A Security Applicable with Deep Learning Algorithm for Big Data Analysis

Ramdas Vankdothu**,** *Research Scholar, Department of Computer Science & Engineering, University College of Engineering(A), OsmaniaUniversity Hyderabad ,Telangana ,India.*
Dr. Mohd Abdul Hameed**,** *Assistant Professor, Department of Computer Science & Engineering ,University College of Engineering(A), Osmania University Hyderabad, Telangana ,India.*

Article Info
Volume 82
Page Number: 8493 – 8500
Publication Issue:
January-February 2020

Abstract:

Big data is a field that discusses ways to investigate, regularly extract information, or differently deal with data sets that are excessively large or difficult to be dispensed with traditional data-processing utilization software. The analysis of big data is the critical challenge to be discussed among all research results because it presents more critical business value in any analytics ecosystem. Classification is a mechanism that designs data are allowing economic and efficient completion of precious analysis. So, there is a need for choosing suitable features for preparing the classifier. That is feasible by combining a Feature Selection process with a classification pattern. So, this analysis work initiates a hybrid method defined HCFS-Hierarchical learning for recognizing relevant Feature Subsets compared to the target class and yielded to the classifier representation to improve the performance. The Privacy need is revealed by the integrity characteristic of the big data. The development of science has encouraged every individual to the protection and utilize big data for analyses of the industry, consumer, medical, bank account, etc. obtained privacy break or interruption in most cases. Also, the data appropriated for big data analytics include limited, or copyright retained data, and there endures data secrecy break or interference. So, there is an essential need to protect privacy with specific principles for safeguarding the fine-tuned private data of every individual from interruption for analytics.

In this survey, we examine how Deep Learningcan be applied to discuss some critical problems in large data analyzes, including the extraction of complex models of large amounts of data, moral indexing, data tagging, rapid data recovery. The extension of the investigation study shows the necessity for feature extraction before classification. Feature selection determines a feature subset from existent feature set associated with the target class, while feature extraction obtains new features from a previous feature set. It improves the performance of classification by preparing the classifier with suitable features. Furthermore, Feature Extraction is similarly employed for the enrichment of the organization performance by implementing new features compared to the target class for practice capable of the classifier. Most of the research work based on previous Deep Learning algorithms like Autoencoders (AEs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNN) are the main approaches implemented. However, there are still have a problem in their complexity, Privacy, and also time-consuming in the current approach to solve and end this issue, this work inducts a schema described Enhanced Local N-ary Ternary Patterns (ELNTP) with MDBN (Modified Deep Belief Network) for multiple big data image set classification and provide security for big data analytics. The ELNTP acts by changing the previous LNTP by modification in the assortment of pixel states for identification and MDBN operates through adjustment of parameters in the DBN approach on activation function selection and weight updating process. The ELNTP and MDBN provide excellent performance in big data heterogeneous image set classification than the previous methods. This investigation work examines the result of privacy in the feature selection method because privacy is compulsory when a user distributes a sample feature for the determination of appropriate characteristics from the databank and vice versa. Further, the addition of secrecy-provided mechanism should not pretend the classification performance. Qualitative evaluation of all the proposed classification methods and Security-preserving mechanism has been created with classification accuracy and operating time, sequentially. Statistical analysis of accuracy assessments and computational time represents that the proposed schemes provide compromising results over previous methods.

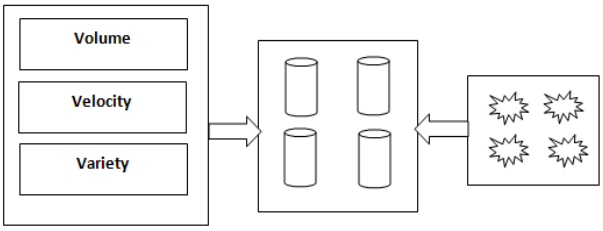
Article History
Article Received: *18 May 2019*
Revised: *14 July 2019*
Accepted: *22 December 2019*
Publication: *07 February 2020*

Keywords:*ELNTP, DBN, MDBN.*

1. INTRODUCTION

Big data analysis and deep learning in data science are both aspects in Data Science and have become a very common industry both publicly and privately, collecting large amounts of domain data, including social statistics, cyber security, phishing detection, And valuable information about retail issues such as medical informatics. Organizations like Google and Microsoft are looking for a lot of knowledge in business sciences and decisions that impact current and future technology. Deep learning algorithms retrieve difficult, high-level concepts, such as defining data in graded learning. Critical abstractions are determined at the level that is derived based on the maximum number of direct concepts displayed at the additional level of the system. The basic feature of deep learning is to study and learn large volumes of unallocated data and to provide it as an indispensable tool for analyzing large data, as untreated data is often categorized. Or not ranked. In this article, we examine how deep learning can be used to analyze some of the key problems in big data analysis, including large volumes of data, meaningful indexing, data labeling, fast Data recovery and data analysis. Data discrimination work We are also studying some features of deep research that further improve the problems of large-scale data analysis, including data flow, high dimensional data, and distributed computing and design development. Need We provide information about future important work by replicating some issues, including sample model models, domain development models, standardization for collecting useful data ideas, developing cement catalogs, semi-structured learning, and dynamic learning.

