also reduces the time required for data gathering, acquisition, analysis and reporting.

Keywords: *Blog Classification, Bagging, Logitboost, Adaboost, BFO and PSO.*

1. Introduction

Within a very short span of time, blogging has addicted innumerable internet users, representing their culture and media of communication. While companies own the traditional social media, and other online web cast including news**,** blogs have two unique characteristics:

1. Primarily, created based on individuals personalized approach.
2. Secondarily, blogs interlink forming localized communities.

Increase of daily bloggers paved way for researchers to performbetter on content analysis. The evolutionary in blogs are based on the characteristics of the blogs. Bloggers use user-friendlymethods to prepare, analyse and characterize the blogs with the help of various search engines.

The face of communications on the Internet is being changed by blogs. It not only helps in publishing the contents, but also enables the user to interact among themselves forming a close-knit community, forming a very dominant medium in the web.The personalized nature of blogs in combination with the capacity to formulate and analyze them immediately bring forward innovative opportunities in the analysis of the pulse of society and the search and speeding up the working of many applications including market research public relations, psychology of the society, and political scenario.

World wide web acts as an intermediate for people to easily update, socialize, interact and share with each other their views. The online is the most popular means to bloggers in sharing new informations in the format of blog reviews, sentiments, events that are created on daily routines. Though all the comments are not informative high quality comments have great impact on the readers opinions about the original post content and their quality are discussed. Therefore, developing an efficient mechanism is highly desirable. For this purpose, sites like slash dot, where users rate comments based on their reliability forms an automated system using machine learning techniques[3]. This classification problem is usually solved by improvising the learning techniques, while learning with high performance needs an increased amount of manually labelled blogs. In contrast, unlabelled blogs can be retrieved easily from internet, these are done using semi supervised mode[2]. The magnitude of blogs is multiplying at avery fast rate and isdiscovered that nearly 2-7% of internet users create blogs and around 11% of these users follow. Technorati's web crawlers states that for about 12,000 new blogs are created in every 7.4 seconds. Because of the popularity and accessibility of blogs content classification system directory is hard to generate due to the tediousness in categorizing blogs into different groups based on its unique nature.

There is no hard-and-fast rule as to what should be written in a blog. From personal to politics, anything can be written, hence writing whatever is there in one's interest**,** even devising a new vocabulary and grammar. A few blogs, in the perspective of attracting more audience, intentionally move away from rules of language and decorum. It is very obvious that blogs are growing in geometric proportion, although specific figures differ from time to time. A situation has arisen such that blogs have become a very active part of the online media. Research on blogging has brought forth the true attitude of the blogging activity, the general behavior of the bloggers and the common practices involved in blogs.

Currently blogs have become the source for the advertisement market. Various applications have been developed for capturing

Copyright © 2018 Neeba E A and Koteeswaran S. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

this, thus many search engines are providing blog search services. The explosion of blogging has thrown light on more research opportunities on text mining, retrieval of information, social studies and other areas which need more exposure. Each and every blog is made of many elements including entries, tgas, lings and others. Entry seems to be the most important element for the content analysis of a blog.

A blog deals with a wide range of issues including professional book keeping or more often frequently-improvised views and personal remarks. Mood classification has attracted an increasing interest in the different research communities, and is used in a variety of activities including implementation of techniques in assisting behavioral scientists and improving the interaction between doctor and patient. As access to internet has become easier and has almost become a very common mode of communication between communities, more number of people share their information, ideas or opinion in their weblogs.

2. Literature Reviews

Sun et al., [3] classified blogs based on the personal versus official aspect. The system adopted a 2-stage method; primary is the training model, where lexicons were constructed automatedly, and in the other scoring as well as ranking were performed respectively. The outcomes of simulations revealed that lexicons hold good for feature selection and mutual information. Sentiment words could add flavor to the contents.

Ayyasamy et al., [4] described a technique to classify blogs automatedly through the usage of fuzzy similitude. The authors performed simulations utilizing TREC data set as well as employed the method to 6 distinct fuzzy similitude metrics. Outcomes of experiments prove that the Einstein fuzzy similitude metrics performed better than the others.

