iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VI **Month of publication:** June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6243

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

Big Data in the Machine Learning Techniques Perspective - A Review

Olawale Adepoju[1], Devaraj Verma C.[2]
[1, 2]*Department of Computer Science and Engineering, Jain (Deemed to be University), India.*

Abstract: In this information age, the decision-makers have been able to access enormous quantities of data, and every day a massive stock of petabytes and zeta bytes of data is generated using today's digital services and modern technology such as the Web, social networking sites, internet of things and cloud computing.
Machine learning techniques for the interpretation and analysis of these data are applied quite effectively. There are however various machine-learning strategies for analysing data, all of which do not work on every data on the same basis. In this paper, we have presented a survey about some machine learning techniques and discussed their application along with the steps involved for it.
Keywords: Machine Learning, Data Mining, Big Data, Internet of things, Cloud Computing

I. INTRODUCTION

Imagine a world with no preservation of information where every individual perspective, every interaction, or angle retrievable about a person or organization is actually lost following use. This will lead to a lack of capacity for companies to collect valuable data and analyse them, as well as new prospects and priorities. This has proved to be fundamental to routine workouts, everything from details on consumers related to sales, to the workplace, and so forth. Data or information is a structure under which every company thrives. Currently, consider the level of details and the growth of information and data collected these days through advances and web technologies. When storage capacities and data collection techniques have increased, monumental knowledge initiatives have proved to be efficiently available. The solution needs to be analysed and given because of the rapid increase of these data, in order to process and derive value and information from such data. Policy-makers will need useful insights into other complex and fast-changing data from daily transactions to consumer experiences and social network data.
This value can be given by big data analysis, which uses state-of-the-art analytical techniques for big data. The purpose of this review is to breakdown a portion of the numerous algorithms for machine learning that can be related to big data, and the possibilities provided by the use of large-data analysis in different option areas. The machine learning system encompasses the pre-processing, learning, and evaluation methods.

II. LITERATURE REVIEW

A. Big Data

A vast variety of researchers around the world use the term big data, which reveals remarkably large (structured or unstructured) datasets. Such data are too big and moving rapidly and modern databases are not fit to manage such massive data. A large number of data sets are obtained from different fields such as sensors, transactional applications, networks, and social media, etc.
Big data has stimulated the future of information technology [1], focused mainly on the third platform, primarily on big data, cloud computing, internet, and social enterprises, from the software and information technology background. Such data are available in petabytes and beyond in structured, semi-structured, and unstructured formats. It is classified between 3V and 4V. 3V is about distance, pace, and variety. By understanding the numerous V's associated with the number, speed, variation, and veracity, the big data phenomenon can be clearly represented.
Three V's were described by Doug Laney for big data [2]. The length, pace, and variety come from this.

1) *Volume:* This explains the quantity of data per second accumulated from Exabyte to zeta bytes in various tools, such as transaction processes, community media, social media, etc. Since 2005 and 2017 more than 130 data from Exabyte have been generated and Internet activities have increased.
2) *Speed:* This function reflects the pace at which data are collected and analyzed to meet demands. The strong and continuous inflows of knowledge, data are collected in real-time and accurately, making it useful and precise for researchers to take decisions.

3) *Variety:* This word applies to the different data we may use. Wide range: The structure, structure, or semi-structure may be used. Data are typically stored as tablets or databases, although there have been problems with data in recent times. Some information is given by e-mails, records, photographs, video, tracking sensors, and much more representing unstructured data.
4) *Veracity:* This explains the accuracy of knowledge. In other words, the data refers to distortion, noise, anomalies, etc.

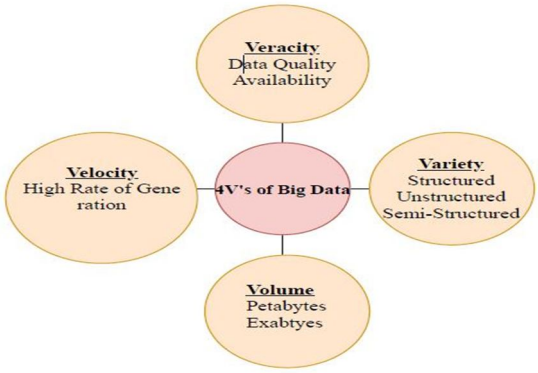


Fig. 1 The 4V's which characterize Big Data

Massive volume, speed, flexibility, and accuracy of information through various traditional and computational insightful policies are the primary target of large-scale analyses of information [3]. Gandomi and Haider investigated part of these extraction techniques for data acquisition [4].

