www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 7 Issue 2 February 2018, Page No. 23656-23662
Index Copernicus Value (2015): 58.10, 76.25 (2016) DOI: 10.18535/ijecs/v7i2.23

Feature Selection for Opinion Mining Using Shuffled Frog Leaping Algorithm

S.M.Hemalatha[1]*, C.S. Kanimozhi Selvi[2]

[1]P.G.Scholar/CSE, Kongu Engineering College,Perundurai, Erode
[2]Professor/CSE, Kongu Engineering College, Perundurai, Erode

Abstract: An essential part of our information-collecting behavior has always been to find out what other people think. With the growing availability and popularity of opinion-rich resources such as online review sites and personal blogs, new opportunities and challenges arise as people now can, and do, actively use information technologies to seek out and understand the opinions of others. Hence, Sentiment Analysis research has increased tremendously in recent times. Sentiment analysis deals with the methods that automatically process the text contents and extract the opinion of the users. In this work, biomedical opinions are extracted from twitter which contains many features needed to classify the opinions. However, such datasets contain many irrelevant or weak correlation features which influence the predictive accuracy of classification. Without a feature selection algorithm, it is difficult for the existing classification techniques to accurately identify patterns in the features. The purpose of feature selection is to not only identify a feature subset from original set of features but also to reduce the computation overhead in data mining .In the proposed feature selection approach, Shuffled Frog Leaping Algorithm (SFLA) algorithm optimizes the process of feature selection and yields the best optimal feature subset which increases the predictive accuracy of the classifier. SFLA is used as a feature selector and generates the feature subset and Naive Bayes, SVM and K-nn classification used to evaluate the feature subset produced. Experimental results show that the Naïve Bayes classification produces better accuracy when the selected features from shuffled frog leaping algorithm are used.

Keywords: Biomedical; Feature Selection; Opinion, SFLA, Sentimental analysis.

1. Introduction

Data mining is the process of examining model in huge data sets that include methods at the intersection of machine learning, statistics, and database systems It is an important process where intelligent methods are applied to collect data models. It is an interdisciplinary subfield of computer science [7][8]. The main aim of the data mining process is to collect information from a data set and convert it into a meaningful structure for further use. Aside from the raw examine phase, it includes database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics ,complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining is the process of "knowledge discovery in databases" process, or KDD [7][8].

Opinion Mining is also called as sentimental analysis. It is a study of people’s thoughts, appraisals, behavior, and emotions toward entities, individuals, issues, events, topics and their variable. The Sentiment is defined as a character, thought, or perception and that is prompted by expression, whereas Opinion is defined as an aspect, judgment formed in the mind about a specific matter [10]. The definitions indicate that an opinion is more of a person view about something, whereas a sentiment is more of a feeling. For example the sentence “I am involved about the present state of the economy” deliberate a sentiment whereas the sentence “I assume the economy is not performing well” expresses an opinion. In most cases opinions imply positive or negative Sentiments [6]. Sentiment analysis and opinion mining is the area of study that examines people’s thoughts, emotion, calculation, behavior from written language. The main aim of sentiment analyses is to identify with the growth of social media such as reviews, forum discussions, Twitter, and social networks media etc. For the first time in human history, users now have a huge volume of opinionated data recorded in digital form for analysis

2. Literature Survey

Hu and Bin(2016) Proposed a feature selection for high dimensional biomedical data using an improved shuffled frog leaping algorithm it contains thousands of features which can be used in molecular diagnosis of disease, the feature selection algorithm, used to accurately identify patterns in the features. The main work of feature selection is to not only identify a feature subset from an original set of features but also reduce the computation overhead in data mining. The proposed work is to search the space of possible subsets to obtain the set of features that maximize the predictive accuracy and minimize irrelevant features in high-dimensional biomedical data. To evaluate the effectiveness of our proposed method we have employed the K-nearest neighbor method with a comparative analysis in which we compare our proposed feature with genetic algorithms, particle swarm optimization, and the shuffled frog leaping algorithm. Experimental results expose our improved algorithm achieves obtain in the identification of similar subsets with classification accuracy