Shi et al., [5] proposed a novel technique of multilabel classification for Chinese blogs. The authors attained automated classification using dempster-shafer theory on semantic word similitude protocol. And test results prove that one of the biggest Social Network Service (SNS) is China. The outcomes showed that the suggested technique gave an excellent performance in multilabelling actual social networking blogs and corpuses.

Bai et al., [6] proposed a novel classification technique called Folksonomy as well as SVM classifier (FSVMC). The FSVMC used SVMs as a Tagagent. This is a program for determining if a certain tag ought to be appended to a web blog page while Folksonomy is dedicated to categorizing web blog posts. Outcomes of experiments indicated the technique is flexible as well as efficient much more than conventional techniques.

Gao et al., [7] presented a technique of microblog data classification utilizing mixed features of chi-square, the authorused features selection as well as the weight computation technique. The outcomes of the experiments showed that the technique proved to be helpful in improving classification of microblog data.

Duan et al., [8] calculated the Semantic Orientation between the particular phrase as well as the polar terms. The SO of the particular blog is computed through the max MI value. Blog texts are sorted as joy and the blog archive gathered by the Monitor and Research Center for National Language Resource achieved a higher enhancement in comparison to conventional techniques.

Sun et al., [9] presented a blogger modelling technique for Chinese microblog sentiments classification. The presented work adopts a multi-dimensional method for accomplishing the features extraction job. In comparison to conventional techniques merged with the interactive characteristic of the blog, blog comments are regarded as the performance feature of microblog emotions expression. The results demonstrated the efficacy of the multi-dimensional features extraction technique on sentiments classification as well as for the classification of the impact of fuzzy Support Vector Machine (SVM) is better than SVM.

Dalal and Zaveri [10] attempted to automatedly classified the text entries written by bloggers on several sport blogs, to the appropriately categorizing the sports through stages such as preprocessing, features extraction, as well as NB classifications. The classification accuracy rendered by this technique by empirical evaluation is around 87% over the tes dataset. Apart from the classification of text entries of sport blogs, it is suggested that the extracted attributes themselves be further sorted under more meaningful titles that result in the creation of semantic resources which lend more comprehension to the job of classification. The semantic resources may be utilized for data mining needs which may emerge in the future.

Chen et al., [11] proposed blogs IR model on the basis of semantic through analyzing the existenttechnology in blog systems as well as focus on semantic retrieval of blog information resources. For implementing semantic descriptions for blog information resources, the authors designed blog ontologies as well as domain subject classification ontologies. With pre-set rules, the authors suggested a class hierarchy tree creating protocol for expanding blog classification items semantic retrieval, as well as execute it in the end through SPARQL query.

3. Methodology

3.1 Proposed PSO feature selection

Particle swarm optimization search technique which is simple and is formulated based on the search technique of birds in search of food. PSO uses this technique where problems are solved using social interaction. This has been applied successfully in solving many search and optimization problems.

To start with the pre-processed web data randomly initialized with a random position and zero velocity. The velocity and the fitness value of each and every particle is different depending on the fitness function of each particle in the swarm, the particle with best fitness function is accredited as *pbest* and the fitness of the global best particle is termed as *gbest*. Finally the clusters are optimized by continuous updation in the function using number of iterations and termination criteria.

Velocity update equation

$$V_i^{t+1} = w \cdot V_i^t + c_1 \cdot r_1 \left(P_{best} - X_i^t \right) + c_2 \cdot r_2 \left(G_{best} - X_i^t \right) \quad (1)$$

Position Update equation

$$X_i^{t+1} = X_i^t + V_i^{t+1} \quad (2)$$

Wherein, parameters are:

$V_i(t)$ - velocity

P_{best} - Personal Best Position

G_{best} - Global Best Position

$X_i(t)$ - Current Position

c_1 & c_2 - Acceleration Constant.

r_1 & r_2 - Random function ranging between [0, 1].

w - Inertia weight.

