International Journal of Engineering & Technology, 7 (4) (2018) 2565-2568

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.15607
Research paper

Categorizing online news articles using penguin search optimization algorithm

D. Nithya [1] *, Dr. S. Sivakumari [2]

[1] *Assistant Professor, Department of Computer Science and Engineering, School of Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore- 641 108*
[2] *Professor and Head Department of Computer Science and Engineering, School of Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore- 641 108*
**Corresponding author E-mail: nithya.apcse@gmail.com*

Abstract

Online news is an emerging channel where the internet users can get news. Analyzing huge volume of online news articles is a challenging one, because online news articles are generated and updated time to time. Big data techniques are used to tackle this problem. In order to classify the news articles into different categories, an approach based on Evolving Fuzzy Systems(EFS) was used. It categories news articles based on the changes in the content of the corresponding articles. However, it has the problem in the selection of threshold value. Moreover Gaussian membership function is used in EFS that describes the closeness to the prototype. Sometimes it is hard to justify. So in this paper, a Penguins Search Optimization Algorithm(PeSOA) is introduced to optimize the pruning threshold value and a bell shaped fuzzy membership function is introduced to define the closeness to the prototype. The optimized pruning threshold is used in term filtering which prune the generated terms based on their frequencies of occurrence throughout the collection. Then the fuzzy rules are generated by EFS where bell shaped fuzzy membership function is used to define the closeness to the prototype. Based on the fuzzy rules the online news articles are categorized.

Keywords: *Bell Shaped Fuzzy Membership Function; Evolving Fuzzy System; Online News; Penguins Search Optimization Algorithm; Web News Mining*

1. Introduction

In an emerging internet world, huge amount of data are generated especially in digital format. News is the most frequently searched content by the internet users using both mobile phones and computers. It has become one of the major channels for internet users to get news and its utilization ratio has remaining high. Such huge volume of data is analyzed by using text mining. Text mining [1], [2] is used to extract relevant information from unstructured type of data. An important application of the text mining [3] which could be also related with the social big data is web news mining [4], [5]. Since social big data is: web news mining. Since news websites are daily overwhelmed with plenty of news articles, the online newspapers generate an important part of the huge amounts of new information produced each day. So an automatic system [6] is required to analysis, treat and classify the web news articles. It is also used to manage the web news article and also for user recommendation tasks.

An approach based on Evolving Fuzzy System (EFS) [7] was used for web news mining. It classified different web news articles into various topics areas based on the text content of the articles. Initially a set of terms associated with each document are produced. Then the generated terms are pruned based on term frequency and inverse document frequency value (tf-idf). The generated terms are removed from the dataset based on the threshold value. Then, number of fuzzy rules is generated according to the text content of the news article. Based on the fuzzy rule the news articles are categorized. However, the fixed pruning threshold value affects the categorization of news articles and the Gaussian membership function hard to justify the closeness of the prototype. So in this paper, Penguin Search Optimization algorithm is introduced to optimize the threshold and a bell shaped fuzzy membership function is used in EFS to improve the accuracy of web news mining.

2. Literature survey

A clustering based K-means and Back Propagation Neural Network [8] was proposed for web news mining. This clustering method was classified the news according to their categories based on the text and contents. K- Means clustering was applied on the BBC news articles which cluster the news according to the categories like politics, sports, normal and movies based on the user defined k values and similarity metric is Euclidean distance. Then, Back Propagation Neural Network was used to classify the BBC news. It was also used to check the performance because sometimes user uploads correct news but the system rejected as it as false data and vice versa. This method saves the time of users. However the user defined k values influence the performance of web news mining.

A Deep learning technique [9] was introduced for online news classification. The classification of news started with a dataset in which the class assignments are known. It predicted the target class for each case in the data. Initially, a training database was created using online news. Then, neural network arc was selected and initialize the weights of neural network. Set a goal for neural network if the goal was met then reselect the training sides. Otherwise change the weight of the neural network. Based on the weights, the online news is classified.

An intelligent system [10] was proposed for classification of inner structures of the online news based on Support Vector Machine (SVM) and Neural Network (NN). This system was started with uploading files and then the features were extracted using various parameters and classifier. Neural Network was applied to train the dataset. Then, testing was done on whole dataset by using a NN and SVM. Finally the data were classified based on the inner clusters of each and every field is financial, entertainment and sports. Still this system needs an improvement in terms of classification accuracy.

