International Journal of Combined Research & Development (IJCRD)
eISSN:2321-225X;pISSN:2321-2241 Volume: 11; Issue: 4; April -2022

A Deep Autoencoder for Multi Classification on Large Datasets in Hadoop Distributed File System

Sirisha Madhuri T
Research Scholar
Department of CSE
School of Engineering and Technology
SPMVV, Tirupati.

Dr. P. Venkata Krishna
Professor
Department of Computer Science
SPMVV, Tirupati

Dr. V. Saritha
Professor & HOD
Department of CSE
School of Engineering and Technology
SPMVV, Tirupati

ABSTRACT

Classification is widely used technique in the data mining domain, where scalability and efficiency are the immediate problems in classification algorithms for large databases. These days large amount of data is generated, that need to be analysed, and pattern have to be extracted from that to get some knowledge. Classification is a supervised machine learning task which builds a model from labelled training data. The model is used for determining the class; there are many types of classification algorithms indeed, such as tree-based algorithms, Naive Bayes and many more. These classification algorithms have their own pros and cons, depending on many factors such as the characteristics of the data. To overcome this, a dimensionality reduction technique known as the Deep Auto Encoder is being used for large dataset classification on Hadoop Distributed File System (HDFS).

Keywords : Classification, Deep Auto Encoder (DAE), Naive Bayes, Dimensionality Reduction.

1. INTRODUCTION

The process of arranging large datasets to detect patterns and describe the relationships to resolve the issues through the analysis of data is known as data mining [1].The rapid and huge development of the produced information, replicated and spendon the basis of society has created the Big Data is one among the wide range of researches [2]. Due to the increment in processed data size, people initiated to resolve issues present in big data based on the help of machine learning methods [3] [4]. Based on the general experience of several decision approaches have been replaced by using the data analysis, data mining and so on because of the big data technology improvement. Moreover, in predicting and analysing the problems of big data performs an important part in several domains including public health, economics, etc. due to its analysis and decision for these fields will based on large amount of data present [5].

To efficiently study from large scale in all types of real implementations including classification and clustering is most important in the generation of big data [6]. Due to heterogeneous data and large data volumes, the data classification became highly difficult [7]. The required information is gathered from large amount of data and the classification is completed before the initialization of actual classification. Supervised classification and unsupervised classification are the major classification methods [8]. Based on this classification techniques, intelligent decision making offered. In this classification, two types of phases are present, learning process phase is the first phase which supplies a large amount of training data sets and examining is performed then rules and patterns are produced. The second phase execution initialize that is calculation or experiment of datasets and archives the classification patterns accuracy [9].

General schemes of the data classification are: support vector machine based data classification [10]. Recently, lots of effective and accurate algorithms were presented for the issues in classification including SVM, decision tree, Naive Bayes method, artificial neural network and Extreme Learning Machine (ELM) [11]. There are significant growths on the neural network enhancement in the field of machine learning. To hierarchically perform the information from layer to layer, neural learning achieves the brain cognitive structure [12] [13]. The approaches of SVM are generally utilized as a technique of binary classification, but recent researches describes that Multiclass Support Vector Machine (MSVM) are mostly

discussed supervised learning algorithms. The input vectors are classified into multiple classes along with trained oracles [14].

For Single hidden layer Feed forward Neural Networks (SFNNs), ELM is a learning algorithm [15]. Randomly the weights among input and hidden layers are allocated is the initial advantage of the ELM algorithm. Random feature mapping is realized by the interrelationamong the input and hidden layers. The output weights among hidden and output layers are trained initially, layer-by-layer back propagated tuning is not required [16]. Moreover, ELM is implemented very easily and for classification is low sensitive to user described parameters. Since, highly quick capacity of learning and best generalization capabilities of ELM, to construct the classification model ELM is utilized [17].**2.**

