JOURNAL OF PURE AND APPLIED MICROBIOLOGY, Aug. 2015. Vol. **9**, p. 17-23
SPECIAL ISSUE ON RECENT RESEARCH CHALLENGES IN BIO-MEDICAL APPLICATIONS

Artificial Bee Colony Optimization for Feature Selection With Furia In Opinion Mining

T. Sumathi[1], S. Karthik[2] and M. Marikkannan[3]

[1]Department of Computer Science and Engineering,
Institute of Road and Transport Technology, Erode,Tamilnadu, India.
[2]Department of Computer Science and Engineering,
SNS College of Technology, Coimbatore, Tamilnadu, India.
[3]Department of Computer Science and Engineering,
Instutute of Road and Transport Technology, Erode, Tamilnadu, India.

(Received: 19 February 2015; accepted: 22 April 2015)

Opinion mining, a sub-discipline of information retrieval and concerns not with what a document is about, but with its expressed opinion. Feature selection is an important step in opinion mining, as customers express product opinions separately according to individual features. Statistical techniques like Correlation based Feature Selection (CFS) have been extensively used for feature selection to reduce the corpus size. Feature selection is NP hard and selecting features using statistical techniques is suboptimal. In this work a novel feature selection technique using Multi Objective Artificial Bee Colony algorithm is proposed. The proposed technique is evaluated using Fuzzy Unordered Rule Induction Algorithm (FURIA) classifier. Results show improved classification accuracy in the proposed technique.

Key words: Opinion Mining, Fuzzy Unordered Rule Induction Algorithm (FURIA), Correlation based Feature Selection (CFS), Artificial Bee Colony (ABC), Feature selection, Inverse Document Frequency (IDF).

Opinion Mining, also called Sentiment analysis is an emerging research area that classifies automatically people's opinion and emotions. Opinions can be in the form of reviews of products, a specific topic or services[1,2]. Opinion mining becomes important as people are moving towards internet based opinions before making a purchase decision.

Feature selection is an activity, which select relevant features based on a particular measurement. Its main purposes are to simplify the training process and to reduce the time of training process. It is important to select a feature selection technique which reduces the number of features without reducing the performance of opinion mining. There are three groups of feature selection techniques i.e. filter, wrapper and embedded[3]. In a filter category, a group of features is selected based on a particular mathematical equation and may be used with any classifier. Contrary to that, the features that are selected in the wrapper and the embedded techniques are bound to a particular classifier. Other than very rigid in term of classifier, the wrapper and embedded techniques normally require high allocation of resource and require longer execution time[4]. Feature selection methods provide a criterion for eliminating terms from document corpus to reduce vocabulary space. Commonly used feature selections used in the literature are[5-9] Information gain, Mutual Information, Chi Square test, Genetic Algortihm.

A classic example of statistical feature extraction technique is the Information gain. It is based on the presence and absence of a term in a

* To whom all correspondence should be addressed.

document a threshold is set and the terms with less information gain is removed. *Odd Ratio*: It is suitable for binary class domain where it has one positive and one negative class for classification. The algorithm is run on each class and the top- n features are taken from the sorted list. *Document Frequency*: Measures the number of appearances of a term in the available number of documents in the corpus and based on the threshold computed the terms are removed. Once the features are selected they can be classified using a rule based classifier as shown in figure 1. Given a training set, the Rule Based Classifier (RBC) uses a Rule Generator to generate a set of rules. This set of antecedents derived from the training set represents the test sample and is used to classify the test sample. The general formats of the rule formed are shown in Figure 1[8].

A rule *r* covers an instance when Left Hand Side (LHS) of the rule covers the class on the Right Hand Side (RHS). Opinion Mining extracts opinions from a set of documents expressed by a specified target. Feature selection in classification is viewed as fundamental in machine learning.

This work uses Artificial Bee Colony (ABC) optimization algorithm for feature selection. Features from IMDb dataset are extracted using Inverse Document Frequency (IDF) and feature selected using the proposed ABC method. The selected features are classified as positive or negative review using FURIA classifier. Section 2 presents the problem formulation. Section 3 describes the methods used and section 4 discusses about experiments and results. Section 5 narrates the conclusion of work.

