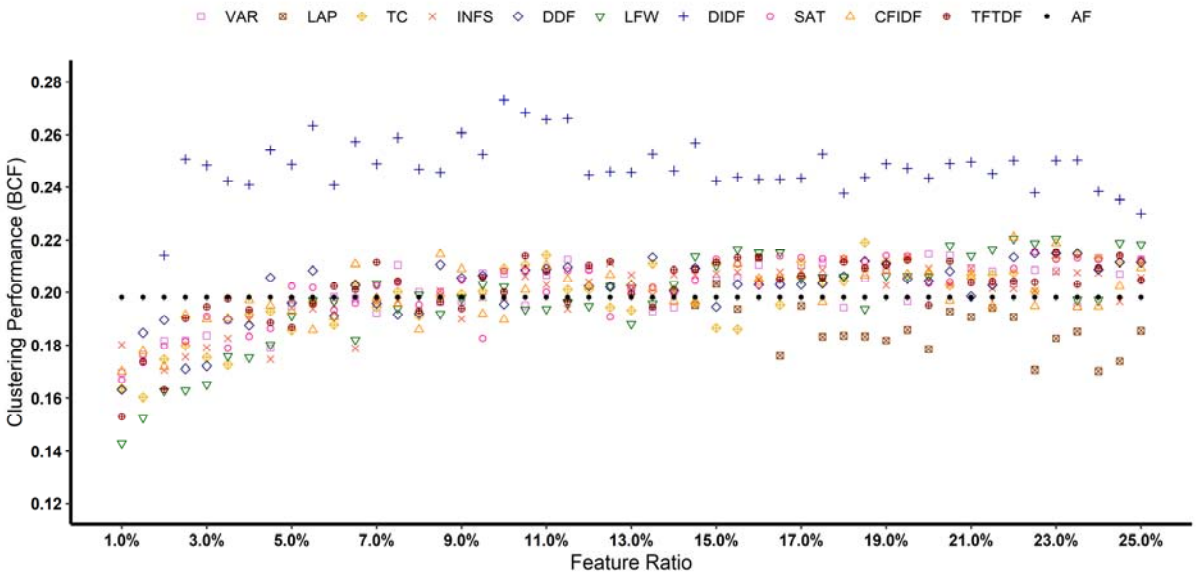

Fig. 2. Comparison of feature selection methods based on Clustering Performance measured using BCF on the dataset D3.

In both the test cases, starting from a feature ratio of 2 %, the proposed method DIDF was able to identify most discriminative features that resulted in a clustering performance better than those of other feature selection methods and base line performance. With the increase in the number of features, proportion of irrelevant features increase, which deteriorates clustering performance.

6. Conclusion

An unsupervised univariate filter feature selection method DIDF was proposed to reduce the ill effects of curse of dimensionality on text clustering performance. DIDF was compared with nine related feature selection methods reported in the literature. Eight datasets, with varied characteristics, were considered for testing the effectiveness of feature selection methods. DIDF was proved to be the most promising method in identifying relevant features for better clustering performance, for sparse datasets with varying degrees of skewness and dispersion in class distribution.

References

[1] Abualigah, L. M. Q. (2019). *Feature selection and enhanced krill herd algorithm for text document clustering,* Springer, Berlin.
[2] Amigó, E.; Gonzalo, J.; Artiles, J.; Verdejo, F. (2009): A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information retrieval, **12**(4), pp. 461–486.
[3] Agarwal, N.; Sikka, G.; Awasthi, L. K. (2020): Enhancing web service clustering using Length Feature Weight Method for service description document vector space representation. Expert Systems with Applications, **161**, pp. 113682.
[4] Bagga, A.; Baldwin, B. (1998): Entity-Based Cross-Document Co referencing Using the Vector Space Model. In 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, **1**, pp. 79–85.
[5] David, W. (2018): Basic Guidelines for Common Business Statistics Metrics. Business Education Innovation, **32**(2), pp. 21–26.
[6] Garcia-Dias, R.; Vieira, S.; Pinaya, W. H. L.; Mechelli, A. (2020). Clustering analysis, *In Machine Learning,* Academic Press, London.
[7] Guyon, I.; Gunn, S.; Nikravesh, M.; Zadeh, L. A. (2008). *Feature extraction: foundations and applications,* Springer, Berlin.
[8] He, X., Cai, D.; Niyogi, P. (2005): Laplacian score for feature selection. Advances in neural information processing systems, **18**, pp. 507–514.
[9] Li, J., Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R. P.; Tang, J.; Liu, H. (2017): Feature selection: A data perspective. ACM Computing Surveys, **50**(6), pp. 1–45.
[10] Liu, L.; Kang, J.; Yu, J.; Wang, Z. (2005): A comparative study on unsupervised feature selection methods for text clustering. IEEE International Conference on Natural Language Processing and Knowledge Engineering, pp. 597–601.

