A Review Paper on Data Mining and Big Data

OPEN ACCESS

Volume : 6

Special Issue : 1

Month : November

Year: 2018

ISSN: 2321-788X

Impact Factor: 3.025

Citation:
Jeyaseelan, C., and M. Muthusrinivasan. "A Review Paper on Data Mining and Big Data." *Shanlax International Journal of Arts, Science and Humanities*, vol. 6, no. 1, 2018, pp. 1–5.

DOI:
https://doi.org/ 10.5281/ zenodo.1614140

C.Jeyaseelan
Associate Professor, Department of Computer Applications
Sri Kaliswari College, Sivakasi

M.Muthusrinivasan
Assistant Professor, Department of Computer Applications
Sri Kaliswari College, Sivakasi

Abstract

In the recent world, big data is the very popular term. Big data is generated from many sources such as social media, digital images or videos and so on. The data mining is very helpful for extracting useful information from big data. It is a process which finds useful patterns from a large amount of data. There are so many techniques fin data mining such as clustering, prediction, and classification and decision tree available for solving the problems of big data. In this paper, we present the overview of data mining and big data issues, challenges and its solutions.

Keywords: Big data, data mining

Introduction

There is a large amount of data is available in the information industry and it is exceeding day by day. The term data mining is firstly originated in the year of 1990 before statisticians used terms like Data Fishing or Data Dredging. The most important purpose of data mining is to find the valuable information from large data sets. In others words, data mining is a process of mining knowledge from data. Data mining used in many modern applications these days such as Market Analysis and Management, Corporate Analysis & Risk Management, Fraud Detection, science exploration, sports, astrology, and Internet Web Surf-Aid and so on.

Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data [1].

Data mining is used for exploring and analyzing large amounts of data to find patterns for big data. The advent of big data, the data mining is more prevalent. Four or five years ago, companies collected all data of transaction stored in a single database. Today, a volume of data is collected have exposed. Marketers can also collect information about every conversation people are having about their brand. It requires the implementation of new processes, technology and governance mechanisms that are collectively being referred to as big data. Today, big data is big business[2].

We can define big data is a process that allows companies to extract information from a large amount of data. Big data is used in data mining techniques because the size of information is larger.

The main purpose of data mining fin either classification or prediction. In classification, sorting a data into groups, e.g., marketers are only interested in those who responded or not those who did not respond to the promotion. In prediction, to predict a value, e.g., marketers are only interested in predicting for those who responded in promotion only.

Algorithms Used in Data Mining for Big Data

A. *Classification trees* A popular data-mining technique that is used to classify a dependent categorical variable based on measurements of one or more predictor variables. The result is a tree with nodes and links between the nodes that can be read to form if-then rules[3].

B. *Logistic regression* A statistical technique that is a variant of standard regression but extends the concept to deal with classification. It produces a formula that predicts the probability of the occurrence as a function of the independent variables[4].

C. *Neural networks* A software algorithm that is modeled after the parallel architecture of animal brains. The network consists of input nodes, hidden layers, and output nodes. Each unit is assigned a weight. Data is given to the input node, and by a system of trial and error, the algorithm adjusts the weights until it meets certain stopping criteria. Some people have likened this to a black–box approach[5].

D. *Clustering techniques like K-nearest neighbors* A technique that identifies groups of similar records. The K-nearest neighbor technique calculates the distances between the record and points in the historical (training) data. It then assigns this record to the class of its nearest neighbor in a data set[6].

Types of Data Mining System

Data mining systems can be categorized according to various criteria the classification is as follows [7]:

A. Classification of data mining systems according to the type of data source mined: In an organization, a huge amount of data is available where we need to classify these data but these are available most of the times in a similar fashion. We ask to sort out these data according to its character (maybe the audio/picture, text formatting and so forth)[8].

B. Classification of data mining systems, according to the data model: There are so many numbers of data mining models (Relational data model, Object Model, Object-Oriented Data Model, Hierarchical data Model/W data model) are available and eevery model we are practicing the different data. Agreeing to these data model the data mining system classifies the information [9].

C. Classification of data mining systems, according to the sort of knowledge discovered: This classification based on the variety of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, categorization, clustering, and so on some systems tend to be comprehensive systems offering several data mining functionalities together [10].

D. Classification of data mining systems, according to excavation techniques used: This classification is according to the data analysis approach used such as machine learning, neural nets, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, and so on The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems. A comprehensive system would offer a broad assortment of data mining techniques to suit different situations and options, and offer different levels of user interaction[11].

Issues, Challenges and Problems of Big Data in Data Mining

A. Problems

The main problems in big data have grown tremendously. This large amount of data is beyond thef software tools to manage. The exploring a large amount of data, exacting useful information from data sets and knowledge is a challenge; sometimes it is a major problem. Also, big data is unstructured, huge size and it is not easy to handle.

B. Issues

The main issues of data mining in big data are follows

a) Poor data quality, e.g., noisy data, dirty data and inadequate size of data.
b) Redundant data is uploaded from various sources such as multimedia files.
c) Security, the privacy of the companies
d) Algorithm of data mining is not effective.
e) Difficult to processing an unstructured data into structured data.
f) Higher cost, less flexibility.

C. Major Challenges

- Big Data Mining Platform
- Big Data Semantics and Application Knowledge
- Information Sharing and Data Privacy
- Domain and Application
- Big Data Mining Algorithms
- Local Learning and Model Fusion for Multiple Information Sources mining from Sparse
- Uncertain, and Incomplete Data
- Mining Complex and Dynamic Data [12]

Solutions

A. Hadoop: It is the open-source software framework for distributed storage of very large datasets on computer clusters. Hadoop provides massive amounts of storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs. Hadoop is widely used in industrial applications with Big Data, including spam filtering, network searching, click stream analysis, and social recommendation. To distribute its products and services, such as spam filtering and to searchg, Yahoo has run Hardtop in 42,000 servers at four data center as of June 2012. Currently, the largest Hadoop cluster contains 4,000 nodes, which is expected to increase to 10,000 with the release of Hadop2.0 [13].