Related works

Soliman *et al*[10] Support Vector Machine (SVM) for classification of Opinions, Akaichi *et al*[11] constructed a sentiment lexicon and opinions were classified using SVM and Naive Bayes. Sima and Vunvulea[12] proposed a solution for extracting opinions and opinion holders from large texts. The goal is to achieve a high level of domain independence by implementing a rule-based approach. The results of proposed system prove an accuracy which was comparable to that of systems using a supervised learning approach, which was domain dependent. Medagoda *et al*[13] investigated opinion mining in three non English language. Lo and Potdar[14] discussed various techniques used in literature for opinion mining. Mazzonello *et al*[15] focused on the application and the comparison of three classification techniques over a text corpus composed of reviews of commercial products in order to detect opinions about them. The chosen domain was about "perfumes", and user opinions composing the corpus are written in Italian language. The proposed approach was completely data-driven: a Term Frequency / Inverse Document Frequency (TFIDF) terms selection procedure has been applied in order to make computation more efficient, to improve the classification results and to manage some issues related to the specific classification procedure adopted.

Wang *et al*[16] proposed threshold-based feature selection techniques, Li and Chen[17] proposed feature selection mechanism for Chinese documents. Liao and Pan[18] presented a new approach combining AIG with WA called AIG-WA. Better performances were obtained when applying this method on Chinese text dataset provided by Fudan Database Center. Samsudin *et al*[19] prposed a mutual information-based technique to perform feature, Baccianella *et al*[20] presented six feature selection techniques for ordinal classification problem.

METHODOLOGY

Feature selection is the supervised algorithm finds a feature set that produces higher classification accuracy. But traditional approaches to feature selection with single evaluation criteria have limited capability about the knowledge discovery and decision support. Figure 2 shows the flow chart for the proposed methodology. The proposed technique is evaluated using IMDb Dataset. The opinions are classified as positive and negative. The preprocessing of the data includes removal of stop words and stemming of words. The features are extracted using Inverse document frequency (IDF) and the proposed feature selection is used for feature selection. The selected features are classified using FURIA. The methods specified in the flow chart are explained in the following sections.

IMDb database

IMDb database provides comprehensive information on movies which have been produced till now[21,22]. Viewers can see the summary and plot of the movie and can give their rating based on a ranking scale with 1 denoting awful and 10 denoting excellent[22,23].

Pre-processing Data

In the preprocessing stage stopwords which are commonly occurring terms like "I", "we', "us', "and" are removed as they are termed noise [24]. In the next stage stemming is used to remove suffixes and prefixes[25, 26]. In this work Porter stemming is used. Once the noise is cleaned we use Inverse Document Frequency to create the feature vector space.

Inverse document frequency (IDF) is an widely used concept in information retrieval. It has been used in various application areas like databases natural language processing, knowledge management, text classification and information retrieval. For term t_i on total number of document D the IDF is obtained using equation (1)

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \quad ...(1)$$

Where $|\{d : t_i \in d\}|$ is the number of documents where the term t_i appears. To avoid divide-by-zero, $1 + |\{d : t_i \in d\}|$ is used.

Table 1. Term Frequency Table

Document 1		Document 2	
Term	Term Count	Term	Term Count
Movie	7	Movie	9
good	2	interest	3
music	4	music	2
actor	6	plot	3

Consider the term frequency table consisting of only two documents as shown in table

The IDF for the term "movie" in document 1 is calculated as in equation (2):

$$idf(this, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \quad ...(2)$$

Here number of documents present is two and also number of documents which holds "this" is two and so the value is computed as

$$idf(this, D) = \log \frac{2}{2} = 0 \quad ...(3)$$

A popular technique for feature selection is the Correlation Based Feature Selection (CFS). It depends on a heuristic for estimating the quality of a subset of features. It considers the usefulness of a feature for predicting the class label and also the intercorrelation among other features in the subset. The correlation of a feature is computed as:

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k + k - (k-1)\overline{r_{ii}}}} \quad ...(4)$$

where r_{zc} is the correlation between the summed k number of features, $\overline{r_{zi}}$ is average of the correlations between features and $\overline{r_{ii}}$ = average inter-correlation between features.