[11] Liu, Z.; Lin, Y.; Sun. M. (2020). Document Representation, *In Representation Learning for Natural Language Processing,* Springer, Singapore.
[12] Liu, T.; Liu, S.; Chen, Z.; Ma, W. Y. (2003): An evaluation on feature selection for text clustering. In Proceedings of the 20th international conference on machine learning (ICML-03), pp. 488–495.
[13] Nalluri, S. P.; Kurra, R. R. (2014): Subspace clustering of text documents using collection and document frequencies of terms. International Review on Computers and Software, **9**(10), pp. 1692–1699.
[14] Nalluri, S. P.; Kurra, R. R. (2015a): Effective Clustering of Text Documents in Low Dimension Space using Semantic Association among Terms. International Review on Computers and Software, **10**(5), pp. 467–474.
[15] Nalluri, S. P.; Kurra, R. R. (2015b): Feature Selection based on Term Frequency and Term Document Frequency for Text Clustering. International Journal of Applied Engineering Research, **10**(10), pp. 26175–26190.
[16] Roffo, G.; Castellani, U.; Vinciarelli, A.; Cristani, M. (2020): Infinite feature selection: a graph-based feature filtering approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, [preprint].
[17] Ray, P.; Reddy, S. S.; Banerjee, T. (2021): Various dimension reduction techniques for high dimensional data analysis: a review. Artificial Intelligence Review, pp. 1–43.
[18] Solorio-Fernández, S.; Carrasco-Ochoa, J. A.; Martínez-Trinidad, J. F. (2020a): A review of unsupervised feature selection methods. Artificial Intelligence Review, **53**(2), pp. 907–948.
[19] Solorio-Fernández, S.; Carrasco-Ochoa, J. A.; Martínez-Trinidad, J. F. (2020b): A systematic evaluation of filter Unsupervised Feature Selection methods. Expert Systems with Applications, **162**, pp. 113745.
[20] Warf, B. (Ed.). (2021). *Geographies of the Internet,* Routledge, New York.
[21] Wu, J.; Xiong, H.; Chen, J. (2009): Adapting the right measures for k-means clustering. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 877–886.
[22] Wang, J.; Zhang, H.; Wang, J.; Pu, Y.; Pal, N. R. (2021): Feature selection using a neural network with group lasso regularization and controlled redundancy. IEEE Transactions on Neural Networks and Learning Systems, **32**(3), pp. 1110–1123.
[23] Zhai, C.; Massung, S. (2016). *Text data management and analysis: A practical introduction to information retrieval and text mining,* ACM; Morgan and Claypool, California.
[24] Zhou, K.; Yang, S. (2020): Effect of cluster size distribution on clustering: a comparative study of k-means and fuzzy c-means clustering. Pattern Analysis and Applications, **23**(1), pp. 455–466.

Authors Profile

Sivaram Prasad Nalluri, is a research scholar in the department of Computer Science and Engineering of Acharya Nagarjuna University, Andhra Pradesh, India. He is also working as a faculty member in the Information Technology department of Bapatla Engineering College, Bapatla, India. He obtained Post Graduate Degree in Computer Science and Engineering from Jawaharlal Nehru Technological University, Hyderabad (India) in 2002. His research interests include Machine Learning, Deep Learning and Cyber Security. He is affiliated with CSI and ISTE as life member.

Rajasekhara Rao Kurra is working as a professor in Computer Science and Engineering department of Usha Rama College of Engineering & Technology, Telaprolu, India. He obtained Ph.D in Computer Science & Engineering from Acharya Nagarjuna University, Guntur (India). His research interests include Machine Learning, Embedded Systems, Software Engineering and Knowledge Management. He is a Fellow of IETE. He is affiliated with IE, ISTE, ISCA, CSI as life member and with ACM as member. He is the Chairman of CSI Vijayawada Chapter, India.