International Journal of Intelligent Computing and Technology
ISSN: 2457-0249 (Online)
Website: https://www.ijict.com Jan (2022) - Vol.5, Iss.2 Pages:15-20

FEATURE SELECTION METHODS TO REDUCE HIGH DIMENSIONALITY IN BIGDATA

A Balasathya[1], S Banumathi[2]
[1]III -BSc,., [2]Assistant Professor, PG Department of Computer Science
Holy Cross College, Trichy, Tamilnadu

Article History- Received: June 2021; Published: Jan 2022

Abstract

Big data is a combination of structured, semi structured and unstructured data collected by organizations that can be mined for information and used in machine learning projects, predictive modelling and other advanced analytics application. Dimensionality in statistics refers to how many attributes a dataset has, healthcare data is notorious for having vast amounts of variables in an ideal world, and this data could be represented in a spread sheet, with one column representing each dimension. This paper presents a study on feature selection method to reduce high dimensionality issue in big data. Big data play an important role as data mining techniques are not capable to handle these big data. Big data is having large, complex and velocity characteristics which are the research area now-a-days. For large volume data, it having large high dimensions need new or modified existing feature selection techniques. In this paper, they discussed difference feature selection methods like filters, wrappers, embedded and hybrid.

Keywords: *Big data, High dimensionality, feature selection, embedded, hybrid*

1. INTRODUCTION

Paul Zikopoulos et al., Mentioned that every day trillions of data are generated across the world and put the information systems facing the emergence of big data phenomenon[2].This vertiginous evolution makes the enterprise confronting the challenge to build its own big data. To achieve the challenge, the enterprise is supposed to embark on big investments in terms of resource and material to process peta bytes of diverse data, this last are sometimes useful and sometimes useless. The problem here is how to optimize data relevancy to extract value from the big data sources [8]. From this reasons, the authors proposed an ETL and Map Reduce Hybrid access based on Data Filtering and processing to form an effective on-demand dimensional Big Data, enabling enterprises to process related data in powerful and effective way according to the stakeholder's needs [1].

Curbera et al., embedded analytics and statistics for big data have emerged as an important topic across industries [3]. As the volumes of data have increased, software engineers are called to support data analysis and applying some kind of statistics to them. This research article presents an overview of embedded data analytics and statistics tools and libraries, both position-only software packages and statistically capable programming languages.

Roy Thomas et al., conducting big data analytics in an organization is not just about using a processing framework (e.g.Hadoop/Spark) to learn a model from data directly in a single file system (e.g.HDFS)[5]. All users are frequently needed to pipeline real time data from other systems into the processing framework, and continually update the learned mode. The processing plan needs to be easily invaluable for different purposes to produce various models. The model and the following model updates need to be collective with a product that may require a real time indicator using the latest trained model. All these need to be mutual among different teams in the organization for different data analytics purposes. In this paper, the author proposed area time data-analytics-as-service architecture that uses Restful web services to wrap and integrate data services, dynamic model training services (supported by big data processing framework), prediction services and the product that uses the models. The challenges in wrapping big data processing systems are addressed as services and other architecturally important factors that affect system efficiency, performance in real time and accuracy of prediction).The proposed system test the architecture using a log-driven system activity anomaly detection system where staleness of data used in model preparation, speed of model update and prediction are important requirements.

Kamal Kc et al., Filter methods are robust in terms of over fitting and showing effectiveness in computation time. The purpose of the feature selection process was to select the required features. The approaches focused on the filter are independent of the algorithm for supervised learning. They are cheaper than the wrapper and embedded methods in computational terms. The high-dimensional data, the filter methods are suitable rather than the wrapper and embedded methods.[6]

2 FEATURE SELECTION: FILTER METHOD, WRAPPER METHOD AND EMBEDDED METHOD

Feature selection means selecting and retaining only the most important features in the model. Feature selection is different from feature extraction. In feature selection, subset of the features whereas in feature extraction, we create a new feature from the existing features [9].

