ISSN:2320-0790

An International Journal of Advanced Computer Technology

ANALYTICS OF BIG DATA

Asst. Prof. Shubhada Talegaon
Parul Institute of Engineering & Technology, Dist. Vadodara, State : Gujarat

Abstract: Big Data analytics has started to impact all types of organizations, as it carries the potential power to extract embedded knowledge from big amounts of data and react according to it in real time. The current technology enables us to efficiently store and query large datasets, the focus is now on techniques that make use of the complete data set, instead of sampling. This has tremendous implications in areas like machine learning, pattern recognition and classification, sentiment analysis, social networking analysis to name a few. Therefore, there are a number of requirements for moving beyond standard data mining technique. Purpose of this paper is to understand various techniques to analysis data.

Keywords: machine learning, pattern recognition, sentiment analysis, social networking analysis

Introduction

Big data is now a reality: The volume, variety and velocity of data coming into your organization continue to reach unprecedented levels. This phenomenal growth means that not only must you understand big data in order to decipher the information that truly counts, but you also must understand the possibilities of what you can do with big data analytics.

What is big data analytics?

Big data analytics is the process of examining big data to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions. With big data analytics, data scientists and others can analyze huge volumes of data that conventional analytics and business intelligence solutions can't touch. Consider this; it's possible that any organization could accumulate (if it hasn't already) billions of rows of data with hundreds of millions of data combinations in multiple data stores and abundant formats. High-performance analytics is necessary to process that much data in order to figure out what's important and what isn't.

Various technique for analysis of Big Data

I Association rule

Introductory Overview:

The purpose of association rule is to detect relationships or associations between specific values of categorical variables in large data sets. This is a common task in many data mining projects as well as in the data mining subcategory text mining.[1] Association rule learning is a method for discovering interesting correlations between variables in large databases. These powerful exploratory techniques have a wide range of applications in many areas of business practice and also research - from the analysis of consumer preferences or human resource management, to the history of language. It was first used by major supermarket chains to discover interesting relations between products, using data from supermarket point-of-sale (POS) systems.

Association rule learning is being used to help:

- Place products in better proximity to each other in order to increase sales.
- Extract information about visitors to websites from web server logs.

- Analyze biological data to uncover new relationships.
- Monitor system logs to detect intruders and malicious activity.
- Identify if people who buy milk and butter are more likely to buy diapers.[2]

II Classification tree analysis

- Classification tree analysis is when the predicted outcome is the class to which the data belongs.[3]
- Classification Trees: Classification trees are used to predict membership of cases or objects in the classes of a categorical dependent variable from their measurements on one or more predictor variables.
- If the target variable is categorical, then a classification tree is generated. To predict the value (category) of the target variable using a classification tree, use the values of the predictor variables to move through the tree until you reach a terminal (leaf) node, then predict the category shown for that node.

Construct a *classification tree*. The decision process used by *classification tree* provides an efficient method can be applied to a wide variety of classification problems.

For example with the help of classification tree decision regarding types of coin will be easy. To devise a system for sorting a collection of coins into different classes (perhaps pennies, nickels, dimes, and quarters) create class according to diameter. Diameter of each coin can be used to devise a *hierarchical* system for sorting coins. Roll the coins on edge down a narrow track in which a slot the diameter of a dime is cut. If the coin falls through the slot it is classified as a dime, otherwise it continues down the track to where a slot the diameter of a penny is cut. If the coin falls through the slot it is classified as a penny, otherwise it continues down the track to where a slot the diameter of a nickel is cut, and so on.

III. Genetic algorithms

A genetic algorithm (GA) is a search technique used in computing to find exact or approximate solutions to optimization and search problems.[5] Genetic algorithms are categorized as global search heuristics. Genetic algorithms are a particular class of evolutionary algorithms (EA) that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover

These mechanisms are used to "evolve" useful solutions to problems that require optimization.
Genetic algorithm is model of evolution of population of artificial individuals emulating Darwinian selection. The driving force behind new batter solution is retention and combination of good partial solution of a problem.

Genetic algorithms are being used to:

- Schedule doctors for hospital emergency rooms
- Return combinations of the optimal materials and engineering practices required to develop fuel-efficient cars
- Generate "artificially creative" content such as puns and jokes
- Which TV programs should we broadcast, and in what time slot, to maximize our ratings?

IV. Machine learning

In recent years, novel application domains have triggered fundamental research on more complicated problems where multi-target predictions are required. Such problems arise in diverse application domains, such as document categorization, recommender systems, tag prediction of images, videos and music, information retrieval, natural language processing, drug discovery, biology, etc. Specific multi-target prediction problems have been studied in a variety of subfields of machine learning and statistics, such as multi-label classification, multivariate regression, sequence learning, structured output prediction, preference learning, multi-task learning, recommender systems and collective learning. [6] Despite their commonalities, work on solving problems in the above domains has typically been performed in isolation, without much interaction between the different sub-communities.