A news article classification framework [11] was proposed for news articles classification. In this framework, the classification was performed using N grams where the textual features were extracted from text and visual features were generated from one representative image. Then these features were given as input to the Random Forest machine learning method which is widely used classification and regression method. It consists of N trees that predict the class label based on the majority voting function. By using the multimodal features i.e., text and visual features in Random Forest classification improved the accuracy of web news article classification. Considering more number of features will improve the classification accuracy.

A method [12] presented was presented for text classifier which automatically categorize the content of web feeds. The text categorization was splitted into seven sub process are read document, Tokenize text, Stemming, Stop words removal, Vector representation of text, Feature Selection and Learning Algorithm. The string representation of content was get into read document phase and then removes words which have little semantic meaning by using stop words removal process. The stem words in the content were obtained in the stemming process. Then the stem words were converted in the text to be categorized into Support Vector Machine (SVM) matrix representation of words which is done in the vector representation phase. Finally the SVM algorithm was applied to categorize the content. But the SVM has the limitation of speed and size in both testing and training.

3. Materials and methods

In this section, the proposed method for optimization of pruning threshold and bell shaped fuzzy membership function is described in detail. The optimized pruning threshold is used in proposed EFS-PeSOA which is the main difference between proposed EFS-PeSOA and existing EFS. With the consideration of web news classification accuracy the PeSOA optimizes the pruning threshold. The pruning threshold is optimized by using Penguins Search Optimization algorithm [13] which based on the hunting behavior of penguins. The penguins collaborates their efforts and synchronize their dives to optimize the pruning threshold in the process of hunting and nutrition. In addition to this, a bell shaped fuzzy membership function is used instead of Gaussian membership function to improve the EFS system. The proposed web news mining method consists of two modules are term extraction and evolving classification.

3.1. Term extraction

In the term extraction module, different topic areas from the web is extracted and summarized each article with a set of terms in which the each term has its corresponding relevance value. Initially the most relevant terms of each article is obtained by using open source tool called RapidMiner. It produces a set of terms associated with each document. In order to consider the relevance of different terms in the framework of all news, prune the generated terms based on their frequencies of occurrences throughout the collection. The relevance of different terms in all collected news article are calculated by using term frequency and inverse document frequency (tf-idf). Then sum the tf-idf of a particular term it is lower than a pruning threshold then it is removed from the dataset. The threshold value is optimized by using penguin search optimization algorithm.

Penguin search optimization algorithm is a rearranged model of social relations, which depends on chasing of penguins. The penguin search optimization algorithm is started by initializing population of penguin which is made of several groups. To each group, a variable number of penguins are allocated based on food availability. Here, the food is termed as pruning threshold value. Each group of penguins starts searching in a specific position (hole) and random levels. Untill the penguins find optimal pruning threshold value different penguins in each group hunt moves randomly. Each penguin looks for pruning threshold in a random way and individually in its group. After a rough number of dives, the location of food and plenty of food is shared by the penguins. At one point of view, one can have 0 to N penguins based on the abundance of food. Then check the number of fish (value of classification accuracy) in hole. If it is not enough for the group, part of the group migrates to another hole. It ensures inter group communication. The group who ate the most fish (high classification accuracy) delivers us the location of rich food represented by the hole and the level.

Penguin Search Optimization Algorithm

Step 1: Generate random population of P solutions (penguins) in groups

Step 2: Initialize the threshold values in the holes and the levels

Step 3: For i = [1] to number of generations

Step 4: For each individual i ∈ P do

Step 5: while oxygen reserves are not depleted do

Step 6: Take a random step

Step 7: Improve the penguin position

$$D_{new} = D_{last} + \text{rand}()|X_{best} - X_{id}| \quad (1)$$

// rand () is a random number of the distribution, D_{last} denotes the current solution, X_{best} denotes the best local solution, X_{id} denotes the final solution and D_{new} denotes the new solution.

Step 8: Update pruning threshold for this penguin

Step 9: End

Step 10: End

Step 11: Update pruning threshold in the holes, levels and the best group

Step 12: Redistributes the probabilities of penguins in the holes and the levels.