B. Machine Learning

Machine learning is a collaborative area of study that brings together insights into several parts of science, including machine intelligence, statistical analysis, and IT. Machine learning is alluded to as the gaining from previous interaction with a few undertakings and implementation steps. The motivation behind the algorithms is to learn "without specifically programming," as Arthur Samuel described in 1959 [5], from the current knowledge. The machine learning approach enables users to uncover a hidden framework and create independent predictions of large data sets.

Such predictions are derived from previous estimates and from repeatable results. A few instances of the day by day life machine learning applications incorporate:

1) *Self-Driving Vehicles:* Such vehicles contain sensors that identify things from a vast region in every direction. The Car has tools to evaluate and process all data collected on the road for secure navigation [6, 7]. Such an instance is indeed the perfect illustration of machine learning implementations.
2) In Recommender systems also, machine learning has made waves. For example Amazon or Netflix Recommender Mechanisms. The recommendation system in Netflix depended on restricted Boltzmann machine and grid factoring [8].
3) Also in Fraud detection, the machine has been very useful, it is one of the most important aspects nowadays. All e-business face this challenge of fraud and network intrusion.

Machine learning problems are basically classified under three classes. They are:

a) Supervised learning involves labeled information to be trained. Supervised learning involves labeled information to be trained. For every information labeled includes data and ideal value for the target yield. The learning algorithm breaks down the trained information and provides a deducted capacity that can be used to map new qualities.
b) For example, clusters evaluation is unlabeled information in the unsupervised learning process. The result or consequence of the data is unknown. It is used for clustering (grouping) problems, for anomaly detection (in banks for irregular transactions), where links between the data are required.
c) The third class, reinforcement, enables an agent to take its result from the input to the ties to the external situation [9]. Reinforced learning is a kind of learning in which individual trains himself to benefit from its experience so that its reward can be amplified. The agent has no guidance of options, however on the off chance that the order is successfully implemented, it is compensated at that stage and else, it receives a negative reward. When the agent has a comparative situation, it must determine if the option depends on the experience of the past. The compensation is numerical validation and tags for the success of an operation.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

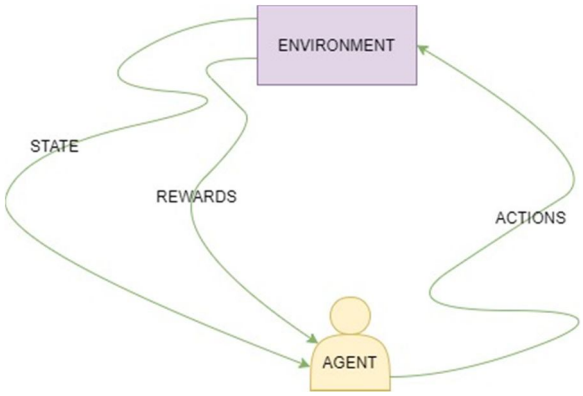


Fig. 2 Reinforcement Learning Curve

Through an information processing perspective, both the supervised and unsupervised training system is favoured for information analysis, and reinforcement strategies are favoured for improving specific choice-making [10].
The outcomes might change the estimation for the selected measure of training. Structuring a learning framework in machine learning includes four plan decisions:

- Preparing Data
- Choose target feature
- Assign representation
- Usage of the correct algorithm of learning

III.METHODOLOGY

While information continues to expand rapidly every day, various insightful learning approaches are being introduced in order to respond to some key data prediction analytical problems. This section demonstrates clearly a variety of machine learning approaches for the big data study.
Several learning techniques are there in machine learning for analysis of big data, some of these techniques' are but not limited to Logistic Regression, Support Vector Machine, Linear Regression, K-Means Clustering, Naïve Bayes, and Artificial Neural Networks.