If the optimized clusters are formed after number of iterations, then the system will move to an end state known as Meet Stopping Criterion. If stopping criteria is not met then velocity and position loops back to fitness evaluation for further optimization. Even though it is an optimized technique, the *gbest* leads to trapping in the local area. Here velocity is defined as the rate of change of motion of a particle with a specific speed and direction of motion. Base on the number of iterations the velocity needs to be continuously optimized.

Every particle has its own inertia (momentum) which stores in its memory their previous best position, thus preventing them from

changing position suddenly. The tendency of the particle to return to its best position is defined by Cognitive Component model.
Social component compares the particle with its neighbour, in this position and velocity of a particle is updated as a group, thus helping the group attain globally best results when compared to their neighbor. Velocity is another important criteria and clamps each velocities in each dimension depending upon its fitness within the region that is searched. [25]. The two possible criteria are:

- Too high velocity moves the particle past the optimal solutions.
- Those too low get stuck in local minima.

3.2 Proposed Bacteria Foraging Optimization (BFO) Feature Selection

The Bacteria Foraging Optimization is a global optimization algorithm which calculates the cost function leading to a better fitness with lesser cost in every iteration. Here, the optimizingvariablesdenotes the position of the bacteria at a favourable range located at each point. At every forward step the bacteria move to a new position and cost function for that particular movement is calculated. Cost function for each movement is calculated giving a better optimization parameter with both cost and fitness optimization. The foraging method of E. Coli bacteria (strategy to search for food) is taken for representation here. This is governed by chemotaxis, swarming, reproduction as well as elimination and dispersal. The first step is chemotaxis that is search for food that is when the bacterium faces a favourable environment and which help their growth, it moves in the direction of the food by swimming and tumbling. When the environment is adverse then they move away from the food. Here, decrease in cost function is depicted by favourable environment and vice versa. In swarming, bacteria move in the form of ringreducing the mean square error to a minimum value.

Swarming and chemo taxis
If s1,s2,s3 are the variables for optimization, they denote the axes of spatial coordinates, assuming fi,j,k,l(s1,s2,s3) denotes position of ith bacteria at a point in thecoordinate system. Furthermore, C(i) denotes the run length of a bacteria. Then, the motion of the ith bacteria in jth chemotaxis isgiven as: assume s1 as m1 and so on.

$$f^{i,j+1,k,l}(m1,m2,m3) = f^{i,j,k,l}(m1,m2,m3) + C(i)\frac{del(i)}{\sqrt{del^T(i)\,del(i)}} \quad (3)$$

where del(i) is the position of bacteria denoted as 3 coordinate system. Every element is an arb

$$f^{i,j+1,k,l}(m1,m2,m3) = f^{i,j,k,l}(m1,m2,m3) + C(i)\frac{del(i)}{\sqrt{del^T(i)\,del(i)}} \quad (3)$$

where del(i) is the position of bacteria denoted as 3 coordinate system. Every element is an arbitrary number ranging between [-1, 1]. If ith bacteria is in a desirable setting that is denoted by lesser value of cost function at the point, it moves to the earlier (j-1 th) chemotaxis step. Else, del (i) is designated a fresh value ranging between [-1,1].
After every step, the bacterium reaches new point in space and each point represents an optimization parameter which is the coordinate axis of the position of the bacterium. At this new location, the fitness of each bacterium is studied which acts as a deciding factor in the subsequent movement of the bacteria. Better fitness represents lessercost function.
Reproduction
Fitness of every bacteria after chemo taxis procedure is the total sum of cost function $\sum_{j=1}^{Nc} P^{i,j,k,l}$, wherein NC is the overall quantity of steps in the chemo taxis procedure. More healthy bacterium represents better set, while the healthy bacteria survive, the other half gets eliminated. Each surviving bacteria splits using binary fission and are situated in the same position, thus maintaining the bacterial population.

Elimination and Dispersal Event
While chemo taxis performs the search procedure that helps in maintaining the population thus keeping the count of bacteria at an optimum level. But this is not sufficient to reach a global minimum point presuming that to be the best fitness position. The bacteria that has the probability Ped (elimination as well as dispersion) is eliminated from current location thus maintaining the population [16].