Step 13: Update best solution

Step 14: End

3.2. Evolving classification

The evolving classification consists of two modules are creation of the evolving fuzzy rules and web news classification. In the creation of the evolving fuzzy rules, fuzzy rules are created by using eClass0 classifier. The tf- idf values are given as input to the creation of the evolving fuzzy rules. Then, the creation of the evolving fuzzy rules potential of the k-th article is calculated which represents a function of the accumulated function between a news article and all other k-1 articles. Update all the prototypes (a data sample that groups several samples which represent a certain class) considering the potential of the k-th news article and insert it as a prototype. To remove existing prototypes, the bell shaped membership function [14] is calculated between a data sample and a prototype. It represents the closeness to the prototype. Based on the fuzzy rules the web news articles are categorized.

4. Results and discussion

In this section, the results of the existing and proposed web news mining methods are analysed in terms of accuracy, precision and recall. For the experimental purpose, five different sets of data are created which combines two or more categories. The five datasets are Health vs. Science (H-Sc), Science vs. Technology (Sc-Te),

Health vs. Science vs. Sports (H-Sc-Sp), Business vs. Health vs. Science vs. Sports (B-He-Sc-Sp), and Arts vs. Business vs. Health vs. Science vs. Sports vs. Travels (A-B-H-Sc-Sp-Tr).

4.1. Accuracy

The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined. Accuracy is calculated as follows:

$$\text{Accuracy} = \frac{\text{True positive + True negative}}{\text{True positive + True negative+ False positive+ False negative}} \quad (2)$$

The following Table1 shows the comparison of accuracy between existing Evolving Fuzzy System (EFS) and proposed EFS with Penguin Search Optimization Algorithm (EFS-PeSOA) method.

Table 1: Comparison of Accuracy

Methods	Dataset				
	H-Sc	Sc-Te	H-Sc-Sp	B-He-Sc-Sp	A-B-H-Sc-Sp-Tr
EFS	72.2	78.3	64.4	52.3	35.7
EFS-PeSOA	80	87.27	70.45	60.7	42.45

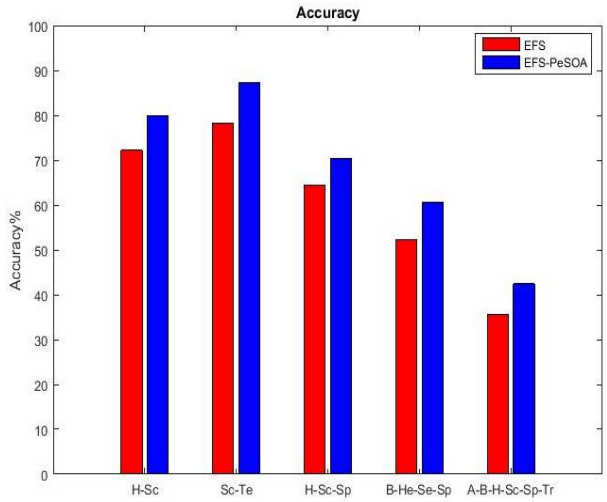

Fig. 1: Comparison of Accuracy.

Fig.1 shows the comparison of accuracy between Evolving Fuzzy System (EFS) and proposed EFS with Penguin Search Optimization Algorithm (EFS-PeSOA). X axis represents the different datasets and Y axis denotes the accuracy in terms of %. For H-Sc set of data, the accuracy of proposed is EFS-PeSOA 10.8% greater than existing EFS. From the Fig.1 it is proved that the proposed EFS-PeSOA has high accuracy than the existing EFS.

4.2. Precision

Precision value is evaluated according to the relevant information at true positive prediction, false positive.

$$Precision = \frac{Truepositive}{(Truepositive+Falsepositive)} \quad (3)$$

The following Table2 shows the comparison of precision between existing Evolving Fuzzy System (EFS) and proposed EFS with Penguin Search Optimization Algorithm (EFS-PeSOA) method.

Table 2: Comparison of Precision

Methods	Dataset				
	H-Sc	Sc-Te	H-Sc-Sp	B-He-Sc-Sp	A-B-H-Sc-Sp-Tr
EFS	72.5	77.5	65.1	53.12	34.7
EFS-PeSOA	81	86.97	69.96	61.13	43.29

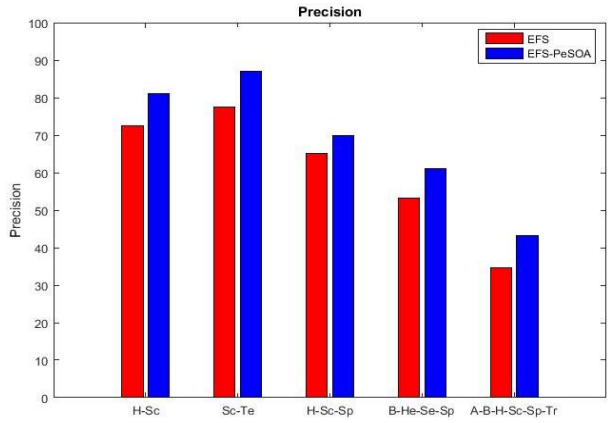

Fig. 2: Comparison of Precision.

