International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Weighted Sentiment Analysis Using Artificial Bee Colony Algorithm

Ruby Dhurve[1], Megha Seth[2]

[1]M.Tech Scholar, Computer Science & Engineering, RCET, Bhilai, Chhattisgarh, India

[2]Assistant Professor, RCET, Bhilai, Chhattisgarh, India

Abstract: ***Data mining is the process of extracting interesting and useful data from different perspective and summarizing into useful information. Sentiment analysis is an application of NLP, data mining and text mining to identify sentiments or mood of the public about particular topic or products or customer reviews. This paper proposes to improve the methods for classification of review and also detect the polarity of reviews using machine learning approach. Objective of the research paper is to select the best features selection methods for sentiments analysis. Bog of noun, bog of words, stop word removal, stemmer are used for feature selection. ABC algorithm is used for classification of text in three classes: negative, positive and neutral. The main aim of the thesis is to compute the result of SVM and ABC classifier. In result nature inspired ABC classifier BON give the better results than the BOW, SVM with both BON and BOW.***

Keywords: Sentiment Analysis; Feature selection; BOW (Bag of Words); BON(Bag of Nouns); Parser; Artificial Bee Colony Algorithm; Support Vector Machine

1. Introduction

Data mining is the process of extracting interesting and useful data from different perspective and summarizing into useful information. Sentiment analysis is an application of NLP, data mining and text mining to identify sentiments or moods of public about particular topic or product. Sentiment Analysis used to identify the attitude, judgment, evaluation or emotional communication of a reviewer or speaker with respect to some topic in document [3]. Sentiment Analysis includes branches of computer science like NLP, Machine Learning, Text Mining and Information Theory and Coding. Sentiment analysis is done on three levels Document Level, Sentence Level, Entity or Aspect Level.

Due to the exponential enhancement in the Internet usage and replacement of public opinions, sentiment analysis becomes an important process in today's life. Sentiment analysis is an application of natural language processing, data mining and text mining to identify sentiments or mood of the public about particular topic or products or customer reviews. Sentiment analysis is a process of extracting information from user's opinions. Every person shares his or her information in social network sites, blogs, product review websites and web forums. Thus, the thoughts of other people provide information that helps in decision making process. But, sentiment analysis is a challenging task because it is very difficult to find the exact sentiment from text as there are so many challenges like entity identification, subjectivity detection in SA.

SA is a task of mining use-full information expressed in the text, reviews or opinion and has received a lot of focus from the research community in NLP in recent years. Identifying words or reviews that carry sentiments is a crucial task in sentiment is a crucial task in SA. The work in this thesis concentrate on improve the classification results of reviews and classified according to the reviews polarity and weighted are calculated. Earlier were used as classification of reviews.

2. Literature Review

Farhadloo.M et al (2013) describe aspect level sentiment analysis considering three classes for sentiment polarity of each sentence (positive, neutral and negative). In the aspect identification step they proposed to not ignore the part-of-speech tags, and instead of clustering with bag of words, employ a clustering over the sentences only using bag of nouns and results show that clustering with BON yields more meaningful aspects than using BOW[1]. The main contribution of this paper is the proposal of a new feature set and score representation that leads to more accurate sentiment analysis by using SVM classifier. This scheme is based upon the three scores (positiveness, neutralness and negativeness) that are learned from the data for each term. Using this new score representation scheme, they improve the performance of 3-class sentiment analysis on sentences by 20% in terms of average f1-score, as compared to previously published research. This indicates that there is still room for feature engineering to improve the performance of classifiers in sentence-level opinion mining, and expect that future research will continue to improve on opinion mining[1]. The implications for practice from this paper are i) much improved performance of the sentiment analysis, and ii) an ability to more accurately extract sentiments from domains with higher granularity of opinions (positive, neutral, and negative).

T.Sumathi et al (2013) used Optimal feature selection for reducing feature subset size and computational complexity thereby increasing the classification accuracy. The ABC algorithm being a powerful optimization technique and is widely used for solving combinatorial optimization problems. Hence, this method is incorporated for optimizing the feature subset selection in this investigation. In this paper, the movie reviews is classified using opinion mining. Experiment results evaluated feature selection techniques based on IDF and proposed ABC. Experimental results show that the

classification accuracy of the classifiers improves in tune of 1.63% to 3.81% for the proposed ABC feature selection.