A. Logistic Regression

The logistics regression is the sufficient statistical analysis for the dichotomous (binary) dependent variables. It is predictive analysis, as in all regression analyses, and also used to characterize data and illustrate the correlation between a binary variable dependent and one or more nominally independent variables, interval variables, or ratio rates. The relationship between both variables is not seen as a straight line by logistic regression. Logistic regression uses the natural logarithm function instead to determine the relationship between the variables and uses testing data to determine the coefficients.
The logistic regression is based on binary data, where the occurrence happens or the event does not happen. The function forecasts future outcomes. In light of other function x, it attempts to figure out whether or not an occurrence occurs. Then y can be either 0 or 1. If the occurrence occurs, y is assigned the value 1, then y can be either 0 or 1. If the occurrence occurs, y is assigned the value 1. In the logistic regression, a model corresponding to data points is found using the sigmoid function. The method gives the data model an 'S 'curve. The curve is between 0 and 1, and when y is binary, it is simple to apply.
There is a threshold of 0.5 for the curve, it is assumed that if the value is higher than 0.5 it is classified as 1 and else classified as 0.
Following is the sigmoid function:

$$y^i = \frac{1}{1+e^{-(z)}}$$

The Sigmoid function is used to find a binary classification probability. In this equation, y is the probability of the output, and z is th e log - odds of the example.

$z = b + m_1 x_1 + m_2 x_2 + m_3 x_3 + \ldots\ldots\ldots\ldots m_n x_n$

Thus b is the intercept of linear regression, m is the values and bias weighted, and x the values shown.
The sigmoid function predicts the likelihood of a certain outcome.

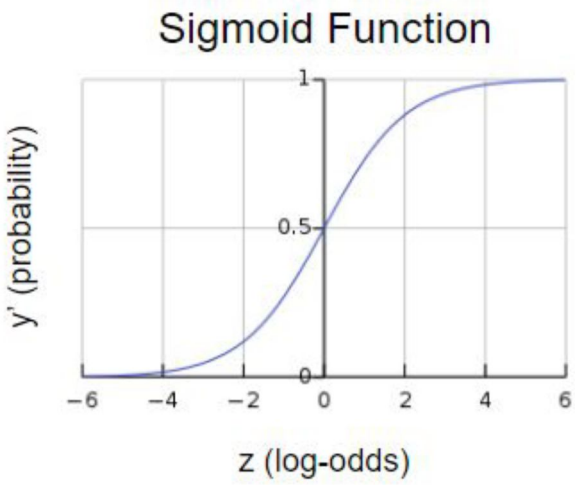


Fig. 3 Diagram of Logistic Regression Curve

B. Support Vector Machine

The essential idea driving Support Vector Machine (SVM) is a hyper-plane used in the decision. By plotting any acquired data value in an n-dimensional space or graph, a support vector algorithm is performed. The total number of data features present here is "n." The value of each data is displayed as a different graph coordinate. Svm is an example of supervised learning techniques that can be implemented in problems of classification and regression. Practically, the grouping procedure isn't straightforward, and, when it depends on the partition utilizing diverse lines it is known as hyper plane classifiers.

The optimum hyper plane is chosen according to the line-distance. Svm Vector help divides the class from one side to the other. The difference is called the margin and the margin is referred to as vectors of support.

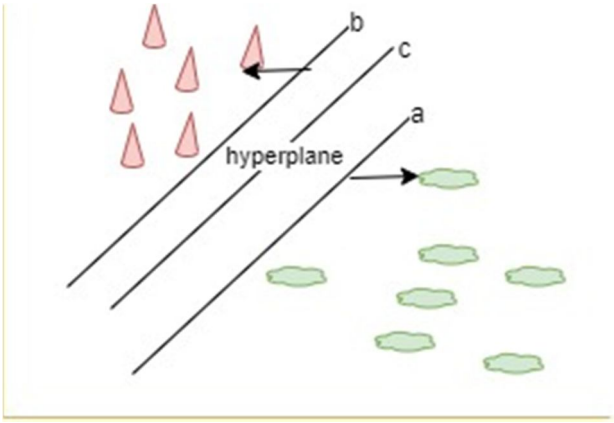


Fig. 4 Support Vector Machine Hyper-plane

C. Linear Regression

This algorithm defines a line that is best matched by determining different metrics for all the inputs variables known as a coefficient(s). This line is better adjusted between inputs (X), output (Y). The main purpose is to get a line that fits the data the best. That most suitable line is the one with a low total forecast error (every data point). Error is the range from the regression line points.

Y (predicted) = b0 + b1 * x

Y (predicted) must be forecasted in view of the data x and discover the values of the b0 and b1 coefficients as part of the linear regression. Another is a predictor and the other is a variable dependent. Yet it does not follow a probabilistic relationship, which is measurable. In the absence of a variable that can be precisely expressed, the relation between the two variables is called probabilistic. For example, a direct, variable-based mathematical solution for conventional least squares and optimization of gradient descent can be used to take linear regression from data. For example, a direct, variable-based mathematical solution for conventional least squares and optimization of gradient descent can be used to take linear regression from data.

