International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 12, December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

Predictive Analysis of Big Data in Data Mining

[1]Arunraj Gopalsamy, [2]Dr.B.Radha

[1]Lead IT Architect, IBM, (India)
[2]Professor, STC College, (India)

ABSTRACT

Big Data and Predictive analysis are highly focused in the area of Data science. Big Data has become important as many organizations and have been collecting massive amounts of domain-specific information, which can contain either useful or un-useful information about problems such as national intelligence, cyber security, fraud detection, marketing, and medical informatics. Healthcare is one of the sector which collects more than 50% of the Big Data and companies are analysing large volumes of data for business analysis and decisions, impacting existing and future technology. However, these companies/sectorsdo not know where and how to start with. Understanding the need and purpose is more important. A key benefit of predictive analysis is the deep learning of massive amounts of unsupervised data, making it a valuable tool for Big Data Analytics where raw data is largely unlabelled and un-categorized. In the present study, we explore how predictive analysis can be utilized for addressing some important problems in Big Data Analytics, including extracting complex patterns from massive volumes of data, semantic indexing, data tagging, fast information retrieval, and simplifying discriminative tasks. We conclude by presenting insights into relevant future works by posing some questions, including defining data sampling criteria, domain adaptation modelling, defining criteria for obtaining useful data abstractions, improving semantic indexing, semi-supervised learning, and active learning.

Keywords: Big Data, Data Mining, Predictive Analysis

I INTRODUCTION

Unfortunately, traditional analytics tools are not well suited to capturing the value hidden in Big Data. The volume of data is too large for comprehensive analysis. The range of potential correlations and relationships between disparate data sources, from back end customer databases through to live web based click streams, are too great for any analyst to test all hypotheses and derive all the value buried in the data. Machine learning is a rather new domain of IT and advanced mathematics, based on new statistical algorithms that could analyze big volume of diverse data sources (image, sound, video, social network, geo-localization, “traditional” structured database, etc.) in near real time. Computers, using these new types of programs, could learn from data for better future use.

Several approaches to machine learning are used to solve problems. The focus will be on the two most commonly used ones – supervised and unsupervised learning—because they are the main ones supported by Mahout. Supervised learning is tasked with learning a function from labelled training data to predict the value of any valid input. Common examples of supervised learning include classifying e-mail messages as spam,

labelling Web pages according to their genre, and recognizing handwriting. Many algorithms are used to create supervised learners, the most common being neural networks, Support Vector Machines (SVMs), and Naive Bayes classifiers. Unsupervised learning, as you might guess, is tasked with making sense of data without any examples of what is correct or incorrect. It is most commonly used for clustering similar input into logical groups. It also can be used to reduce the number of dimensions in a data set to focus on only the most useful attributes, or to detect trends.

Common approaches to unsupervised learning include k-Means, hierarchical clustering, and self-organizing maps. Apache Mahout is a new open source project by the Apache Software Foundation (ASF) with the primary goal of creating scalable machine-learning algorithms that are free to use under the Apache license. Mahout contains implementations for clustering, categorization, and evolutionary programming. Furthermore, where prudent, it uses the Apache Hadoop library to enable Mahout to scale effectively in the cloud.

II PROBLEM ENCOUNTERED WITH BIG DATA PROCESSING

As we know that we are residing in an era of big data where the data from science, Internet, finance etc. are top big data drivers. Till now no unification of big data have been proposed by the researchers. The misconceptions with big data are that it cannot be processed using pre-processing techniques. This era should help in handling the enormous data which is produced in multiple areas and this is currently a challenge in IT Industry. The three important aspects pertaining to the processing of big data are: 1. Storing and managing Data, 2. Big Data Analysis and Computation, 3. Data Security.

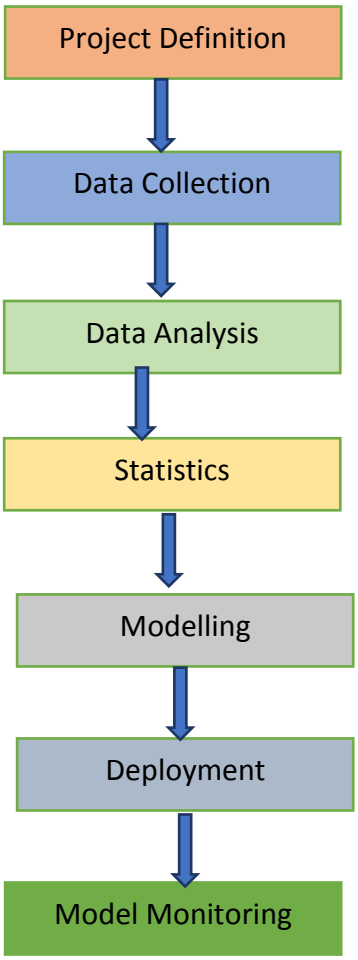
III PREDICTIVE ANALYSIS

Predictive analytics is an area of statistics that deals with extracting information from data and using it to predict trends and behavior patterns. Often the unknown event of interest is in the future, but predictive analytics can be applied to any type of unknown whether it be in the past, present or future. For example, identifying suspects after a crime has been committed, or credit card fraud as it occurs. The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting them to predict the unknown outcome. It is important to note, however, that the accuracy and usability of results will depend greatly on the level of data analysis and the quality of assumptions.