Z.Guppu et al (2010) describes Artificial bee colony (ABC) algorithm invented recently by Karaboga is a biological-inspired optimization algorithm, which has been shown to be competitive with some conventional biological-inspired algorithms, such as genetic algorithm (GA), differential evolution (DE) and particle swarm optimization (PSO). However, there is still an insufficiency in ABC algorithm regarding its solution search equation, which is good at exploration but poor at exploitation. Inspired by PSO, we propose an improved ABC algorithm called gbest-guided ABC (GABC) algorithm by incorporating the information of global best (gbest) solution into the solution search equation to improve the exploitation. The experimental results tested on a set of numerical benchmark functions show that GAB algorithm can outperform ABC algorithm in most of the experiments.

3. Problem Identification

Sentiment analysis can be performed at different levels of granularity with different levels of detail. We perform sentiment analysis on text or reviews which typically consists of one or more sentences. Supported by this observation, the type of granularity we study is the sentence level. Other granularity levels can be the document level, word level or the phrase level. The level of detail typically goes into determining the weight of sentiments, which is what we investigate as well. A more detailed approach could be to determine the emotion expressed in addition to the polarity.

From the all review papers we conclude that there is various feature selection and classification methods are used for classify the reviews of customers in different classes.

3.1 Problem Identification

As the size of digital information grows exponentially, large volumes of raw data need to be extracted. Nowadays, there are several methods to customize and manipulate data according to our needs. The most common method is to use Data Mining (DM). DM has been used in previous years for extracting implicit, valid, and potentially useful knowledge from large volumes of raw data. The extracted knowledge must be accurate, readable, comprehensible, and ease of understanding. Furthermore, the process of data mining is also called as the process of knowledge discovery which has been used in most new inter-disciplinary area such as database, artificial intelligence statistics, visualization, parallel computing and other fields. We found that many optimization algorithms have been used for classification tasks. From the best of our knowledge, previous researches on ABC algorithm have focused on optimization but none of them is for classification tasks [19]. The nature method ABC algorithm is used for classification of the SA in three classes or polarity of sentences or sentiments.

3.2 Solution of the Problem

In many research paper have describe the various methods for selecting for feature set using bag of word, bag of noun, parts-of-speech, stemmer etc has very complex to select the feature set from sentiment reviews of customer. We have use BOW, BON and Stanford parser for feature selection and ABC algorithm is for classification to get the best optimum result of reviews that easily classify the three classes or polarity of reviews. It reduce the error rate for classify the sentiment of text or reviews. At last we compare the result of SVM binary classification and ABC algorithm used for classification and results show ABC algorithm with sphere benchmark function give the best optimum result for classification of sentiments.

4. Methodology Used

This section explains the methodology for the problem discussed in previous chapter. In SA, we have numbers of sentences or reviews of documents. All reviews document may convey opinion or may not. Sentiments polarity of reviews can be classified at this method. Many researchers are trying to calculate the sentiment weighted by different classification algorithm and get the best result of polarity of sentiments. In that method we use the supervised learning approach for the system. An overview of sequential steps and techniques commonly used in sentiment classification approaches, as shown in figure1.

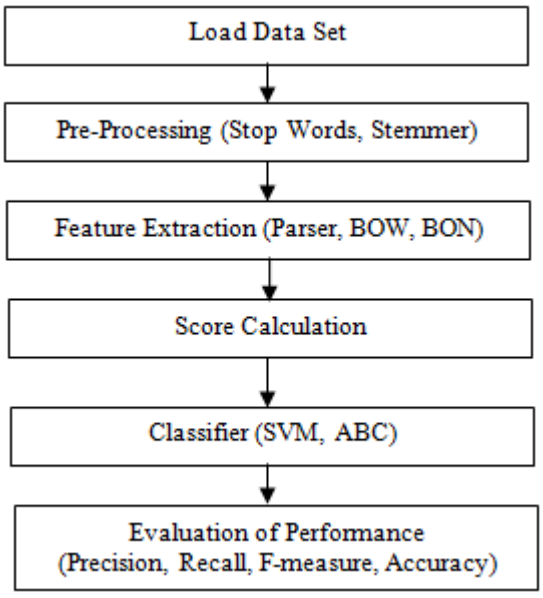

Figure 1: Working of Proposed Model

4.1 Load Data Set

Load the data which are to be classified the sentiment of sentences or reviews of customer. The loaded data that should be customer reviews in negative, positive or neutral. That loaded data can be classified according to training data which are trained by the training data sets. The training data sets are collected from the website with address here: http://www.cs.uic.edu/~liub/FBS/Reviews-9-products.rar. We consider only Nokiya 6600 reviews for our project.

