OPTIMIZING BIG DATA PROCESSING THROUGH MACHINE LEARNING-BASED REDUNDANCY REDUCTION TECHNIQUES

Pagidimarri Krishna
Research Scholar
Shri JJT University
Rajasthan.

Dr. Prasadu Peddi
Guide
Shri JJT University
Rajasthan.

Dr. Manendra Sai Dasari
Co-Guide
Shri JJT University,
Rajasthan.

Abstract

As the volume and complexity of data continue to grow exponentially, optimizing big data processing has become a critical challenge for organizations seeking to derive actionable insights from vast datasets. Machine learning (ML) techniques are playing a pivotal role in enhancing the scalability, efficiency, and accuracy of big data analytics. This paper explores various ML approaches for optimizing big data processing, focusing on methods such as distributed machine learning, deep learning, and parallel processing. We discuss how these techniques can be applied to reduce the computational cost, improve data storage management, and accelerate decision- making in big data environments. Additionally, we examine the integration of advanced ML models with cloud-based and edge computing frameworks, which facilitate real-time data analysis and reduce latency. Through case studies and empirical results, we highlight the effectiveness of these approaches in a variety of industries, including finance, healthcare, and e-commerce. This paper also addresses key challenges, such as data privacy, model interpretability, and system scalability, and proposes future directions for research in optimizing big data processing through ML.

Keywords: *Big Data, Machine Learning, Data Processing, Scalability, Efficiency, Distributed Learning, Deep Learning, Parallel Processing, Cloud Computing, Real-Time Analytics*

INTRODUCTION

Machine learning is used when humans are unable to interpret or understand a humungous amount of data. It teaches computers how to operate with data efficiently. It is a tool that helps transform the data into analytical information and helps extract information from the data to predict future happenings. Every data-related problem is unique, and the same type of approach cannot be employed for each issue. Therefore, Machine learning employs a variety of algorithms for data analysis and prediction. It can read data in numerous ways. The type of method used is determined by what exactly you are searching for from the data, the number of variables involved, the best model to use, and other parameters to consider. Machine learning is a field of study that gives computers the ability to learn without being explicitly programmed, aiming to understand computational mechanisms by which experience can lead to improved performance. It is a highly interdisciplinary field building upon ideas from many different kinds of domains. In the past decades, machine learning has covered almost every domain of our life which is so pervasive that you probably use it dozens of times a day without knowing it. It is primarily influencing the broader world through its implementation in a wide range of applications, which has brought great impact on the science and society. A great number of machine learning algorithms have been proposed in

the last decades, such as neural network, decision tree, support vector machine, k-nearest-neighbor, genetic algorithms, Q-learning, etc. They have been used in diverse domains such as pattern recognition, robotics, natural language processing, and autonomous control systems.

LITERATURE REVIEW

Aws I. Abu Eid (2024) This article delves into the intricate relationship between big data, cloud computing, and artificial intelligence, shedding light on their fundamental attributes and interdependence. It explores the seamless amalgamation of AI methodologies within cloud computing and big data analytics, encompassing the development of a cloud computing framework built on the robust foundation of the Hadoop platform, enriched by AI learning algorithms. Additionally, it examines the creation of a predictive model empowered by tailored artificial intelligence techniques. Rigorous simulations are conducted to extract valuable insights, facilitating method evaluation and performance assessment, all within the dynamic Hadoop environment, thereby reaffirming the precision of the proposed approach.

Ashish Mishra (2023) The focus of this paper revolves around the examination of flow of ternary hybrid nanofluid, specifically the Al2O3–Cu-CNT/water mixture, with buoyancy effect, across three distinct geometries: a wedge, a flat plate, and a cone. The study takes into account the presence of quadratic thermal radiation and heat source/sink of non-uniform nature. To develop the model, the Cattaneo–Christov theory is utilized. The equations governing the flow are solved by applying similarity transformations and employing the "bvp4c function in MATLAB" for numerical analysis and solution. Conventional methods for conducting parametric studies often face challenges in producing significant conclusions owing to the inherent complex form of the model and the method involved.