Predictive analytics is often defined as predicting at a more detailed level of granularity, i.e., generating predictive scores (probabilities) for each individual organizational element. This distinguishes it from forecasting. For example, "Predictive analytics—Technology that learns from experience (data) to predict the future behavior of individuals to drive better decisions. In future industrial systems, the value of predictive analytics will be to predict and prevent potential issues to achieve near-zero break-down and further be integrated into prescriptive analytics for decision optimization, Furthermore, the converted data can be used for closed-loop product life cycle improvement which is the vision of the Industrial Internet Consortium.

Predictive Analysis Process Flow

1. **Define Project:** Define the project outcomes, deliverable, scope of the effort, business objectives, identify the data sets that are going to be used.
2. **Data Collection**: Data mining for predictive analytics prepares data from multiple sources for analysis. This provides a complete view of customer interactions.
3. **Data Analysis**: Data Analysis is the process of inspecting, cleaning and modelling data with the objective of discovering useful information, arriving at conclusion
4. **Statistics**: Statistical Analysis enables to validate the assumptions, hypothesis and test them using standard statistical models.
5. **Modelling**: Predictive modelling provides the ability to automatically create accurate predictive models about future. There are also options to choose the best solution with multi-modal evaluation.
6. **Deployment**: Predictive model deployment provides the option to deploy the analytical results into everyday decision making process to get results, reports and output by automating the decisions based on the modelling.
7. **Model Monitoring**: Models are managed and monitored to review the model performance to ensure that it is providing the results expected.



Benefits of Predictive Analysis

1. **Improve efficiency in production** - Using predictive analytics, companies can effectively forecast for inventory and required production rates, while also using past data to estimate potential production failures. They can then use this to prevent the same errors from occurring
2. **Gain advantage over competitors** - Tapping into the customer data you have available can present you with insightful information as to why customers chose you over your competitors, highlighting unique selling points that you can then further promote to enhance leads.
3. **Reduce risk** – Reducing risk by means of making sensible, effective decisions
4. **Detect fraud** - By tracking changes in this behaviour on a site or network, it can easily spot anomalies that may indicate threat or fraud.
5. **Better marketing campaigns** - Predictive analytics' bread and butter is sifting through data to provide you with educated predictions on what to expect next.
6. **Meet consumer expectations** - Let predictive analytics do the behind the scenes work to help you form a better picture of who your customers are and what they want, so you can provide an offering tailored specifically for them.

IV TOOLS AND TECHNIQUES FOR COMPUTING BIG DATA

Predictive Analysis and Big Data are the hand on hand technologies through which more services can be given efficiently and effectively. The machine learning algorithms are the basis of designing and developing data Analytics to provide predictions.

A. Open Source Solutions would be needed for the predictive Analysis on the Big Data
    1. Visualization tools and systems
    2. QoS-based healthcare application provisioning frameworks.
    3. QoS optimization techniques for Big Data
    4. Techniques for preserving security and privacy

Radoop

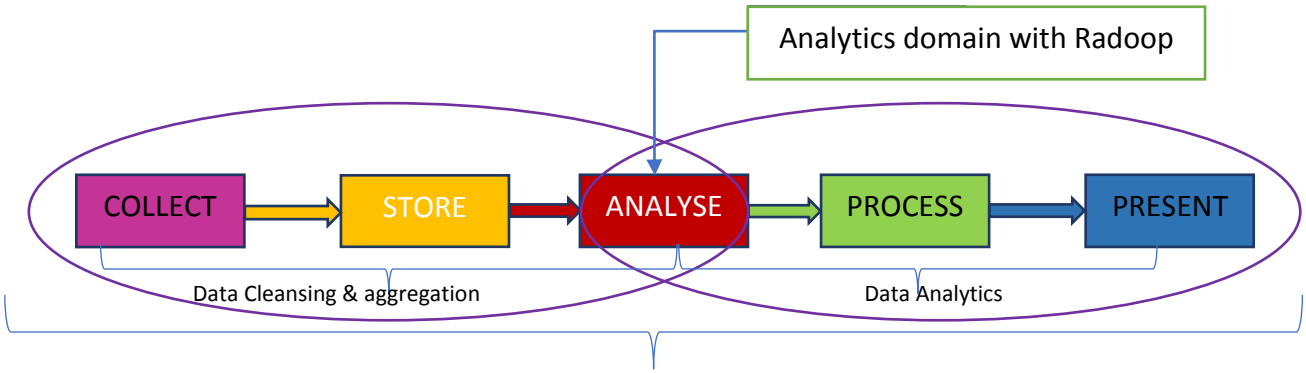
Rapid Miner Radoop is a code-free environment for designing advanced analytic processes that push computations down to your Hadoop cluster.RapidMiner Radoop provides an easy-to-use graphical interface for analyzing data on a Hadoop cluster with a running Hive server. This introduction provides a quick description of the software and the capabilities of the solution for processing and analyzing big data. Radoop helps in envisioning our data in better way in terms of performing the predictive analytics of the big data and which would support us in terms of bringing the value to data and ensuring the expectations are met.