Deep learning algorithms is a promising way to find high levels of complex data automatically summarized. These algorithms create a hierarchical structure and represent layers of learning and data representation, where higher level (more abstract) features are defined in terms of lower level features (less abstract). The classification of deep learning algorithms is driven by artificial intelligence that mimics the core sensory areas of the deep learning process and modern brainstorming, which automatically extracts the basic data on the features. Deep learning algorithms are beneficial when it comes to learning from a large number of uncensored data, and data is usually learned through layers. In the deep learning algorithm, the process of extracting the representation of the central idea data is automated [1]. Deep learning algorithms use a lot of data to automatically extract complex presentations. These algorithms are primarily driven by artificial intelligence, which is the overall goal of capturing the ability of the human brain to observe, analyze, learn, and make decisions, especially sim extremely complex problems.



Multi-source big data collectingDistributed Big Data Storingintra/inter big data processing

Fig.1 Common Architecture of Big Data

2. REVIEW OF LITERATURE

The different stages of the large data classification problem are feature selection, feature extraction, deep learning, and privacy. Comprehensive review of current methods is critical to producing an effective classification algorithm.

The large volume of unstructured data presents an immediate challenge to the traditional computing environment and requires scalable storage and distributed strategies for data consulting and analysis. However, this large volume of data is a positive feature of the statistical data. Many companies, such as Facebook, Yahoo, and Google, already have huge amounts of data and have recently begun to reap the benefits.

Carine Azevedo et al. [2]It suggests a new method for dynamic selection of properties to be applied in classifier ensembles. This method specifies the best subsets of attributes for an individual instance or set of instances of the input data set. Therefore, each test instance is compiled by using a subset of unique properties in the compilation process. The main objective of this document is to expand the method of selecting dynamic properties that have been proposed for individual workbooks, and to adjust this approach to use in workbook collections. To validate our proposed method, a pilot analysis is performed to explore the effectiveness of the approach compared to the current common methods. Our findings indicated performance gains when comparing the proposed method with existing common methods.

Jian Shen et al. [3]As jian shen, the current traffic classification is an essential challenge for network management and management. Choosing features is an effective way to minimize dimensions and reduce redundant information. For a precise designation of traffic at a lower price for evaluations, a method is proposed to identify a subset of mixed characteristics in the sliding block base, which is flexible according to the classification performance. Besides, a gradual convergence strategy is designed based on methods of selecting subsets of mixed characteristics. The approach combines all the properties selected. To discover the value of the relationship between all selected functions, an additional selection round is added to the original algorithm rule. Three sets of tests examine offers. Our theoretical analysis and empirical observations reveal that the proposed method consumes fewer grades with similar or better academic performance in different sizes of different sizes. Also, the additional convergence strategy improves rating accuracy.

Carine A et al. [4]A study on the impact of evaluation criteria and similarities in the method of selecting Feature Selection (FS). The main objective of this survey is to assess the significance of these parameters in the FS method that has been analyzed. This method will be evaluated using eight different configurations for these two important parameters. Basically, various correlation measures will be chosen as criteria for evaluation and distance measurements as similarity measures. In addition, to evaluate the effect of these parameters, a pilot analysis will be performed, where different configurations will be evaluated to determine the best configuration. Next, the best configuration result will be compared with the methods of extraction and selection of current properties, applied to different classification problems. The results shown in this document indicate that the suggested method with best configuration has better performance results than current methods, in most cases.

Abdussalam [5]In modeling prediction, feature selection techniques are an important step in data processing before the prediction model is created. Selecting the most important input properties is important to increase the accuracy of prediction, data reduction and training time. In this work, some techniques for selecting and analyzing features are compared. Next, the feature selection techniques are used as a filter before predicting the price of electricity and comparing its effect on the accuracy of prediction and the Mean Absolute Percentage Error (MAPE) for each specific subset.