4.2 Pre-Processing

Pre-processing of data is the process of preparing and cleaning the data of dataset for classification. Here is the hypothesis of having the data properly pre-processed: to reduce the nose in the text should help improve the

performance of the classification result and speed up the classification process, thus in aiding in real time sentiment analysis.

4.2.1 Stop Words Removal

A stop list is the name commonly given to a set or list of stop words. It is typically language specific, although it may contain words. A search engine or other natural language processing system may conation a variety of stop list, one par language, or it may conation single stop list that is multilingual. Some of the more frequently used stop words for English include "a " , "of " , " the" , " I" , "it", "you" , and "and" these are generally regarded as 'functional words' which do not carry meaning. When assessing the contents of natural language, the meaning can be conveyed more clearly by ignoring the functional words. Hence it is practical to remove those words which appear too often that support no information for the task.

4.2.2 Stemmer

It is the process for reducing derived words to their stem, or root from. Stemming program is commonly referred to as stemmer or stemming algorithm.

The stemming is the process for finding the root words or is the procedure of describing relevant tokens into a single type. For example "He teach us in an interesting manner" This sentence after stemming is converted into "teach interest manner" thus, by using stem (root) word, the comparison of sentence word with number of positive/negative words becomes easy.

In our project we use the porter stemming algorithm: the porter stemming algorithm is a process for removing the commoner morphological and inflexional endings from words in English. Its main use is as part of term normalization process that is usually done when setting up Information Retrieval System. The algorithm was originally described in Porter, M.F, 1980, An algorithm for suffix stripping, program,14(3):130-137, It has since seen reprinted in Spark Jones. Karen, and Peter Willet, 1997.

4.3 Feature Extraction

Many statistical feature selection methods for document level classification can be used for SA is a need to convert text into feature that represents most important features. Such features are nothing but sentiment bearing words that frequently occur in data. For feature extraction we use the parser that parse the whole documents, BOW and BON are also used for feature selection to give the classification results.

4.3.1 Parser

The Stanford Parser: A statistical parser- A natural language parser is program that works out the grammatical structure of sentences, for instance, which groups of words go together and which words are the subjective or objective of verb. The Stanford Parser is used to parse the sentence. Stanford Parser parse the loaded data that are to be classified and give result with noun, pronoun, adjective, adverb each word is unified by the parser and unwanted words are remove.

4.3.2 BOW

Bag of Words is document that can be made from the parse data that contain all the words from the parse data. And also make the bag of word matrix with the frequencies. In BOW mat frequencies of all the words are calculate and that are used for score calculation of the any loaded data. That BOW is help for classification of the sentence.

4.3.3 BON

Bag of Noun is a document that can be made from the parse data that can contain only the Noun of parse data. That is also made the bag of noun matrix with the frequencies of the each noun in the text data and help in calculating the score and classification of the sentiments.

4.4 Score Calculation

In score calculation the word frequencies are calculated from the positive, negative and neutral words that are extracted from the data base, and their frequencies and related frequency are calculated. Score of sentences or text documents is loaded in the system is calculate the frequencies of positive words by common positive words from BOW data mat,. BOW and BON frequencies of the sentences are calculated, according to that calculation classification can be predicted for loaded sentiments in the system.

4.5 Classifier

Sentiment polarity is vague with regard to its conceptual extension. There is not clear boundary between the concepts of "positive", "neutral" and "negative". Classification is done on the feature extracted for every sentiment. There is classifier for each entity to be evaluated for a set of faculty. For classification we have chosen two classifiers: Support Vector Machine (SVM) and ABC algorithm.

4.5.1 SVM Classifier

Support Vector Machine is a supervised learning technique, which is basically used for binary classification. Current research showed that the SVM is the most accurate method foe classification.

SVM classifiers are widely used in sentiment classification. SVM classifiers are widely used in sentiment classification problem. For classifying document we are using separate SVM classifier for each aspect. As we computing sentiment of document from aspect level SA, we choose one aspect from the set and classify that document using SVM classifier of that particular aspect.

