Research Article

ACCENTS Transactions on Information Security, Vol 5(18)
ISSN (Online): 2455-7196
http://dx.doi.org/10.19101/TIS.2020.517009

An efficient clustering mechanism in big data framework for data preprocessing and management

Shweta Kumari[1*], Kailash Patidar[2], Rishi Kushwah[2] and Gaurav Saxena[2]
M.Tech Scholar, Department of Computer Science Engineering, Sri Satya Sai University of Technology & Medical Sciences, Sehore, Madhya Pradesh, India[1]
Assistant Professor, Department of Computer Science Engineering, Sri Satya Sai University of Technology & Medical Sciences, Sehore, Madhya Pradesh, India[2]

Abstract

An efficient data handling mechanism has been applied based on epoch-based k-means associated fuzzy clustering (EKFC). In the first phase weights have been assigned to individual data segment presented based on the classification key metrics. It has been assigned automatically. Then weight preprocessing has been done in such manner to prune the unwanted weights. It has been pruned in such way to filter the weights which are not scalable. Then epoch-based k-means associated fuzzy clustering (EKFC) approach has been applied for data arrangement. First different epochs have been considered for the calculation of initial seeds values. These seeds have been considered after considering 100 epochs. After 100 epochs seeds have been determined. These seeds values have been used as the initial centroid for the k-means clustering. After the complete validation similar clusters from the two clustering approaches have been considered. In the next phase operational clustering has been performed. In the final phase threshold ranking has been performed. It has been performed for the final classification based on the above clusters. It will arrange in the order of threshold values. It will be used for the determination of the priority of the task in the big data environment. The results are found to be prominent in terms of classification accuracy.

Keywords

Big data, EKFC, Epochs, K-means, Fuzzy C-means.

1.Introduction

The term big data means the handling of high volume of data in terms of specific data requirement and arrangement [1]. The major aim is the data arrangement task which can be handled in such a way to provide velocity, variety and high volume [2−5]. In today’s world the amount of data generated is very high so by the use of big data framework, different computational aspects can be applied successfully [4−6]. The main capabilities of this environment are data processing with high volume, time saving and the better efficiency [8, 9].

Big data environment can be better understood by the following attributes [8−12]:

- Volume: The amount of data processed or generated comes under the volume category.
- Variety: It is the type of data under structured and unstructured category.
- Velocity: data or information processed or generated speed comes under the velocity category.
- Value: knowledge processing under the process of the above generation and processing in different levels.

The main objective is to apply epoch-based k-means associated fuzzy clustering (EKFC) for data management in big data environment.

2.Literature survey

In 2020, Li et al. [13] discussed about big data management in case of mineral resources. They have discussed about the system design process. It has been discussed in terms of resource management in terms of big data. The GIS technology has been discussed based on big data. It covers the resource management aspect along with the geological exploration. Their approach has improved the ability of production management.

[*] Author for correspondence

In 2020, Li [14] discussed about the applicability of big data for the improvement of postgraduate management. Their approach is based on big data management. Their results show that the applicability of the approach in terms of management of postgraduates in case of colleges and universitie.

In 2020, Chaves et al. [15] discussed about data mining technologies in terms of exploration of large volume of data. They have discussed different aspects of big data along with the applicability in business analytics. They have also discussed regarding the subject which may be capable in achieving positive ideal for the used and the companies.

In 2020, Chen et al. [16] discussed about the big data in terms of strategic management. They have discussed a novel strategic method based on big data for the cultural industry. Their results support the approach.

In 2020, Chunlei et al. [17] discussed the big data in terms of workload evaluation. They have suggested the major drawback is the lack of relevant models in the same domain. They have proposed natural language processing with machine learning algorithms. They have obtained the reference value and then the workload evaluation has been performed. Their result is found to be efficient in the identification and the prediction of attribute class impact factors.

In 2020, Deng et al. [18] discussed about the enterprise management in case of big data. They have discussed the complete characteristics along with the reengineering process. It is beneficial for the enterprise in case of big data.