Qinbao Song et al. [6]The FAST-recommended method, which is the method for selecting features to be accelerated by the assembly. It works in two stages. Using the overall theory of graph theory, the features are first subdivided into groups. Specific attributes for the target group are selected from each group in the second stage for sub-categories of attributes. Fast static assembly policy has the ability to create a subset of sufficient and independent properties. Minimal extension tree assembly technology is used to ensure high speed performance. Experimental analysis was performed to evaluate the performance and efficiency of the fast algorithm.

Chen et al. [7]Provides **Marginalized denoising autoencoders**(mSDA), SDA is effectively measured to obtain high-resolution data and graphics faster than devices. Its approach underscores the noise in SDA training and, therefore, does not require gradual deviations or other optimization algorithms to learn the parameters. There are hidden nodes in the layers of the SDA module, the solution of which can be closed very fast. In addition, SDA only has two free parameters, which control the amount of noise and the number of piles, which greatly simplifies the model selection process.

Coates et al. [8]Take advantage of the relatively cheap computing power of a set of GPU servers. Specifically, they develop their own system (using neural networks) based on the Community off the Shelf High Performance Computing (COTS HPC) and provide high-speed communications infrastructure for integrated computing.The system is capable of training 1 billion network-connected networks in just 3 devices within two days, and can extend the network to over 11 billion parameters using only 16 devices and where DistBeliefCan be compared to Compared to the computing resources used in Distribute, the COTSHPC distributed network system is generally available to a wider audience, making it a viable alternative to other deep learning specialists who use large-scale models.

Chopra et al. [9]Proposed to Learn useful data (predictions) from unstructured data, keeping in mind the available account information for changing the distribution between training data and tests. Proposed a model for deep learning (based on neural networks). The aim is to learn a lot of intermediate hierarchical presentations on the way to interpolation between the training and diagnostic fields. In terms of identifying things, their study shows improvement compared to other methods. Previous studies raise the question of how to enhance learning representations and ability to generalize deep data patterns, and it has been suggested that the ability to integrate patterns is a key requirement in big data analysis, where often the distribution of information changes and the data destinations.

Bengio et al. [10]given About some of the features that make data representation better for performing discrimination. Describe the open-ended question about setting standards for better representation of data in deep learning. Compared to the most common learning algorithms in which categorization error is commonly used as an important measure for training modeling and learning patterns, deep learning algorithms using big data Similar quality development is inappropriate for training, because most of the big data is analyzed. Include learning from highly controlled data. Although monitoring of data availability may be useful in some large data areas, it is not yet a matter of determining the quality of abstraction and summarizing statistics in abstraction in large data analysis. -In addition, setting the standards required to produce a good representation of the data raises a question. Good data representation is effective in data indexing and / or mapping.

3. PROBLEM STATEMENT

Big data frequently Dealing with unnamed data. In some cases, the extraction of features and the selection of appropriate features have attractive procedures in which researchers fail. The problem of heterogeneous data is required, preferably data comparable to medical imaging classification. The aim of the researchers is to increase the accuracy of the classification in a moderate manner rather than to provide basic attention to data privacy. The feature Selection and Feature Extraction classification methods are provided for large volume of data to provide better classification accuracy. Most of the Previous Deep learning algorithms implemented in the existed work. One of the previous methods is Autoencoder (AEs) classification method is a special neural network structure, with an input layer, an output layer, and a hidden layer. Adjust the weights of the hidden layer through training to allow the input and output value to be entered as soon as possible. This classification improves performance typically, but there are still challenges in this AEs normally it fails miserably out of sample. To determine the above-listed problem, the alternative method for AEs is lightened CNN (Convolutional neural networks) executed to obtain ultimate accuracy, CNN designs manage to be difficult or complicated local facial piece collection, which happens in a waste of time and place. To mitigate this problem, a lightened CNN structure to learn a compressed embedding for face description in Deep learning methods.But they have some constraints in this approach are noted below:

1. High computational cost.
2. Two primary issues are computation time and overfitting.
3. If you don't have a good GPU, they are quite slow to train (for complex tasks).
4. They use to need a lot of training data.
5. First and foremost, large volumes of data exhibit a relevant, challenging problem for deep learning.
6. Do not have sufficient in the way of the robust theoretical framework

And to determine above-listed difficulties in a lightened CNN, A deep learning design based on Recurrent Neural Networks (RNN) is performed, particularly long short-term memory (LSTM), They have gained a lot of interest in Automatic Speech recognition (ASR). Although some success stories have been reported, RNN training remains a major challenge, especially with limited training data and listed some limitations below:

1. loss of flexibility nevertheless
2. If the network is pretty deep, each training step is going to take much longer
3. As for the previous study, their research does not explicitly encode the frequency shift of the data among the source domain and the destination domains
4. Big data frequently maintain a large number of samples (inputs), vast differences of class types (outputs), and much large dimensionality. These features instantly lead to running-time complexity and representation complexity.
5. Different challenging problem incorporated with the significant velocity is that data are usually non-stationary, i.e., data sharing is increasing over time. Substantially, non-stationary data are generally classified into pieces with data from a small-time period.