Features of Radoop

1. Easy to maintain and develop Visual Programming Environment
2. Automatic Execution of Analytic Workflows into Hadoop (run the process where the data is)
3. Purely functional operators for data access, data preparation and modelling. The technology becomes transparent.

4. Supports Kerberos authentication
5. Supports data access authorization employing Apache Sentry & Apache Ranger
6. Supports HDFS encryption to seamlessly integrate with data security policies
7. Supports Hadoop impersonation
8. Transparent data exchange between local memory and cluster
9. Push any RapidMiner operator or sub-process (including extensions) down to Hadoop and execute in a parallel way
10. Smart optimization of processes by grouping requests and reusing Spark containers as much as possible
11. Visualization of sampled Hadoop data within Studio

Integration of Data warehouse with Radoop

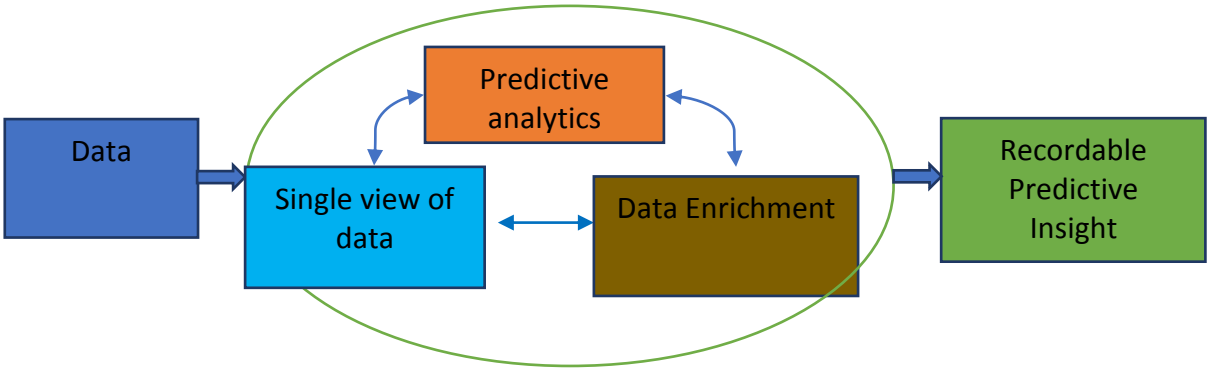


UNIFIED DATA MANAGEMENT

Opportunity of Radoop

- Avoid expensive data movement
- Leverage convenient data transformation
- For variety of data several data connectors (approx. 1000) are used.

Predictive Analytics Architecture



V CONCLUSION

The present paper elaborates the concept of Predictive Analysiswith Radoopand its functionalities. Our conclusion in the present paper is based on the fact thatpredictive analysis is a tool which is made for handling Big data and meets with the expectation of ever growing demands of the data. Any general integration on this work has not yet been known but some work has been done on integrating Hadoop functions to RapidMiner.

REFERENCES

[1]. MeghnaUtmal - Taxonomy on the integration of Hadoop and Rapid Miner for Big Data Analytics, IEEE Journal

[2]. A. Rishika Reddy - Predictive Big Data Analytics in Healthcare, IEEE Journal

[3]. JosepLluísBerral; Nicolas Poggi; David Carrera; Aaron Call; Rob Reinauer; Daron GreenALOJA: A Framework for Benchmarking and Predictive Analytics in Hadoop Deployments

[4]Tapan Chowdhury; Susanta Chakraborty; S. K. SetuaKnowledge extraction from big data using MapReduce-based Parallel-Reduct algorithm

[5]. Witten, I.H., Frank, E.: "Data Mining: Practical machine Learning tools and techniques", 2nd edition,Morgan Kaufmann, San Francisco(2005).

[6]. Daniel T.Larose, Chantal D.Larose "Data Mining and Predictive Analytics", 2edition, MISL-WILEY

[7]. Dean Abbott: "Applied Predictive Analytics: Principles and Techniques for The Professional Data Analyst" MISL-WILEY

[8]. Alcala-Fdez et al, "KEEL: A software tool to Assess Evolutionary Algorithms to Data mining Problems", Soft computing 13:3,pp 307- 318(2009).