Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 9, August 2021: 1534 – 1540

Research Article

Feature Driven Clustering Big Data Covid-19 Analytics

Govindraju G.N[1], B.K.Raghavendra[2], Raghavendra S.[3], Santosh Kumar J.[4]

Abstract

The rise in internet, IoT and web-services specially with blogs has enhance the demand of BigData, which demands robust and highly-efficient system to analytics, which will serve real time and accurate distributed data. The framework which will distribute the data for storing and computation that is parallelized computing have been found key driving force behind the BigData analytics; however, the system always lacks in optimal data pre-processing, feature sensitive computation and more importantly feature learning makes major at-hand solutions inferior, especially in terms of time and accuracy. The proposed model hypothesizes that an analytics solution with BigData features or characteristics must have the ability to process humongous, heterogenous, structured, unstructured and semi-structured multi-dimensional features to yield time-efficient and accuracy analytical outputs. To process analytical task our proposed model at first employs tokenization, followed by Word2Vec based semantic feature extraction using CBOW (Bag of words) and N-Skip-Gram methods. Unlike other clustering models, we propose a improved multi-objective GA (Genetic algorithm) IMOGA to serve dual purposes, first to improve centroid and second optimize the clusters. Our proposed model applied Euclidian distance information to perform centroid optimization, while Silhouette coefficient was applied to perform cluster validation and its optimization. Eventually, the optimal amalgamation of tokenization, Word2Vec word-embedding or feature extraction, and IMOGA K-Means clustering in parallel to the Spark distributed data framework exhibited better performance in terms of execution time and clustering. Our proposed model was found more effective with Skip-Gram Word2Vec feature extraction. Simulation results with a publicly available COVID-19 data exhibited better performance than existing K-Means based MapReduce distributed data frameworks.

Keywords: BigData Analytics, Spark Distributed Framework, Improved Multi-Objective Genetic Algorithm, K-Means Clustering, Word2Vec, Word-Embedding.

[1]Research scholar, Computer Science and Engineering Department, BGSIT, B G Nagar Mandya, VTU Belagavi, Karnataka.

[2]Professor and Head, Computer Science and Engineering Department, BGSIT, B G Nagar Mandya, VTU Belagavi, Karnataka.

[3]Associate Professor, Computer Science and Engineering Department, Christ Deemed to be University, Kengeri campus,Karnataka.

[4]Research scholar, Computer Science and Engineering Department, BGSIT, B G Nagar Mandya, VTU Belagavi, Karnataka.

1. INTRODUCTION

The high-pace rise in advanced computing, software technologies, internet and allied information and communication technologies, and web-services has inculcated socio-economic and scientific transition to make timely and optimally accurate decisions [1-3].

In sync with above stated key features of real-world BigData analytics demands, machine learning based approaches have always gained dominance [2][9][10][19][21]; though, their own dependency on allied pre-processing, feature extraction, selection and mapping has remained a challenge [2]. It indicates that a robust BigData analytics model can be realized only with optimal set of pre-processing, feature extraction, feature selection and classification systems [2][3]. Any lack of these key computing components might force the analytics model to perform inferior giving rise to false-positive and hence inaccurate outcome [2]. Moreover, applying a machine learning model with classical computing (data) environment might undeniably cause local-minima and convergence and hence can impact overall efficacy and veracity [2]. Though, to cope-up with high pace up-surge in BigData demands, machine learning driven models turn-out to be viable [2][9][13][19][21]. Noticeably, most of the classical machine learning based analytics models employ pre-processing, feature extraction (from the heterogenous, unstructured or unannotated data), feature selection and classification; however, their corresponding complexity and its impact on computational time has never been considered. It raises questions on their generalization, especially towards the contemporary analytics purposes [2]. Ironically, the selection of an optimal set-of computing elements in certain analytics paradigm has remained a challenge [2][3][9][19] that drives academia-industries to develop more efficient solution. Considering aforesaid facts, in this research the key emphasis is made on designing a state-of-art new and robust BigData analytics paradigm with better pre-processing, feature extraction, feature selection and classification system to meet contemporary analytics demands. Exploring in depth, it can be found that regression, reinforcement leaning and clustering are the key machine learning methods used these days towards BigData analytics. However, clustering based methods because of its ease of computation, higher scalability and minimum computational overhead enable (it) to be used in varied BigData analytics purposes [2]. Amongst the major machine learning, especially clustering algorithms, K-Means algorithms have been applied extensively towards BigData analytics purposes [20]. However, the random cluster centroid deployment, lack of cluster validation has remained an open challenge for K-Means based BigData analytics [2][21].