2. RELATED WORK

For polysemic object, the multi-classification learning provided a multi-dimensional view, based on this achievement it became popular research topic recently in machine learning. FangfangLuoet al. [18] had proposed the kernel extreme learning machine and implemented to the problem of multi-label classification (ML-KELM). Operations of iterative learning were avoided by this approach. The transformation to binary multi-label vector from the original value of ML-KELM network were solved by design of self-adaptive threshold function. ML-KELM had the optimal solution of least square of ELM and less amount of parameters required for adjustment, stable running, rapid convergence speed and great generalization performance. The experiments for extensive multi-label classification were performed on datasets of various scale. The experiment results demonstrated that ML-KELM provided an outstanding performance in large scale dataset along with better feature of dimension instance.

Using deep learning, feature learning of large scale task driven from big data allowed by deep learning.YueDenget al. [19] had proposed the fuzzy deep learning paradigm and buried shortcomings of fixed representation. The concepts of fuzzy learning into deep learning were introduced. From both fuzzy system and neural representations, the proposed system of fuzzy were a hierarchical deep neural network which derived the information. Later, the information knowledge gathered from these combined two views, which data representation to be classified is produced. In comparison with other non-fuzzy and shallow approaches, the presented fuzzy deep learning method provided high performance.

AlexandrosIosifidiset al. [20] had proposed an ELM by using classification scheme which exploited information of geometric class. By network hidden layer outputs and ELM arbitrary dimension spaces, this work formulated and exploited presented scheme data descriptions in resolved feature space. The performance of system improved by using the geometric class information exploitation. The presented scheme evaluated in publicly available datasets and compared its performance with recent proposed one class extreme learning machine algorithm and one-class classifiers. The experimental results were demonstrated that the presented method provided great performance than the other schemes.

For graph classification, from the large set of graph objects the sub graph mining were a major problem. ZhanghuiWanget al. [21] had proposed a discriminative sub graph mining scheme on the basis of ELM-filter technique within the scalable MapReduce computing model. The collection of parts randomly partitioned among worker nodes and each worker applied a quick pattern evolutionary scheme to mine a set of discriminative sub graphs with the help of ELM-Filter strategy in its partition. Great training accuracy of ELM produced by the set of discriminative sub graphs. The experimental results on both real and synthetic datasets were provided that the presented method outperformed the other approaches in terms of classification accuracy and runtime efficiency.

Migel DTisseraet al. [22] had proposed an approach for synthesising deep neural networks with the help ofELM as supervised auto encoders. The error rate of classification increasingly enhanced along with addition of auto encoding ELM modules in a stack on the basis of standard datasets for multi-class image classification (MNIST, CUFAR-10 and Google street view house numbers (SVHN)). Moreover, 99.19% of MNIST test mages were accurately classified which exceeded the better error rates presented for standard 3 layer ELMs. Simultaneously this approach offered a

significantly faster training algorithm based on that best performance were achieved.

3. METHODOLOGY

The classifier design contains the important steps such as pre-processing, appropriate classification selection and feature selection strategies. A DAE (Deep Auto Encoder) based SVM (Support vector machine) classifier with selected features is presented for the above mentioned steps for large dataset classification. The dimensionality is reduced with principal component analysis (PCA). DAE extracts the features for optimal classification and which reduce the training time for classification. The DAE training includes the meta-heuristic based Adam optimization algorithm which reduces the computational complexity. The deep learning classifier uses real world datasets such as CIFAR, MNIST to show the performance of proposed system. Different classification problem and different classifiers are used for comparative analysis. The results indicated that proposed approach improves classification performance by discarding redundant or unimportant features and reduces the dimensionality of dataset. The results strongly suggest that the proposed method can speed up the computation time of a proposed algorithm and simplify the classification tasks. Finally the results are visualized using MATLAB tool.

4. EXPECTED RESULTS

The research can be implemented using MATLAB tool. The performances shall be evaluated in terms of accuracy, runtime, Speedup and error rate.