Xia Li (2012) developed a hybrid scheme that associate the merits of a global explore algorithm, the Shuffled Frog-Leaping expose Optimization (EO) and that strong robustness and fast convergence for high-dimensional continuous function optimization. A Modified Shuffled Frog-Leaping Algorithm (MSFLA) is an algorithm that examines that increases the leaping rule by correctly extending the leaping step size and accumulation a leaping inertia component to account for social characteristic. To further enhance the local search ability of MSFLA and speed up convergence. It is characterized by alternating the coarse-grained Cauchy mutation and the fine-grained Gaussian mutation. Compared with other algorithms the MSFLA mainly used for benchmark examples, the hybrid MSFLA-EO is exposing to be a good and strong choice for solving high-dimensional continuous function optimization problems. It possesses excellent performance in terms of the mean function values, the success rate and the Fitness Function Evaluations (FFE), which is a rough measure of the complexity of the algorithm.

3. Feature Selection

Feature selection (Ram Swami M. and Bhaskaran R., 2009) was found to be effective in enhancing learning efficiency, increasing predictive accuracy and reducing complexity of learned results. The main goal of the Feature selection in supervised learning that produces higher classification accuracy With a subset of input variables by terminating features, which are irrelevant or of no predictive information [5][11][15].

Figure.1. presents a unified view of a feature selection process. A common feature selection technique as introduced which contains two steps: feature selection, and model fitting and performance calculation [5][11]. The feature selection model contains three steps: (i) initiate a candidate set accommodate with a subset of the prime features via specific research strategies; (ii) calculating the candidate set and appraisal the fitness of the features in the candidate set. Based on the calculation, few features in the candidate set may be removed or added to the privileged feature set according to their importance; and (iii) considering whether the present set of best features are good enough using specific stopping criterion.

If it is, a feature selection algorithm will evacuate the set of best features, otherwise, it insists until the stopping criterion is met. In the techniques of initiating the candidate set and calculating it, a feature selection algorithm may cause the information from the training data, current selected features, target learning model, and given preceding knowledge to convoy their search and calculation. Once a set of features is selected, it can be used to clarify the training and test data for model fitting and prediction The performance achieved by a specific learning model on the test data and it can be used to as an indicator for calculating the effectiveness of the feature selection algorithm for that learning model. In this techniques determining on how and when the utility of selected features is evaluated, different strategies can be adopted, which broadly classified into three categories: filter, wrapper and embedded models

3.1 Filter Approach

This model evaluates the features according to heuristics based on the general characteristics of the data. The feature selection method is applied independently in addition to the data mining algorithm [5]. Feature relevance score is computed and those features with fewer score are removed. The remaining subset of features is given as input to the classification algorithm. Filter techniques are suitable for high dimensional datasets that are computationally simple and fast. Also, it is separated from the mining algorithm so feature selection demand to be performed only once with different classifiers can be used for evaluation.

3.2 Wrapper Approach

In these wrapper approach uses learning algorithm it determines how good a given attribute subset is [11]. In this model a search method in the space is defined with possible feature, and various subsets of features are accomplished and calculated.

Compare to filter approach the wrapper approach contribute to much slower than the filter approach, as the data mining algorithm is applied to each attribute subset considered by the search[5][15]. The main advantages of these techniques it involves the cooperation between feature subset exploration and model selection, and the ability to take into account feature dependencies. A common disadvantage of this approach is that they have a higher risk of over fitting compare to filter approach with very high computationally intensive.

3.3 Embedded Approach

It is a process in which finding for the best subset of features is constructing into the classifier conception with combined space of feature subsets and hypotheses [5][11]. This approach is similar to wrapper approach because it is specific algorithm. The Advantage of the embedded approach includes the interaction with the classification model, compare to wrapper approach it is less computationally intensive.

It is based on filter approach. Here the feature selection method is separate from the classification algorithm. The subset of feature selected from the shuffled frog leaping algorithm is presented to the classification algorithm and the opinions are classified [15].