2.RELATED WORK

This section discusses some of the key literatures pertaining to machine learning based BigData analytics, parallelization in BigData analytics etc.

3. RESEARCH QUESTIONS

The proposed research effort made to intended to achieve a justifiable answer for the following research questions.

RQ1:*Is data-sensitive pre-processing and low-dimensional feature extraction be efficient towards BigData analytics?*

RQ2:*Is Word2Vec be efficient to ensure computationally-efficient BigData analytics?*

RQ3:*Is strategic implementation of heuristic models such as Improved Multi-Objective IMOGA based GA be efficientin optimal clustering for BigData analytics?*

RQ4:*Is IMOGA-K Means clustering algorithm with Spark distributed framework be efficient for BigData analytics?*

Thus, the answer for the above questions we have robust method ofBigData analytics serving timely as well as reliable decision support.

4. Our Contribution

The proposed research emphasison enhancing each comprising step including data pre-processing, feature extraction, feature sensitive clustering and cluster optimization. In our proposed BigData analytics model we focused on exploiting the efficacy of the different technologies such as machine learning, semantic feature embedding, evolutionary computing and Apache Spark distributed framework to design a state-of-art new and robust BigData analytics model. Noticeably, being a BigData analytics problem, which is expected to undergo Big data Vs-specific demands, we designed our proposed analytics model in such manner that it could address all allied aspects including large volume of data, multi-dimensional features, unstructured data etc., while accomplishing timely (i.e., signifying velocity), and accurate (say, veracity) analytics outcome. The strategic implementation of the overall proposed model encompassed the following key phase-wise processing elements.

1. *Data Collection*
2. *Pre-processing or Tokenization*
3. *Latent Semantic Feature Extraction*
4. *IMOGA K-Means algorithm-based Clustering for Prediction and/or Prescription task.*

The detailed discussion of the overall proposed model is given in the subsequent sections.

As already stated, the proposed IMOGA K-Means model has been applied over CORD-19 dataset, targeted to provide query driven clustered or segmented document support so as to help doctors, scientists, and researchers for making optimal and timely decisions. The simulation results and allied inferences are given in the subsequent sections.

5.Results and Discussions

Realizing the fact that the majority of the contemporary Big Data analytics model either applies classification systems, regression or the clustering models to perform data analysis and resulting prediction or prescription tasks.

In order to assess the performance of the proposed Big Data analytics model, we performed intra-model analysis as well as inter-model (performance) analysis. Here, intra-model assessment was specially performed to assess the performance with the different cluster configuration, nodes count, etc. While, the inter-model assessment targets to compare the performance by our proposed model with other existing approaches. The detail of these results and allied inferences is given as follows:

A. Intra-Model Assessment

In this performance assessment method, we focus on assessing the results by varying cluster counts and the node counts. Noticeably, it is always hypothesized that with lower cluster counts the feature can be limited and hence clustering can be easier; however, under heterogeneous and large dataset, (with likelihood of large feature types), retaining lower cluster count causes reduced Silhouette coefficient. This is because, there can be the probability that in a specific K cluster, the data-elements mapped or assigned might belong to another cluster (as well). In shows inappropriate clustering, and hence states inferior solution. On the contrary, increasing cluster count often gives a clustering model to assign data elements in the different clusters even based on very minute feature difference. It hypothesizes to exhibit higher Silhouette coefficient. Considering, these facts, to assess the performance of our proposed IMOGA K-Means clustering model, we varied the number of clusters and assess Silhouette coefficients.

