CHALLENGES WITH UNSTRUCTURED BIG DATA ANALYSIS USING MACHINE LEARNING APPROACH: A REVIEW

Devendra Kumar Mishra
Assistant Professor, Dept. of CSE, ASET, Amity University (M.P)
Email:dkmishra@gwa.amity.edu

Abstract

Big Data contains large-volume, complex, growing data sets with multiple, sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains.The unstructured data files have text and multimedia content, E-mail messages, videos,photos,audio files,presentations,web pages and various other types of business documents.This data is growing at an unprecedented pace Unstructured data is so valuable, however, organizations are trying to find out efficient ways to extract meaning of it that can translate it into connections and patterns. With the variety, speed and volume of data flowing through agencies' databases, it has become more and more difficult to find patterns that lead to meaningful conclusions. At the same time, agencies need to find ways to make sense of all of this data. While rules-based analytics tools like business intelligence where conclusions are generated based on preset business rules , these systems are limited by their very rules. The rules are set by humans, so they are limited by the way they are designed and directed. They also aren't set up to deal with the volume and variety of data available to agencies today. Gather the most value from large, disparate sets of information, both structured and unstructured, requires a newer approach. In this paper I am going to analyze the complexity associate with unstructured data and also describing machine learning approach that allows a system to analyze variables simultaneously, along with how they interconnect, to form patterns. It is well-suited to complex problems involving multiple variables, and does extremely well with large volumes of unstructured data including images, text, audio, sensor data and more. this approach help organizations not only for discover patterns, but also make more accurate predictions over time as it incorporates more data points.

Keywords: Big Data, Structure Data, Unstructured Data, Machine Learning

I. INTRODUCTION:

Big data is a set of datasets which are so large and complex. Data sets are increasing day by day and transfer, sharing, storage ,capturing of those data are the main challenges in Big Data. Data mining is a process of discovering patterns from a large data set. Big data contains the information that comes from various, autonomous and heterogeneous sources and have growing relationships . There are different types of data like Relational Data, Semi-structured Data, Text Data, Streaming Data, Graph Data, etc. present in Big data. Data from various websites are growing with every second ,so the data become large day by day. Social networking sites generate large amount of data. These data can be based on an event ,on a particular topic, or usual contents.Twitter users convey their opinions in the form of tweets. In a day, more than 600 million tweets are produced. The number of active members on Facebook on a day is around 800 million. The comments or posts produced by the number of users will be more than this. Apart from this there are a number of other sites available those are producing large amounts of data [1].Online shopping sites can improve the brand , color , type and delivery locations of products from opinions. There is a vast amount of data available in Big data. Analyzing this amount of data and extracting information from it that may be useful for the organization for Market Analysis. The process of extracting the

information or knowledge of the huge set of data is known as Data Mining. There are mechanisms to filter unwanted messages from the online social networking wall [2]. The ability to analyze ,manage, visualize, summarize in scalable manner is a difficult task. Storing this amount of data without using it is simply a waste of storage space and time. Data should be processed to extract some useful information or knowledge from it .

II. UNSTRUCTURED BIG DATA

unstructured data is the opposite of structured data. Data typically existing in relational database is called structured data. It can be smoothly mapped into pre-designed fields. unstructured data is not relational and doesn't fit into these kinds of data models defined in prior. Frequently the unstructured data files comprise multimedia and text content. E-mail messages, videos documents forward processing, web pages, audio files, photos, presentations, and various other types of business documents can be considered as unstructured data.

For any organization 80 to 90 percent of the data comes under the category of unstructured data and the figure is continuously increasing. It is a common consideration among many establishments that their unstructured data stores include information that is likely to help them in making better business decisions.

When analyze social networking sites like Facebook, data are of different forms like images ,text, relations, etc. One person will be associated to more than one person. These relationships can be represented as graphs. From these varied sources, the discovery and extraction of useful information will be difficult. Twitter also contains such information. Blogs and News sites are content based. So that large amount of contents will be there to store. It is not possible to store such huge information on a single PC. It leads to enlarged storage and cost. Table 1 show the characteristics of unstructured Big data

Table-1: UNSTRUCTURE DATA CHARACTERISTICS

	UNSTRUCTURE BIG DATA
Complexity	High
Data Format	Different data Format
Mining	Difficult
Size	Large (growing)
Algorithms	Difficult to apply
Aggregation of data	Difficult

III. FRAMEWORK FOR UNSTRUCTURE DATA MINING

In Unstructure data we generally find text data. mining of text can be viewed as consisting of two phases: Text refining that transforms free-form text documents into a chosen intermediate form, and knowledge distillation that deduces patterns or knowledge from the intermediate form.