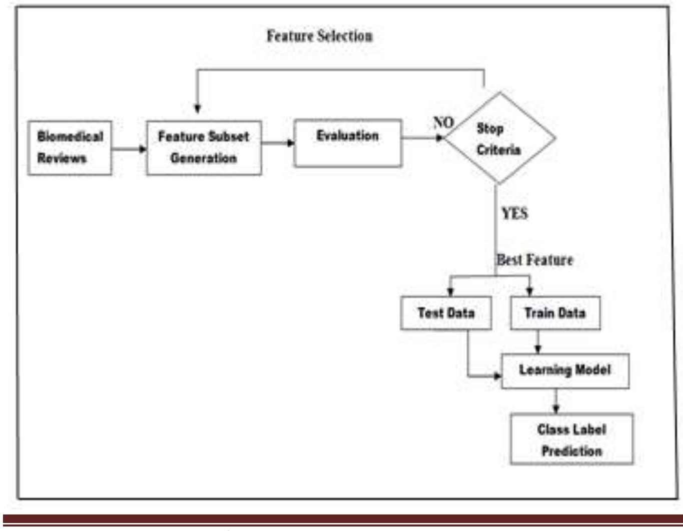

Figure 1: Unified View of Feature Selection Process

4. Shuffled Frog Leaping Algorithm

The SFLA is a nature-inspired algorithms based on the swarm intelligence. It is a mimetic meta-heuristic algorithm, is modeled based on the behaviors of the social frogs. The purpose of the frogs is to find the maximum food with minimum step. SFLA is used for partition data clustering on benchmark problems. Shuffled frog leaping is a relatively new heuristic algorithm which is first introduced by Eusuff and it is a comparatively novel meta-heuristic, which makes use of the cultural environment information named as memes among the population to perform its evolution A meme is an idea, behavior, or style that spreads from person to person within a culture and it can be passed from one to another mind through the way of writing, speech, gestures, rituals, or other aspect with a mimicked thought[1][3][13][14][4]

In the shuffled frog leaping algorithm frogs are divided into several groups and each subgroup is called memeplex [2][9]. In each memeplex uses the instruction of memes generated by the cooperation and competition among the population [12][9]. After the examining in each memeplex, the frogs from all memeplex are shuffled and then frogs are rearranged forming new memeplex, which makes searching process less possible to be trapped in local optimum[1][3][13][14][4].

4.1 Initialization

A randomly distributed population is created, just like different frog in the search space [4][3]. The population size can be determined by the user according to the actual situation. Generally, the population size is an even value between 20 and 100.

4.2 Fitness Function

A fitness function is associated with each frog that represents the degree of fitness of the solution. Considering the real-time classification problem, the objective of feature subset selection is to use fewer features to achieve the same or better performance

Therefore, the fitness evaluation involves two concerns. One is the accuracy of the validation data, and the other is the number of features used. If two subsets with a different number of features achieve the same performance, the subset with fewer features is preferred.

4.3 Evolution

Sort the population in order of decreasing function. Store the sorted population in array so the first position belongs to the best frog Separate the frog into memeplexes .each memeplexes contain n frog Choose a sub-memeplexes from current memeplexes. The sub-memeplexes selection strategy depends on performance of the frogs in current memeplexes. The frogs which have better fitness value have more selection chance than the worse frogs. X_b as the best frog and X_w as the worst frog in sub-Memeplexes are marked [13].

4.4 Perform local and global search

In SFLA algorithm perform two search local search means performed within the memeplexes i.e. subgroup. Local search performed within the only subgroup. It solves the benchmark problems. When using ISFLA it improves the local search space. Global is performed in whole memeplexes i.e. group It also solve benchmark problems

4.5. Updation*:*

If new frog is better than the worst frog in the memeplexes replace the worst frog otherwise apply (1) and (2) replace worst frog [13]

$$Di=rand (X_b - X_w) \quad (1)$$

$$X_{new}(w) = X_{old} (D_{max} \leq i \leq D_{min}) \quad (2)$$

5. System Design and Module Description

The main steps of the proposed feature selection method are illustrated as given in Figure 2. Each step is described as follows