Fig.2 shows the comparison of precision between Evolving Fuzzy System (EFS) and proposed EFS with Penguin Search Optimization Algorithm (EFS-PeSOA). For H-Sc set of data, the precision of proposed is EFS-PeSOA 11.7% greater than existing EFS. X axis represents the different datasets and Y axis denotes the precision. From the Fig.2 it is proved that the proposed EFS-PeSOA has high precision than the existing EFS.

4.3. Recall

The recall value is evaluated according to the classification of web news articles at true positive prediction and false negative prediction.

$$Recall = \frac{Truepositive}{(Truepositive+Falsenegative)} \quad (4)$$

The following Table3 shows the comparison of recall between existing Evolving Fuzzy System (EFS) and proposed EFS with Penguin Search Optimization Algorithm (EFS-PeSOA) method.

Table 3: Comparison of Recall

Methods	Dataset				
	H-Sc	Sc-Te	H-Sc-Sp	B-He-Sc-Sp	A-B-H-Sc-Sp-Tr
EFS	72.2	78.3	64.4	52.3	35.7
EFS-PeSOA	80	87.27	70.45	60.7	42.45

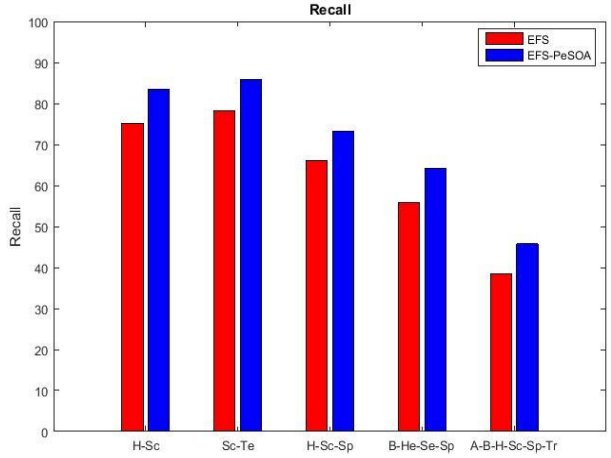

Fig. 3: Comparison of Recall.

Fig.3 shows the comparison of recall between Evolving Fuzzy System (EFS) and proposed EFS with Penguin Search Optimization Algorithm (EFS-PeSOA). X axis represents the different datasets and Y axis denotes the recall. For H-Sc set of data, the recall of proposed is EFS-PeSOA 10.8% greater than existing EFS.

From the Fig.3, it is proved that the proposed EFS-PeSOA has high recall than the existing EFS.

4.4. F-measure

F-measure computes the mutual value of precision and recall as the harmonic mean of precision and recall. It is computed as follows,

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The following Table2 shows the comparison of precision between existing Evolving Fuzzy System (EFS) and proposed EFS with Penguin Search Optimization Algorithm (EFS-PeSOA) method.

Table 4: Comparison of F-Measure

Methods	Dataset				
	H-Sc	Sc-Te	H-Sc-Sp	B-He-Sc-Sp	A-B-H-Sc-Sp-Tr
EFS	72.3	77.9	64.7	52.7	35.2
EFS-PeSOA	80.4	87.1	70.2	60.9	43

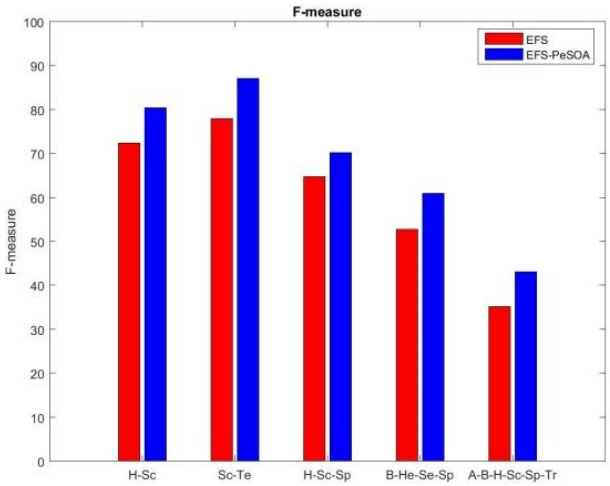

Fig. 4: Comparison of F-Measure.