It also requires more computing resources and even fewer edges. When you have a large set of decision trees, it is difficult to have an intuitive understanding of the relationship in input data. With the analysis of Problem statement and listed limitations, our proposed work will introduce better security provided Deep learning algorithm for Big data analytics.

4. RESEARCH OBJECTIVES AND APPROACH

Based on the measurements from the Problem Statement, the investigation issues which require vast research improvement are expressed as objectives of this dissertation and given below.

- Investigate and examine the different security-preserving techniques for increasing performance.
- To utilize a suitable Feature extraction and Feature selection technique
- To implementing the most useful feature subset to Classifier.
- To achieve deep learning, by reducing the Time complexity using “ELNTP-MDBN”

Method for large dimensional heterogeneous image set administration.

- To better classification accuracy.
- To provide privacy to data used for analysis using Deep learning algorithms
- To dispense with multiple data sources with various data formats.
- To enhanced classification modeling, the more excellent Quality of samples produced by the generator
- Probabilistic patterns, and the invariant property of data representations

5. RESEARCH METHODOLOGY

The Proposed ELNTP-MDBN method for large dimensional heterogeneous image set administration is shown during the flow diagram in below Fig.2. It has two key steps, classified, Enhanced Local N-ternary pattern and modified DBN for feature selection and extraction. The Gaussian filtering is employed for avoiding independent and dependent noise, and noise removal are conducted on the heterogeneous image set. The refined image is normalized for classifying the environment from the foreground. Later the ELNTP is applied for classifying the images, and it is utilized for every section in obtaining the features. Initially, the image type is organized, and the necessary elements for the classification of natural and unusual layers are extorted depending on prepared features. Then the features compared to the testing image are Frequently investigated using the trained characteristics. Finally, the activation function modified in DBN classifier effectively classifies the natural and the unnatural levels. Learning is characterized by immediate supervised and unsupervised training utilizing the mean square failure rate, and the performance of the organization is improved using backpropagation with the least execution time.

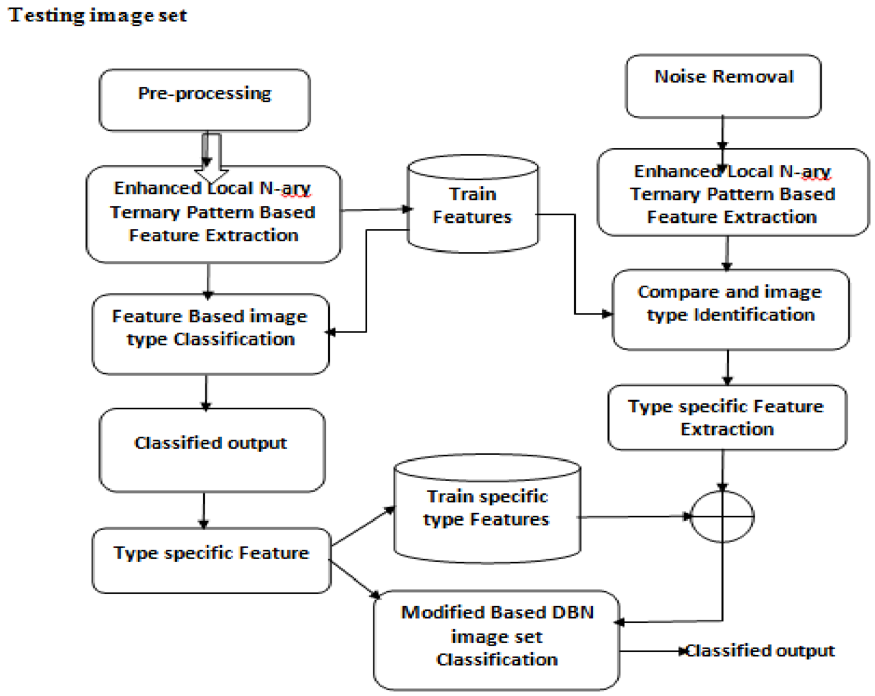

Testing image set
Pre-processing
Noise Removal
Enhanced Local N-ary Ternary Pattern Based Feature Extraction
Train Features
Enhanced Local N-ary Ternary Pattern Based Feature Extraction
Feature Based image type Classification
Compare and image type Identification
Type specific Feature Extraction
Classified output
Train specific type Features
Type specific Feature
Modified Based DBN image set Classification
Classified output