Proposed Feature Selection

In this work, Multi Objective Artificial Bee Colony (MO ABC) meta heuristic algorithm is used for feature selection and creates the best optimal feature subset to increase the classification accuracy and reduce the Root Mean Squared Error (RMSE). A novel fitness function based on number of features, classification accuracy and RMSE is proposed. The objective of feature selection is to

Table 2. Summary of Results

Methods	RMSE	Classification accuracy	Precision	Recall
IDF – FURIA	0.2312	92.25	0.92335	0.9225
CFS feature selection - FURIA	0.2184	92.75	0.9275	0.9275
ABC feature selection - FURIA	0.1906	93.75	0.9376	0.9375
CFS feature selection – Proposed FURIA	0.1546	94.5	0.945	0.945
ABC feature selection – Proposed FURIA	0.1428	95.25	0.9525	0.9525

a) Increase classification accuracy
b) Reduce RMSE
c) Reduce the number of feature vectors

To ensure the above three objectives a novel multi objective fitness function is proposed

Let f_i be the number of feature vector used i=1,2.....,n
F be the total number of feature vector
δ be the RMSE
C_r be the normalized classification accuracy
The proposed fitness function is given by

$$\min g(x) = \frac{f_i \delta e^{-(C_r)^2}}{F} \quad ...(5)$$

The above objective function is evaluated using ABC algorithm which is modified to handle binary values. The steps in ABC [29] is given by

Select random food sources (a combination of feature vectors)
Evaluate using equation 5
Produce new *feasible solutions*
Start greedy selection process
Compute fitness and probability *of selection for next iteration*
Abandon poor food sources
Initiate onlooker bee for searching food source using local search
Initiate scout bee to search new food source
Memorize best food source
Repeat step 3 through 8 until termination criteria met

Fuzzy Unordered Rule Induction Algorithm (FURIA)

Fuzzy Unordered Rule Induction Algorithm (FURIA) uses principles of fuzzy logic for classification and is an extension of the popular algorithm RIPPER. For an instance x_i, the degree of membership can be found using equation (6)[30,31]:

$$\mu_r F(x) = \prod_{i=1...k} i_i^F(x_i) \quad ...(6)$$

Rule(r_n): *(Condition)* $\rightarrow y$
where
Condition is a **conjunction of attribute tests**
$(A_1 = v_1)$ **and** $(A_2 = v_2)$ **and** ... **and** $(A_n = v_n)$
y is the **class label**
LHS: rule antecedent or condition
RHS: rule consequent

Fig. 1. Format of Rule formed for classification

Using only relevant training data for a single antecedent the data can be partitioned into two subsets and the quality measured using equation (7) and (8)

$$D_T^i = \{x = (x_1 ..... x_k) \in D_T^i \mid I_j^F(x_j) > 0 \text{ for all } j \neq i\} \subseteq D_T \quad ...(7)$$

$$pur = \frac{p_i}{p_i + n_i} \quad ...(8)$$

Where

$$p_i \overset{def}{=} \sum_{x \in D_{T-}^i} \mu_{A_i}(A), \; n_i \overset{def}{=} \sum_{x \in D_{T-}^i} \mu_{A_i}(A)$$

The fuzzy rules $r_1^{(j)} ... r_k^{(j)}$ have learned for the class λ_j the support of this class is defined by equation (10):

$$S_j(x) \overset{def}{=} \sum_{i=1,...,k} \mu_{r_i^{(j)}}(x).CF(r_i^{(j)}) \quad ...(10)$$

where the certainty factor of the rule is defined as shown in equation (11):

$$CF(r_i^{(j)}) = \frac{2\frac{|D_T^{(j)}|}{D_T} + \sum_{x \in D_T^i} \mu_{r_i^{(j)}}(x)}{2 + \sum_{x \in D_T} \mu_{r_i^{(j)}}(x)} \quad ...(11)$$

EXPERIMENTAL RESULTS

Performance of the features selected using proposed optimized ABC method for classifying opinions is evaluated and compared without feature selection and CFS based feature

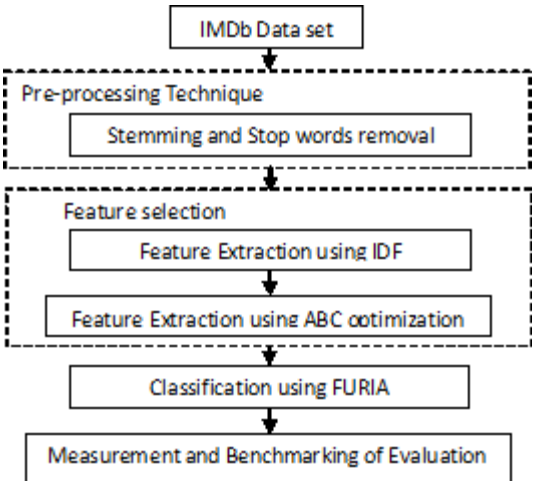

Fig. 2. Flow chart for Proposed Methodology

selection. The features are classified using FURIA. The experimental results achieved for classification accuracy, Root Mean Square Error (RMSE), Precision and Recall are shown in table 2.