Noura AlNuaimi (2022) Organizations in many domains generate a considerable amount of heterogeneous data every day. Such data can be processed to enhance these organizations' decisions in real time. However, storing and processing large and varied datasets (known as big data) is challenging to do in real time. In machine learning, streaming feature selection has always been considered a superior technique for selecting the relevant subset features from highly dimensional data and thus reducing learning complexity. In the relevant literature, streaming feature selection refers to the features that arrive consecutively over time; despite a lack of exact figure on the number of features, numbers of instances are well-established.

Sarker, I.H. (2021) In the current age of the Fourth Industrial Revolution, the digital world has a wealth of data, such as Internet of Things (IoT) data, cybersecurity data, mobile data, business data, social media data, health data, etc. To intelligently analyze these data and develop the corresponding smart and automated applications, the knowledge of artificial intelligence (AI), particularly, machine learning (ML) is the key. Various types of machine learning algorithms such as supervised, unsupervised, semi-supervised, and reinforcement learning exist in the area. Besides, the deep learning, which is part of a broader family of machine learning methods, can

intelligently analyze the data on a large scale. In this paper, we present a comprehensive view on these machine learning algorithms that can be applied to enhance the intelligence and the capabilities of an application.

Bin Hou (2013), With the rapid development of wireless communication industry, shortage situation of spectrum resource is increasingly significant. It has become an important topic to study cognitive radio spectrum allocation algorithm that is of higher spectrum utilization ratio, less system power consumption and better algorithm efficiency. Analyzes spectrum allocation models based on genetic algorithm, and then puts forward new improved genetic algorithm. The algorithm adopts niche crowding operation to avoid individual inbreeding. It adaptively adjusts crossover and mutation probability to keep them always in the appropriate state. It provides more equal individual competition opportunity by hierarchical measures, which can effectively avert premature convergence to local optimal solution.

Support Vector Machine (SVM Algorithm)

SVM is also most widely used algorithm content, state-of the-art machine learning technique is relying to Support on the Vector Machine to complete the data analysis work. It is mainly used for classification. SVM works on the principle of margin calculation, in order to improve final data analysis result in the actual analysis process multiple set analysis samples of data get collected to the determine the sample data of the boundary value. It basically, draws margins between the classes. The margins are drawn in such a fashion that the distance between the margin and the classes is maximum and hence, minimizing the classification error

Boosting Algorithm

Boosting is a technique in ensemble learning which is used to decrease bias and variance it is new type of machine algorithm content. The main advantages of this algorithm consist of accurate processing of data information and improve the accuracy of the final processing result. Boosting creates a collection of weak learners and converts them to one strong learner. In this algorithm function prediction optimized therefore speeding up the processing of data information and at the same time AdaBoost is also an important guarantee for the expansion of the boosting algorithm.

Big Data Application Process

The big data application process usually includes four links: big data acquisition, big data processing, big data analysis, and big data presentation. It is worth noting that, based on the continuity and spiral of organizational activities and decision-making, for the operation of enterprises, big data application is more often not a one-way process, but a cyclical process of acquisition-processing-analysis-presentation. Scholars mainly focused on two aspects of researches: i) one was to design a kind of distributed parallel computing framework or platform for fast dealing with big data, such as MapReduce, Dryad, Graph lab, Hadoop, Haloop, and Twister, etc. ii) the other was to propose a sort of new algorithms to solve a class of determined big data problems. the authors developed a low-complexity subspace learning to handle the incomplete streaming big data.

Big Data Analysis

Big data analysis is the core part of big data application. The most important thing is to look at the essence through the phenomenon by big data analysis, that is, to discover and even predict changes in consumer demand and preference, grasp customer needs in a timely manner, and improve customer satisfaction; to find out the problems in current corporate activities, improve work processes, optimize resource allocation, etc. The most common directions of big data analysis include data mining and predictive analysis.

Big Data Presentation

The results of big data analysis usually need to be presented directly and concisely in a certain form rather than in complex and difficult to understand tables or text descriptions. Big data presentation is also called big data visualization. Its tools are generally divided into three categories: interactive visualization tools, configuration visualization tools, and programming visualization tools. The presentation of big data is very important to reflect the value of big data application in itself, but only a means. For example, when a beverage company analyzes and predicts consumer preferences for beverage categories, the core conclusions of the analysis may be displayed using only a pie chart. Of course, with the diversification of business and popularization of data. visualization smart tools, the presentation of big data is more often in the form of smart dashboards, which requires enterprise employees to have professional skills.