4.5.2 ABC Classifier

The Artificial Bee Colony (ABC) algorithm is one of most popular stochastic, swarm based algorithm proposed by Karaboga in 2005 inspired from the foraging behavior of honey bees. ABC has been applied to solve several problems in various fields and also many researchers have attempted to improve ABC's performance by making some modifications.

Each employed bee is moved to food source area to determine a new food source in the neighbourhood of the current one, and its nectar amount evaluated. If nectar amount of new source is higher, then the bee forgets the first source and memorizes the new one. Onlookers are placed on food

sources by using probability based selection process. As nectar amount in food source increases, probability value with which it is preferred by onlooker's increases similar to natural selection process in evolutionary algorithms.

The artificial bee colony contains three groups: scouts, onlooker bees and employed bees. The bee carrying out random search is known as scout. The bee which is going to the food source which is visited by it previously is employed bee. The bee waiting on the dance area is an onlooker bee. The onlooker bee with scout also called unemployed bee.

In the ABC algorithm, the collective intelligence searching model of artificial bee colony consists of three essential components:

- Employed bees
- Unemployed foraging bees
- Food sources

The employed and unemployed bees search for the rich food sources around the hive. The employed bees store the food source information and share the information with onlooker bees. The number of food sources is equal to the number of employed bees and also equal to the number of onlooker bees. Employed bees whose solutions cannot be improved through a predetermined number of trials (that is "limit") become scouts and their solutions are abandoned. Analogously in the optimization context, the number of food sources in ABC algorithm represents the number of solutions in the population. The position of a good food source indicates the position of a promising solution to the optimization problem and the quality of nectar of a food source represents the fitness cost of the associated solution.

In the ABC algorithm, the maximum number of cycles was taken as 2000. The percentages of onlooker bees and employed bees were %50 of the colony and the number of scout bees was selected to be one. The increase in the number of scouts encourages the exploration process while the increase of onlookers on a food source encourages the exploitation process.

In ABC algorithm six benchmark functions are used to optimize the numerical value.

i. Schaffer function.
ii. Rosenbrock function.
iii. Sphere function.
iv. Griewank function.
v. Rastrigin function.
vi. Ackley function.

ABC algorithm being a powerful optimization technique and is widely used for solving combinatorial optimization problems. Hence, this method is incorporated for optimizing the feature subset selection in this investigation. Using sentiment analysis, features from text are extracted, and classified those providing opinions/sentiments about text/data/documents through Support Vector Machine classifier. The ABC algorithm is used for classification to improve the accuracy of the classifier with BOW and BON feature.

Sphere benchmark function is used for optimization of best result of the classification.

$$f(x) = \sum_{i=1}^{D} x_i^2 \quad \text{.......................................} (1)$$

Where the initial range of x is $[-100,100]^D$. The minimum solution of the sphere function is $\vec{x} = [0,0...,0]$ and $f(x)$.
Fitness functions have been designed for the extraction of best classification results of weight of sentiments.

4.6 Evaluation of Performance

Evaluation of performance can be calculated by precision, recall, F-measure, accuracy.

4.6.1 Precision

In a classification task, the precision for a class is the number of true positives (i.e. the number of items correctly belonging to the positive class) divided by the total number of elements belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class).

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \quad ...(2)$$

4.6.2 Recall

Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \quad ....(3)$$

4.6.3 F-measure

A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score.

$$F-measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad ......(4)$$

4.6.4 Accuracy

The accuracy is the proportion of true results among the total number of cases examined. To make the context clear by the semantics, it is often referred to as the "rand accuracy". It is a parameter of the test.

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

5. Result and Discussion

The proposed software efficiently calculates the weight of sentiments. The software will give the better accuracy, result with weighted sentiment, improved classification of sentiment.

Hence very limited work has been done the field of increasing

the accuracy of the classification of sentiment opinion's or reviews. Optimal feature selection is used for reducing feature subset size and computational complexity thereby increasing the classification accuracy. The ABC algorithm being a powerful optimization technique and is widely used for solving combinatorial optimization problems.

Hence, this method is incorporated for optimizing the feature subset selection in this investigation. Experiment results evaluated feature selection techniques based on BOW, BON and for classification ABC algorithm with sphere benchmark function is used. Experimental results show that the classification results of the sentiment using ABC algorithm as compared to SVM in figure 2.