In 2020, Du [19] discussed about the management of book materials. It has been discussed in terms of degree of information management. The management of materials and the book data has been analyzed in terms of big data visualization and knowledge discovery. It has been initiated in terms of data management, technology and electronic libraries.

In 2020, In [20] discussed about the improvement in terms of entrepreneurial practice ability. It has been discussed in terms of college students' Huizhou merchants. The environment considered here are big data. They have considered network sharing platform. They have discussed the cross regional exchange of the entrepreneurship.

In 2020, Li et al. [21] discussed about the complex macroeconomic environment. It has been discussed in terms of trust product allocation. The environment suggested is the big data. They have suggested that for financial innovation big data may found to be useful. They have extended their impact and suggested that China's trust industry needs to actively adopt big data technology.

In 2020, Ha and Back [22] discussed about the use of social network services (SNS). It has been discussed in terms of data collection, data processing and data analysis. They have suggested that it takes material resources as well as time. They have proposed data filtering algorithm. It has been used for the garbage data filtering purpose. Their filtering accuracy found to be improved due to recursive learning approach. They have achieved 70% accuracy.

In 2020, Kesheng et al. [23] discussed about the big data methods. It has been discussed for the behavior patterns understanding. This can be done in terms scientific quantitate analysis. They have considered data mining techniques. It has been used for the study and analysis of student campus network behavior analysis. They have used data support operation for the student affairs based on the scientific development.

In 2020, Neves and Cruvinel [24] discussed regarding the semantic model. It has been discussed for the digital database’s structuring. The data has been originated from big data. They have examined different structure. These structures belong to structured, semi-structured and unstructured data. It has been used for agricultural risk management. Their approach is beneficial for Data Mart and Data Warehouse (NoSQL). Their configuration covers agriculture data which involves multiple modes operating environment.

In 2020, Youzhuo et al. [25] discussed bout the big data. It has been discussed in terms of big data and IT industry. They have suggested a retrieval system. It is based on Lucene architecture. It has been used based on the integrated architecture. It has been used in big data framework. It has been used in terms of distribution management system. Their proposed application has the capability of information retrieval in the fastest way and it is also convenient.

In 2020, Zhang [26] discussed about the big data prevention technology. They have used the combination of big data technology along with the financial market. It is useful in monitoring the financial market. They have suggested the combination of big data along with the financial market analysis.

In 2020, Zeng [27] suggested the application of big data. It has been used for military logistic support. It has been used for logistics support. It has been based on big data. It has been used in military logistics support.

In 2018, Pillmann et al. [28] discussed about the mobile Internet of Things for the sensor utilization and vehicle enabling. It has been addressed with the big data marketplace which has been leverage which exploits the crows-sourced sensor data. It has been used for the harmonized data model. They have developed common vehicle information model. It has been developed based on generic representation for the whole data value and processing chain. It consists of sensor data though different components and data value process chain in case of big data environment.

In 2020, Yang [29] proposed a big data-based application for the financial field. It has been processed for the analysis the possible risks. Their approach is based on big data. It has been used for the application platform. They have suggested that their application is useful in the personal privacy data and promote the standardization of big data.

3.Methods

An efficient data handling mechanism has been applied based on epoch-based k-means associated fuzzy clustering (EKFC). This approach is divided into following phases:

- Weight assignment
- Weight preprocessing
- EKFC
- Operational clustering
- Threshold classification

Weight assignment

In the first phase weights have been assigned to individual data segment presented based on the classification key metrics. It has been assigned automatically. It has been done based on the key values of the content. In our approach content based key value assignment has been done.

Weight preprocessing

Weight preprocessing has been done in such manner to prune the unwanted weights. It has been pruned in such way to filter the weights which are not scalable. It means if the weights are 1-10 then the minimum and maximum range value is in between 1 and 10. For lower value 1 is assigned and for the higher value 10 is assigned.

EKFC

In this phase EKFC approach has been applied for data arrangement. First different epochs have been considered for the calculation of initial seeds values. These seeds have been considered after considering 100 epochs. After 100 epochs seeds have been determined. These seeds values have been used as the initial centroid for the k-means clustering. The same value termination approach has been used for k-means clustering. These clusters are then validated by fuzzy c-means approach. After the complete validation similar clusters from the two clustering approaches have been considered.