Insufficient understanding and lack of big data thinking

There seems to be a natural conflict between the "big" data and the "small" and medium-sized enterprises. Therefore, most of the SMEs, from managers to employees, have insufficient understanding of big data and believe that only big companies have the big data, need the big data, and can apply the big data. In fact, with the development of advanced technologies, big data and its applications have penetrated into all walks of life and integrated into the development of industries and enterprises. The lack of big data thinking and understanding has caused some SMEs hardly to learn, digest, absorb and innovate quickly in the rapidly changing market, which lead to a difficulty to accurately respond and adjust to the market regardless of their products and services, or production, sales, promotion and other operational activities, or even the overall planning and business model design of organizations, thus missing some development opportunities.

RESEARCH METHODOLOGY

The chapter presents a comparative study of different classification ML algorithms to observe their performances on highly imbalanced complex dataset. Several challenges of ML algorithms occurs during mining imbalanced dataset has been discussed. Numerous popular techniques to reduce the complexity of an imbalanced dataset were reviewed along with their limitations to solve the problem. Some of them like US and OS is applied to sample imbalanced dataset and reduce its complexity. Number of experiments along with their results is produced to check the performance of US and OS to reduce dataset complexity along with it change in the behavior of ML algorithms has also been observed. The term big data, which was first coined in 1990 deals with the study of large and complex datasets. Data

storage, data capturing, data analysis, data querying, data visualization and data transfer are some of the challenges of big data. It can do wonder only if most important information can be extracted through it. In order to dig valuable information from the enormous peak of Big data use of predictive analytics, user behavior analytics and other big data analytics are in trend. Big data can prevent disease, detect crime, and help in business, financial services, etc. by analyzing new correlation and pattern. ML is a branch of computer science that is used to uncover the hidden pattern from large and complex data. Machine learning is a technique through which model is trained to learn from data and hence it is widely used in almost every field in finding a valuable pattern from big data. This technique does not require human interruption for producing result. Modern time businesses are aware of the fact that, big data is influential only if useful information is collected from it with help of an appropriate machine learning algorithm.

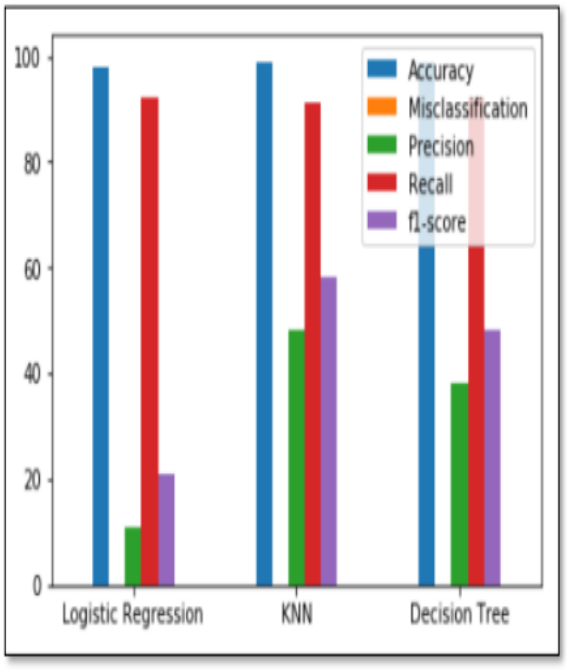
RESULT AND DISCUSSIONS

HPRT is applied to highly imbalanced credit card transaction dataset. It is consists of two days transaction of European cardholders having two classes, Fraud (0.17%) and Non Fraud (99.83%). The proposed algorithm can be applied for large and big dataset as a pre-processing algorithm for reducing the complexity of an imbalanced dataset. This algorithm automatically detects the level of imbalance in a dataset and then reduces it automatically in various steps. Then three different machine learning classifiers (LR, KNN and DT) are used to construct a model for the balanced dataset. Confusion matrix and various other measures evaluate the model which shows KNN as the best model with (99%) accuracy, precision (0.48%), Recall (0.91%), F1-Score (0.58%) and misclassification (0.014%). Confusion matrix confirms that out of 37741 test data only 55 times model predicted wrong result. Performance of decision tree is also satisfactory with 64 wrong predictions having an accuracy (99%), precision (35%), Recall (92%), F1-Score (48%), Misclassification (0.016%).

