BIG DATA PREDICTION USING EVOLUTIONARY TECHNIQUES: A SURVEY

[1]S.Banumathi, [2]A.Aloysius
[1]Assistant professor, [2]Assistant professor
[1, 2] Department of Computer Science,
[1, 2] St Joseph's College, Trichy-2.

Abstract- Big Data is an emerging concept that describes innovative techniques and technologies to analyze large volume of complex datasets that are exponentially generated from various sources and with various rates. Big Data analytics, since dealing with Big Data are big challenges for the applications. Big Data analytics is the ability of extracting useful information from such huge datasets. Predictive analytics is one of types of big data analytics which is learns from experience and predict the future behavior or patterns. This paper presents a review on prediction based evolutionary algorithms which identifies the relevant accurate algorithm based on problem.

Index Terms— Big Data, Big Data Analytics, Big Data prediction, Machine learning algorithms.

I. INTRODUCTION

In the modern world we are overwhelmed with data, with companies such as Google and Facebook dealing with petabytes of data .Google processes more than 24 petabytes of data per day, while Facebook, a company founded a decade ago, gets more than 10 million photos per hour.

The presence of "big data", or this massive amount of increasing data, offers both an opportunity as well as a challenge to researchers.

The term "big data" was created to define the collection of large amounts of data in structured, semi-structured, or unstructured formats in large databases, file systems, or other types of repositories, and the processing of this data in order to produce an analysis and synthesis of the trends and actions in real or almost real-time[3]. Out of the above amounts of data, the unstructured data needs more real-time analysis and bears more valuable information to be discovered, providing a more in-depth understanding of the researched subject. It is also the unstructured data which incurs more challenges in collecting, storing, organizing, classifying, analyzing, as well as managing [1].

A lot of evolution has been made in developing the capability to process, store, and analyze big data. In addition to the big data computing capability, the rapid advances in using intelligent data analytics techniques drawn from the emerging areas of artificial intelligence (AI) and machine learning (ML) provide the ability to process massive amounts of diverse unstructured data that is now being generated daily to extract valuable actionable knowledge. The proper knowledge extraction from the variety of resources needs mining, machine learning and natural languages processing techniques [2]. There are four types of analytics namely prescriptive, predictive, diagnostic, and descriptive. According to Gartner, most of the organization had used predictive compares to other types.

II. PREDICTIVE ANALYTICS

Predictive analytics refers to a technology that learns from experience to predict the future behavior of individuals in order to drive better decisions. The need to devise a new tool for predictive analytics for both types increases [15]. Predictive analytics encompasses data science, machine learning, predictive and statistical modeling and outputs empirical predictions based on given input empirical data. The underlying premise is that future can be predicted on the basis of the past experience. Predictive analytics finds its application in various humanitarian development fields ranging from healthcare to education. The fig.1 depicts the types in prediction based evolutionary algorithms.

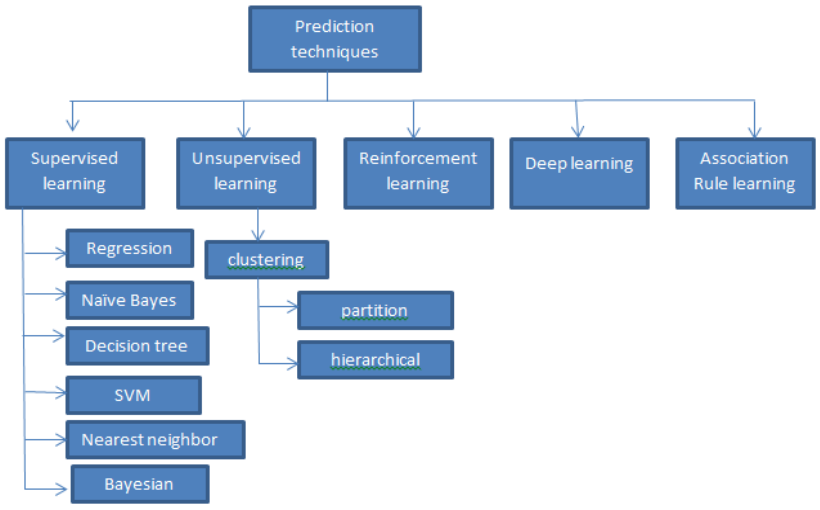

Fig: 2 predictive analytics techniques

III. MACHINE LEARNING

Machine learning (ML), a sub-field of artificial intelligence (AI), focuses on the task of enabling computational systems to learn from data about how to perform a desired task automatically.

The goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use. The machine learning task involves with statistical and probabilistic methods [5]. It process training data and produce a predictive model. Once the predictive model is created, the models provide the outcomes. The data adaptive machine learning methods can be identified throughout the science world [4].