Fig.2 Proposed Work Flow

MDBN

The deep learning component of machine learning technology has many algorithms produced for the classification method. The MDBN process operates with two levels, the first one for layer-wise feature extraction, and the second one is for the regeneration of weight tuning. Frequently the DBN is practiced to learn feature description from both labeled and unlabeled data. It connects the unsupervised method of learning for pre-training and the supervised process of learning for fine-tuning. The DBN structure includes RBMs (restricted Boltzmann machines) to evaluate the reconstruction of weights by layers in the first step. RBM consists of two layers in a function, the nodes in one layer are fully bound by the nodes available in the other layers. But no connection is available between nodes in a similar layer. In the second step, DBN uses the repetitive method of back-up to adjust the weights. The repetition of unsupervised and supervised methods of learning improves the performance of classification.

The paper focuses on two key issues: (1) how deep learning can help solve specific problems in analyzing large data; and (2) how to increase deeper learning areas to reflect some of the challenges associated with comprehensive data analysis. In the first topic, we explore deep learning to analyze the large data identified, including learning large amounts of data, indexing, semantic tasks, and labeling.

ALGORITHM: Modified Deep Belief Networks

Data:

D: training dataset

R: vector of RBMs comprising the MDBN

mean-field?: Boolean indicating whether to use the mean-field value
when propagating values to the next RBM

query-final?: Boolean indicating whether to query the hidden layer of
the final RBM (used in preparation for building a DNN)

```
data ← D;
foreach rbm in R do
rbm ← TrainRBM(rbm, data);
if not last RBM or (last RBM and query-final?) then
data ← Propagate(rbm, data, mean-field?);
```

Here,

DNN→Deep Neural Networks

RBM→Restricted Boltzmann Machine

Describes the basic procedure for MDBN training without supervision. The procedure for training the MDBN class is similar to the algorithm 1, but the SoftMax tag is linked to the data set when the final RBM is trained

6. CONCLUSION

In this approach, the objective of the task of analysis is the need for deep learning-based security algorithm for large data analysis. This paper given description aboutMost recent works in Deep learning techniques like Autoencoder (AE), CNNs, and RNN. Extensive data require the strength of the growing employer to distribute them volume, veracity, variety, velocity, and value attributes.With the help of these features, the implementation of a security-based deep learning algorithm for big data analytics is presented in this work. The analytics task used in this research work is the big data classification.The action of classifying data with the issues and difficulties opened up by the big data environment is critical and challenging. In this paper, the ELNTP and MDBN deep learning scheme proposed to improve both running time and classificationaccuracy.And it will provide outstanding performance in big data heterogeneous image set classification than the previous methods. Further implementation provided security-based feature selection to improve performance. The proposed classification methods and Security-preserving mechanisms provide better classification accuracy and running time compare to previous algorithms.

REFERENCES

1. Chunkai Zhang,Lin Yao, 2017,” Feature selection for high dimensional imbalanced class data based on F-measure optimization” pp.278-283.
2. Romulo de Oliveira Nunes, Carine Azevedo Dantas,2018, "Dynamic Feature Selection for Classifier Ensembles“, pp. 468-473.
3. Jian Shen, Xiaoyan Zhang, 2017, "Sliding Block-Based Hybrid Feature Subset Selection in Network Traffic", pp. 18179-18186.

4. Carine A, Anne M. P. Canuto,2017, "Investigating the Impact of Similarity Metrics in an Unsupervised-Based Feature Selection Method", pp. 55-60, 2017.
5. Abdussalam Mohamed, 2016,"Effective input features selection for electricity price forecasting", pp. 1-5, 2016.
6. Q. Song, G. Wang, 2013," A fast clustering-based feature subset selection algorithm for high-dimensional data "pp. 1–14.
7. Chen M, Weinberger KQ, 2012,"Marginalized denoising autoencoders for domain adaptation".
8. Coates A, Andrew N,2013," Deep learning with COTS HPC systems", pp.1337–1345.
9. Chopra S, Gopalan R,2013, "Dlid: Deep learning for domain adaptation by interpolating between domains".
10. BengioY, Vincent P ,2013, "Representation learning: A review and new perspectives. Pattern Analysis and Machine Intelligence", pp. 1798–1828.