From figure 3, it is observed that the proposed ABC feature selection improves the efficiency of FURIA by 1.61% when compared to IDF and by 1.07% when compared to CFS. It is

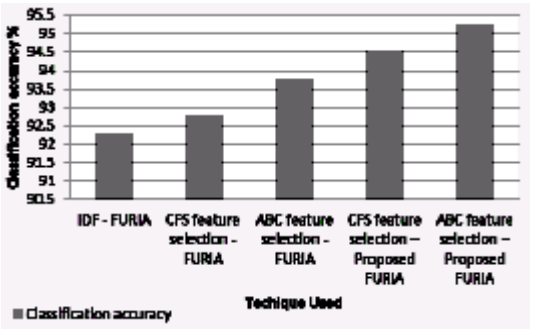

Fig. 3. Classification Accuracy

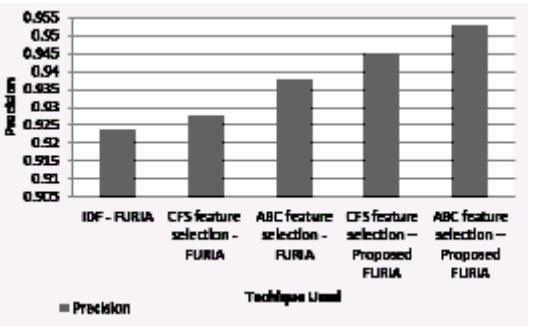

Fig. 4. Precision

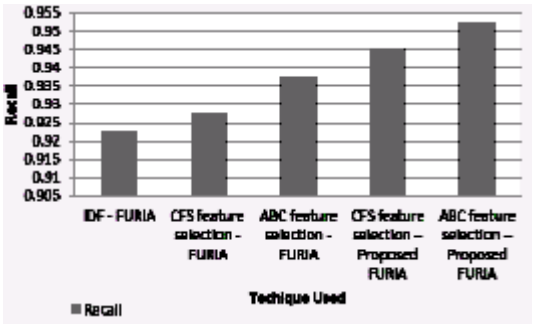

Fig. 5. Recall

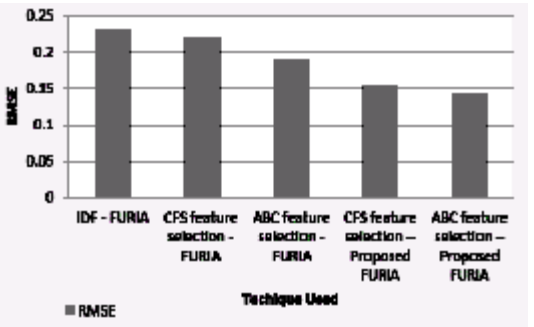

Fig. 6. RMSE

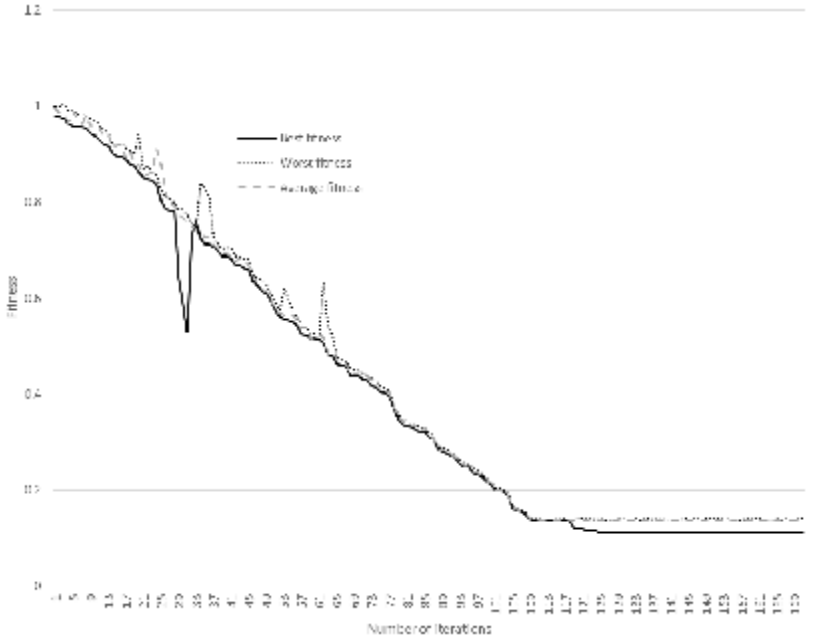

Fig. 7. Fitness Function of the Proposed Feature Selection Technique

also seen that the proposed ABC feature selection with FURIA method improved the classification accuracy by 3.2% when compared with conventional FURIA and improved by 2.66% when compared to FURIA with CFS feature selection.

From figure 4, the proposed ABC feature selection with FURIA method improved the precision value by 3.16% when compared to the conventional FURIA and improved by 2.7% when compared to FURIA with CFS feature selection. Figure 5 shows that the proposed ABC feature selection with FURIA method improved the recall value by 3.25% when compared with conventional FURIA and improved by 2.7% when compared to FURIA with CFS feature selection. The fitness function of the proposed feature selection technique is shown in figure 7. It can be seen that the proposed algorithm performs consistently across multiple runs with minimal deviation in the fitness value showing the ruggedness of the proposed feature selection technique.

CONCLUSION

In this work, a novel feature selection technique using Artificial Bee Colony algorithm for movie review opinion classification is proposed. A novel objective function was proposed which not only reduces the number of features but also improves the classification accuracy and decreases the Root Mean Squared Error. The proposed technique was evaluated on the popular IMDb movie data set. The reviews are pre-processed by stemming and removal of stop words. The proposed technique was evaluate with CFS based feature selection using FURIA for classification. Experimental results show that the classification accuracy of the classifiers improves up to 3.25% with the proposed ABC feature selection method.

REFERENCES

1. Liu, B. "Sentiment analysis and opinion mining", *Synthesis Lectures on Human Language Technologies*, 2012; **5**(1), pp.1-167.
