International Journal of Advanced Science and Technology
Vol. 29, No. 9s, (2020), pp. 4150-4155

# Hybrid Optimization Algorithm for Providing Big Data Classification

M. Kavitha Margret,
*Assistant Professor, Sri Krishna College of Technology, India*
S. Siamala Devi,
*Associate Professor, Sri Krishna College of Technology, India*
E. GoldenJulie,
*Department of Computer Science and Engineering,Anna University Regional Campus .Tirunelveli, India*
D. Vijayanandh,
*Assistant Professor,Hindusthan College of Engineering and Technology, India*
Y. Harold Robinson,
*School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India*

***Abstract***

*The larger amount of databases is required classification for providing efficiency in big data applications. The parallel computing approach utilizes the classification whenever the uncertainty occurs in big data. The information entropy causes through the partitioning issues while compressing the data to a restricted amount of memory. The computational time is high in the process of iterations in to big data classification and the available division in the big data reduces the threshold level. The randomly assigned classification involves the identification of the optimal path for every attribute. The computation of information gain for every attribute will enhance the computation time. The optimization based scheduling will reduce the computation time. The hybrid optimization algorithm uses to obtain the optimal path with attribute based randomly assigned and scheduling process. The experimental results show that the proposed technique has the enhanced optimization compared with the other techniques.*

*Keywords: big data, classification, optimization, accuracy, computation time*

## Introduction

Big data is the collection of data in huge volumes that provides more than sufficient information related to the field. Big data has caused the most of the developing industries to use advanced techniques to analyze the large size datasets. There are three challenging problems with big data causing researchers to find solutions [1]. The first problem is the velocity problem that creates huge amount of data to be handled at a high speed. The next problem is the variance in data as the data are collected from different sources and are formatteddifferently [2]. The last problem is the memory problem that is the major issue among the three as the storage and processing requires large memory. These problems paved way for wider researches to be performed to provide efficient analysis of data. Extreme Learning Machine approach has been presented as an efficient method to analyze the large amount of data [3]. ELM is generally used for the classification and regression of data with a single layer of hidden nodes. The machine learning approach has fast learning speed and best generalization scheme [4]. The approach efficiently solves the problems but cannot be termed as the best method due to the reason that it cannot be effective for all types of datasets. The complex datasets requires wide analysis techniques to efficiently determine the attributes. The cloud computing techniques like MapReduce can be used to analyze the big data without any scalability problems [5]. The method uses distributed computing to train multiple models with large data blocks and combine them using ensemble algorithms. As the method seems efficient the big data can be processed by implementing MapReduce based machine learning approaches [6].

The primary objective of this paper is the determination of optimal cut-points and the efficient scheduling of computation tasks to different nodes in the ELM tree. An attribute has many cut-points which all are send to host nodes to compute the information gain [7]. This approach receives furtherperiod to determine the best possible results of an attribute due to using all the cut-points in the computation process. Hence it is necessary to select the optimal way to reduce the computation time. The approach, instead of selecting all cut-points, randomly selects the cut-points and computes the information gain [8]. After some iteration, the cut-points with best information gain can be found by the optimization algorithms [9]. The computation of the scheduling algorithm is used to schedule these computation tasks to efficient host nodes by estimating



ISSN: 2005-4238 IJAST

which node can process which tasks with higher efficiency using the optimization algorithms [10]. The node data capacity is considered while allocating the tasks and only few randomly selected data are computed [11]. Thus the computation tasks can be performed efficiently with less time and reduced computation complexity [12].

**Related Works**

The elastic extreme learning machine also known as Elastic ELM based on the for the effective big data classification [13]. This approach performs better than normal ELM which has weak learning ability for the updating large scale training dataset. Moreover the matrix multiplication part in the ELM is the most computation expensive part [14]. The matrix multiplication problem is resolved by calculating it by incrementing, decrementing or correctionalcalculation [15]. The intermediate matrices are calculated and then the old matrices are compared with them to provide effective learning for the updating large datasets. The problem with the approach is utilized the weight vector for rapid learning which may reduce the convergence speed [16].