Results of HPRT based ML Models during Experiment I

Matrices & Measures			LR	KNN	DT
Accuracy			0.98%	0.99%	0.99%
Precision			0.11%	0.48%	0.38%
Recall			0.92%	0.91%	0.92%
F1-Score			0.21%	0.58%	0.48%
Misclassification			0.018%	0.014%	0.016 %
Confusion metrics					
Total Test Samples 37741	predicted Negative	predicted positive	36962 | 681 8 | 90	36760 3 | 4 15 | 83	37592 | 51 13 | 85
Actual Negative	TN	FP			
Ac	F	TP			

tual Negative	N				

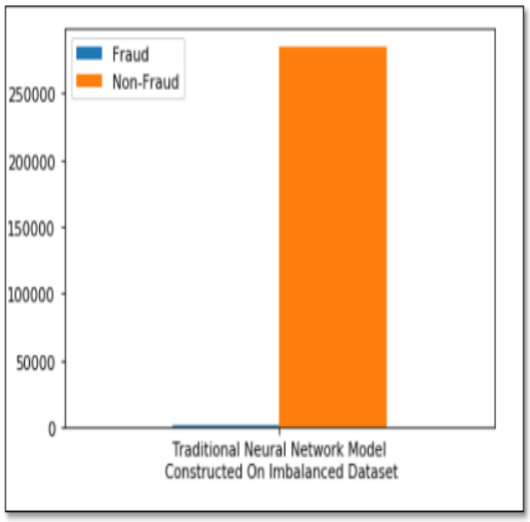


Performance Comparison of HPRT based ML Models during Experiment I

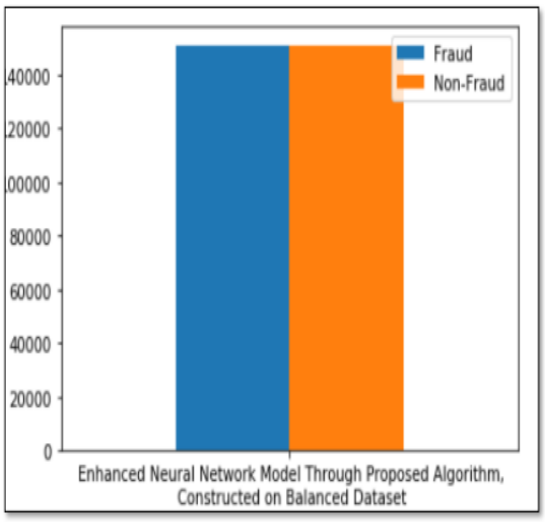
HPRT, An enhance Hybridization Pre-processing and Resampling machine learning approach for optimizing the complexity of an imbalanced dataset by dropping redundancy and outliers from majority class and then adding synthetic feature using SMOTE OS in minority class has been proposed in this study. HPRT successfully overcomes the disadvantage of traditional sampling methods i.e. random US, OS and SMOTH OS. Random US causes huge loss of information, whereas random OS causes over fitting. HPRT is applied to imbalance sample dataset to balance it automatically. Among all the three machine learning algorithm, applied to the balanced dataset. The KNN algorithm performance was better but it is not competent for big dataset because it is an in-memory algorithm. KNN fails to construct good models with dataset having high volume and high dimensionality. Big data computation cost is very high with KNN algorithm, as the distance is calculated for every data points present in the dataset. Therefore, the model constructed by the KNN algorithm is not at all suitable for big data. HPRT based robust machine learning algorithm can be perfect combination for optimizing balance and complex big data. Therefore HPRT based neural network in used in a next section for of big data classification.