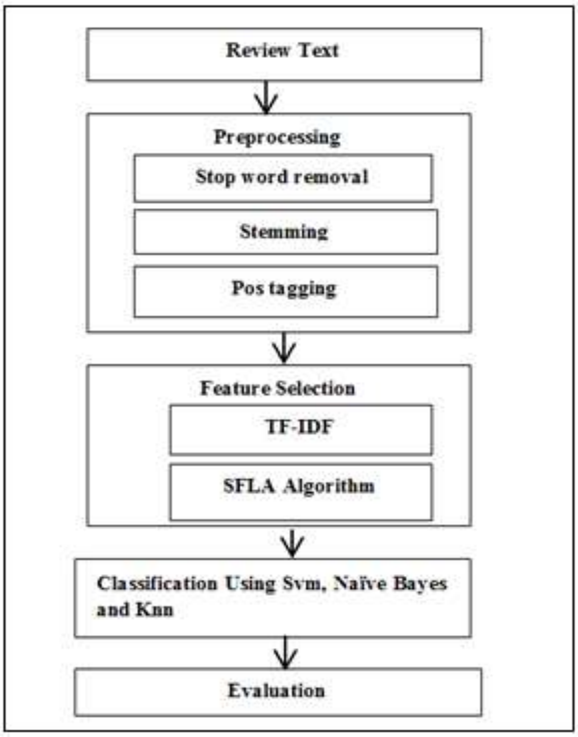

Figure 2: System Design

The proposed system extracts a number of positive and negative opinions. The modules of the proposed system are

1. Data Preprocessing
2. Feature selection
3. Classification

5.1 Data Preprocessing

Data preprocessing is a data mining method that converts raw data into an understandable format with the help of Data Preprocessing Steps Figure 3. Real-world data is often containing unaccomplished data, and deprivation in certain attitude or trends with contains many errors. It is used to eradicate unpredictable information in reviews.

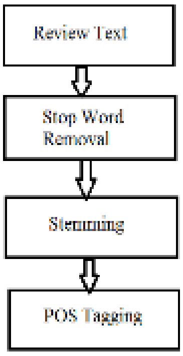

Figure 3: Working on Data Preprocessing Steps

5.1.1 Stop Word Removal

Most frequently used words in English are not useful in text mining. Such words are called stop words. Stop words are language specific functional words which carry no information. It may be of types such as pronouns, prepositions, conjunctions.

Stop word removal is used to remove unwanted words in each review sentence. Words like is, are, was etc. Reviews are stored in text file which is given as input to stop word removal. Stop word is removed by checking against stop words list.

5.1.2 Stemming

Stemming is used to form root word. A stemming algorithm reduces the words "fishing", "fished", and "fisher" to the root word, "fish". There are many algorithms like n-gram analysis, Affix stemmers and Lemmatization algorithms. Porter Stemmer algorithm is used to form root word for given input reviews and store it in text file.

5.1.3 POS Tagging

The Part-Of-Speech of a word is a linguistic category that is defined by its syntactic or morphological behavior. Common POS categories in English grammar are a noun, verb adjective, adverb, pronoun, preposition, conjunction, and interjection. POS tagging is the province of specifying (or tagging) each word in a sentence with its applicable part of speech. POS tagging is an important phase of opinion mining, it is necessary to resolve the features and opinion words from the reviews text. POS tagging can be done manually or with the help of POS tagger. Manual POS tagging of the reviews takes more times to process. Here, POS tagger is used to tag all the words of reviews. Stanford tagger is used to tag each word in a review sentence. Finally, nouns are collected and stored in a text file.

5.2 Feature Selection

The features are selected according to the relevance of the feature. The TF-IDF measure is used for mapping text to numeric.