The new approach effectively reduces the amount of instances and thus increases the classification which greatly reduces the storage requirements and the noise sensitive measures. Inorder to reduce models in larger datasets, the MapReduce based structure is introduced with strategies to integrate multiple solutions and thus avoids drop in classification accuracy rates [17]. This approach can be applied to various applications which require conceptual work integration [18]. But the architecture of the big data systems has to suit for integration and hence the reference architecture models for the classification of the products and services has been presented for the efficient test case classification in big data [19]. The Online Sequential related approach for the effective classification of the ensemble in the peer-to-peer networks [20]. The enhanced tree can be easily implemented for distributed ensemble learning and classification [21]. A two-layer index structure is also presented to efficiently support peer selection for the introduction Quad tree for ensemble learning. The drawback in the approach is that it cannot be utilized for high dimensional ensembles [22].

The clustering performance of the implemented approach can be analyzed by studying in state-of-the-art methods [23]. An approach has been implemented for clustering the Self-Organizing Maps (SOM) [24]. The approach forms two problems and proposes two clustering methods based on the features of the ELM approach with the knowledge [25]. The optimization of the big data after classification has to be dealt with the two stage query processing optimization [26]. As optimization is out of scope for our research, the classification is considered with the classifier trains the configuration and prediction parameters and classifies the big data [27].

**Proposed Technique**

The computation of different parameters of attributes in a larger datasets is needed to be distributed to different computing nodes inorder to process the data in less time. The information gain split ratio and gain ratio calculations are done to estimate the optimal cut points for each attribute. These calculations are performed in the host computing nodes by allocating all the data associated with each cut-point to the available nodes anddetermining the better results. This approach consumes more time as all the cut-points are analyzed and also the fact that only few cut-points will be optimal. Hence, in our approach, the available nodes are analyzed with the available tasks using the optimization algorithms. The node to which if a task is allocated can perform better would have better data capacity. Thus the nodes with better data capacity are optimally selected and the tasks are allocated randomly to the nodes. The random selection is such that not all the cut-point computations are carried out and only few cut-points are computed. This reduces the number of iterations of computation. Fig. 1 demonstrates the classification of big data with oracle solution.


Fig. 1 Oracle solution for big data



ISSN: 2005-4238 IJAST

The genetic algorithm assigns the population of solutions and selects the nodes with high data rate using cross over and mutation of the chromosomes. The nodes with high data rate are assigned with larger tasks or even more than one task can be assigned to it. Similarly, the nodes with low data handling are assigned with suitable tasks. A group of system is taken as the computing nodes; the tasks to be assigned are also taken as T. The information gain is computed using Eq. (1)

$$Info_{Gain}(\gamma, cut_{xy}) = Info(\gamma) - I(cut_{xy}) \quad (1)$$

The information gain computes $\{Info_{Gain_{1j}}, \dots, Info_{Gain_{Mj}}\}$are assigned to N computing nodes. The ratio is computed using Eq. (2).

$$Ratio\ (\gamma, Att_j) = \frac{Info_{Gain}(\gamma, cut_{xy})}{Info_{Split}(\gamma, cut_{xy})} \quad (2)$$

The information split is computed using Eq. (3)

$$Info_{Split}(\gamma, cut_{xy}) = \left(\frac{cut_{xy}}{|\gamma|}\log_2 |\gamma|\right) \quad (3)$$

Algorithm – Hybridoptimization

Begin Procedure

Align all the available nodes and tasks

Computing nodes = $\{No_1, \dots No_n\}$

Compute Tasks = $\{Ta_1, \dots Ta_n\}$

$Fitness_{value} = Computation_{time}$

Initialize every position

choose random variable for computation

update the value for computation

End Procedure

**Results and Discussion**

The experiments are conducted for big data classification for the proposed technique with the exiting method. The simulation has been performed using the system with Pentium 4 Xeon, 3.06GHz CPU, 5l 2 MB RAM, and Red Hat Linux9.0 operating system.Fig. 2 demonstrates the classification accuracy which is compared with the existing method and it shows that the proposed technique has the best classification accuracy.

