An Enhanced Cos-Neuro Bio-Inspired Approach for Document Clustering

Vaishali Madaan, Kundan Munjal, Sahil Verma, N. Z. Jhanjhi, and Aman Singh

Abstract Data mining is a dynamic and attractive research domain that has become known to discover information from the vast amount of constantly created data. Clustering is an unsupervised approach to data mining in which a group of similar items is assembled in one cluster. The quality of documents retrieved within a lesser amount of time has always been a fundamental problem in web document clustering. The authors introduce similarity technique-based K-means clustering using bee swarm optimization and artificial neural networks in this work. The artificial neural network helps classify the best centroid location based on the similarity index of the document and according to the trained structure of ANN to organize the best cluster number to test queries. The quality of papers returned is improved significantly with lesser execution time and improved efficiency through the projected method.

Keywords Cosine similarity · K-means clustering · Bee swarm optimization · Artificial neural network

V. Madaan
Maharishi Markandeshwar University, Mullana, India

K. Munjal
Apex Institute of Technology, Chandigarh University, Mohali, India

K. Munjal
University College of Engineering, Punjabi University, Patiala, India

S. Verma (✉)
Department of Computer Science and Engineering, Chandigarh University, Mohali, India
e-mail: sahilverma@ieee.org

N. Z. Jhanjhi
Taylor's University, Subang Jaya, Malaysia
e-mail: NoorZaman.Jhanjhi@taylors.edu.my

A. Singh
School of Computer Science and Engineering, Lovely Professional University,
Phagwara, India

S.-L. Peng et al. (eds.), *Intelligent Computing and Innovation on Data Science*,
Lecture Notes in Networks and Systems 248,
https://doi.org/10.1007/978-981-16-3153-5_54

6 Conclusion and Future Work

There are many document classification applications in the industrial world already presented by researchers. Still, they faced several problems like best centroid localization, less uniqueness of clustered data; clustering time is more, etc. The proposed research work enhances the clustering mechanism and validates the clustering using artificial neural network (ANN). To solve the problem mentioned above, cosine similarity and BSO and ANN are used to determine the text documents' resemblance. The neural network is a computation algorithm used to classify the problem iteratively. From the experiment, it has been concluded that the average value of precision, recall, average classification error, and *F*-measure for 1000 number of the document are 0.82, 0.79, 3.2%, and 0.814, respectively. The value of the proposed *F*-measure is high compared to the existing *F*-measure value, where the *F*-measure value of the proposed work is increased by 6.4% from the current work. This indicates that the classification accuracy of the ANN structure is high. The average run time taken for 100 iterations was 0.244 s which remained constant. Finally, the comparison is made where our approach outperforms the existing system with an accuracy of 94.77%. Proposed work is proved effective in improving the quality of returned documents as the *F*-measure parameter value is increased efficiency. The future investigation includes evaluating the whole process using other evolutionary approaches, comparing the work done with different state-of-art methods. Work can be enhanced to reduce the run time by trying to use other classification approaches.

References

1. Vijayalakshmi B et al (2020) An attention based deep learning model for traffic flow prediction using spatio temporal features towards sustainable smart city. IJCS, Wiley, Hoboken, pp 1–14
2. Batra I et al (2020) Hybrid logical security framework for privacy preservation in the green internet of things. MDPI-Sustainability 12(14):1–15
3. Batra I et al (2019) A lightweight IoT based security framework for inventory automation using wireless sensor network. IJCS, Wiley, Hoboken, pp 1–16
4. Hearst MA (1999, June) Untangling text data mining. In: Proceedings of the 37th annual meeting of the association for computational linguistics, pp 3–10
5. MacQueen J (1967, June) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol 1, no 14, pp 281–297
6. Nazeer KA, Sebastian MP (2009, July) Improving the accuracy and efficiency of the k-means clustering algorithm. In: Proceedings of the world congress on engineering, vol 1. Association of Engineers, London, pp 1–3
7. Kapil S, Chawla M (2016, July) Performance evaluation of K-means clustering algorithm with various distance metrics. In: 2016 IEEE 1st international conference on power electronics, intelligent control and energy systems (ICPEICES), pp 1–4, IEEE

8. Shafeeq A, Hareesha KS (2012) Dynamic clustering of data with modified k-means algorithm. In: Proceedings of the 2012 conference on information and computer networks, pp 221–225
9. Tan S (2005) Neighbor-weighted k-nearest neighbor for unbalanced text corpus. Expert Syst Appl 28(4):667–671
10. Zheng Z, Wu X, Srihari R (2004) Feature selection for text categorization on imbalanced data. ACM SIGKDD Explor Newsl 6(1):80–89
11. Del Castillo MD, Serrano JI (2004) A multistrategy approach for digital text categorization from imbalanced documents. ACM SIGKDD Explor Newsl 6(1):70–79
12. Cui X, Potok TE, Palathingal P (2005, June) Document clustering using particle swarm optimization. In: Proceedings 2005 IEEE swarm intelligence symposium, 2005. SIS 2005, pp 185–191, IEEE
13. Bharathi A, Deepan kumar E (2014) Survey on classification techniques in data mining. Int J Recent Innov Trends Comput Commun 2(7):1983–1986
14. Djenouri Y, Belhadi A, Belkebir R (2018) Bees swarm optimization guided by data mining techniques for document information retrieval. Expert Syst Appl 94:126–136
15. Karaboga D, Ozturk C (2011) A novel clustering approach: Artificial Bee Colony (ABC) algorithm. Appl Soft Comput 11(1):652–657
16. Lenc L, Král P (2016, April) Deep neural networks for Czech multi-label document classification. In: International conference on intelligent text processing and computational linguistics. Springer, Cham, pp 460–471
17. Zheng J, Guo Y, Feng C, Chen H (2018) A hierarchical neural-network-based document representation approach for text classification. Math Probl Eng
18. Datta D et al (2020) UAV environment in FANET: an overview. Applications of cloud computing: approaches and practices. CRC Press, Taylor & Francis Group, Boca Raton, pp 1–16