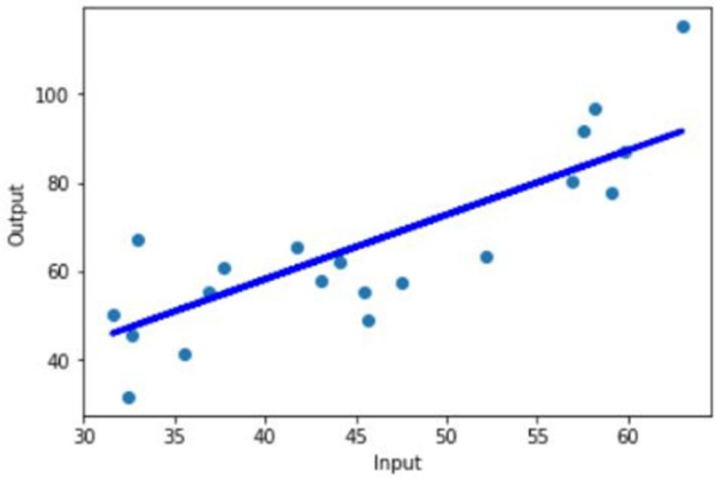


Fig. 5 A Regressor Line on Data

D. Artificial Neural Network

A computational model is a network of artificial neurons (ANN), based on biological neural network architectures and features. Data that passes via the network impacts the ANN structure as the neural network, based on the input and output, improves learning. Artificial Networks (ANN) are neuronal systems with a multi-layer relation.

These are made up of an input layer, several hidden layers, and an output layer. Each node in a layer is connected to each other node in the subsequent layer. By increasing the number of hidden layers, we are increasing the network.

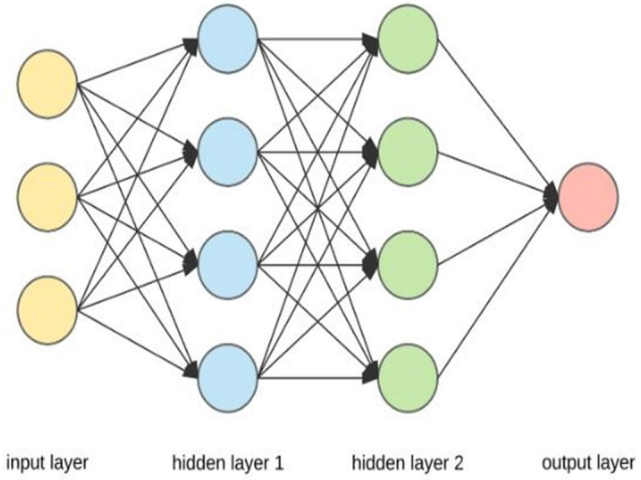


Fig. 6 A representation of Artificial Neural Network

IV. CONCLUSIONS

Machine Learning is vital in tackling the challenges facing a generation of data, data retrieval, and large data processing, which is mainly aimed at turning knowledge into a real impetus for improving businesses and consistent analysis. We talked about big data in this article and also displayed a few machine learning attributes. It has shown several applications of machine learning in our everyday life.

