engineering proceedings

MDPI

Proceeding Paper

Survey on Preprocessing Techniques for Big Data Projects †

Ignacio D. Lopez-Miguel

Centro de Postgrado, Universidad Internacional Menéndez Pelayo, C/ Isaac peral, 23, 28040 Madrid, Spain; lopezmiguelignacio@posgrado.uimp.es

† Presented at the 4th XoveTIC Conference, A Coruña, Spain, 7–8 October 2021.

Abstract: In the era of big data, a vast amount of data are being produced. This results in two main issues when trying to discover knowledge from these data. There is a lot of information that is not relevant to the problem we want to solve, and there are many imperfections and errors in the data. Therefore, preprocessing these data is a key step before applying any kind of learning algorithm. Reducing the number of features to a relevant subset (feature selection) and reducing the possible values of continuous variables (discretisation) are two of the main preprocessing techniques. This paper will review different methods for completing these two steps, focusing on the big data context and giving examples of projects where they have been applied.

Keywords: preprocessing; big data; feature selection; discretisation

check for updates

Citation: Lopez-Miguel, I.D. Survey on Preprocessing Techniques for Big Data Projects. *Eng. Proc.* **2021**, *7*, 14. https://doi.org/10.3390/engproc2021007014

Academic Editors: Joaquim de Moura, Marco A. González, Javier Pereira and Manuel G. Penedo

Published: 7 October 2021

1. Introduction

With the irruption of the "big data" phenomenon, massive amounts of data are generated daily. These data are normally available in a raw format and need to be treated before acquiring any knowledge from them. This step in the big data chain is usually referred to as preprocessing and there exists a wide range of techniques [1].

The main approaches to preprocess big data are discretisation and feature selection. The former transforms continuous data to a limited set of values. Feature selection aims to reduce the number of attributes [1,2].

The remainder of this paper is organised as follows: Section 2 introduces the different preprocessing techniques, dividing them into feature selection and discretisation. For each of these techniques, a classification with different examples for each category is presented. Section 3 concludes the paper and suggests a future line of research.

2. Data Preprocessing

Different feature selection and discretisation techniques are presented in this section based on big data projects where they have been applied.

2.1. Feature Selection

The different feature selection techniques for big data mining can be classified into filter methods, wrapper methods, and embedded methods [1].

2.1.1. Filter Methods

Features are selected according to the value of different metrics, usually certain statistical criteria.

In the context of text mining, it is common to use the bag-of-words approach so that each word is taken as a unique feature. Chi-squared was used to filter the most relevant terms in a text mining algorithm to estimate credit score at Deutsche Bank [3].

In relation to text mining as well, in [4] tweets are analysed in order to figure out the impact of their sentiment on stock market movements. The authors also use a filter method to select the most relevant features—Fisher score.

Based on Chi-squared and the GUIDE regression tree, Loh [5,6] presents a technique to perform feature selection in a large genomic dataset.

Other filter methods include Information Gain [7], correlation [8], variance similarity [9], and Dispersion ratio [10].

Some work has been done to adapt these methods to the big data context, such as in [11], where a framework to parallelise and scale some of these algorithms is introduced.

2.1.2. Embedded Methods

Feature selection is performed in the process of fitting a model to a given dataset.

SVM-RFE (Supported Vector Machine Recursive Feature Elimination), introduced in [12] to analyse DNA microarrays has shown its power in several applications, such as in bioinformatics [13].

The Feature Selection-Perceptron (FS-P) [14] technique has been used in a proton (^{1}H) magnetic resonance spectroscopy (MRS) database to select features that could better predict brain tumours.

Based on a more complex neural network, the embedded method BlogReg is introduced in [15], where it is applied to data collected from the sensors of a robot.

2.1.3. Wrapper Methods

Wrapper methods refer to an iterative process in which a subset of features is evaluated at a time.

A wrapper method based on the decision tree C4.5 has been used for many years [16]. However, developments based on this method are still ongoing, such as the one from [17], which is applied to healthcare data (Medical Internet of Things).

Another wrapper method is based on the SVM algorithm [18]. It has been widely used since its creation, such as in [19], predicting arrhythmias from cardiac data.