Fig.4 shows the comparison of F-measure between Evolving Fuzzy System (EFS) and proposed EFS with Penguin Search Optimization Algorithm (EFS-PeSOA). X axis represents the different datasets and Y axis denotes the F-measure. For H-Sc set of data, the F-measure of proposed is EFS-PeSOA 11.2% greater than existing EFS. From the Fig.4 it is proved that the proposed EFS-PeSOA has high F-measure than the existing EFS.

5. Conclusion

In this paper, penguin search optimization algorithm is introduced to optimize the pruning threshold value which is used in term extraction process of web news mining. The extracted terms are used to create fuzzy rules using eclass0 classifier which is done in the evolving classification process. The bell shaped membership function is used instead of Gaussian bell function to represent the closeness of the prototype. The evolving classification returns a set of fuzzy rules which is used to categorize the web news. The experimental results show that the proposed web news mining method has better accuracy, precision and recall than the existing web news mining method.

References

[1] Karam R, Puri R, Bhunia S (2016), Energy-efficient adaptive hardware accelerator for text mining application kernels. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 24(12), pp. 3526-3537. https://doi.org/10.1109/TVLSI.2016.2555984.

[2] Li Y, Algarni A, Albathan M, Shen Y, Bijaksana MA (2015), Relevance feature discovery for text mining. *IEEE Transactions on Knowledge and Data Engineering*, 27(6), pp. 1656-1669. https://doi.org/10.1109/TKDE.2014.2373357.

[3] Dang S, & Ahmad PH (2014), Text mining: techniques and its application. *International Journal of Engineering & Technology Innovations*, 1(4), 22-25.

[4] Nithya, D, Sivakumari, S (2017), State of the Art of Web News Mining. *International Journal of Computer Engineering and Applications*, 10(8), 122-129.

[5] Nithya, D, Sivakumari, S (2017), A Study on Web Mining Tools. *International Journal of Research in Electronics and Computer Engineering*, 5(2), 135-137.

[6] Wanjari YW, Mohod VD, Gaikwad DB, & Deshmukh SN (2014), Automatic news extraction system for Indian online newspapers. In *third International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)* IEEE, pp. 1-6.

[7] Iglesias JA, Tiemblo A, Ledezma A, Sanchis A (2016), Web news mining in an evolving framework. *Information Fusion*, 28, 2016; 90-98. https://doi.org/10.1016/j.inffus.2015.07.004.

[8] Kaur S, Rashid EM (2016), Web news mining using back propagation neural network and clustering using K-Means algorithm in Big data. *Indian Journal of Science and Technology*, 9(41), pp. 1-8. https://doi.org/10.17485/ijst/2016/v9i41/95598.

[9] Kaur S, Khiva NK (2016), online news classification using Deep Learning Technique. *International Research Journal of Engineering and Technology (IRJET)*, 3(10), pp. 558-563.

[10] Sharma N, Kaur P (2015), Categorize Online news using Various Classification Techniques. *International Journal of Advanced Research in Computer Science & Technology (IJARCET)*, 4 (2), pp. 337-340.

[11] Liparas D, HaCohen-Kerner Y, Moumtzidou A,Vrochidis S, Kompatsiaris I (2014), News articles classification using Random Forests and weighted multimodal features. In *Information Retrieval Facility Conference Springer, Cham*, pp. 63-75.

[12] Longe HOD (2014), A Text Classifier Model for Categorizing Feed Contents Consumed by a Web Aggregator. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 5(9), pp. 95-100. https://doi.org/10.14569/IJACSA.2014.050915.

[13] Gheraibia Y, Moussaoui A (2013), Penguins search optimization algorithm (PeSOA). In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems Springer, Berlin, Heidelberg*, 2013; 222-231.

[14] Huang W, Li Y (2012), Bell-Shaped Probabilistic Fuzzy Set for Uncertainties Modeling. *Journal of Theoretical & Applied Information Technology*, 46(2), pp. 875-882.