Recalling the previous discussion that most of the heuristic methods including GA might undergo local minima and convergence problem, which can be more severe with humongous, heterogenous, unlabeled, high-dimensional data, we examined fitness achievement with the different feature sets. To achieve it, we applied IMOGA K-Means clustering algorithm with both CBOW and SG features. To test convergence probability, we considered smaller number of generation (Fig. 1 and Fig. 2). Here, we considered the total number of generations as 10 and estimated fitness value over CBOW (Fig. 1) and SG features (Fig. 2). Observing the results (Fig. 1 and Fig. 2), it can easily be found that the proposed SG feature (with n=1) achieves higher fitness value swift and therefore can avoid a large redundant computation even over large search space. On the other hand, the fitness estimation over CBOW (Fig. 1) indicates that due to relatively large semantic features learning over it turns out to be difficult, which could be alleviated by taking smaller window size. Note, we had considered the corpus size as 500, while the window size was considered as 5. On the contrary, SG method applied N=1, and therefore learning over more segmented features in SG is easier. Consequently, SG feature with N=1 exhibited better; however, it might be computationally heavier, which can be reduced by taking N=2, 3, or 6 (or higher value).

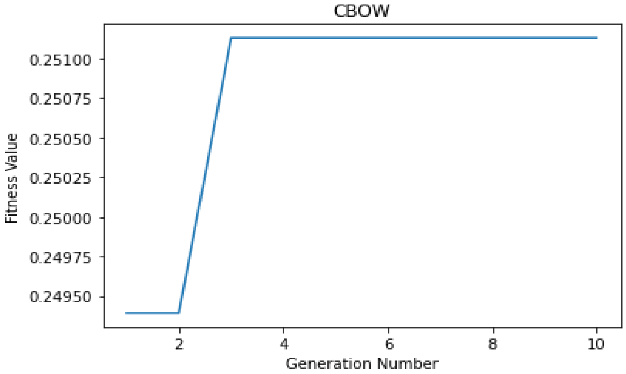

Fig. 1 Fitness value Vs. Generations in CBOW feature learning and clustering

Thus, observing performance (say, intra-model assessment), it can be stated that the proposed IMOGA K-Means based analytics model with SG feature achieves more accurate and time-efficient clustering which can be vital for the BigData analytics problems.

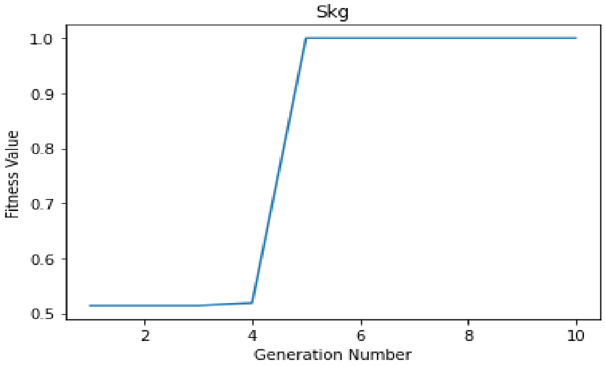

Fig. 2 Fitness value Vs. Generations in SG feature learning and clustering

B. Inter-Model Assessment

To assess the performance of our proposed heuristically transitioned and semantic feature driven clustering environment for bigdata analytics, we compared its performance with other existing BigData distributed frameworks or allied algorithms. Noticeably, exploring in depth it can be found that the different researchers have applied the different datasets to assess their performance and therefore simulating over the same benchmark is difficult. Considering this fact, we examined performance of the different existing methods as well as out proposed IMOGA K-Means model in terms of execution time.