Machine learning has many applications including decision making, predicting and it is a key enabling technology in the deployment of data mining and big data techniques in the diverse fields of healthcare, science, engineering, business and finance. The tasks can be categorized into the following major types:

IV. SUPERVISED LEARNING

In this class of ML, the learning task is to generalize from a training set, which is labeled by a "supervisor" to contain information about the class of an example, so that predictions can be made about new, yet unseen, examples. Supervised algorithm used for Classification and Regression (binary and multi-class problem) anomalie detection (one class problem). If the output (or prediction) belongs to a continuous set of values then such a problem is called regression, while if the output assumes discrete values then the problem is called classification. The followings are some of classification techniques.

a. Regression Analysis

Regression analyses mainly focus on finding relationship between a dependent variable and one or more independent variable. Predict the value of dependent variable based on one or more independent variable. The regression model basically divided uni variate and multivariate and that is further divides into linear and nonlinear. [9]

b. Naive Bayes Classifiers

Naïve Bayes Classifiers are based on Baye's Theorem that assumes independence among features given a class. All the attributes are analysed individually giving all of them equal importance [14]. The very surprised feature of Naive bayes is extremely fast to run large sparsed data set [6].These has been widely used for the Internet traffic classification: e.g., naive Bayesian classification of the Internet traffic.

c. Decision Trees (DT)

Decision Trees define as popularly used intuitive method that can be used for learning and predicting about target features both for quantitative target attributes as well as nominal target attributes. It is directed tree with root node which has no incoming edges, and all other nodes with exactly one incoming edges, known as decision nodes. In the training stage each internal node split the instance space into two or more parts with optimizing the performance. After that every path from the root node to the leaf node form a decision rule. The main advantage of DT is their intuitive interpretation which is crucial even network operators have to analyze and interpret the classification method and results [7].

d. Support Vector Machines

Support Vector Machines (SVM) is a widely used supervised learning technique that is remarkable for being practical and theoretically sound, simultaneously. The approach of SVM is rooted in the field of statistical learning theory, and is systematic: e.g., training a SVM has a unique solution (since it involves optimization of a concave function).

A support vector machine is a Classification method. An SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

The support vector machine has been developed as robust tool for classification and regression in noisy, complex domains. The two key features of support vector machines are generalization theory, which leads to a principled way to choose hypothesis and, kernel functions, which introduce non-linearity in the hypothesis space without explicitly requiring a non-linear algorithm [8].

V. UNSUPERVISED LEARNING

The basic method in unsupervised learning is clustering. In clustering, the learning task is to categorize, without requiring a labeled training set, examples into 'clusters' on the basis of perceived similarity. This clustering is used to find the groups of inputs which have similarity in their characteristics. Intuitively, clustering is akin to unsupervised classification while classification in supervised learning assumed the availability of a correctly labeled training set, the unsupervised task of clustering seeks to identify the structure of input data directly.

Recommendation services to meet the requirements of users, and time technology for analysis of clustering processing is growing in importance, along with the big data analysis technologies [11].

VI. REINFORCEMENT LEARNING

Reinforcement learning differs from the supervised learning methods commonly used to fine-time deep networks. Reinforcement learning problems require learning from evaluations of a learning agent's behavior, or reinforcements, rather than from correct, known outputs from a training set of data [12].

Simply, this is a reward/ punishment based ML technique. In this technique a learner, based on an input received, performs some action, potentially affecting the environment around it. This action is then rewarded or punished. The nature of the mapping from the actions taken by the learner to rewards/ punishments, in general, is probabilistic in nature. The eventual goal of a learner is to discover such an optimal mapping (or policy), from its actions to the rewards/ punishments, so that the average long-term reward is maximized.

VII. DEEP LEARNING

Deep learning algorithms use a huge amount of unsupervised data to automatically extract complex representation. These algorithms are largely motivated by the field of artificial intelligence, which has the general goal of emulating the human brain's ability to observe, analyze, learn, and make decisions, especially for extremely complex problems. Work pertaining to these complex challenges has been a key motivation behind Deep Learning algorithms which strive to emulate the hierarchical learning approach of the human brain. Models based on shallow

learning architectures such as decision trees, support vector machines, and case-based reasoning may fall short when attempting to extract useful information from complex structures and relationships in the input corpus. In contrast, Deep Learning architectures have the capability to generalize in non-local and global ways, generating learning patterns and relationships beyond immediate neighbors in the data [10].

Deep learning is in fact an important step toward artificial intelligence. It not only provides complex representations of data which are suitable for AI tasks but also makes the machines independent of human knowledge which is the ultimate goal of AI and provide superior performance. It extracts representations directly from unsupervised data without human interference.