We conducted two experiments for constructing Neural network model for classification of frauds and non-frauds from credit card transaction dataset – (a) in a first experiment an imbalanced dataset is a feed to traditional neural network classifiers. The first model is consisting of imbalance dataset as input to traditional MLP architecture (1 input layer, 1 hidden layer, and one output layer. Due to imbalance nature of the dataset model does not perform well during classification of frauds and non-frauds. The model achieves high accuracy with low F1-Score. Confusion matrix present lot many false negative i.e. many fraudulent transactions detected as non-fraudulent. This type of model possess loss to finance industry. Traditional neural network model results are: accuracy (99%), Recall (78%), Precision (75%) and recall (39%) rate. Confusion matrix results display 71 misclassifications out of that 60-time model predict frauds as genuine which is very costly for any model. Therefore model was not further analyzed and dropped with an intention to construct efficient MLP based classifier. (b)In a

second experiment, the neural network was combined with the proposed model. Our proposed model was first applied to an imbalance dataset to make it balance and then the neural network is used for building a predictive classification model. The second model result was outstanding, predicting both frauds and non-frauds correctly with very less misclassification. Accuracy matrix shows almost 99% accuracy with good precision (100%), recall (85%) and F1-score (89%) rate. Confusion matrix produces a good result with very less number of misclassification (21 times).



Imbalance Dataset -Experiment I



Balance Sample Dataset with the application of HPRT - Experiment II

CONCLUSION

Machine learning-based redundancy reduction techniques are revolutionizing the way we handle and process data. By leveraging advanced algorithms like dimensionality reduction, clustering, and autoencoders, ML models can identify and eliminate redundancies more efficiently and effectively than traditional methods. Whether it's for optimizing data storage, improving performance in real-time systems, or enhancing network efficiency, these techniques are becoming integral to a wide range of applications. As ML technologies continue to evolve, we can expect even more innovative and powerful methods for redundancy reduction, offering greater scalability, adaptability, and efficiency across industries.

REFERENCE:

1. *Aws I. Abu Eid (2024), "Sports Prediction Model through Cloud Computing and Big Data Based on Artificial Intelligence Method", Journal of Intelligent Learning Systems and Applications, Volume 16, Issue 2, Pages 53-79. Doi: 10.4236/ jilsa. 2024.162005.*
2. *Ashish Mishra (2023), "Development of Machine Learning Algorithm for Assessment of Heat Transfer of Ternary Hybrid Nanofluid Flow Towards Three Different Geometries: Case of Artificial Neural Network", Heliyon, ISSN 2405-8440, Volume 9, Issue 11, https://doi.org/10.1016/j.heliyon.2023.e21453.*
3. *Noura AlNuaimi (2022), "Streaming Feature Selection Algorithms For Big Data: A Survey", Applied Computing and Informatics, ISSN 2210-8327, Volume 18, issue.1/2, Pages 113-135. DOI 10.1016/j.aci.2019.01.001*
4. *Sarker, I.H. (2021)," Machine Learning: Algorithms, Real-World Applications and Research Directions.", SN COMPUT. SCI., ISSN 2661-8907, Volume 2, Issue*

160, https://doi.org/10.1007/s42979-021-00592-x

5. *Bin Hou (2013), "Cognitive Radio Spectrum Allocation Strategy Based on Improved Genetic Algorithm", Communications and Network, ISSN 2352-8648, Volume 5, Issue 3C, Pages 22-26. Doi: 10.4236/cn.2013.53B2005.*
6. *Yang You (2015)," Scaling Support Vector Machines on Modern HPC Platforms", Journal of Parallel and Distributed Computing, ISSN 0743-7315, Volume 76, Pages 16-31, https://doi.org/10.1016/j.jpdc.2014.09.005 .*
7. *Francis E.H (2001), "Tay Application of Support Vector Machines in Financial Time Series Forecasting", Omega, ISSN 0305-0483, Volume 29, Issue 4, Pages 309-317, https://doi.org/10.1016/S0305-0483(01)00026-3.*
8. *Ali Emrouznejad (2016), "Big Data Optimization: Recent Developments and Challenges" Springer, ISSN 2197-6511, Volume 18, DOI:10.1007/978-3-319-30265-2*
9. *Sarker, I.H. (2021)," Machine Learning: Algorithms, Real-World Applications and Research Directions.", SN COMPUT. SCI., ISSN 2661-8907, Volume 2, Issue 160, https://doi.org/10.1007/s42979-021-00592-x*
10. *Reem Almutiri (2016),"ASurvey of Machine Learning for Big Data Processing", Journal on Big Data, ISSN 2196-1115, Volume 4, Issue 2, Pages 97-111. https:// doi.org/10.32604/jbd.2022.028363.*