6.Conclusion

This research work primarily focused on developing an enhanced distributed framework for BigData analytics, especially under typical large data with heterogenous, unstructured a multi-dimensional feature. Realizing the fact that merely employing distributed framework such as MapReduce and Spark can't yield optimal performance, until it doesn't address the key issues of data-heterogeneity, humongous size with unannotated data structure etc. In sync with this inference, this research proposed a state-of-art new and robust Spark distributed framework was developed with semantic word-embedding, and evolutionary computing assisted lightweight clustering algorithm. Here, the key intend was to enhance data quality and analytics-oriented suitability followed by time-efficient and accurate clustering to make optimal decisions. To achieve it, the proposed model at first tokenized the input data, which is a COVID-19 related data repository containing almost 4.5 lakhs of articles and statistical document related to COVID-19's symptoms, vaccines, current measures, health complications etc. Once tokenizing the inputs, a relevant dictionary was constructed. Subsequently, for each data instance, word-embedding methods CBOW and N-Skip Gram were applied to extract the feature. Noticeably, unlike N-skip gram (SG) which generated the context word for a given target words,. Subsequently, the extracted features were passed to the Spark framework which was programmed in such manner that it employed parallelized function of IMOGA for data clustering-based prediction and/or prescription. More specifically, we applied IMOGA for the centroid optimization in K-Means algorithm as well as clustering optimization. Here, IMOGA was applied in parallel to perform centroid optimization as well cluster verification, using Euclidean distance and Silhouette Coefficients, respectively. Noticeably, the enhancement in cluster-centroid ensured that the data-elements are properly clustered based on inter-cluster as well as intra-cluster information. On the contrary, Silhouette coefficient enabled cluster-verification guaranteed that each data elements have been placed to the most suitable and appropriate cluster and hence accomplished optimal clustering without imposing additional computational cost and time. Summarily, the use of SG at one hand enabled optimal set of feature extraction, while IMOGA-K Means algorithm guaranteed optimal clustering-based prediction or prescription for aforesaid COVID-19 BigData problem. The parallelized implementation of SG driven IMOGA K-Means enabled time-efficient analytics to serve real-world demands. Interestingly, with higher clusters and node configuration the proposed distributed framework exhibited better time efficiency. Moreover, higher Silhouette coefficient affirms optimality of clustering and hence its reliability towards real-world analytics purposes. Noticeably, the proposed Spark BigData analytics model was applied to perform content prediction and prescription over COVID-19 pandemic related data, and therefore it can be vital for the different tasks related to vaccine related research, query-driven data retrieval and visualization etc. However, the proposed model can be applied for any BigData analytics problems including text analytics, heterogenous data processing and allied analytics tasks.

7.COPYRIGHT

REFERENCES

[1] R. Krikorian. (2010). Twitter by the Numbers, Twitter. [Online]. Available: http://www.slideshare.net/raf_krikorian/twitter-by-the-numbers?ref=http://techcrunch.com/2010/09/17/twitter-seeing-6-billion-api-callsper-day-70k-per-second/

[2] A. L' Heureux, K. Grolinger, H. F. Elyamany and M. A. M. Capretz, "Machine Learning with Big Data: Challenges and Approaches," in IEEE Access, vol. 5, pp. 7776-7797, 2017.

[3] ABI. (2013). Billion Devices Will Wirelessly Connect to the Internet of Everything in 2020, ABI Research. [Online]. Available: https://www.abiresearch.com/press/more-than-30-billion-devices-willwirelessly-conne/

[4] W. Raghupathi and V. Raghupathi, ``Big data analytics in healthcare: Promise and potential," Health Inf. Sci. Syst., vol. 2, no. 1, pp. 1_10, 2014. [4] O.Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, ``Efficient machine learning for big data: A review," Big Data Res., vol. 2, no. 3, pp. 87_93, Sep. 2015.

[5] M. A. Beyer and D. Laney, "The importance of `big data': A definition," Gartner Research, Stamford, CT, USA, Tech. Rep. G00235055, 2012.

[6] V. Mayer-Schönberger and K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think. Boston, MA: Houghton Mifflin Harcourt, 2013.