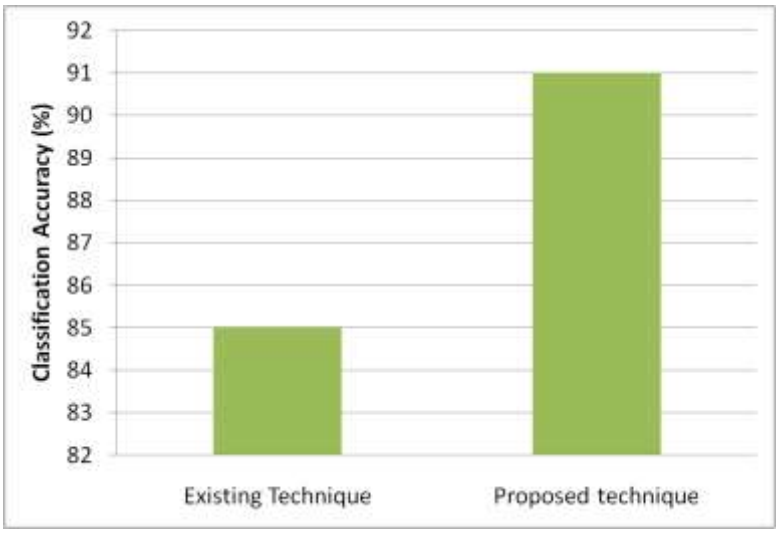


Fig. 2 Classification Accuracy



ISSN: 2005-4238 IJAST

The computation time is calculated for every process and the accuracy is generated.It is further evaluated for determining which optimization based method is better. The cut-points of every attribute are estimated from the population of cut-point solutions using the optimization algorithms. From the analysis, it is clear that the optimization based approach provides high accuracy of big data classification. Fig. 3 demonstrates the computation time for the comparison of proposed technique and the existing technique and the simulation result shows that the proposed technique has the minimum amount of computation time that the existing technique.

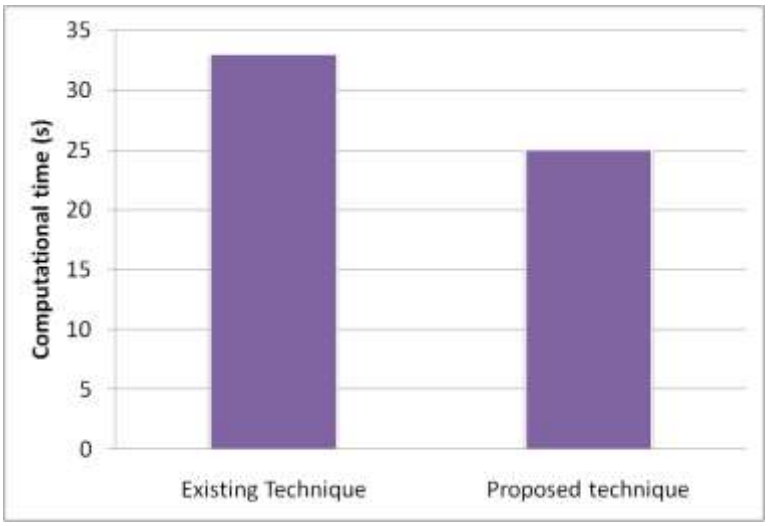


Fig. 3 Computation time

The computation of gain parameters of each cut-point with the data associated with them takes more time in general. The proposed method using hybrid optimization technique has better computation time than other approaches.

**Conclusion**

The big data classification is a wider area of research that provides vast information to various fields. In this study, the over-partitioning problem in using the decision tree technique is overcome by the proposed approach. But the problem of high computation time and optimal cut-point selection is not considered at a centre stage. Hence, an approach is proposed with the efficient hybrid optimization algorithm to establish the optimal cut-points. The approach, instead of selecting all cut-points, randomly selects the cut-points and computes the information gain. The computation of the information gain takes more time and hence a scheduling algorithm is used to schedule these computation tasks to efficient host nodes by estimating which node can process which tasks with higher efficiency based on the node data capacity using the optimization algorithms. Thus,the optimal cut-points are identified and the computation time is also reduced considerably especially using the Hybrid optimization approach in the proposed method.