Operational Clustering

In the next phase operational clustering has been performed. In this phase different operational values have been used for the clustering validation. After the complete validations the clusters have been moved for the final classification.

Threshold classification

In the final phase threshold ranking has been performed. It has been performed for the final classification based on the above clusters. It will arrange in the order of threshold values. It will be used for the determination of the priority of the task in the big data environment. The flowchart of the working process is shown below in *Figure 1*.

4.Results

In this section the comparison is shown based on E-K-means and FCM clustering algorithms. It has been found that the variations found in our method is minor so the approaches applied is capable in proper clustering. *Figure 2* shows the comparative accuracy for the complete set applying EKFC. *Figure 3* shows the average accuracy for the complete set applying EK. *Figure 4* shows the average accuracy for the complete set applying FCM. The results clearly show that the variations are minor in case of both the algorithms. FCM is found slightly better. Otherwise the performance is approximately same in data management thresholding. EK shows the epoch K-means and FCM for fuzzy c-means algorithm.

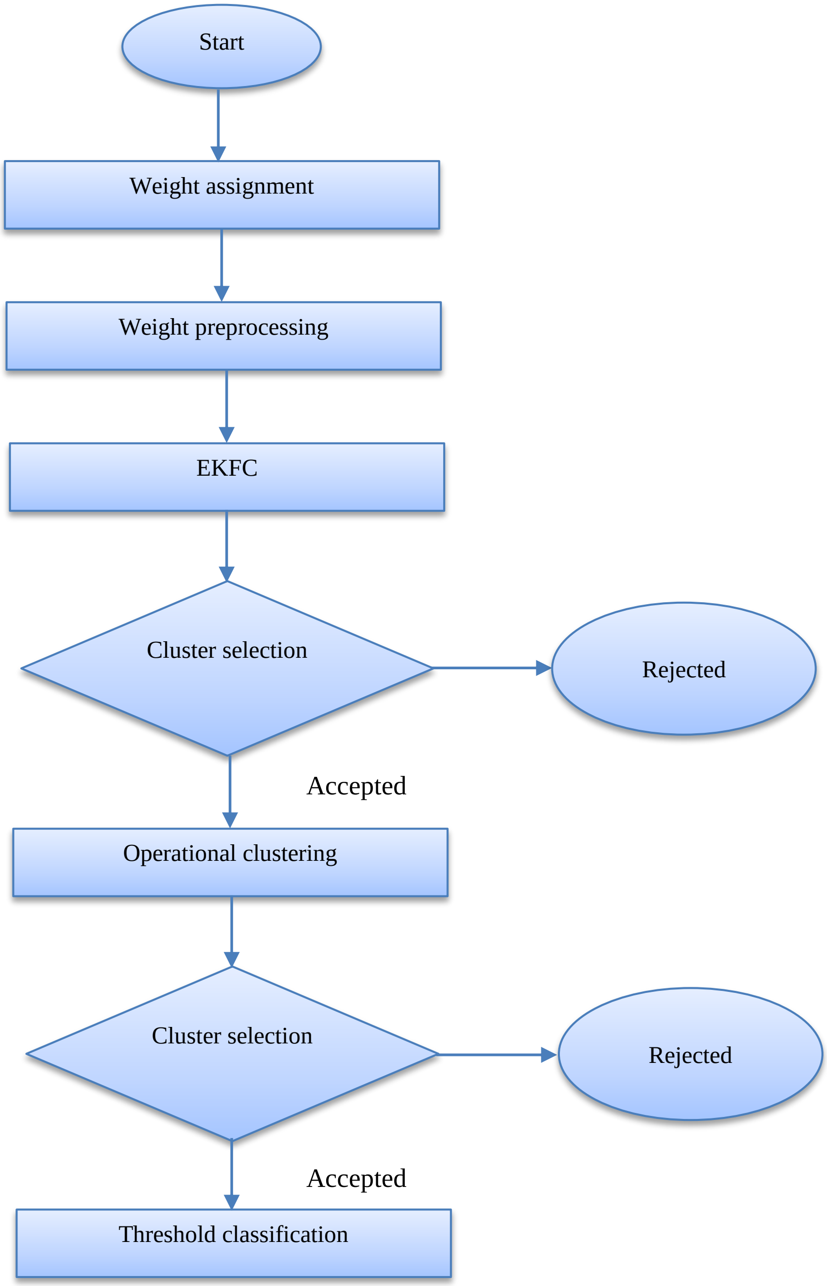

