IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Review study of Big Data Analytics

Shrutika Sanjiv Padalkar
Department of IT GMVCS Tala University of Mumbai

Ashwini Dattatray Salunke
Department of IT GMVCS Tala University of Mumbai

Prof A.A.Amburle
Assistant professor
GMVCS & GMVIT University of Mumbai

Abstract :- Big data is a new driver of the world economic and societal changes. The world's data collection is reaching a tipping point for major technological changes that can bring new ways in decision making, managing our health, cities, finance and education. While the data complexities are increasing including data's volume, variety, velocity and veracity, the real impact hinges on our ability to uncover the `value' in the data through Big Data Analytics technologies. Big Data Analytics poses a grand challenge on the design of highly scalable algorithms and systems to integrate the data and uncover large hidden values from datasets that are diverse, complex, and of a massive scale. Potential breakthroughs include new algorithms, methodologies, systems and applications in Big Data Analytics that discover useful and hidden knowledge from the Big Data efficiently and effectively.

Keywords :- Introduction to big data, Big data analytics, Analytical theory and methods.

INTRODUCTION

The quantity of data created by humans is quickly increasing every year as a result of the introduction of new technology, gadgets, and communication channels such as social networking sites.Big data is a group of enormous datasets that can't be handled with typical computer methods. It is no longer a single technique or tool; rather, it has evolved into a comprehensive subject including a variety of tools, techniques, and frameworks.Quantities, letters, or symbols on which a computer performs operations and which can be stored and communicated as electrical signals and recorded on magnetic, optical, or mechanical media.

I. INTRODUCTION TO BIG DATA

The "Internet of Things" and its widely ultra-connected nature are leading to a burgeoning rise in big data. There is no dearth of data for today's enterprise. On the contrary, they are mired in data and quite deep at that. That brings us to the following questions:

1. Why is it that we cannot forego big data?
2. How has it come to assume such magnanimous importance in running business?
3. How does it compare with the traditional Business Intelligence (BI) environment?
4. Is it here to replace the traditional, relational database management system and data warehouse environment or is it likely to complement their existence?"

II. BIG DATA ANALYTICS

Big Data Analytics is...

1 Technology-enabled analytics: Quite a few data analytics and visualization tools are available in the market today from leading vendors such as IBM, Tableau, SAS, R Analytics, Statistica, World Programming Systems (WPS), etc. to help process and analyze your big data.

2. About gaining a meaningful, deeper, and richer insight into your business to steer it in the right direction. understanding the customer's demographics to cross-sell and up- sell to them, better leveraging the services of your vendors and suppliers, etc.

3. About a competitive edge over your competitors by enabling you with findings that allow quicker and better decision- making.

4. A tight handshake between three communities: IT, business users, and data scientists.

5. Working with datasets whose volume and variety exceed the current storage and processing capabilities and infrastructure of your enterprise.

III. ANALYTICAL THEORY AND METHODS

1. NAIVE BAYES

Naive Bayes is a probabilistic classification method based on Bayes' theorem. Bayes' theorem gives the relationship between the probabilities of two events and their conditional probabilities.

A naive Bayes classifier assumes that the presence or absence of a particular feature of a class is unrelated to the presence or absence of other features. For example, an object can be classified based on its attributes such as shape, color, and weight.

The input variables are generally categorical, but variations of the algorithm can accept continuous variables, There are also ways to convert continuous variables into categorical ones. This process is often referred toas the discretization of continuous variables. For an attribute such as income,the attribute can be converted into categorical values as shown below.

- Low Income: income < $10,000
- Working Class: $10,000 < income < $50,000
- Middle Class: $50,000 < income < $1,000,000
- Upper Class: income >$1,000,000

The output typically includes a class label and its corresponding probability score. The probability score is not the true probability of the class label, but it's proportional to the true probability.

Because naive Bayes classifiers are easy to implement and can execute efficient. Spam filtering is a classic use case of naive Bayes text classification. Bayesian spam filtering has become a popular mechanism to distinguish spam e-mail from legitimate e-mail.

Naive Bayes classifiers can also be used for fraud detection. In the domain of auto insurance, for example, based on a training set with attributes such as driver's rating, vehicle age, vehicle price, historical claims by the policy holder, police report status, and claim genuineness, naive Bayes can provide probability- based classification of whether a new claim is genuine.

2. BAYES' THEOREM

The conditional probability of event C occurring, given that event

A has already occurred, is denoted as P(C|A), which can be found using the formula in Equation 5-6.

( )

( )

( ) P

P A

P A C C A Ç

= ……(5-6)

Equation 5-7 can be obtained with some minor algebra and 70

substitution of the conditional probability ( )

( ) ( ) ( )

. P

P A

P A C P C

C A =..... (5-7)

Where c is the class label C Î{c1, c2 , ... cn} and A is observed

attributes { } 1 2 , , ... m A = a a a Equation 5-7 is the most common form of

the Baye's theorem.

Mathematically, Bayes' theorem gives the relationship between the probabilities of C and A, P(C) and P(A), and the conditional

probabilities of C given A and A, given C, namely P(C/A) and P(A/C)

3. DIAGNOSTICS

Unlike logistic regression, naive Bayes classifiers can handle missing values. Naive Bayes is also robust to irrelevant variables that are distributed among all the classes whose effects are not pronounced.

The model is simple to implement even without using libraries.The prediction is based on counting the occurrences of events, making the classifier efficient to run. Naive Bayes is computationally efficient and is able to handle high- dimensional data efficiently. .In some cases naive

Bayes even outperforms other methods. Unlike logistic regression, the naive Bayes classifier can handle categorical variables with many levels. Recall that decision trees can handle categorical variables as well, but too many levels may result in a deep tree. The naive Bayes classifier overall performs better than decision trees on categorical values with many levels.

Compared to decision trees, naive Bayes is more resistant to overfitting, especially with the presence of a smoothing technique.

One problem of the Laplace smoothing is that it may assign too much probability to unseen events. To address this problem, Laplace smoothing can be generalized to use e instead of 1, where typically

e Î[0,1] see equation 5-8. ( )

[ ]

( )

P`` ( )

x

count x x

count x e

e

+

=

+ ……(5.8)

Smoothing techniques are available in most standard software packages for native Bayes classifiers. However, if for some reason (like performance concerns) the native Bayes classifiers needs to be coded directly into an application, the smoothing and logarithm calculations should be incorporated into the implementation.

Despite the benefits of naive Bayes, it also comes with a few disadvantages. Naive Bayes assumes the variables in the data are conditionally independent. Therefore, it is sensitive to correlated variables because the algorithm may double count the effects.

As an example assume that people with low income and low credit tend to default. If the task is to score "default" based on both income and credit as two separate attributes, naive Bayes would experience the double-counting effect on the default outcome, thus reducing the accuracy of the prediction.

Although probabilities are provided as part of the output for the prediction, naive Bayes classifiers in general are not very reliable for probability estimation and should be used only for assigning class labels.

Naive Bayes in its simple form is used only with categorical variables. Any continuous variables should be converted into a categorical variable with the process known as discretization.

CONCLUSION

Hence we Had studied About Big data analytics and analytical theory and methods.

REFERENCES:-

1) Big Data- The definitive guide to the revolution in business analytics-Fujitsu

2) Data Science & Big Data Analytics Discovering, Analyzing, Visualizing and Presenting Data EMC Education Services Published by John Wiley & Sons, Inc