REFERENCES

[1] Petrova, Ekaterina, et al. "In search of sustainable design patterns: Combining data mining and semantic data modelling on disparate building data." Advances in Informatics and Computing in Civil and Construction Engineering. Springer, Cham, 2019. 19-26.

[2] Elkano, Mikel, et al. "CHI-BD: A fuzzy rule-based classification system for Big Data classification problems." Fuzzy Sets and Systems 348 (2018): 75-101.

[3] Zou, Huasheng, and Zhiyuan Jin. "Comparative Study of Big Data Classification Algorithm Based on SVM." 2018 Cross Strait Quad-Regional Radio Science and Wireless Technology Conference (CSQRWC). IEEE, 2018.

[4] Jin, Shangzhu, Jun Peng, and Dong Xie. "Towards MapReduce approach with dynamic fuzzy inference/interpolation for big data classification problems." 2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC). IEEE, 2017.

[5] Chen, Cen, et al. "Extreme Learning Machine and Its Applications in Big Data Processing." Big Data Analytics for Sensor-Network Collected Intelligence. 2017. 117-150.

[6] Deng, Zhenyun, et al. "Efficient kNN classification algorithm for big data." Neurocomputing 195 (2016): 143-148.

[7] Bikku, Thulasi. "A Novel Multi-Class Ensemble Model for Classifying Imbalanced Biomedical Datasets." IOP Conference Series: Materials Science and Engineering. Vol. 225. No. 1. IOP Publishing, 2017.

[8] Wang, Hai, et al. "Towards felicitous decision making: An overview on challenges and trends of Big Data." Information Sciences 367 (2016): 747-765.

[9] Koturwar, Praful, SheetalGirase, and DebajyotiMukhopadhyay. "A survey of classification techniques in the area of big data." arXiv preprint arXiv:1503.07477 (2015).

[10] Wang, Jia, Shuai Liu, and Houbing Song. "Fractal Research on the Edge Blur Threshold Recognition in Big Data Classification." Mobile Networks and Applications 23.2 (2018): 251-260.

[11] Triguero, Isaac, et al. "MRPR: A MapReduce solution for prototype reduction in big data classification." neurocomputing150 (2015): 331-345.

[12] Zhou, Lina, et al. "Machine learning on big data: Opportunities and challenges." Neurocomputing 237 (2017): 350-361.

[13] Jia, Feng, et al. "Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data." Mechanical Systems and Signal Processing 72 (2016): 303-315.

[14] Bishwas, Arit Kumar, Ashish Mani, and Vasile Palade. "Big data classification with quantum multiclass SVM and quantum one-against-all approach." Contemporary Computing and Informatics (IC3I), 2016 2nd International Conference on. IEEE, 2016.

[15] Iosifidis, Alexandros. "Extreme learning machine based supervised subspace learning." Neurocomputing 167 (2015): 158-164.

[16] Huang, Zhiyong, et al. "An efficient method for traffic sign recognition based on extreme learning machine." IEEE transactions on cybernetics 47.4 (2017): 920-933.

[17] Xin, Junchang, et al. "Elastic extreme learning machine for big data classification." Neurocomputing 149 (2015): 464-471.

[18] Luo, Fangfang, et al. "A multi-label classification algorithm based on kernel extreme learning machine." Neurocomputing260 (2017): 313-320.

[19] Deng, Yue, et al. "A hierarchical fused fuzzy deep neural network for data classification." IEEE Transactions on Fuzzy Systems 25.4 (2017): 1006-1012.

[20] Iosifidis, Alexandros, et al. "One-class classification based on extreme learning and geometric class information." Neural Processing Letters 45.2 (2017): 577-592.

[21] Wang, Zhanghui, et al. "Extreme learning machine for large-scale graph classification based on MapReduce." Neurocomputing 261 (2017): 106-114.

[22]Tissera, Migel D., and Mark D. McDonnell. "Deep extreme learning machines: supervised autoencoding architecture for classification." Neurocomputing 174 (2016): 42-49.