Feature Selection Methods:

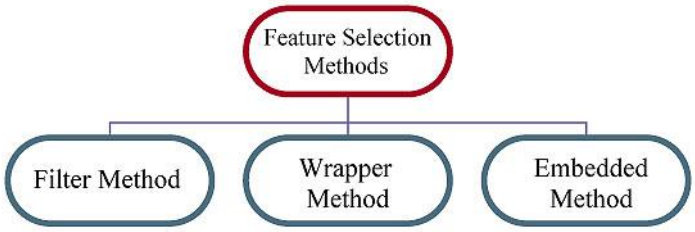

Fig 1 Feature Selection Methods

Filter Method:

In this method, features are filtered based on general characteristics (some metric such as correlation) of the dataset such correlation with the dependent variable. Filter method is performed without any predictive model. It is faster and usually the better approach when the number of features is huge. Avoids over fitting but sometimes may fail to select best features [9].

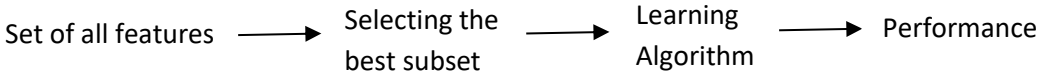

Fig 2 Filter Methods

Wrapper Method:

In wrapper method, the feature selection algorithm exits as a wrapper around the predictive model algorithm and uses the same model to select best features .Though computationally expensive and prone to over fitting, gives better performance.

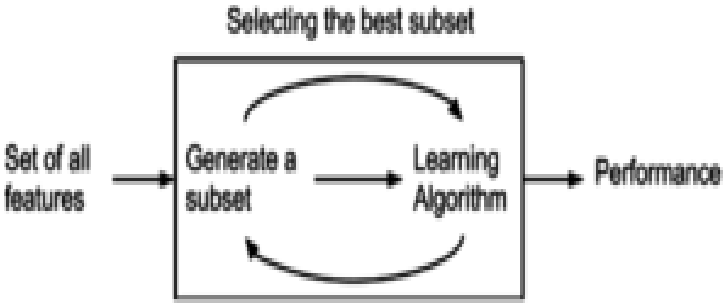

Fig 3 Wrapper Methods

Embedded Method:

In embedded method, feature selection process is embedded in the learning or the model building phase. It is less computationally expensive than wrapper method and less prone to over fitting

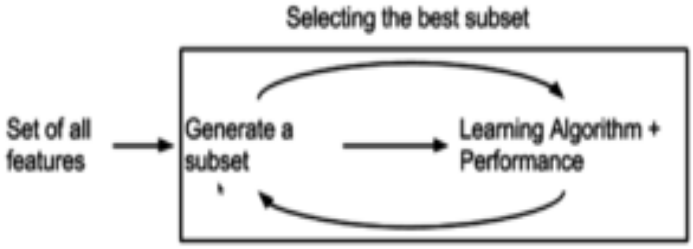

Fig 4 Embedded Methods

Hybrid Method:

A hybrid feature selection method is proposed for classification in small sample size data sets. Filter and wrappers are used together seen in the below figure 5. First, applied filter to remove features which reduce size and then wrapper is used which work fast because of small data and get good accuracy with more small features set. In big data, some research is finding hybrid approach.