FSSEM (Feature Subset Selection wrapped around EM clustering) [20] is also a wrapped method, and a popular stepwise approach for regression problems [21].

2.2. Discretisation

Discretisation is the step where continuous variables are transformed into categorical ones [2]. There exist multiple classifications for discretization techniques, but here one of unsupervised and supervised discretisation is chosen [2].

2.2.1. Unsupervised

Unsupervised discretisation methods do not take into account the target of the learning algorithm when the features are discretised.

Equal width interval discretisation and equal frequency interval discretisation need to be adapted in the context of big data streaming as done in [22].

In [23], k-means [24] discretisation is used to transform the target for road detection.

Other methods based on k-means algorithm have been proposed, such as Cokmeans and Bikmeans, used in [25] in the context of microarrays.

2.2.2. Supervised

Supervised discretisation does take into account the target of the learning algorithm.

One of the most popular methods is based on entropy [26]. This algorithm was parallelised in [27].

Chi-squared is the basis for ChiMerge [28], ChiSplit [29], and Khiops [30]. They were parallelised in [31] to work for big data problems.

The previously presented approaches are univariate, but there also exist supervised multivariate discretisation (SMD) techniques, such as the one in [32].

3. Conclusions

Due to extension limitations, this paper has only given some feature selection and discretisation techniques, mentioning some up-to-date examples of where they are used. There is growing interest in adapting these techniques so that they can perform efficiently in the big data context. In this direction, a future line of work is to create a comprehensive and complete taxonomy of the up-to-date feature selection and discretisation techniques, performing experimental results in the big data context.

Funding: This research received no external funding.

References

1. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. Recent advances and emerging challenges of feature selection in the context of big data. *Knowl.-Based Syst.* **2015**, *86*, 33–45. [CrossRef]
2. Dash, R.; Paramguru, R.; Dash, R. Comparative analysis of supervised and unsupervised discretization techniques. *Int. J. Adv. Sci. Technol.* **2011**, *2*, 29–37.
3. Hristova, D.; Probst, J.; Eckrich, E. Ratingbot: A text mining based rating approach. *ICIS* **2017**, *8*, 1–20.
4. Abbes, H. Tweets Sentiment and Their Impact on Stock Market Movements. Master's Thesis, École de gestion de l'Université de Liège, Liège, Belgium, 2016.
5. Loh, W.Y. Regression trees with unbiased variable selection and interaction detection. *Stat. Sin.* **2002**, *12*, 361–386.
6. Loh, W.Y. Variable Selection for Classification and Regression in Large p, Small n Problems. In *Probability Approximations and Beyond*; Springer: New York, NY, USA, 2012; Volume 205, pp. 135–159.
7. Azhagusundari, B.; Thanamani, A.S. Feature selection based on information gain. *Int. J. Innov. Technol. Explor. Eng. (IJITEE)* **2013**, *2*, 18–21.
8. Hall, M. Correlation-Based Feature Selection for Machine Learning. Ph.D. Dissertation, University of Waikato Hamilton, Hamilton, New Zealand, 1999.
9. Nassuna, H.; Eyobu, O.S.; Kim, J.H.; Lee, D. Feature selection based on variance distribution of power spectral density for driving behavior recognition. In Proceedings of the 2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA), Xi'an, China, 19–21 June 2019; pp. 335–338.
10. Fong, S.; Biuk-Aghai, R.P.; Si, Y.W. Lightweight feature selection methods based on standardized measure of dispersion for mining big data. In Proceedings of the 2016 IEEE International Conference on Computer and Information Technology, Nadi, Fiji, 8–10 December 2016; pp. 553–559.
11. Morán-Fernández, L.; Bolón-Canedo, V.; Alonso-Betanzos, A. Centralized vs. distributed feature selection methods based on data complexity measures. *Knowl.-Based Syst.* **2017**, *117*, 27–45. [CrossRef]
12. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
13. Lin, X.; Li, C.; Zhang, Y.; Su, B.; Fan, M.; Wei, H. Selecting feature subsets based on svm-rfe and the overlapping ratio with applications in bioinformatics. *Molecules* **2018**, *23*, 52. [CrossRef]
14. Mejia-Lavalle, M.; Sucar, L.; Arroyo-Figueroa, G. Feature selection with a perceptron neural net. In Proceedings of the International Workshop on Feature Selection for Data Mining, Hong Kong, China, 18–22 December 2006; pp. 131–135.
15. Kaya, E.; Morani, K. The Improvement Achieved Using Blogreg Feature Selection Algorithm in a Developed Artificial Neural Network Classification. *Int. J. Sci. Res. Eng. Technol. (IJSET)* **2019**, *13*, 28–31.
16. Langley, P. Selection of relevant features in machine learning. *Proc. AAAI Fall Symp. Relev.* **1994**, *97*, 245–271.
17. Lee, S.J.; Xu, Z.; Li, T.; Yang, Y. A novel bagging c4.5 algorithm based on wrapper feature selection for supporting wise clinical decision making. *J. Biomed. Informat.* **2018**, *78*, 144–155. [CrossRef] [PubMed]
18. Maldonado, S.; Weber, R. A wrapper method for feature selection using support vector machines. *Inf. Sci.* **2009**, *179*, 2208–2217. [CrossRef]
19. Mustaqeem, A.; Anwar, S.; Majid, M.; Khan, R. Wrapper method for feature selection to classify cardiac arrhythmia. In Proceedings of the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju, Korea, 11–15 July 2017; Volume 2017, pp. 3656–3659.
20. Dy, J.G.; Brodley, C.E. Feature subset selection and order identification for unsupervised learning. In Proceedings of the Seventeenth International Conference on Machine Learning, San Francisco, CA, USA, 2 October 2000; pp. 247–254.
21. Pace, N.; Briggs, W. Stepwise logistic regression. *Anesthesia Analgesia* **2009**, *109*, 285–286. [CrossRef]
22. Sisovic, S.; Brkic Bakaric, M.; Matetic, M. Reducing data stream complexity by applying count-min algorithm and discretization procedure. In Proceedings of the 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService), Bamberg, Germany, 26–29 March 2018; pp. 221–228.
23. Xiao, L.; Dai, B.; Liu, D.; Zhao, D.; Wu, T. Monocular road detection using structured random forest. *Int. J. Adv. Robot. Syst.* **2016**, *13*, 101. [CrossRef]

24. MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Berkeley, CA, USA, 1967; pp. 281–297.
25. Li, Y.; Liu, L.; Bai, X.; Cai, H.; Ji, W.; Guo, D.; Zhu, Y. Comparative study of discretization methods of microarray data for inferring transcriptional regulatory networks. *BMC Bioinform.* **2010**, *11*, 520. [CrossRef] [PubMed]
26. Fayyad, U.; Irani, K. Multi-interval discretization of continuous-valued attributes for classification learning. *IJCAI* **1993**, *13*, 1022–1027.
27. Ramírez-Gallego, S.; García, S.; Mourino-Talin, H.; Martinez, D. Distributed entropy minimization discretizer for big data analysis under apache spark. In Proceedings of the 2015 IEEE Trustcom/BigDataSE/ISPA, Helsinki, Finland, 20–22 August 2015; pp. 33–40.
28. Kerber, R. Chimerge: Discretization of numeric attributes. In Proceedings of the Tenth National Conference on Artificial Intelligence, AAAI'92, San Jose, CA, USA, 12–16 July 1992; pp. 123–128.
29. Bertier, P.; Bouroche, J.M. *Analyse des données Multidimensionnelles*; PUF: Paris, France, 1975.
30. Boulle, M. Khiops: A statistical discretization method of continuous attributes. *Mach. Learn.* **2004**, *55*, 53–69. [CrossRef]
31. Zhang, Y.; Yu, J.; Wang, J. *Parallel Implementation of chi2 Algorithm in Mapreduce Framework*; Springer: Cham, Switzerland, 2014; pp. 890–899.
32. Jiang, F.; Zhao, Z.; Ge, Y. A supervised and multivariate discretization algorithm for rough sets. In *Rough Set and Knowledge Technology*; Yu, J., Greco, S., Lingras, P., Wang, G., Skowron, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 596–603.