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

T. Sumathi¹, S. Karthik² and M. Marikkannan³

¹Department of Computer Science and Engineering, Institute of Road and Transport Technology, Erode, Tamilnadu, India.
²Department of Computer Science and Engineering, SNS College of Technology, Coimbatore, Tamilnadu, India.
³Department of Computer Science and Engineering, Institute 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. 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. 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. 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 Information gain, Mutual Information, Chi Square test, Genetic Algorithm.

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
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 19.

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 discussed 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*\(^{19}\) proposed 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\textsuperscript{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\textsuperscript{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\textsuperscript{24}. In the next stage stemming is used to remove suffixes and prefixes\textsuperscript{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 a 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)

$$\text{idf}_{t_i} = \log \frac{N}{|\{d : t_i \in d\}|}$$  \hspace{1cm} \text{(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

<table>
<thead>
<tr>
<th>Term</th>
<th>Document 1</th>
<th>Document 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movie</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>good</td>
<td>2</td>
<td>interest 3</td>
</tr>
<tr>
<td>music</td>
<td>4</td>
<td>music 2</td>
</tr>
<tr>
<td>actor</td>
<td>6</td>
<td>plot 3</td>
</tr>
</tbody>
</table>

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):

$$\text{idf} (\text{movie}, D) = \log \frac{N}{|\{d : \text{movie} \in d\}|}$$  \hspace{1cm} \text{(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

$$\text{idf} (\text{this}, D) = \log \frac{2}{2} = 0$$  \hspace{1cm} \text{(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:

$$r_{zi} = \frac{k \overline{r}_{zi} - (k-1) \overline{r}_{ii}}{\sqrt{k + k - (k-1) \overline{r}_{ii}}}$$  \hspace{1cm} \text{(4)}

where $r_{zi}$ 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

<table>
<thead>
<tr>
<th>Methods</th>
<th>RMSE</th>
<th>Classification accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDF – FURIA</td>
<td>0.2312</td>
<td>92.25</td>
<td>0.92335</td>
<td>0.9225</td>
</tr>
<tr>
<td>CFS feature selection - FURIA</td>
<td>0.2184</td>
<td>92.75</td>
<td>0.9275</td>
<td>0.9275</td>
</tr>
<tr>
<td>ABC feature selection - FURIA</td>
<td>0.1906</td>
<td>93.75</td>
<td>0.9376</td>
<td>0.9375</td>
</tr>
<tr>
<td>CFS feature selection – Proposed FURIA</td>
<td>0.1546</td>
<td>94.5</td>
<td>0.945</td>
<td>0.945</td>
</tr>
<tr>
<td>ABC feature selection – Proposed FURIA</td>
<td>0.1428</td>
<td>95.25</td>
<td>0.9525</td>
<td>0.9525</td>
</tr>
</tbody>
</table>
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  
\( \delta \) be the RMSE  
\( C_n \) be the normalized classification accuracy  

The proposed fitness function is given by

\[
\min g(x) = \frac{\delta + \sqrt{\frac{1}{C_n}}}{F}
\]  

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)

\[
\mu_i F(x) = \prod_{i=1}^{k} t_i^x (x_i)
\]  

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)

\[
D_r = \{ x = (x_1, ..., x_n) \in D_r | I_j^x (x_i) > 0 \text{ for all } j \neq 1 \} \subseteq D_r
\]

\[
pr_i = \frac{P_i}{P_i + n_i}
\]

Where

\[
P_i = \sum_{x \in D_r} \mu_i (A) \quad n_i = \sum_{x \in D_r} \mu_i (A)
\]

The fuzzy rules \( r_i^{(j)} \) have learned for the class \( \lambda_i \), the support of this class is defined by equation (10):

\[
S_j(x) = \sum_{i \in \lambda_i} \mu_i^{(j)} (x) CF(r_i^{(j)})
\]

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

\[
CF(r_i^{(j)}) = \frac{2 | D_r \cap \sum_{x \in D_r} \mu_i^{(j)} (x) |}{2 + \sum_{x \in D_r} \mu_i^{(j)} (x)}
\]

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
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

![Classification Accuracy](image1)

**Fig. 3. Classification Accuracy**

![Precision](image2)

**Fig. 4. Precision**

![Recall](image3)

**Fig. 5. Recall**

![RMSE](image4)

**Fig. 6. RMSE**

![Fitness Function](image5)

**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 evaluated 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


23. Srikanth, S V., Ayesha, R., Motoki S., Chandrasekaran R L., and Anand A. Predicting Indian Movie Ratings on IMDB.