Figure 1 Flowchart of the complete procedure

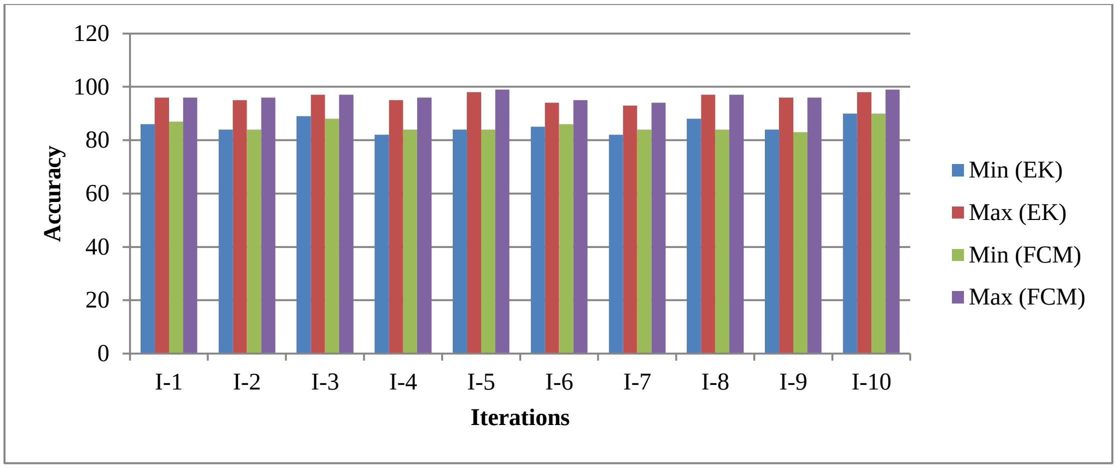

Figure 2 Comparative accuracy for the complete set applying EKFC

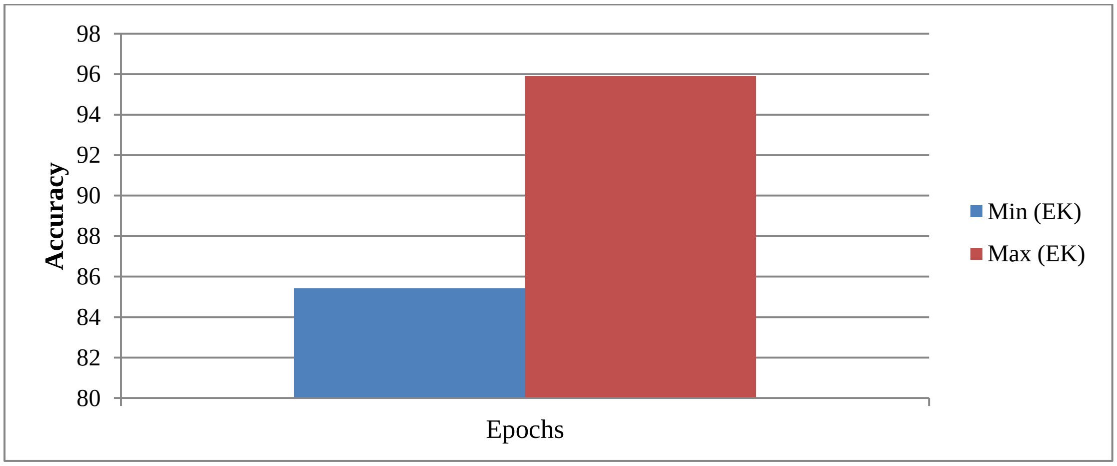

Figure 3 Average accuracy for the complete set applying EK

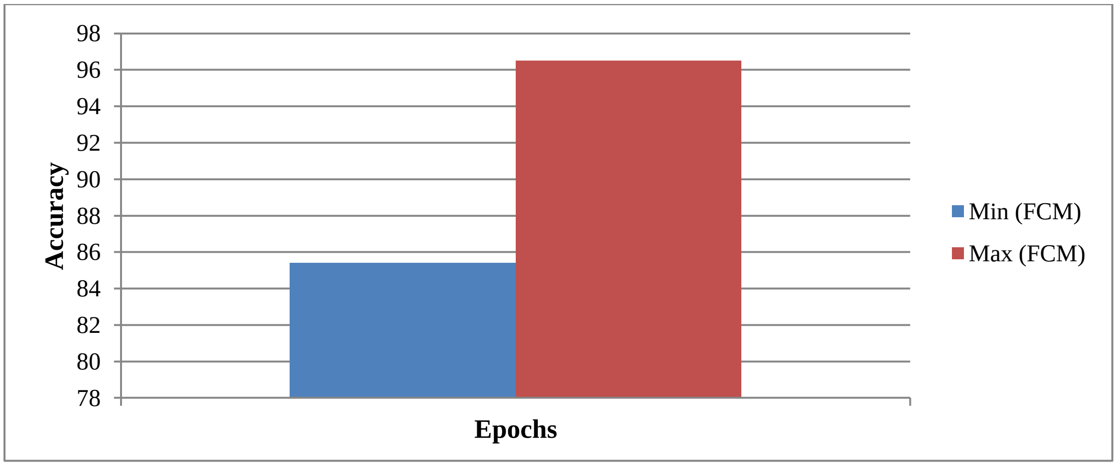

Figure 4 Average accuracy for the complete set applying FCM

5.Conclusion

An efficient data handling mechanism has been applied based on epoch-based k-means associated fuzzy clustering (EKFC). This approach consists of weight assignment where the weight values have been assigned for each process or for the data segment. Then EKFC algorithm has been applied for the further clustering. The matching clusters have been passed for the further process. Then operation clustering has been applied on the obtained clusters and the similar clusters from the both epochs k-means and FCM have been processed. Then final threshold-based classification has been applied to obtain the better results. The results clearly show that the variations are minor in case of both the algorithms. FCM is found slightly better.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Tsai CW, Lai CF, Chao HC, Vasilakos AV. Big data analytics: a survey. Journal of Big data. 2015; 2(1):1-32.

[2] Sivarajah U, Kamal MM, Irani Z, Weerakkody V. Critical analysis of Big Data challenges and analytical methods. Journal of Business Research. 2017; 70:263-86.