3.3 Classifiers

3.3.1 Bagging

Bagging is a technique to improve outcomes of machine learning classification protocols. The two potential classes are classification protocol generates a classifier $H: D \to \{-1,1\}$ on the base of a training dataset of sample descriptions D. Bagging technique builds a set of classifiers H_m, m=1… M in the training dataset. The classifiers are merge as one compound classifier and the weighted combination of individual classifier predictions are as follows:

$$H(d_i) = sign\left(\sum_{m=1}^{M} \alpha_m H_m(d_i)\right) \quad (4)$$

A sample di is considered to be the class for which most classifiers vote. This is known as the theory of classifiers voting. Variables α_m, m=1,…,M are defined to be more accurate classifiers those have strong effect on the end prediction rather than lesser accurate classifiers. The precision of base classifiers $H_m$could be greater than the precision of arbitrary classifications. This is why the classifiers H_m are known as weak classifiers [12].

3.3.2 AdaBoost

AdaBoostcombined classifiers, all of which can only be somewhat inaccurate. Within the context of learning procedure, weak learners are trained in a sequential manner one at a time. Everycycle a weak classifier is inaccurate for classifying the samples, forming the weak hypothesis. Once a particular quantity of iterations are over, the resulting weak hypotheses are merged in a linear manner into one prediction rule, the combined hypothesis.
The generalized AdaBoost protocol for binary classification uses a vector Dt over the datasets. At cycle t, the aim of AdaBoost is the weak hypothesis $h_t : X \to R$ with adequately lower error with respect to the weights D_t. In this scenario, weak hypotheses $h_t(x)$ makes the real-valued confidence-rated predictions. Initially, the distribution D_t is uniformly initialized, which means that all instances are equally weighted. The boosting protocol increasesand decreases with the weights $D_t(i)$ when $h_t(x_i)$ makinggood or badprediction of instances, with variations proportional $h_t(x_i)$ the end hypothesis, $f_T : X \to R$, calculates its prediction utilizing the weighted vote on weak hypothesis $f_T(x) = \sum_{j=1}^{T} \alpha_j h_j(x)$. The rule may be given as [13]:

$$D_{t+1} = \frac{D_t(i) * e^{-\alpha_t y_i h_t(x_i)}}{Z_t} \quad (5)$$

$$= \frac{e^{-\sum_{i=1}^{t} \alpha_i y_i h_i(x_i)}}{M * \prod_{j=1}^{t} Z_j}$$

$$= \frac{e^{y_i f_i(x_i)}}{M * \prod_{j=1}^{t} Z_j}$$

$$= \frac{e^{-mrg(x_i, y_i, f_i)}}{M * \prod_{j=1}^{t} Z_j}$$

3.3.3 Logitboost

The additive model by a greedy stage-wise process is as follows:

$$\frac{\partial L_i}{\partial F_{i,k}} = -\left(r_{i,k} - p_{i,k}\right), \quad \frac{\partial^2 L_i}{\partial F_{i,k}^2} = p_{i,k}\left(1 - p_{i,k}\right). \tag{6}$$

Assuming the sum-to-zero constraint on a "base class" the resultant derivatives do not dependent on the base option:

0: $r_{i,k} = 1$, *if* $y_i = k$, $r_{i,k} = 0$ else.

1: $F_{i,k} = 0$, $p_{i,k} = \frac{1}{K}$, $k = 0$ *to* $K-1$, $i = 1$ *to* N

2: For m=1 to M Do

3: For k=0 to K-1, Do

4: Calculate $w_{i,k} = p_{i,k}\left(1 - p_{i,k}\right)$.

5: Calculate $z_{i,k} = \frac{r_{i,k} - p_{i,k}}{p_{i,k}\left(1 - p_{i,k}\right)}$.

6: Fit the function $f_{i,k}$ by a weighted least-square of $z_{i,k}$ to x_i with weights $w_{i,k}$.