2. Pak, A., & Paroubek, P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *In LREC*, 2010; **10**: pp. 1320-1326.
3. Samsudin, N., Puteh, M., Hamdan, A. R., and Nazri, M. Z. A. "Immune Based Feature Selection for Opinion Mining", *In Proceedings of the World Congress on Engineering*, 2013; **3**.
4. Abbasi, A., Chen, H., and Salem, A., "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums", *ACM Transactions on Information Systems*, 2008; **26**(3): pp.12.
5. Selvam, B., and Abirami, S. "A survey on opinion mining framework", *International Journal of Advanced Research in Computer and Communication Engineering*, 2013; **2**(9): pp. 3544-3549.
6. Agarwal, B., and Mittal, N. "Categorical probability proportion difference (CPPD): A feature selection method for sentiment classification", In 24th International Conference on Computational Linguistics, 2012; p. 17.
7. Gopi, S. C., Reddy, V. S. N., Chandana, M., and Kumar, T. A. "An Efficient Method to Solve Optimization Problem in Feature Selection", *International Journal of Advanced Research in Computer Science and Software Engineering*, 2012; **2**(4): pp.283-286.
8. Chandrakala, S., and Sindhu, C. "Opinion Mining and sentiment classification a survey", *ICTACT journal on soft computing*, 2012; **3**(1): pp. 420-427.
9. Sharma, A., and Dey, S. "Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis", *IJCA Special Issue on Advanced Computing and Communication Technologies for HPC Applications*, 2012; **3**: pp.15-20.
10. Soliman, T. H. A., Elmasry, M. A., Hedar, A. R., and Doss, M. M. "Utilizing support vector machines in mining online customer reviews", *In Proceeding IEEE International Conference on Computer Theory and Applications:* 2012; pp. 192-197.
11. Akaichi, J., Dhouioui, Z., and Lopez-Huertas Perez, M. J. "Text mining facebook status updates for sentiment classification", In Proceeding *IEEE International Conference on System Theory, Control and Computing*, 2013; pp. 640-645.
12. Sima, I. M., and Vunvulea, M. "A rule-based, domain independent approach for opinion and holder identification", *In Proceeding IEEE International Conference on Intelligent Computer Communication and Processing*,2013; pp. 55-62.
13. Medagoda, N., Shanmuganathan, S., and Whalley, J. "A comparative analysis of opinion mining and sentiment classification in non-English languages", *International Conference on Advances in ICT for Emerging Regions*, 2013;

pp.144-148.