5.2.1 TF: Term Frequency

It measures how frequently a term occurs in a document. Since every document disparate in length, it is possible that a term would appear many times in long document compared to shorter one. Thus, the term frequency is often divided by the document length as a way of normalization:

$$\text{TF} = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}}$$

5.2.2 IDF: Inverse Document Frequency

It measures how important a term is. While computing TF, all terms are considered equally relevant. However it is known that specific terms, such as "is", "of", and "that", may present many times but have little considerable. Thus we need to weigh down the frequent terms while scale up there is ones, by computing the following:

$$\text{IDF=LOG}\ \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}}$$

The cosine of the angle between them TF-IDF values of the extracted nouns is found and stored in a matrix.

5.2.3 Cosine Similarity

Cosine similarity is a measure of how identical two documents are likely to be in phrase of their subject matter. It is a measure of vector space between two documents that evaluate them.

5.3 Steps of SFLA Based Feature Selection

The main steps of the SFLA Based feature selection method are illustrated as given in Figure 4.

Step1: Extract Biomedical Data From Twitter
Step 2: Apply Preprocessing
  i) Stop Word Removal
  ii) Stemming
  iii) Pos Tagging
Step3: Extract Noun Using Pos Tagging
Step4: Calculate TF-IDF and Cosine Similarity for each Noun
Step 5: Arrange nouns in ascending order of their cosine similarity values
Step6: Separate the Noun into groups (memeplexes)
Step7: Divide the extracted Nouns equally into memeplexes
Step8: Substitute Position value as Cosine Similarity (ie. X)
Step 9: Repeat until Maximum iteration is reached
  Calculate Cost Function using the formula (X. ^2)
  Find Local Best value and Local Worst value in each memeplex
  IF Local Best value is better than the Local Worst value
  THEN Substitute Local Best value=Local Worst Value
  ELSE Randomly Generate Value using (1) and (2)[13]
Step10: Find the Global Best value from the Local Best value
Step11: Return the Optimal Best value

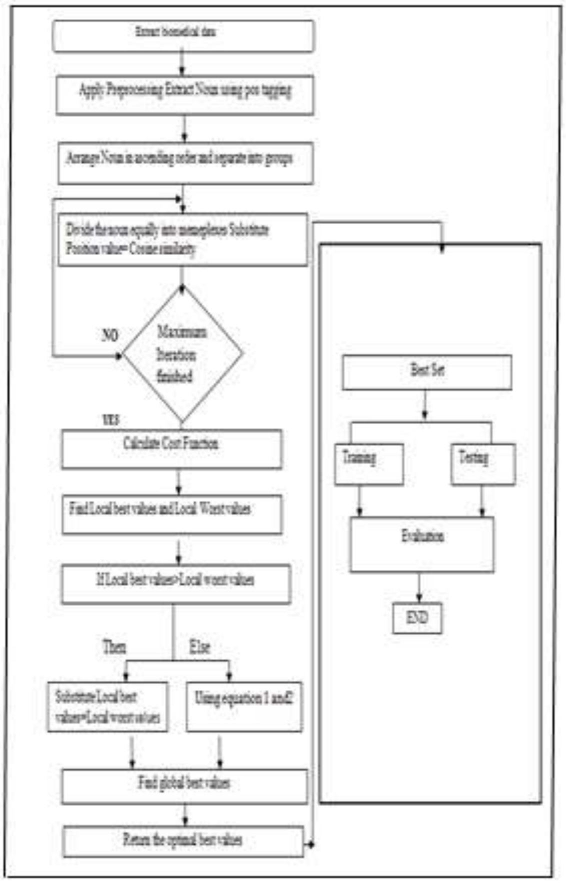
Extract biomedical data
Apply Preprocessing Extract Noun using pos tagging
Arrange Noun in ascending order and separate into groups
Divide the noun equally into memeplexes Substitute Position value= Cosine similarity
NO
Maximum Iteration finished
YES
Calculate Cost Function
Find Local best values and Local Worst values
If Local best values>Local worst values
Then
Else
Substitute Local best values=Local worst values
Using equation 1 and2
Find global best values
Return the optimal best values
Best Set
Training
Testing
Evaluation
END

Figure 4: Feature Selection Process with Classifier

5.4 Classification using SVM, Naive Bayes and K-nn

The Support Vector Machines (SVM) is a data classification method that separates data using hyper planes. SVM technique is generally used for data which has non-regularity which means, data whose distribution is unknown Naive Bayes is a simple method and that is mainly used for building classifiers models that assign class labels to problem instances, signified as vectors of feature values, where the class labels are strained from specific limited set.