[3] Hussin SK, Omar YM, Abdelmageid SM, Marie MI. Traditional machine learning and big data analytics in virtual screening: a comparative study. International Journal of Advanced Computer Research. 2020; 10(47):72-88.

[4] Kaur S, Baghla S. Harnessing supremacy of big data using hadoop for healthy human survival making use of bioinformatics. International Journal of Advanced Technology and Engineering Exploration. 2018; 5(48):460.

[5] Mtey MM, Dida MA. Towards interoperable e-Health system in Tanzania: analysis and evaluation of the current security trends and big data sharing dynamics. International Journal of Advanced Technology and Engineering Exploration. 2019; 6(59):225-40.

[6] Bertot JC, Gorham U, Jaeger PT, Sarin LC, Choi H. Big data, open government and e-government: issues, policies and recommendations. Information polity. 2014; 19(1, 2):5-16.

[7] Pouyanfar S, Yang Y, Chen SC, Shyu ML, Iyengar SS. Multimedia big data analytics: a survey. ACM Computing Surveys (CSUR). 2018; 51(1):1-34.

[8] Dhas JJ, Vigila SM, Star CE. Forecasting of stock market by combining machine learning and big data analytics. In international conference on soft computing systems 2018 (pp. 385-95). Springer, Singapore.

[9] Omollo R, Alago S. Data modeling techniques used for big data in enterprise networks. International Journal of Advanced Technology and Engineering Exploration. 2020; 7(65):79-92.