System classifies the weight of sentiments from both SVM binary classifier and ABC nature classifier. Actual polarity or weights of sentiments or reviews are known by us. We test the system with 20 sentiments will give the correct weight or not. Out of 20 sentiments SVM classifier with BOW features and ABC classifier with BOW features gives the 12 sentiments are correctly classified, SVM classifier with BON feature give the 11 sentiments are correctly classified and at last ABC classifier with BON features classified the 14 sentiments correctly.

Table 1: Classification Results

Classifier	SVM (BOW)	SVM (BON)	ABC (BOW)	ABC (BON)
Precision	1	1	1	1
Recall	0.60	0.55	0.60	0.70
F-measure	0.75	0.70	0.70	0.82
Accuracy	0.60	0.55	0.60	0.70

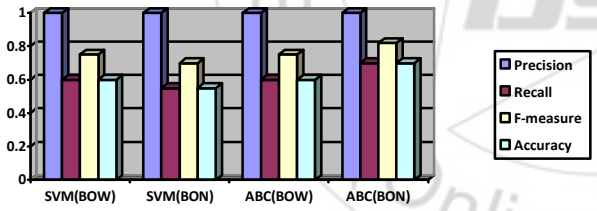

Figure 2: Performance of ABC and SVM classifier on Evaluation

Develop software improves the accuracy of classification using ABC algorithm with BOW and BON features that improve classification accuracy. At last the comparison between SVM (BOW), SVM (BON), ABC (BOW) and ABC (BON) the BON feature selection with ABC classifiers give the better accuracy than the BOW with ABC and SVM classifier feature selection.

6. Conclusion and Future Work

6.1 Conclusion

Sentiment analysis is the process identifying customer sentiments and emotional states. The feelings of the customer can be expressed in positive, negative and neutral ways. Mostly, parts of speech are used as feature to extract the sentiment of the text but we use BOW and BON as a feature of Datasets.

The system computes sentiments polarity. These sentence splits in BOW and for more accurate BON are extracted from the parse data set. Classification with feature BON give the best result as compare to BOW in any classification algorithm.

ABC algorithm is nature inspired algorithm and it give the best optimum solution for the polarity detection of sentiments of reviews of customer. The system has been tested on 20 sample reviews. It has been observed that system has identified sentiments for reviews of people.

The Artificial Bee Colony (ABC) algorithm is a new searching algorithm under Swarm Intelligence technology. Many approaches, methods and goals have been tried out for SA. From the best of our knowledge, previous researches have never applied the ABC algorithm for SA classification. We have compared the proposed algorithm with SVM classifier algorithms which selected from data mining software tools: "Matlab". It has been proved that the proposed ABC data mining algorithm can obtain the better result for weighted (polarity) sentiments. Therefore, we can conclude that the proposed ABC algorithm for SA can obtain competitive result against SVM algorithms and can be considered as useful and accurate classifier.

6.2 Future Work

The proposed system works for simple and short reviews. It can be extended to work for big reviews also. BOW and BON file and their frequencies can be improved by adding more sentiments and reviews in data set. The proposed system use Stanford Parser. If any better parser is available in future classification accuracy of SA has been improved. The system works for short data set. It can be extended to for all type of reviews that is challenges of SA. The proposed system is not web based, so it can be extended as web based application in future. Limitation of the system is not work properly because the data base not so strong, it based on small knowledge base system according to our knowledge base system give can classify the weight of sentiments. It does work well for our data base.

References

[1] Farhadloo. M, Rolland. E, "Multi-Class Sentiment Analysis with Clustering and Score Representation" 13th International Conference on Data Mining Wokshop, IEEE, 2013.
[2] S. Njolstd. P, S. Hoysaeter. L, Wei. W and Atle Gull. J. "Evaluating Feature Sets and Classifiers for Sentiment Analysis of Financial News, " IEEE/WIC/ACM International Joint Conferences on Web Intelligence(W1) and Intelligent Agent Technologies(IAT), IEEE, 2014.
[3] Valakunde. N, Patwardhan. M, "Multi-Aspect and Mutli-Class Based Document Sentiment Analsis of Educational Data Cater in Accreditation Process" International