A Support Vector Machine, Naive Bayes, and Knn classifier are built using the training data subset of best features selected from the SFLA algorithm. The model is then evaluated with the test data

6. Result and Discussion

Twitter account has been created and 1500 tweets were extracted with hash tag "breast cancer". In the pre-processing step, retweets were deleted. Tweets are labeled manually as positive and negative tweets. 70% of the tweets are given for training and 30% are given for testing. SVM, Naive Bayes and K-nn classification are performed in two phases. In the first phase, the original tweets after pre-processing are given as input to the classifier. In the second phase, the optimal features are only considered for classification and the results are evaluated.

Table 1 illustrates the accuracy of the algorithms with and without feature selection process

Table 1: Comparison of the Method

	Without Feature Selection	With Feature Selection
SVM	0.67	0.74
Naïve Bayes	0.77	0.88
K-nn	0.52	0.43

Figure 5: shows the comparing the without and with Feature Selection method using SVM , Naive Bayes and Knn classifier

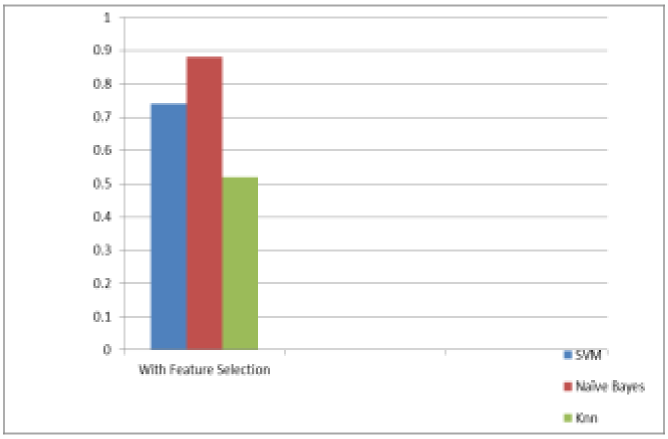

Figure 5: Without Feature Selection in different classifier

The accuracy of the classifier without feature selection in Naive Bayes classifier is 10% higher than the SVM classification without feature selection.

Figure 6 shows the comparing the without and with Feature Selection method using SVM and Naive Bayes classifier

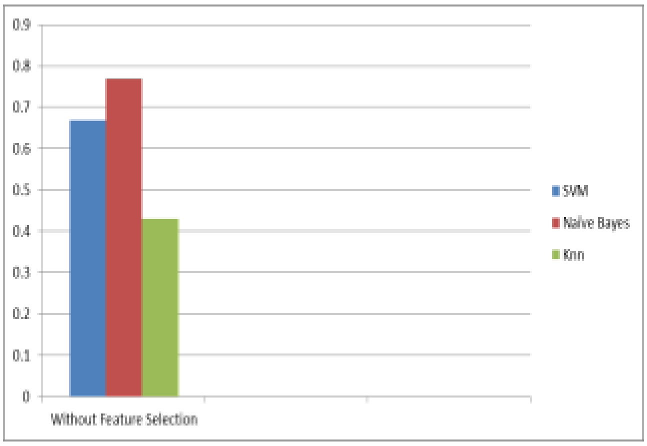

Figure 6: With Feature Selection in different classifier

The accuracy of the classifier without feature selection in Naive Bayes classifier is 14% higher than the SVM classification without feature selection.

The accuracy of the classifier without feature selection in K-nn classifier is 0.09% higher than the K-nn classification without feature selection.

7. Conclusion and Future Work

The purpose of feature selection is to not only identify a feature subset from original set of features but also to reduce the computation overhead in data mining. Without a feature selection algorithm, it is difficult for the existing classification techniques to accurately identify patterns in the features.