Deep learning (DL) is an ML technique that comprises deep and complex architectures. These architectures consist of multiple processing layers, each capable of generating non-linear response corresponding to the data input. These layers consist of various small processers running in parallel to process the data provided. These processors are called neurons. DL has proved to be efficient in pattern recognition, image and natural language processing. DL finds its applications in very broad spectrum of applications ranging from healthcare to the fashion industry, with many key technology giants like Google, IBM and Facebook deploying DL techniques to create intelligent products.

VIII. ASSOCIATION RULE LEARNING

It is a method for discovering interesting relations between variables in large databases. In this, we seek to learn about associations between the features present in examples. Unlike classification (supervised learning), which strictly and discretely tells the class of an example, relations or associations among various variables in an example database are considered in association rule learning. We take an example case mentioned in where a weather dataset is considered. The usual classification problem would be to tell whether, based on the values of given weather features or attributes (like temperature, outlook and wind conditions) in the dataset, a game would be played or not. If, however, we consider association learning perspective then (instead of always telling about the status of the game) different rules among different features or variables can also be considered. As an example a rule can be established that if the outlook is sunny and the game is being played then the day is going to be non-windy. This type of learning technique can be particularly important for farmers in planning their activities for the best possible crop productions.

IX. CONCLUSION

The amount of data has been increasing and data set analyzing become more competitive. Predictive analytics is the combination of analytics techniques and decision optimizations. A descriptive survey and analysis was performed to provide an overview of evolutionary algorithms. There are various techniques used for prediction in order to improve the prediction accuracy. Most of the solutions provided by the Decision Tree and Bayesian network proves to better than other techniques. However, the techniques accuracy differed based on nature of data and where it from.

REFERENCES

[1] Bogdan Ionescu, Dan Ionescu, Cristian Gadea, Bogdan Solomon and Mircea Trifan ,"An Architecture and Methods for Big Data Analysis", springer 2014,vol356,pp 491-514.

[2] Neha khan, Mohd Shahid Husain, Mohd Rizwan Beg, "Big data classification using evolutionary techniques: a survey", IEEE conference on Engineering and Technology (ICETECH), 2015.

[3] Seref sagiroglu, Duygu sinanc, "big data: a review", IEEE,2013.

[4] M. I. Jordan , T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects", sciencemag.org, vol349, Issue 6245, 2015.

[5] G.Vaitheeswaran, L. Arockiam, "Machine Learning Based Approach to Enhance the Accuracy of Sentiment Analysis on Tweets", International Journal Of Advanced Research In Computer Science And Management Studies, volume4,issue 5,2016.

[6] Enric Junque, De Fortuny, David Martens, Foster Provost,"Modeling With BigData:Is Bigger Really Better?",Doi: 10.1089/Big.2013.0037 ,Mary Ann Liebert, Inc.,Vol1, No 4,December 2013.

[7] Wei Dai , Wei Ji, " A Map reduce implementation of c4.5 Decision Tree algorithm", International Journal of Data base theory and Applications, Vol.7, No.1, 2014.

[8] Mrs.P.Sheela Rani, S.Shalini, J.Rukmani@keerthika, A.Shanthini, "Energy efficient scheduling of map reduce for evolving big data applications", International journal of advanced research in computer and communication engineering, vol.5, issue.2, 2016.

[9] Ramya MG, Chetan Balaji, Girish L, " Environment change prediction to adapt climate smart agriculture using big data analytics", IJARCT.ORG, 2015.

[10] Maryam M Najafabadi,Flavio Villanustre,Taghi M Khoshgoftaar,Naeem Seliya,Rall Wald Email aut,"Deep learning applications and challenges in big data analytics", springer, 2015.

[11] Se-Hoon Jung, Jong-Chan Kim, Chun-Bo Sim," Prediction Data Processing Scheme using an Artificial Neural Network and Data Clustering for Big Data", IEEE Explorer, 2015.

[12] Charles W. Anderson∗, Minwoo Lee† and Daniel L. Elliott, "Faster Reinforcement Learning After Pretraining Deep Networks to Predict State Dynamics", IEEE Explorer, 2015.

[13] Yisheng Lv,Yanjie Duan, wenwen kang, Zhengxi Li, Fei-Yue Wang, "Traffic flow prediction with Big data: A Deep Learning approach", IEEE Trancscations on Intelligent transportation systems, vol.16, no.2, 2015.

[14] krithika verma, Pradeep kumar Singh, "An insight to softcomputing based prediction techniques in software", I.J.Modern education and computer science,2015.

[15] Amir Gandomi, Murtaza Haider, "Beyond the hype: big data concepts, methods, and analytics", International Journal of Information Management, 2015.