International Journal of Research in Engineering, Science and Management
Volume 4, Issue 4, April 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

Data Mining: The Queen of Analytics

Bibin John Jacob[1*], Ashwin Philip[2], Sanal Jacob[3]
[1,2,3]*Department of Computer Applications, Saintgits College of Applied Sciences, Kottayam, India*

***Abstract*: This paper presents an overview on data mining.**

***Keywords*: Artificial, Data, Intelligence, Mining.**

1. Introduction to Data Mining

One of the most innovative breakthrough of technology, aiding us with effective date collection, data warehousing as well as computer processing. Also, called 'Knowledge Discovery in Data', is becoming popular in the realm of commerce and business activities in general. In every aspect we can think of, let it be online shopping, swiping a card at a grocery store, being in social media platforms, even in governmental networking such as online records, bank transactions, data being downloaded in a database. Being a digital epoch in which most organizations are processing and analyzing data, Data Mining has proved to be very efficient and useful. Data Mining has made it easier to discover or extract the data we want from a colossal amount of data. We cannot say it is the same as searching a desirable data. Data Mining makes it possible to find interesting patterns in the field of machine learning and databases and now, in the field of Artificial Intelligence (AI) as well. Too much information and too little knowledge, is a problem, now rectified with the introduction of Data Mining, as it helps extract the needed information only.

2. How Can We Extract Desired Data from Such Large Databases?

Mining uses some certain specific tools for that matter, the tools being,

1. Oracle Data Mining: is mainly designed for data warehousing, supporting business intelligence activities, mostly in analytics. It is written in C, C++ and assembly language.
2. Orange: is data visualizing using visual programing front end for data analysis. It is written in Python, C, C++ and Cython.
3. IBM: International Business Machines, written in JAVA.
4. Teradata: assists in analysis and decision making. Interesting thing about teradata is it is written in various languages, such as SQL, R, Python, etc.
5. Rapidminer: Assist indirectly in data mining as it provides an interface that allows users to interact with devices via graphical icons and audio indicators.

3. Where is Data Mining Applicable?

In banks, the software is used for fraud detection by analyzing past transactions patterns. Card marketing, i.e. delivering targeted ads to customers via online bank statement by identifying customer's segment car issuers and improvising profitability more effectively. In telecommunication, data mining is used for call detail record analysis by identifying similar used patterns, customer loyalty, maintain customer purchase. In retail, it is used for performing basket analysis to improve stocking and promotion, sales forecasting, database marketing etc.

Like every other software, data mining also has its own cons.

Achieving certain data means, violating user privacy. Indicating a lack of security and privacy for its users. It may possibly misinterpret the user's satire as a positive sentiment leading to mining of incorrect information.

A. Data Mining in Cyber Crime Detection

It is truly a benefit that we have so many opportunities and easier ways to access anything and everything at the end of our fingertips. But surely there are problems and misuses associated with it. Networks or such social platforms may be used as a commission of crimes.

Fortunately, data mining has been of significant importance in deception detection. Various data classification and regression algorithms are used for this purpose. Some of them are:

1) ARIMA (Autoregressive Integrated Moving Average Model)

These models help in better understanding the data or to forecast data. It is used along with artificial neural networks to determine a metric safety value. Given a time series of data where t is an integer index and are real numbers, then an ARMA (p', q) model is given by:

$$\left(1-\sum_{i=1}^{p'}\alpha_i L^i\right)X_t=\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t$$

Every ARIMA modal with d>0 is not wide sense stationary. Therefore, is a generalized form the equation can be written as,

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t=\delta+\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t$$

2) Artificial Neural Networks

Are network modals used to estimate functions that can depend on large number of inputs and are generally unknown.

*Corresponding author: bibin.jacob18@gmail.com

For example, a neural network for handwritten recognition is defined by a set of input neurons which may be activated by the pixels of an input image. Then they are weighed and transformed by a function that activates these and then determine which character was read.