[10] Atat R, Liu L, Wu J, Li G, Ye C, Yang Y. Big data meet cyber-physical systems: a panoramic survey. IEEE Access. 2018; 6:73603-36.

[11] Shobha K, Nickolas S. Time domain attribute based encryption for big data access control in cloud environment. ACCENTS Transactions on Information Security. 2017; 2(7):73-7.

[12] Liebowitz J, editor. Big data and business analytics. CRC press; 2013.

[13] Li D, Gong Y, Tang G, Huang Q. Research and design of mineral resource management system based on big data and GIS technology. In 5th IEEE international conference on big data analytics (ICBDA) 2020 (pp. 52-6). IEEE.

[14] Li J. Research on the management mode of graduate students in colleges and universities based on big data. In international conference on intelligent transportation, big data & smart city (ICITBS) 2020 (pp. 617-20). IEEE.

[15] Chaves A, Moura Í, Bernardino J, Pedrosa I. The privacy paradigm: an overview of privacy in business analytics and big data. In 15th iberian conference on information systems and technologies (CISTI) 2020 (pp. 1-6). IEEE.

[16] Chen C, Zuo R, Ni H. Research on the application of big data in the field of strategic management of cultural industry. In international conference on intelligent transportation, big data & smart city (ICITBS) 2020 (pp. 795-8). IEEE.

[17] Chunlei Z, Yin J, Qianli X. The workload assessment of national grid big data projects based on content recommendations and text classification. In 5th international conference on cloud computing and big data analytics (ICCCBDA) 2020 (pp. 482-90). IEEE.

[18] Deng L, Ye S, Zhou X. Research on two-way integration business process reengineering under big data. In IEEE 5th information technology and mechatronics engineering conference (ITOEC) 2020 (pp. 1699-703). IEEE.

[19] Du Q. Research on the application of big data in book management. In 12th international conference on measuring technology and mechatronics automation (ICMTMA) 2020 (pp. 773-5). IEEE.

[20] In Gao X. Research on the entrepreneurship practice of college huizhou merchants based on big data. In international conference on intelligent transportation, big data & smart city (ICITBS) 2020 (pp. 604-7). IEEE.

[21] Li D, Gong Y, Ren M, Li D. The Research and design of trust business management and analysis system based on big data technology. In 2020 5th international conference on big data analytics (ICBDA) 2020 (pp. 68-72). IEEE.

[22] Ha IK, Back BH. Effective garbage data filtering algorithm for SNS big data processing by machine learning. In 2020 international conference on artificial

intelligence in information and communication (ICAIIC) 2020 (pp. 520-4). IEEE.

[23] Kesheng L, Yikun N, Zihan L, Bin D. Data mining and feature analysis of college students' campus network behavior. In 5th IEEE international conference on big data analytics (ICBDA) 2020 (pp. 231-7). IEEE.

[24] Neves RA, Cruvinel PE. Model for semantic base structuring of digital data to support agricultural management. In 14th international conference on semantic computing (ICSC) 2020 (pp. 337-40). IEEE.

[25] Youzhuo Z, Yu F, Ruifeng Z, Shuqing H, Yi W. Research on lucene based full-text query search service for smart distribution system. In 3rd international conference on artificial intelligence and big data (ICAIBD) 2020 (pp. 338-41). IEEE.

[26] Zhang T. Research on big data's prevention technology of financial systemic risk. In international conference on intelligent transportation, big data & smart city (ICITBS) 2020 (pp. 663-6). IEEE.

[27] Zeng Y. Analysis on the influence and countermeasures of big data in military logistics support. In international conference on intelligent transportation, big data & smart city (ICITBS) 2020 (pp. 648-51). IEEE.

[28] Pillmann J, Sliwa B, Wietfeld C. The AutoMat CVIM-a scalable data model for automotive big data marketplaces. In 19th IEEE international conference on mobile data management (MDM) 2018 (pp. 284-5). IEEE.

[29] Yang R. Research on the risk and supervision method of big data application in financial field. In international conference on intelligent transportation, big data & smart city (ICITBS) 2020 (pp. 695-8). IEEE.