Intermediate form can be semi-structured such as the conceptual graph representation, or structured such as the relational data representation.

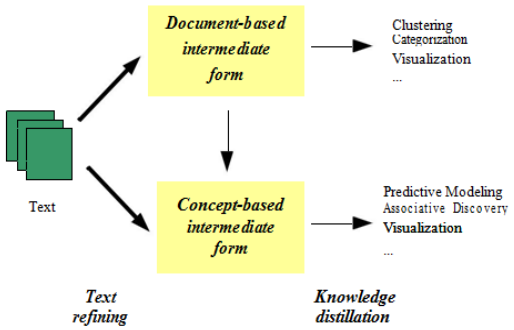

Figure 1

Figure 1 showing a mining framework for Text In this Text refining converts unstructured text documents into an intermediate form . that can be document-based or concept-based. Knowledge distillation from a document-based intermediate form deduces patterns or knowledge across documents.[3]

IV. MACHINE LEARNING APPROACH

There are several area where we can apply Machine Learning approach, the most important of which is data mining. People are often making mistakes during analyses or, possibly, when trying to establish relationships between multiple features. This makes it difficult for them to find solutions to certain problems. Machine learning can often be successfully applied to these problems, improving the efficiency of systems and the designs of machines. The main focus will be on the two most commonly used ones — supervised and unsupervised learning Supervised learning is tasked with learning a function from labeled training data in order to predict the value of any valid input. Common

examples of supervised learning include recognizing handwriting,classifying email messages as spam, and labeling Web pages according to their type. Many algorithms are used to create supervised learners, the most common being neural networks, Support Vector Machines (SVMs), and Naive Bayes classifiers. Unsupervised learning, 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 in order 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[4][5].

In this paper, I have concentrated on the supervised machine learning approaches only. Table 2 compare different supervised learning approach on parameter that are challenging to Big data.

Table:-2 Comparison between supervised learning approach

Parameter	Decision Trees	Neural Networks	Naïve Bayes	kNN	SVM
Accuracy in general	2	3	1	2	4
Speed of classification	4	4	4	1	4
Dealing with discrete/binary/ continuous attributes	4	3	3	3	2
Dealing with danger of overfitting interdependent attributes	2	1	3	3	2
Explanation ability/transparency of knowledge/classifications	4	1	4	2	1
Model parameter handling	3	1	4	3	1
Tolerance to redundant attributes	2	2	1	2	3
Attempts for incremental learning	2	3	4	4	2
Tolerance to noise	2	2	3	1	2
Speed of learning	3	1	4	4	1

Where 4 shows best and 1 shows poor

V.CONCLUSION

This paper present diffenent supervised machine learning approach with the parameter that are challenges in case of Big data. if data are unstructured then complexity increase more and more. The key question arise when dealing with Machine Learning approach that which learning algorithm is superior to others, but under which conditions a particular method can significantly outperform others on a given application problem. After analyze the strengths and limitations of each method, we can integrated two or more algorithms together to solve a problem . The objective is to utilize the strengthes of one method to complement the weaknesses of another.

REFERENCES

[1] Chau, Michael, and Hsinchun Chen. "A machine learning approach to web page filtering using content and structure analysis." *Decision Support Systems* 44, no. 2 (2008): 482-494

[2] Vanetti, Marco, Elisabetta Binaghi, Elena Ferrari, Barbara Carminati, and Moreno Carullo. "A System to Filter Unwanted Messages from OSN User Walls." *Knowledge and Data Engineering, IEEE Transactions on* 25, no. 2 (2013): 285-297.

[3] Hearst, M. A. (1997) Text data mining: issues, techniques, and the relationship to nformation access. Presentation notes for UW/MS workshop on data mining, July 1997.

[4] S. B. Kotsiantis" Supervised Machine Learning: A Review of Classification Techniques" Informatica **31** (2007) 249-268 .

[5] Jiang Zheng, Aldo Dagnino "An Initial Study of Predictive Machine Learning Analytics on Large Volumes of Historical Data for Power System Applications " IEEE International Conference on Big Data 2014