Machine learning is the modern science of finding patterns and making predictions from data based on work in multivariate statistics, data mining, pattern recognition, and advanced predictive analytics. It includes software that can learn from data. It gives computers the ability to learn without being explicitly programmed, and is focused on making predictions based on known properties learned from sets of "training data." [7]

Machine learning methods are vastly superior in analyzing potential customer churn across data from multiple sources such as transactional, social media, and CRM sources. High performance machine learning can analyze all of a Big Data set rather than a sample of it

Machine learning is being used to help:

- Distinguish between spam and non-spam email messages
- Learn user preferences and make recommendations based on this information
- Determine the best content for engaging prospective customers
- Determine the probability of winning a case, and setting legal billing rates
- Which movies from our catalogue would this customer most likely want to watch next, based on their viewing history?

V. Regression Analysis

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable,

Regression analysis is used to determine different levels of customer satisfactions and how they affect customer loyalty and how service levels can be affected by

At a basic level, regression analysis involves manipulating some independent variable (i.e. background music) to see how it influences a dependent variable (i.e. time spent in store). It describes how the value of a dependent variable changes when the independent variable is varied. It works best with continuous quantitative data like weight, speed or age.

Regression analysis is being used to determine how:

- Levels of customer satisfaction affect customer loyalty [8]
- The number of supports calls received may be influenced by the weather forecast given the previous day
- Neighborhood and size affect the listing price of houses
- *How does your age affect the kind of car you buy?*

VI Sentiment Analysis

Sentiment analysis or opinion mining; computational study of opinion (sentiments, emotion) expressed in text. Sentiment analysis determine if sentence or document express positive negative or neutral sentiments towards some objects.[9]

Sentiment analysis is being used to help:

- Improve service at a hotel chain by analyzing guest comments.
- Customize incentives and services to address what customers are really asking for.
- Determine what consumers really think based on opinions from social media.
- How well our new is is return policy being received? [10]

VII Social Network Analysis

Social network analysis [SNA] is the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities. The nodes in the network are the people and groups while the links show relationships or flows between the nodes. SNA provides both a visual and a mathematical analysis of human relationships. Management consultants use this methodology with their business clients and call it Organizational Network Analysis

The variety of data refers to the multitude of sensors and data sources collecting natural language and text, images and video, geo-spatial data, and time series. All of these contain potential relationships from person to person, from person to object, and from object to object. Moreover, relationships like "friendship" overlap with other interactions like "communication" so that one type of relationship may predict the other under some conditions.

Social network analysis is a technique that was first used in the telecommunications industry, and then quickly adopted by sociologists to study interpersonal relationships. It is now being applied to analyze the relationships between people in many fields and commercial activities. Nodes represent

individuals within a network, while ties represent the relationships between the individuals.[11]

Social network analysis is being used to:
- See how people from different populations form ties with outsiders
- Find the importance or influence of a particular individual within a group
- Find the minimum number of direct ties required to connect two individuals
- Understand the social structure of a customer base

Conclusion

When any business wants to discover interesting correlations, categorize people into groups, optimally schedule resources, or set billing rates, a basic understanding of the seven techniques mentioned above can help Big Data work. In the age of big data when companies are awaiting a breakthrough vehicle to bring modern data analysis and prediction will really help if it is develop on cloud.

References

[1] http://www.statsoft.com/Textbook/Association-Rules /
[2] http://www.firmex.com/blog/7-big-data-techniques-that-create-business-value
[3] http://en.wikipedia.org/wiki/Decision_tree_learning
[4].http://www.obgyn.cam.ac.uk/cam-only/statsbook/stclatre.html
[5] http://mydatamine.com/when-to-use-genetic-algorithm-for-data-mining-task/
[6] http://www.ngdata.com/machine-learning-and-big-data-analytics-the-perfect-marriage/
[7] http://www.wired.com/2014/03/use-data-tell-future-understanding-machine-learning/
[8] http://www.bigdata-startups.com/data-mining-techniques-create-business-value/
[9] http://blogs.sas.com/content/anz/2012/04/25/the-value-of-high-performance-analytics/
[10] http://www.cogno-sys.com/big-data/big-data-sentiment-analysis/
[11] http://www.insna.org/what_is_sna.html
[12]http://www.gurufocus.com/news/279031/ibm-takes-a-giant-leap-to-promote-big-data-analysis