In this work, biomedical opinions are extracted from twitter which contains many features needed to classify the opinions. However, such datasets contain many irrelevant or weak correlation features which influence the predictive accuracy of classification. SFLA is implemented as a feature selector and the feature subset is generated and SVM, Naive Bayes, K-nn classifier is used to evaluate the feature subset produced.

SFLA algorithm optimizes the process of feature selection and yields the best optimal feature subset which increases the predictive accuracy of the classifier. Experimental results show that the naïve Bayes and K-nn classification produces an increase in accuracy of 4% when compared to SVM classification with selected set of features from shuffled frog leaping algorithm are used.

Future work will concentrate on improving the accuracy which can be done with many natural inspired algorithms like Cuckoo Search, Paddy field Optimization and other nature-inspired optimization algorithms with SVM and other classification methods.

References

[1]Amiri, Babak, Mohammad Fathian, and Ali Maroosi. "Application of Shuffled Frog-Leaping Algorithm on Clustering." The International Journal of Advanced Manufacturing Technology 45.1, 199-209, 2009.

[2] Hany Hasanien, Senior Member, “Shuffled Frog Leaping Algorithm for Photovoltaic Model Identification,” IEEE Transactions on Sustainable Energy, vol. 6, pp. 509-515, 2015

[3] Hu, Bin,. "Feature selection for optimized high-dimensional biomedical data using the improved shuffled frog leaping algorithm." IEEE/ACM transactions on computational biology and bioinformatics (2016).

[4] Javad Ebrahimi, Syed Hossein Hosseinian, and B. Gevorg Gharehpetian, “Unit Commitment Problem Solution Using Shuffled Frog Leaping Algorithm”, IEEE Applied Mathematics and Computation, vol. 218, pp.9353-9371, 2012.

[5]Jorge Vergara. A.Pablo Estevez, “A Review of feature selection methods based on mutual information,” Neural Computation & Application, vol.24, pp. 175-186, 2014.

[6]Liu, Bing. "Sentiment analysis and Opinion mining." Synthesis lectures on human language technologies 5.1, 1-167, and 2012.

[7]C. Kalaichelvi, and K. Selvi. "Frequent itemsets generation using collective support threshold for associative classification." National Conference on Recent Trends in Communication and Signal Processing. Vol. 2009.

[8] C.S .Kanimozhiselvi, and A. Tamilarasi. "Mining of High Confidence Rare Association Rules with Automated Support Thresholds." European Journal of Scientific Research 52.2, 2011.

[9] Ling Wang, Chen Fang, “An effective Shuffled frog-leaping algorithm for multi-mode resource-constrained project scheduling problem,” Information Sciences, vol. 181, pp. 4804-4822, 2011.

[10] Annamalai, R., J. Srikanth. "Accessing the Data Efficiently using Prediction of Dynamic Data Algorithm." International Journal of Computer Applications 116, no. 22 2015.

[11]Yong Zhang, Dunwei Gong, Ying Hu ,and Wanqiu Zhang, “Feature selection algorithm based on bare-bones particle swarm optimization,” Neuro computing,vol. 148, pp. 150-157, 2015.

[12]Xia Li, Jianping Luo, Minrong Chen and Na Wang, “An Improved Shuffled Frog-leaping Algorithm with External Optimization for continuous optimization,” Information Sciences, vol. 192, pp. 143-151, 2012.

[13] Samuel, G. Giftson, and C. Christober Asir Rajan. "A Modified Shuffled Frog Leaping Algorithm for Long-Term Generation Maintenance Scheduling." Proceedings of the Third International Conference on Soft Computing for Problem Solving. Springer, New Delhi, 2014.

[14] Mohan P, Chelliah S. An Authentication Technique for Accessing De-Duplicated Data from Private Cloud using One Time Password. International Journal of Information Security and Privacy (IJISP). 2017 Apr 1;11(2):1-0.

[15] Yong Zhang, Dunwei Gong, Ying Hu, Wanqiu Zhang, “Feature selec-tion algorithm based on bare bones particle swarm optimization,” Neurocomputing, vol. 148, pp. 150-157, 2015

.