B. Cloudera

Cloudera is similar to Hadoop with extra services. It helps in business, to allow people in the companies is easy to access the data from the larger database. It also provides data security which is highly important for storing sensitive and personal information.

C. MongoDB

MongobDB is the modern approach to databases. It is the very good approach for managing data that changes frequently or data or unstructured. Common use cases include storing data for mobile apps, product catalogs, real-time personalization, content management and applications delivering a single view across multiple systems.

Map Reduce is the hub of Hadoop and is a programming paradigm that enables mass scalability across numerous servers in a Hadoop cluster. In this cluster, each server contains a set of internal disk drives that are inexpensive. To enhance performance, Map Reduce assigns workloads to the servers in which the processed data are stored. Data processing is scheduled based on the cluster nodes. A node may be assigned a task that requires data foreign to that node[14].

Conclusion

Today, all the IT professionals, engineers and researchers are working on big data. Big data is a termf concerning about large volume of complex data sets. To solve problems of big data challenges, many researchers proposed a different system models, techniques for big data. The high-performance computing paradigm is required for data mining to solve the problem of big data. We conclude that there are still chances to improve the algorithms and techniques for data mining. I n this paper, big data are facing lots of challenges, issues and providea solutions to handle the big data.

References

Andrews, R., Diederich, J. & Tickle, A. (1995) *"Survey and critique of techniques for extracting rules from trained artificial neural networks,"* Knowledge-Based Systems, vol. 8, no. 6, pp. 373-389.

Armstrong, J. S. (2001) Principles of Forecasting: A Handbook for Researchers and Practitioners, Norwell, MA: Kluwer Academic Publishers.

Carpenter. G. A. & Tan. A.W. (1995) *"Rule Extraction: From Neural Architecture to Symbolic Representation,"* Connection Science, vol. 7, no. 1, pp. 3-27.

Chen. M, Mao. S, Liu. Y. (2014) Big data: a survey. Mobile Networks and Applications. vol. 19, no. 2, pp. 171–209.

Duch, W., Adamczak. R. & Grabczewski, K. (2000) *"A new methodology of extraction, optimization and application of crisp and fuzzy logical rules,"* IEEE Trans Neural Networks, vol. 11, no.2, pp. 1-31.

Fu, L. M. (1994) *"Rule Generation from Neural Networks,"* IEEE transactions on systems, man and Cybernetics, vol. 28, no. 8, pp. 1114-1124.

Horikawa, S. et al., (1992) *"On Fuzzy Modeling Using Fuzzy Neural Networks with the Back - propagation Algorithm,"* IEEE Trans Neural Networks, vol. 3, pp. 801-806.

Kasabov, N. (2001) *"On -line learning, reasoning, rule extraction and aggregation in locally optimized evolving fuzzy neural networks,"* Neurocomputing, vol. 41, no.1-4, pp. 25-45.

Maire, F (1997) *"A Partial Order for the M -of-N Rule Extraction Algorithm,"* IEEE Trans Neural Networks, vol. 8, no. 6, pp. 1542-1544.

Mitra, S., Pal. S. K. & Mitra, P. (2002) *"Data mining in soft computing framework: A survey,"* IEEE Trans Neural Networks, vol. 13, pp. 3-14.

Opitz. D. & Shavlik, J. (1996) *"Actively searching for an effective neural-network ensemble,"* Connection Science, vol. 8, pp. 337-353.

Refenes, A. N., Zapranis, A. & Francis, G. (1994) *"Stock Performance Modelin g Using Neural Networks: A Comparative Study with Regression Models,"* Neural Network, 5, pp. 961-970.

Setiono, R. (2000) *"Extracting M -of-N Rules From Trained Neural Networks'* 'IE, EE Trans Neural Networks, vol. 11, no. 2, pp. 512-519.

Thrun, S. (1995) *"E xtracting rules from artificial neural networks with distributed representations."* In Tesauro, G.; Touretzky, D.; and Leen, T., eds., Advances in Neural Information Processing Systems (NIPS) 7. Cambridge, MA: MIT Press, pp. 505-512.

Tickle, A., Andrews, R., Golea. M. & Diederich, J. (1998) *"The Truth Will Come To Light: Directions and Challenges in Extracting the Knowledge Embedded Within Trained Artificial Neural Networks,"* IEEE Trans Neural Networks, vol. 9, no. 6, pp. 1057-1068.

Xingquan Zhu, Ian Davidson, *"Knowledge Discovery and Data Mining: Challenges and Realities,"* ISBN 978-1-59904-252, Hershey, New York, 2007.

Web Sources

http://www.ijarcs.info/index.php/Ijarcs/article/view/3789
http://www.warse.org/IJETER/static/pdf/Issue/NCTET2015sp32.pdf
https://pdfs.semanticscholar.org/ac4c/fc4fc86b5b680632cb4d80ddbc64717a44c3.pdf
https://www.coursehero.com/file/29187567/A-Review-Paper-On-Data-Mining-And-Big-Datapdf/
https://www.coursehero.com/file/p78kvp1s/To-distribute-its-products-and-services-such-as-spam-filtering-and-searching/
https://www.dummies.com/programming/big-data/engineering/data-mining-for-big-data/
https://www.import.io/post/all-the-best-big-data-tools-and-how-to-use-them/