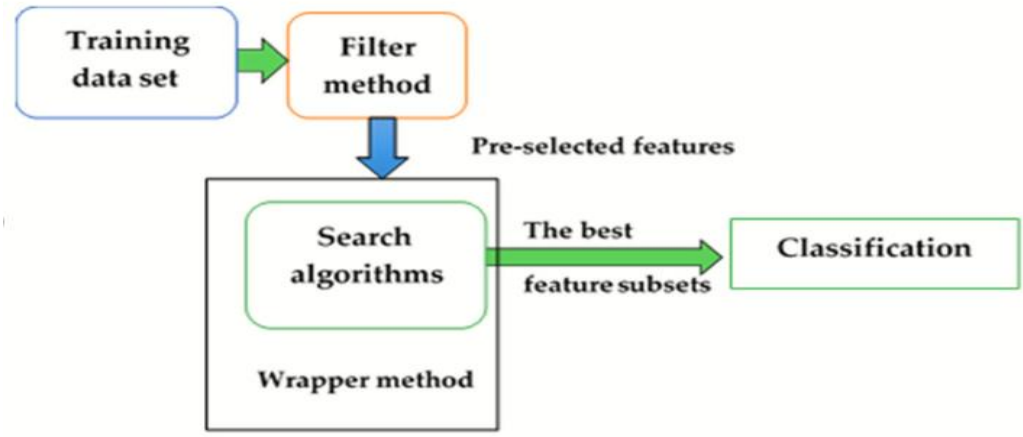

Fig 5 Hybrid Method

3 COMPARATIVE STUDY ON VARIOUS FEATURE SELECTION METHODS

Reference	Method	Advantage	Disadvantage
PaulZikopoulos et al., [2]	Hybrid filtering method	Very reliable. Can produce high gains [10].	Required many components. Very expensive.
Curberaet al.,[3]	Data analysis preprocessing method	Make it easier to interpret and use Changing the raw data	Inconsistent and superfluous data.

		into a clean data set.	
YangHui etal.,[5]	Embedded and wrapper method	Simple and efficient. Fast and accurate. Highly modifiable [10].	Not fast enough. Dataset must be pre-analysed. Complex implementation
KamalKcetal.,[6]	Filtering method	Fast and useful as preprocessor. Captures dependencies.	Ignores dependencies. Slow for large datasets. [10].

The above mentioned table shows that existing system process method with their advantages and disadvantages.

CONCLUSION

As growing digital era, knowledge is additionally growing in each moment. This huge knowledge is massive in volume, complex and speedy arrival want analysis method in effective manner, huge knowledge have several problems and among them high dimensional knowledge is one issue. The literature on feature selection techniques is very vast encompassing the applications of machine learning and pattern recognition. Comparison between feature selection algorithms can only be done using a single dataset since each underlying algorithm will behave differently for different data. A typical database management system is unable to process as much information. Filter method is faster and useful when there is more number of features. Wrapper method gives better performance while the embedded method lies in between the other two methods.

REFERENCES

1. Chaiken,Ronnie, "HD: easy and efficient parallel processing of massive datasets" Proceedings of the VLDB Endowment, Vol.1, Issue. 2, 2019, pp.1265-1276.
2. PaulZikopoulos,Chris Eaton, “Hybrid Understanding big data Analytics for enterprise class hadoop and streaming data”, McGraw-Hill Osborne Media, 2019.
3. Curbera,Francisco,"Unraveling the Web services web an introduction to embedded data, WSDL,and UDDI."s IEEE Internet computing, Vol.6, Issue.2,2017, pp.86-93.
4. Dean, Jeffrey, Sanjay Ghemawat, "Wrapped simplified data processing on large clusters", Communications of the ACM“, Vol.51, Issue.1, 2018, pp.107-113.
5. Fielding, RoyThomas, “ Architectural styles and the design of network-based software architectures” Diss. University of California, Irvine, 2019.
6. “Microsoft Azure HD Insight”, http://azure.microsoft.com/enus/services/hdinsight/.
7. Erich Gamma, Richard Helm, Ralph E.Johnson, John V lissides, ”Design patterns: elements of reusable object oriented software”, vol.2, Issue.06.2017.
8. Banumathi. S and Aloysius. A, “Big Data Prediction Using Evolutionary Techniques: A Survey”, Journal of Emerging Technologies and Innovative Research, Vol. 3, Issue.9, Pg. 89-91, Sep 2016.
9. https://www.datasciencesmachinelearning.com/2019/10/feature-selection-filter-method-wrapper.html

How to cite this article:
A Balasathya, S Banumathi, "Comparative Study on Feature Selection Methods to Reduce High Dimensionality in Bigdata", International Journal of Intelligent Computing and Technology (IJICT), Vol.5, Iss.2, pp.15-20, 2022