Understanding of Data Preprocessing for Dimensionality Reduction Using Feature Selection Techniques in Text Classification

Varun Dogra, Aman Singh, Sahil Verma, Kavita, N. Z. Jhanjhi, and M. N. Talib

Abstract The volume of textual data in digital form is growing with each day. For arranging these textual data, text classification has been used. To achieve efficient text classification, data preprocessing is an important phase. It prepares information for machine learning models. Text classification, however, has the issue of the high dimensionality of space for features. Feature selection is a technique for data preprocessing widely used on high-dimensional data. By feature selection techniques, this high dimensionality of feature space is solved and increases text classification efficiency. Feature selection explores how a list of features used to create text classification models may be chosen. Its goals include reducing dimensionality, deleting uninformative features, reducing the amount of data available to classifiers for learning, and enhancing classifiers' predictive performance. The different methods of feature selection are presented in this paper. This paper also presents the advantages and limitations of feature selection methods.

Keywords Text classification · Data preprocessing · Feature selection · Dimensionality reduction

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

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

Kavita
e-mail: kavita@ieee.org

N. Z. Jhanjhi
School of Computer Science and Engineering, Taylor's University, Subang Jaya, Malaysia
e-mail: NoorZaman.Jhanjhi@taylors.edu.my

M. N. Talib
Papua New Guinea University of Technology, Lae, Papua New Guinea
e-mail: muhammad.talib@pnguot.ac.pg

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_48

References

1. Magnini B, Pezzulo G, Gliozzo A (2002) The role of domain information in word sense disambiguation. Natural Language Engineering
2. Ramisetty S, Verma S (2019) The amalgamative sharp wireless sensor networks routing and with enhanced machine learning. J Comput Theor Nanosci 16(9):3766–3769
3. Zhang ML, Zhou ZH (2007) ML-KNN: a lazy learning approach to multi-label learning. Pattern Recogn 40(7):2038–2048
4. Batra I, Verma S, Malik A, Ghosh U, Rodrigues JJ, Nguyen GN, Mariappan V (2020) Hybrid logical security framework for privacy preservation in the green internet of things. Sustainability 12(14):5542
5. Armanfard N, Reilly JP, Komeili M (2015) Local feature selection for data classification. IEEE Trans Pattern Anal Mach Intell 38(6):1217–1227
6. Pölsterl S, Conjeti S, Navab N, Katouzian A (2016) Survival analysis for high-dimensional, heterogeneous medical data: exploring feature extraction as an alternative to feature selection. Artif Intell Med 72:1–11
7. Lei S (2012) A feature selection method based on information gain and genetic algorithm. In: 2012 international conference on computer science and electronics engineering, vol 2. IEEE, pp 355–358
8. Jin X, Xu A, Bie R, Guo P (2006) Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. In: International workshop on data mining for biomedical applications. Springer, Berlin, Heidelberg, pp 106–115
9. Youn E, Koenig L, Jeong MK, Baek SH (2010) Support vector-based feature selection using Fisher's linear discriminant and support vector machine. Expert Syst Appl 37(9):6148–6156
10. Wang Y, Wang XJ (2005) A new approach to feature selection in text classification. In: 2005 international conference on machine learning and cybernetics, vol 6. IEEE, pp 3814–3819
11. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238
12. Labani M, Moradi P, Ahmadizar F, Jalili M (2018) A novel multivariate filter method for feature selection in text classification problems. Eng Appl Artif Intell 70:25–37
13. Granitto PM, Furlanello C, Biasioli F, Gasperi F (2006) Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemom Intell Lab Syst 83(2):83–90
14. Rani P, Verma S, Nguyen GN (2020) Mitigation of black hole and gray hole attack using swarm inspired algorithm with artificial neural network. IEEE Access 8:121755–121764
15. Chen H, Jiang W, Li C, Li R (2013) A heuristic feature selection approach for text categorization by using chaos optimization and genetic algorithm. Mathematical Problems in Engineering
16. Guyon I, Bitter HM, Ahmed Z, Brown M, Heller J (2003) Multivariate non-linear feature selection with kernel multiplicative updates and gram-schmidt relief. In: BISC flint-CIBI 2003 workshop, Berkeley, pp 1–11
17. Urbanowicz RJ, Olson RS, Schmitt P, Meeker M, Moore JH (2018) Benchmarking relief-based feature selection methods for bioinformatics data mining. J Biomed Inform 85:168–188
18. Vickers NJ (2017) Animal communication: when i'm calling you, will you answer too? Curr Biol 27(14):R713–R715
19. Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12(1):55–67
20. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J Royal Stat Soc Ser B (Stat Methodol) 67(2):301–320