**References**

[1]. Shinde, G., Deshmukh, S.N.: Sentiment TFIDF feature selection approach. Int. J. Comput. Commun. Eng. (2016)
[2]. A.C. Pandey, D.S. Rajpoot, M. Saraswat, Twitter sentiment analysis using hybrid cuckoo search method, Inf. Process. Manag. 53(4) (2017) 764–779.
[3]. Harold Robinson, Y,.Santhana Krishnan. R., Golden Julie, E., Raghvendra Kumar, Le Hoang Son, PhamHuy Thong: Neighbor Knowledge- based Rebroadcast Algorithm for minimizing the Routing overhead in Mobile Ad-hoc Networks, Ad Hoc Networks, 93, 1-13, (2019).
[4]. C.-M. Huang, C.-H. Shao, S.-Z. Xu, and H. Zhou, ``The social Internet of Thing (S-IOT)-based mobile group handoff architecture and schemes for proximity service," IEEE Trans. Emerg. Topics Comput., vol. 5, no. 3, pp. 425437, Jul./Sep. 2017.
[5]. Zhou, Z., Jin, X.L., Vogel, D.R., Fang, Y., Chen, X., 2011. Individual motivations and demographic differences in social virtual world uses: an exploratory investigation in second life. Int. J. Inf. Manag. 31 (3), 261–271.