3) Bayesian Network Algorithm

Bayes network is a probabilistic graphical modal representing a set of random variables and their conditional dependencies via a DAG (Directed Acyclic Graph).

The network can be used to link certain crime scenes with a certain known criminal by taking all the data about the crime scene and comparing with the data of the respective crime scene of all criminals in custody and an inference system is built and if it matches then, the profile of the criminal and that of the crime scene are linked.

4. System Analysis by Data Mining

A. Functional Requirements

- Generating sets of data: is done by getting crime data from institutions according to what parameter we may want to predict and detect.
- Cleaning the data set: this stage is where we organize the collected data into such a form, easy to run analysis and get accurate results.
- Analyzing the data: the data is analyzed and the results are collected and stored, this data is then handed over to the police agencies for further investigation.

B. Nonfunctional Requirements

- Performance requirements: depends on how quickly the system will be able to run the analysis and prepare the crime based patterns. This is an important necessity in order to get accurate results.
- Safety and security requirements: the system must be safe and not susceptible to attacks.
- Data and integrity requirements: this step ensures that the system processes the data efficiently and the output produced is accurate.
- Availability requirements: the system should be available to use whenever necessary.
- Portability requirements: the system should be run on any device with the same efficiency, accuracy and speed. And the system should be run on any software.
- Maintainability requirements: like any other systems, this system also needs to be maintained and repaired for errors and bugs regularly, so that the system can function smoothly.
- Software quality assurance attributes: The software should be reliable, efficient, compatible. It should be able to work in unusual situations as well.

5. Data Modeling and Design

We have classified data modeling into various levels to understand how it actually works. There are 3 levels.

1. *Level 0:* In this level, the crime data is taken from the law enforcement agencies and is used by our software to predict the crime pattern of the desired field in order to get the results.

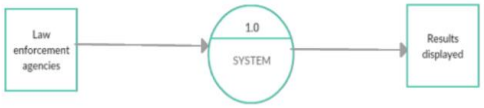

Fig. 1. Level 0

2. *Level 1:* In this level, we can see the strategy of working of our system.

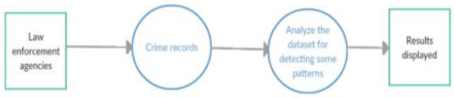

Fig. 2. Level 1

3. *Level 2:* In this level, the raw data that was obtained from the law enforcement agencies will be pre-processed and data-mining algorithms will be applied to predict the value of the target variables which will be displayed as the result. This is the most detailed level of all.

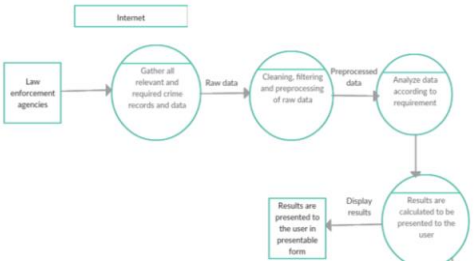

Fig. 3. Level 2

6. Conclusion

With the advancement of life comes its respective responsibilities in loads. And as we know, technological and software advancement is the biggest of them all. The amount of data a person has to handle has also shot up in huge amounts, where the development and use of data mining and its tools become significantly important. Although the few functions of the software can be graded up so as to be completely dependable and error free for a more trustworthy and efficient usage.

Acknowledgement

I'd like to sincerely thank our professor Mr. Sanal Jacob, for being supportive and for guiding us in the right direction. I'd also like to thank the entire Computer Science department of MG University for giving me this opportunity and for supporting us in all ways.

References

[1] Deep belief networks http://www.scholarpedia.org/article/Deep_belief_networks
[2] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," in Neural Computation, vol. 18, no. 7, pp. 1527–1554, 2006.
[3] Charu C. Aggarwal, "Data Mining: The Textbook," 2015.
[4] Data mining, https://en.wikipedia.org/wiki/Data_mining
[5] Cybercrimes deception detection, https://en.wikipedia.org/wiki/Cybercrime