[7] H. V. Jagadish et al., ``Big data and its technical challenges," Commun. ACM, vol. 57, no. 7, pp. 86_94, 2014.

[8] M. James, C. Michael, B. Brad, and B. Jacques, Big Data: The Next Frontier for Innovation, Competition, and Productivity. New York, NY: McKinsey Global Institute, 2011.

[9] J. Singh, "Real time BIG data analytic: Security concern and challenges with Machine Learning algorithm," 2014 Conference on IT in Business, Industry and Government (CSIBIG), Indore, India, 2014, pp. 1-4.

[10] M. Mohammadi, A. Al-Fuqaha, S. Sorour and M. Guizani, "Deep Learning for IoT Big Data and Streaming Analytics: A Survey," in IEEE Communications Surveys & Tutorials, 2018, vol. 20, no. 4, pp. 2923-2960.

[11] P. Bellini, F. Bugli, P. Nesi, G. Pantaleo, M. Paolucci and I. Zaza, "Data Flow Management and Visual Analytic for Big Data Smart City/IOT," 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, Leicester, United Kingdom, 2019, pp. 1529-1536.

[12] T. Chardonnens, "Big Data analytics on high velocity streams: specific use cases with Storm", Software Engineering Group, Department of Informatics, University of Fribourg, Switzerland, 2013.

[13] E. A. Mohammed, B. H Far, and C. Naugler, "Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends", BioData Mining 2014, pp. 7:22.

[14] I. A. Ajah, H F. Nweke, "Big Data and Business Analytics: Trends, Platforms, Success Factors and Applications", Big Data and Cognitive Computing, 2019, 3, pp. 1-32.

[15] R. Narasimhan and T. Bhuvaneshwari, ``Big data: A brief study,'' Int. J. Sci. Eng. Res., vol. 5, no. 9, pp. 350_353, 2014.

[16] F. J. Ohlhorst, Big Data Analytics: Turning Big Data into Big Money, vol. 15. Hoboken, NJ: Wiley, 2012.

[17] A. Kaplunovich and Y. Yesha, "Consolidating billions of Taxi rides with AWS EMR and Spark in the Cloud: Tuning, Analytics and Best Practices," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 4501-4507.

[18] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, ``Spark: Cluster computing with working sets,'' in Proc. 2nd USENIX Conf. Hot Topics Cloud Comput., 2010, p. 10.

[19] C.-T. Chu et al., ``Map-reduce for machine learning on multicore,'' in Proc. 20th Conf. Adv. Neural Inf. Process. Syst. (NIPS), 2006, pp. 281-288.

[20] A. C. Onal, O. BeratSezer, M. Ozbayoglu and E. Dogdu, "Weather data analysis and sensor fault detection using an extended IoT framework with semantics, big data, and machine learning," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 2017, pp. 2037-2046.

[21] H. Chiroma et al., "Progress on Artificial Neural Networks for Big Data Analytics: A Survey," in IEEE Access, vol. 7, pp. 70535-70551, 2019.

[22] Y. Yong and G. Xin_cheng, "A new minority kind of sample sampling method based on genetic algorithm and K-means cluster," 2012 7th International Conference on Computer Science & Education (ICCSE), Melbourne, VIC, Australia, 2012, pp. 126-129

[23] T. Ma, T. Wang, D. Yan and J. Hu, "Improved genetic algorithm based on K-Means to solve path planning problem," 2020 International Conference on Information Science, Parallel and Distributed Systems (ISPDS), Xi'an, China, 2020, pp. 283-286.

[24] Q. Liu, X. Liu, J. Wu and Y. Li, "An Improved NSGA-III Algorithm Using Genetic K-Means Clustering Algorithm," in IEEE Access, vol. 7, pp. 185239-185249, 2019.

[25] L. Minqiang, L. Jianwu and K. Jisong, "Genetic algorithms for auto-clustering in KDD," in Journal of Systems Engineering and Electronics, vol. 11, no. 3, pp. 53-58, Sept. 2000.