ISSN: 2005-4238 IJAST

[6]. Harold Robinson, Y., Golden Julie, E.: MTPKM: Multipart Trust Based Public Key Management Technique to Reduce Security Vulnerability in Mobile Ad-Hoc Networks, Wireless Personal Communications, 109, 739–760 (2019).
[7]. Wang, G., Gunasekaran, A., Ngai, E. W., & Papadopoulos, T. (2016). Big data analytics in logistics and supply chain management: Certain investigations for research and applications. International Journal of Production Economics, 176, 98–110.
[8]. Liang, Po-Wei, and Bi-Ru Dai. Opinion Mining on Social Media Data. Mobile Data Management (MDM), 2013 IEEE 14th International Conference on. Vol. 2. IEEE, 2013.
[9]. Subramanian Balaji, Yesudhas Harold Robinson, Enoch Golden Julie,:GBMS: A New Centralized Graph Based Mirror System Approach to Prevent Evaders for Data Handling with Arithmetic Coding in Wireless Sensor Networks, Ingénierie des Systèmesd'Information, 24, 5, 481-490 (2019).
[10]. K. Shvachko, H. Kuang, S. Radia, R. Chansler, The hadoop distributed file system, in: 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies, MSST, IEEE, 2010, pp.1–10.
[11]. Riquelme, F., González-Cantergiani, P.: Measuring user influence on Twitter: a survey. Inf. Process. Manage. 52(5), 949–975 (2016).
[12]. Harold Robinson, Y., Rajaram, M.: Energy-aware multipath routing scheme based on particle swarm optimization in mobile ad hoc networks , The Scientific World Journal, 1-9 (2015).
[13]. Chard, K., Caton, S., Rana, O., Bubendorfer, K., 2010. Social Cloud: Cloud Computing in Social Networks, the IEEE 3rd International Conference on Cloud Computing, Miami, USA, 5–10 July.
[14]. Iqbal, R., Grzywaczewski, A., Halloran, J., Doctor, F., Iqbal, K., 2017. Design implications for task-specific search utilities for retrieval and reengineering of code. Int. J. Enterp. Inf. Syst. 11 (5), 738–757 (Taylor and Francis, pp 1751–7575).
[15]. R. Santhana Krishnan, E. Golden Julie, Y. Harold Robinson, S.Raja, Raghvendra Kumar, Pham HuyThong, Le Hoang Son, Fuzzy Logic based Smart Irrigation System using Internet of Things, Journal of Cleaner Production, Volume 252, 10 April 2020, 119902
[16]. G. Han, L. Zhou, H. Wang, W. Zhang, and S. Chan, ``A source location protection protocol based on dynamic routing in WSNs for the social Internet of Things," Future Generat. Comput. Syst., vol. 82, pp. 689697, May 2018.
[17]. Cheng, X., Meng, B., Chen, Y., Zhao, P., Li, H., Wang, T., Yang, D.: Dynamic table: a layered and configurable storage structure in the cloud. In: Bao, Z., et al. (eds.) WAIM 2012. LNCS, vol. 7419, pp. 204–215. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33050-6 21
[18]. Lakshminarayanan, K.; Santhana Krishnan, R.; Golden Julie, E.; Harold Robinson, Y.; Kumar, R.; Son, L.H.; Hung, T.X.; Samui, P.; Ngo, P.T.T.; Tien Bui, D. A New Integrated Approach Based on the Iterative Super-Resolution Algorithm and Expectation Maximization for Face Hallucination. Appl. Sci. 2020, 10, 718. doi: 10.3390/app10020718
[19]. X. Zhang, S. Ding, and Y. Xue, ``An improved multiple birth support vector machine for pattern classification," NeuroComputing, vol. 225, pp. 119128, Feb. 2017.
[20]. Salloum, S. A., Al-Emran, M., Monem, A. A., &Shaalan, K. (2017). A survey of text mining in social media: Facebook and twitter perspectives. Advances in Science, Technology and Engineering Systems Journal, 2(1), 127–133.
[21]. Balaji, S., Golden Julie, E.,Harold Robinson, Y., Raghvendra Kumar, Pham Huy Thong, Le Hoang Son : Design of a security-aware routing scheme in Mobile Ad-hoc Network using Repeated Game Model, Computer Standards & Interfaces, 66, (2019)
[22]. A. Hogenboom, B. Heerschop, F. Frasincar, U. Kaymak, and F. de Jong. (2014) "Multi-lingual support for lexicon-based sentiment analysis guided by semantics," Decision support systems, vol. 62, pp. 43–53.
[23]. A. Rasooli and D. G. Down, ``A hybrid scheduling approach for scalable heterogeneous hadoop systems," in Proc. SC Companion High Perform. Comput.,Netw., Storage Anal. (SCC), Nov. 2012, pp. 12841291.
[24]. Santhana Krishnan, R., Golden Julie, E., Harold Robinson, Y., Raghvendra Kumar, Le Hoang Son, Tong Anh Tuan, Hoang Viet Long, Modified Zone Based Intrusion Detection System for Security Enhancement in Mobile Ad-hoc Networks, Wireless Networks, 1-15, (2019)
[25]. A. Ahmad et al., "Toward modeling and optimization of features selection in big data based social Internet of Things," Future Generat. Comput. Syst., vol. 82, pp. 715726, May 2017.



ISSN: 2005-4238 IJAST

[26]. Samariya, D., Matariya, A., Raval, D., Babu, L. D., Raj, E. D., &Vekariya, B. (2016). A hybrid Approach for big data Analysis of cricket fan Sentiments in twitter. Paper presented at the proceedings of international conference on ICT for sustainable development.
[27]. Y. Harold Robinson, I. Jeena Jacob, E. Golden Julie, P. EbbyDarney, "HadoopMapReduce and Dynamic Intelligent Splitter for Efficient and Speed transmission of Cloud-based video transforming", IEEE - 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 400-404, IEEE.
[28]. Sreeja N.K., SankarAPattern matching based classification using Ant Colony Optimization based feature selection ,2015,Applied Soft Computing Journal,vol31,(2818),91-102
[29]. Bhuvaneswari K., Rauf H.A.Edgelet based human detection and tracking by combined segmentation and soft decision,2009,International Conference on Control Automation, Communication and Energy Conservation, INCACEC 2009,5204487



ISSN: 2005-4238 IJAST