7: $F_{i,k} = F_{i,k} + \nu \frac{K-1}{K}\left(f_{i,k} - \frac{1}{K}\sum_{k=0}^{K-1} f_{i,k}\right)$

8: End

9: $p_{i,k} = \exp\left(F_{i,k}\right) \Big/ \sum_{s=0}^{K-1} \exp\left(F_{i,s}\right)$

10: Stop

The diagonal approximation is necessary if base learners are to be executed using regression tree. For industrial applications, utilizing tree as weak learnersseems to be the general practice [14].

4. Experiments and Related Results

The classification accuracy, average precision and average recall for BFO and PSO feature selection including Bagging and Logitboost classifiers are shown below:

Table 1: Classification Accuracy

Techniques	Classification Accuracy
PSO - Bagging	0.8768
BFO- Bagging	0.8811
PSO- Logitboost	0.8874
BFO- Logitboost	0.9158

Table 2: Average Precision

Techniques	Average Precision
PSO - Bagging	0.8637
BFO- Bagging	0.8699
PSO- Logitboost	0.8767
BFO- Logitboost	0.9075

Table 3: Average Recall

Techniques	Average Recall
PSO - Bagging	0.8755
BFO- Bagging	0.875
PSO- Logitboost	0.885
BFO- Logitboost	0.9137

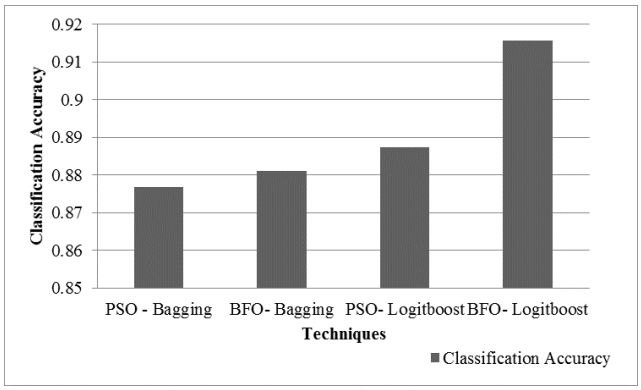


Figure 1 Classification Accuracy

In table 1 as well as figure 1, it is observed that the classification accuracy has enhanced by 0.49% in comparison to PSO-Bagging as well as BFO-Bagging.

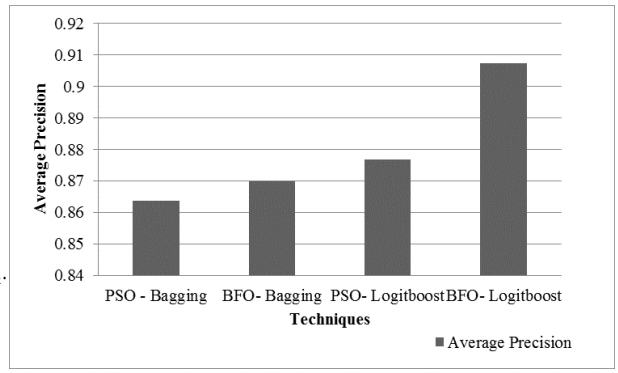


Figure 2 Average Precision

In table 2 as well as figure 2, it is observed that the average precision has enhnaced by 3.45% in comparison to PSO-Logitboost as well as BFO-Logitboost.

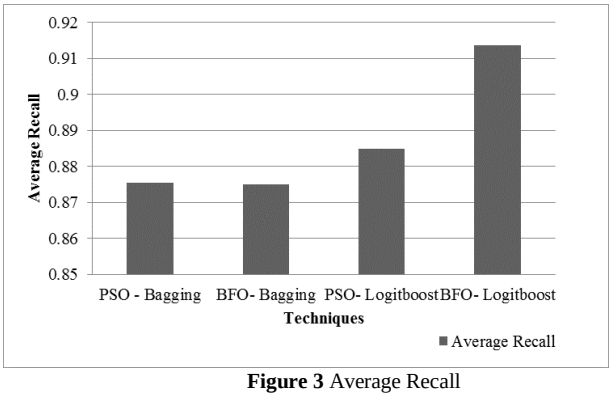


Figure 3 Average Recall

In table 3 as well as figure 3, it is observed that the average recall has enhanced by 1.14% in comparison to BFO-Bagging as well as PSO-Logitboost.