conference on Cloud & Ubiquitous Computing & Emerging Technologies, IEEE, 2013.
[4] Liu. L, Nie. X, Wang. H, "Toward a Fuzzy Domain Sentiment Ontology Tree for Sentiment Analysis" 5th International Congress on Image and Signal Processing(CISP 2012), IEEE, 2012.
[5] Sumathi. T, Karthik. S, Marikkannan. M, "Artificial Bee Colony Optimation For Feature In Opinion Mining ", Journal of Theoretical and Applied Information Technology, Elsevier, 2012.
[6] Basari. A, Hussin. B, Ananta. I, Junta Zeninarja, "Opinion Mining of Movie Review using Hybrid Method of support Vector Machine and Particle Swarm Optimization" Malaysian Technical Universities Conference on Engineering & Technology 2012, MUCET 2012 Part 4-Information And Communication Technology, Elsevier, 2013.
[7] ManjuS. R, Kalaiman. E, R. Bhavani, "Product Aspect Ranking Using Sentimentic Oriented Sentiment Classifier" IJERA, 2014.
[8] Singh. N, Ghalib. M. R, "An Effective E-Commerce Management using Mining Techniques" International Journal of scientific and Research Publications, Volume 3, Issue 8, IJSPR, 2013.
[9] R. Shikalgar. N, Badgujar. D, "Online Review Mining For Forecasting Sales" IJRET, 2013.
[10] Vigneshkumar K, Gnanavel S, "Mining Online Reviews for Predicting Sales Performance in Movie Domain" TIJCSA, 2013.
[11] S. Modha. J Prof & Head S. Pandi. G , J. Modha. S, "Automatic Sentiment Analysis for Unstructured Data", IJARCSSE, 2013
[12] Patil. G, Galande, V, Kekam. V, Dange. K, "Sentiment Analysis Using Support Vector Machine" IJIRCCE, 2014.
[13] Basiri. M. E, Naghsh-NilchiA. R, AND Ghasem-Aghaee. N, "Sentiment Prediction Based On Dempster-Shafer Theory Of Evidence" HINDAWI, 2014.
[14] Tripathi. G, S. N, "Opinion Mining: A Review" IJICT, 2014.
[15] Joshi. N, Itkat. S, "A Survey on Feature Level Sentiment " IJCSIT, 2014.
[16] Saif. H, He. Y and Alani. H, "Semantic Sentiment Analysis of Twitter", ISWC, 2012.
[17] Patni. S, Wadhe. A, "Reviews Paper on Sentiment Analysis is - Big Challenge" IJARCSMS, 2014.
[18] Varghese. R, "A Survey On Sentiment Analysis and opinion mining" association rules", IJRET, 2013.
[19] Shukran. M, Yeh. W, Wahid. N, Zaidi. A, "Artificial Bee Colony based Data Mining Algorithms for Classification Task" Vol. 5, No 4, August 2011.
[20] Akay. B, Karaboga. D, "A modified Artificial Bee Colony algorithm for real-Parameter optimization " ELSEVIER, 2010.
[21] Murugan. R, Mohan. M, "MODIFIED ARTIFICIAL BEE COLONY ALGORITHM FOR SOLVING ECONOMIC DIPATCH PROBLEM" ARPN, 2012.
[22] Khaze. S, Maleki. I, Hojjatklhah. S, Bagherinia. A, "Evaluation The Efficiency of Artificial Bee Colony and The Fireefly Algorithm in Solving The Continuous Optimization Problem " IJCSA, Vol. 3, No. 4, 2013.
[23] Joshi. N, Itkat. S, "Feature Selection with Chaotic Hybrid Artificial Bee Colony Algorithm based on Fuzzy (CHABCF)" ISPACS, 2013.
[24] Guppu Zhu, Sam Kurong, ELSEVIER, Applied Mathematics and computation 217(2010) 3166-3173, "Gbest- guided artificial bee colony algorithm for numerical function optimization."

Author Profile

Ruby Dhurve received the B.E. degree from Chhattigarh Swami Vivekanand Technical University, Bhilai (C.G.) India in Computer Science & Engineering in the year 2013. She is currently pursuing M.Tech. Degree in Computer Science Engineering with specialization in Computer Science & Engineering from CSVTU Bhilai (C.G.), India. Her research area includes Data Mining and Text Mining etc.

Megha Seth is currently Assistant professor in Department of Computer science & Engineering RCET, Bhilai (C.G.) India. She completed her B.E and M.Tech. in Computer Science and Engineering Branch , experience nine year read RCET Bhilai. Her research area includes Data mining, Image processing, Computer Network, AI & NN etc. She has published many Research Papers in various reputed National & International Journals, Conferences, and Seminars.