14. Lo, Y. W., and Potdar, V. "A review of opinion mining and sentiment classification framework in social networks", *In Proceeding IEEE International Conference on Digital Ecosystems and Technologies,* 2009; pp. 396-401.
15. Mazzonello, V., Gaglio, S., Augello, A., and Pilato, G. "A Study on Classification Methods Applied to Sentiment Analysis", *In Proceeding IEEE International Conference on Semantic Computing,* 2013; pp. 426-431.
16. Wang, H., Khoshgoftaar, T. M., and Van Hulse, J, "A comparative study of threshold-based feature selection techniques", *In Proceeding IEEE International Conference on Granular Computing,* 2010; pp. 499-504.
17. Li, Y., and Chen, C. "Research on the feature selection techniques used in text classification" , *In Proceeding IEEE International Conference on Fuzzy Systems and Knowledge Discovery,* 2012; pp. 725-729,.
18. Liao, Y., and Pan, X. "A Novel Feature Selection Approach and Feature Weight Adjustment Technique in Text Classification", *In Proceeding IEEE International Conference on Software Engineering Research, Management and Applications,* 2009; pp. 41-44.
19. Samsudin, N., Puteh, M., and Hamdan, A. R. "Bess or xbest: Mining the malaysian online reviews", *In Proceeding IEEE International Conference on Data Mining and Optimization,* 2011; pp. 38-43.
20. Baccianella, S., Esuli, A., and Sebastiani, F. "Feature selection for ordinal text Classification", *Journal of Neural Computation,* 2013; **26**(3): pp.557-591.
21. Jotheeswaran, J., and Kumaraswamy, Y. "Opinion mining using decision tree based feature selection through manhattan hierarchical cluster measure", *Journal of Theoretical & Applied Information Technology,* 2013; **58**(1): pp.72-80.
22. Avancha, S., Kallurkar, S., and Kamdar, T. "Design of Ontology for the Internet Movie Database (IMDb)", *Semester Project, CMSC,* 2001; **771**.
23. Srikanth, S V., Ayesha, R., Motoki S., Chandrasekaran R L., and Anand A. Predicting Indian Movie Ratings on IMDB.
24. Fumani, M. R. F. Q., and Ramachandra, C. S. "The concept of stopwords in persian chemistry articles: a discussion in automatic indexing", *International Journal of Theoretical Linguistics,* **1**(1): pp.100–103.
25. Dragut, E., Fang, F., Sistla, P., Yu, C., and Meng, W. "Stop word and related problems in web interface integration", *Proceedings of the VLDB Endowment,* 2009; **2**(1); pp.349-360.
26. Jivani, A. G. "A Comparative Study of Stemming Algorithms", *International Journal of Computer Technology and Applications,* 2011; **2**(6): pp.1930-1938.
27. Karakos, D., Dredze, M., Church, K., Jansen, A., and Khudanpur, S. "Estimating document frequencies in a speech corpus", *In Proceeding IEEE Workshop on Automatic Speech Recognition and Understanding,* 2011; pp. 407-412.
28. TSai, P. W., Pan, J. S., Liao, B. Y., and Chu, S. C. "Enhanced artificial bee colony optimization", *International Journal of Innovative Computing, Information and Control,* 2009; **5**(12): pp. 5081-5092.
29. Schiezaro, M., and Pedrini, H. "Data feature selection based on Artificial Bee Colony algorithm", *EURASIP Journal on Image and Video Processing,* 2013; **1**: pp.1-8.
30. Palanisamy, S., and Kanmani, S. "Artificial Bee Colony Approach for Optimizing Feature Selection", *International Journal of Computer Science Issues,* 2012; **9**(3): pp.432-438.
31. Gasparovica, M., and Aleksejeva, L. "Using fuzzy unordered rule induction algorithm for cancer data classification", *Proceedings of 17th International Conference on Soft Computing,* 2011; pp 141-147.
32. Rahman, M. M., and Davis, D. N. "Fuzzy unordered rules induction algorithm used as missing value imputation methods for K-Mean clustering on real cardiovascular data", *In Proceedings of the World Congress on Engineering,* 2012; **1**: pp.391.