5. Conclusion

Blogs are hard for categorizing by both humans as well as machines, as they are written capriciously. For any particular blog, there is no particular group which fulfils the judgment of the classifier. In this work, it addresses the job of distinguishing the two genres of blogs, informative as well as affective content. The bagging method is an adequate mean for improving efficiency of standardized machine learning protocols. Result show that the classification accuracy has improved by 0.49% compared to PSO-Bagging and BFO-Bagging.

References

[1]. Jin .W, Joty.S, Carenini. G and R,"Detecting Informative Blog Comments using Tree Structured Conditional Random Fields".

[2]. Ikeda.D, Takamura. H and Okumura. M, "Semi-Supervised Learning for Blog Classification, in AAA, pp. 1156-1161, July, 2008.

[3]. Sun. X, Li. S, Xu. W, Chen. G and Guo. J, "Machine learning based blog classification personal vs. official facet", *Computer Research and Development,3rd International Conference,* vol. 1, pp. 31-34, IEEE, March, 2013.

[4]. Ayyasamy. R. K, Tahayna. B, Alhashmi. S. M., " concept based modeling approach for blog classification using fuzzy similarity",*Fuzzy Systems and Knowledge Discovery, Eighth International Conference, v*ol. 2, pp. 1007-1011, IEEE, July 2011.

[5]. Shi, C., Li, J., Chen, J., & Chen, X. (2013, August). Chinese SNS blog classification using semantic similarity. In *Computational Aspects of Social Networks (CASoN), 2013 Fifth International Conference on* (pp. 1-6). IEEE.

[6]. Bai. R, Wang. X, and Liao. J, "Folksonomy for the blogosphere: Blog identification and classification", *Computer Science and Information Engineering, 2009 WRI World Congress,* Vol. 3, pp. 631-635.

[7]. Gao. X, Liu. L, and Hou. S, "A method of micro-blog information classification based on mixed characteristics", *Ninth International Conference on Natural Computation, p*p. 853-857, IEEE, July 2013.

[8]. Duan, X., He, T., & Song, L. (2010, August). Research on sentiment classification of Blog based on PMI-IR. In *Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on* (pp. 1-6). IEEE.

[9]. Sun. X, Li. Cand Ren. F, "Modeling blog user for Chinese microblog sentiment classification with fuzzy SVM, *IEEE 3rd International Conference on Cloud Computing and Intelligence Systems*, pp. 130-134, November, 2017.

[10]. Dalal, M. K., & Zaveri, M. A. (2012, January). Automatic text classification of sports blog data. In *Computing, Communications and Applications Conference (ComComAp), 2012* (pp. 219-222). IEEE.

[11]. Chen, Y., Guo, W., & Zhao, X. (2010, July). A Semantic Based Information Retrieval Model for Blog. In *Electronic Commerce and Security (ISECS), 2010 Third International Symposium on* (pp. 257-260). IEEE.

[12]. Machová, K., Barcak, F., & Bednár, P. (2006). A bagging method using decision trees in the role of base classifiers. *Acta Polytechnica Hungarica*,*3*(2), 121-132.

[13]. Lee, W. L., Lommatzsch, A., & Scheel, C. (2007, October). Feed Distillation Using AdaBoost and Topic Maps. In *TREC* (Vol. 7, p. 21).

[14]. Li, P. (2012). Robust logitboost and adaptive base class (abc) logitboost.*arXiv preprint arXiv:1203.3491*.

[15]. Ghorpade-Aher. J and Bagdiya. R, "A Review on Clustering Web data using PSO", I*nternational Journal of Computer Applications, 108*, 2014.

[16]. Sharma, V., Pattnaik, S. S., & Garg, T. (2012). A Review of Bacterial Foraging Optimization and Its Applications. In *National Conference on Future Aspects of Artificial intelligence in Industrial Automation, NCFAAIIA* (pp. 9-12).