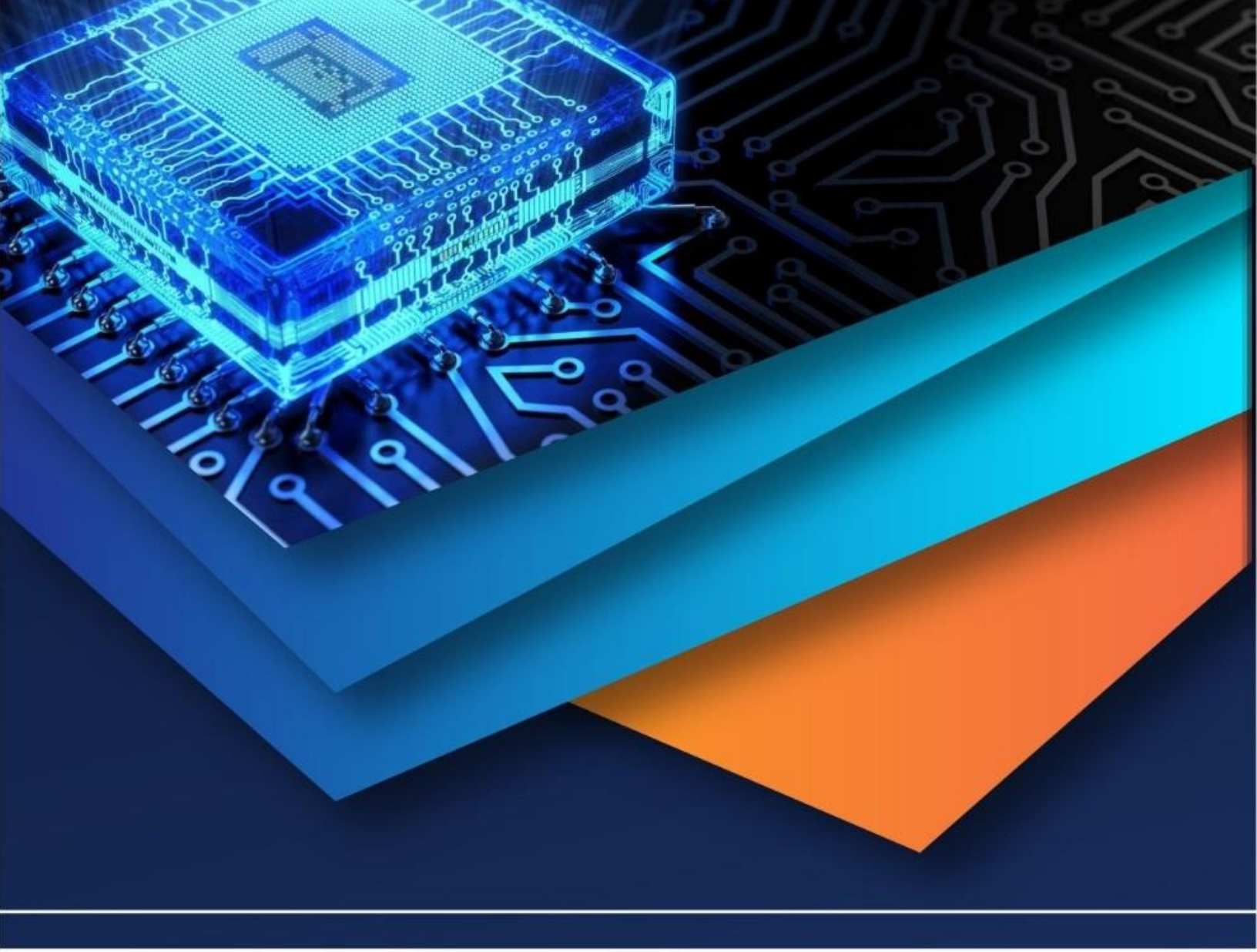
The potential reach of data analysis is to see how Machine Learning is made easier for non-specialists in multiple environments to communicate with an alternative type of knowledge.

V. ACKNOWLEDGMENT

The authors are grateful to Jain University, Department of Computer Science and Engineering, for their in-valuable support and feedback on this work.

REFERENCES

[1] X. Jin, B. W.Wah, X. Cheng and Y. Wang, Significance and challenges of big data research, Big Data Research, 2 (2) (2015), pp.59-64. https://doi.org/10.1016/J.BDR.2015.01.006.
[2] Doug Laney, 3D Data Management: Controlling Data Volume, Velocity, and Variety, META Group Research Note, 2001, 6: 70.
[3] M. K.Kakhani, S. Kakhani and S. R.Biradar, Research issues in big data analytics, International Journal of Application or Innovation in Engineering & Management, 2 (8) (2015), pp. 228-232.
[4] A. Gandomi and M. Haider. Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, 35 (2) (2015), pp. 137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007
[5] Big Data History and Current Considerations, http://www.sas.com/en_us/insights/big-data/what-is-big-data.html
[6] Machine Learning, https://en.wikipedia.org/wiki/Machine_learning
[7] Poczter, S. L., & Jankovic, L. M. (2013). The Google Car: Driving Toward a Better Future? Journal of Business Case Studies (JBCS), 10(1), 7-14. https://doi.org/10.19030/jbcs.v10i1.8324
[8] Xavier Amatrian. How does the Netflix movie recommendation algorithm work? 2014, https://www.quora.com/How-does-the-Netflix-movie-recommendation-algorithm-work (accessed 27 December 2014)
[9] C.M. Bishop, Pattern recognition and machine learning, Springer, New York, 2006.
[10] J. Qui, Q. Wu, G. Ding, Y. Xu and S. Feng, A survey of machine learning for big data processing. EURASIP Journal on Advances in Signal Processing, Springer, 2016, 67 (2016). https://doi.org/10.1186/s13634-016-0355-x.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY