Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018)
IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2

Partition Tuning based Bagging technique to Skew Handling

Kanak Meena
Indira Gandhi Delhi Technical University for Women,CSE
kanak556@gmail.com

Dr. Devendra K.Tayal
Indira Gandhi Delhi Technical University for Women,CSE
dev_tayal2001@yahoo.com

Abstract: In this paper, we have proposed a partitioning technique to handle big data. MapReduce is very sensitive to data skewness so there is a need to develop such technique which handle the data skewness and do not affect the performance of the task. To overcome this problem we have proposed a partitioning algorithm name as PTBSH (Partition Tuning based Bagging technique to Skew Handling). It ensures the even distribution of data with the help of frequency keys and bagging. Proposed algorithm has been compared with known existing methods in terms of data skewness, data locality and runtime. Experiments have been performed on seven data sets which have been extracted from the UCI repositories.

Keywords: Data Skewness, Hadoop, MapReduce, Bagging, Big Data.

I. INTRODUCTION

The large amount of data generated by the healthcare [1-2] industry and this enormous amount of data badly needs some method to handle it efficiently. Medical data contains large amount of veracity in data. Healthcare is an exceedingly data escalated industry where information are driven by record keeping, consistence and regularity requirements, and patient care. These different data information incorporate radiology pictures, clinical records, human genetics records, and population data genomic sequence. The utilization of huge information in healthcare offers many attractive opportunities while posing significant challenge [1-3]. In any case, traditional data [4] processing and analytical calculations can't fulfill the necessities of big healthcare data and cloud computing. Luckily, advancement in information management, such as parallel computational models as MapReduce [4-19], can be connected to process and investigate different and large scale datasets. However, big data are so expensive, large and complex that they can't be overseen under traditional strategies [4-12]. When we deal with big data so many problems occur one is data skewness due to complexity of data and uneven distribution of data among the nodes [11-16]. There are various parameters of big data which makes it complex and difficult understand, briefly defining the parameters are as follows:-

1. **Volume**: Huge amount of data, hence volume is the foremost characteristic of big data.
2. **Variety**: Variety demonstrates the various sorts of structured and unstructured data in various configurations [4]. Not at all like traditional databases which included information as it were tabular, spread-sheets and relational model whereas big data includes Email, PDFs, text, images, videos etc.
3. **Velocity**: It refers to the speed of data creation/generation, processing and retrieval of information.
4. **Variability**: Inconstancy in big data refers to the varieties in a solitary information type. Big data has numerous varieties of information [4], however varieties in every assortment brings about variability.
5. **Veracity**: It refers more to the provenance or reliability of the information source, it's specifically, and that it is so significant to the investigation based on it. Information of the data's veracity thusly causes us better comprehend the dangers related with examination and business choices in view of this specific data sets [4-6].
6. **Validity**: Its meaning as,to the exactness of the data, it tells how much precise or correct an data is for its usage in a particular task or execution. While veracity concentrates more on the meaningfulness of the outcomes [5].
7. **Volatility**: It is in the context with the measure of span for which we require the data. There might be some which has a lifetimes span, yet there may be situations where data may not be applicable after some time, and is no longer of any value.
8. **Visualisation**: It is a standout amongst the most important characteristics [6]. Showing the bits of knowledge in type of visuals for example, diagrams, charts, graphs etc. is constantly more

preferable and best rather than relational tables and reports.

9. **Value**: Value can be found in huge data, including understanding your clients better, focusing on them and target them as per their need [6-10].

As healthcare is very important aspect of knowledge, there is lot of scope of improvement for the retrieval of information. In this paper we are presenting the partition technique which ensures the even distribution of data to handle the data skewness name as Partition Tuning based Bagging [8-10] technique to Skew Handling (PTBSH). There are different types of skewness viz map side and reduce side, here, we focus to reduce the skewness at reducer side only. Bagging means Bootstrap aggregation [8-10], used to classify the data with good accuracy & it is a technique used to provide weight to the task and help in prioritizing the task as per our need. Several studies have been accounted for that they bagging and boosting techniques in the field of healthcare are very effective [6]. These studies have connected distinctive ways to deal with group the information with high classification accuracy (Refer section III for more details.)

This paper has been divided in six section which are dedicated to the specific tack as follows; section II is detailing the literature review of the proposed work, section III is dedicating to the proposed technique, section IV briefing the experimental details, section V shows the results of the experiments in graphical manner and section VI outlines the conclusion of the proposed works.

II. RELATED WORK

This section is dedicated to the literature review in respect of the proposed work. We consider a situation when computation load is unbalanced among the map side and reduce side is considered as map skew and reduce skew. Hadoop[4-7,20] is an framework which has hash partitioning as a default partition method. But it does not ensure the equal distribution of the data among nodes and data skewness is also very high. Skewtune[11-12] is the technique which automatically mitigates the skewness but it is user defined only. Skew Reduce [12] it is a technique which handle the skewness but manually, user need to process everything manually. LEEN [13] it is a fairness scheme which ensure the even distribution of data before the shuffle phase. But it achieves only 40% performance enhancement and it cannot handle the very large amount of data, causes unbalanced distribution and high skewness in output. Closer [19] it is a single stage partitioning method and handle the data skewness effectively but its disadvantage; it is very time consuming and seriously affecting the performance. Sampling-based partitioning in MapReduce for skewed data [14] is a technique based on pre-sampling partitioning method to deal with skewed data but its overhead time during the shuffling phase is very high. PTSH [6] is a technique based on two stage partition scheme; it also uses the virtual partitioning scheme. Shortcoming of this technique are very time consuming, overheads are high in shuffling phase due to two stage strategy. It cannot handle the alternate combination of the attributes. It handles only continuous relation of the attributes. To overcome all the drawback of these techniques we have proposed the partition techniques name as Partition Tuning based Bagging technique to Skew Handling (PTBSH). It is an extension of the PTSH and in respect of MapReduce framework. Proposed technique handles the alternate as well as continuous combinations of the attributes.

III. PROPOSED TECHNIQUE

This section has been discussed about the proposed technique for the medical data. As it is increasing enormously day by day and to manage that data is getting very difficult. This massive data management needs scalable solutions. So we have designed a technique to handle the data skewness problem for medical data. This technique is an extension of the existing technique PTSH [6]. But this approach has some restrictions and shortcomings: Partitions [21-24] must be handling in a continuous and distributed mode only. It cannot distribute the data evenly as well as it cannot assign the weight to the data but with the help of bagging it can distribute the data with weight. Their overheads are high due to the long shuffling time. To overcome the shortcoming of the existing techniques we have proposed partitioning algorithm.

For example : (R_1,R_2,R_3,R_4, R_5) repartitioning will be handled in continuous mode only, so results will be (R_1,R_2), (R_3,R_4) &(R_5) or (R_1,R_2,R_3),(R_4, R_5) it cannot be like (R_1, R_4), (R_2,R_5), (R_3). This technique only focuses on continuous partitioning scheme. Continuous partitioning can be ensured by virtual repartitioning in spilled files. This scheme cannot handle the asynchronous mode of the attributes. The communication cost of this scheme is also very high during the shuffling phase. PTSH cannot handle the alternative pairing of the attributes from the data. Its time complexity is also very high. But the proposed technique can work in any manner either continuous or alternative. As well as, it ensures the less overhead

because the shuffling time is less. As we have performed on different datasets. For the accurate classification we have used the concept of bootstrap aggregation. This algorithm is not restricted up to medical data only. Further we have presented the pseudo code of the proposed algorithm.

ALGORITHM

This is the algorithm which is used to design this partitioning technique PTBSH. Where K(key), A(sequence of continuous combination),B(Sequence of the alternate combination), D (d_1, d_2.....d_i) presents the subsequence of the continuous combination of the attributes whereas C(c_1, c_2.....c_i) presents the subsequence of the alternate combination of the attributes. Sum represents by "S".

1. Data : Q :(Q_1,Q_2….Q_n), K
2. Output: V: an index of subsequence
3. $W_g = \frac{1}{G}$ for all g= 1,....G
4. Low← max{d_i}
5. High← $\sum_n^1 di + 1$
6. Num ← 1
7. $\theta \widehat{aggr} \leftarrow 2$
8. While (low<high) do
9. mid ← low + $\frac{(high - low)}{2}$ +W_g
10. for each ($d_i \in A$) & ($c_i \in B$) do
11. S ← S+ d_1 + c_i + $\theta \widehat{aggr}\, \iota$
12. If (S > mid) then
13. Num ++
14. S ← d_l ;
15. V ← $V\ U(i)$; end
16. If (num ≤ K) then
17. Low← $mid + 1 + \widehat{\theta g}$
18. end
19. end
20. return V

IV. EXPERIMENTAL STUDY

This section provides the information about the performed experiments for the proposed approach. We have used the 9 node MapReduce cluster set for the analysis of experiments. Datasets for this experiment has been taken from UCI repositories, below table 1, contains the information about them:

Name of Data Set	No of Attributes	No of Instances	Types of Attribute
Poker Hand	11	1025010	Categorical, Integer
KDD CUP_1999	42	4000000	Categorical, Integer
KDD CUP_1998	481	191779	Categorical, Integer
Iris	4	150	Real
Cover Type	54	581012	Categorical, Integer
Breast Cancer	10	286	Categorical, Integer
Heart Diseases	75	303	Categorical, Integer

Table 1: Dataset details used for experiments

Measure of data Skewness

In this paper, we have used the method of coefficient of variance to calculate the data skewness [4-15] by following the ZipF distribution law [16]. A few distribution of information, for example, the Bell Curve, are symmetric. This implies that the right and the left parts of the distribution of data are immaculate identical representations of each other (mirror image). Not every distribution of information is symmetric. We have observed that data skew emerges out of the physical properties of items and hotspots on subsets of the entire domain.

$$\text{Cov} = \frac{stdev}{mean} \times 100\% \ldots (1)$$

Due to the use of fairness and accurate classifying techniques name as, bagging for the distribution of the data at reducer side. The coefficient of variance is achieved by the PTBSH is better than other compared methods. As shown in below table 2:-

Used Method	Cov
Hadoop	77%
Closer	25%
LEEN	15%
PTSH	11%
PTBSH	9.5%

Table 2. Calculation of the data skewness

Measure of Data Locality

Data locality measure [6-8] is the important parameter for the evaluation of performance of partition scheme. In this paper it is the sum of the frequencies of keys in nodes with the sum of mean aggregation, which are partitioned to that of the

frequencies and aggregation during the shuffle time. Used formula for the calculation has been mentioned below:

$$\text{Locality}_{\min} = \frac{\sum_{i=1}^{K} min_{1\le j\le n}}{\sum_{i=1}^{K} FK_i} FK_i^j + \widehat{\theta_{aggr}} \quad .....(2)$$

$$\text{Locality}_{\max} = \frac{\sum_{i=1}^{K} max_{1\le j\le n}}{\sum_{i=1}^{K} FK_i} FK_i^j + \widehat{\theta_{aggr}} \quad .....(3)$$

Where $min_{1\le j\le n} FK_i^j + \widehat{\theta_{aggr}}$ indicates the minimum frequency of key (k_i) in data node n^i , $\widehat{\theta_{aggr}}$ is the mean of aggregation and $max_{1\le j\le n} FK_i^j + \widehat{\theta_{aggr}}$ is the maximum frequency key k_i in data node n^i , $\widehat{\theta_{aggr}}$ is the mean of aggregation value. This table3 shows the result range of locality measure in different methods of partitioning on used dataset KDD CUP_1999.

Used Method	Data Locality Range
Hadoop	5%
Closer	1-15%
LEEN	1-18%
PTSH	1-14%
PTBSH	1-13%

Table 3. Calculation of the data locality

The detail performance of the existing techniques and the proposed technique with their runtime on the different dataset which has been discussed in below table 4:-

Dataset	Methods	Node with Minimum Load- Runtime (Sec)	Node with Maximum Load- Runtime (Sec)
Poker Hand	Hadoop	10	90
	PTSH	25	60
	PTBSH	37	45
KDD CUP_1999	Hadoop	6	101
	PTSH	36	83
	PTBSH	42	65
KDD CUP_1998	Hadoop	21	179
	PTSH	62	135
	PTBSH	82	90
Iris	Hadoop	18	226
	PTSH	81	185
	PTBSH	105	105
Cover type	Hadoop	21	245
	PTSH	76	211
	PTBSH	101	134
Breaset Cnacer	Hadoop	4	45
	PTSH	15	30
	PTBSH	19	21
Heart Diseases	Hadoop	11	76
	PTSH	24	45
	PTBSH	37	15

Table 4. Shows the details of the runtime on different datasets

V. RESULTS ANALYSIS

This section presents the graphical presentation of the experimental results. Different graphs have been created by the values obtained during the experiments. It is a graph which has been designed by using the table 2, it shows the coefficient of variance of different techniques by using the KDD CUP _1999 datasets. Where we have found our proposed method shows the least value of data skewness among the existing techniques. In fig 2, shows the relationship of the data locality and datasets with respect to the different techniques. How they are working on the used datasets for the experiments. Figure3 shows the runtime of different techniques on used datasets.

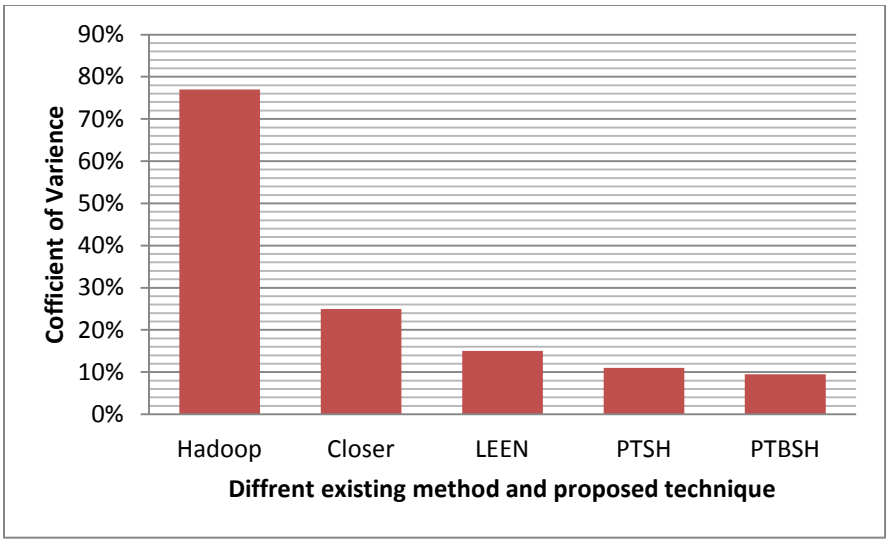

Fig 1. It is a graph which shows the coefficient of variance of different techniques

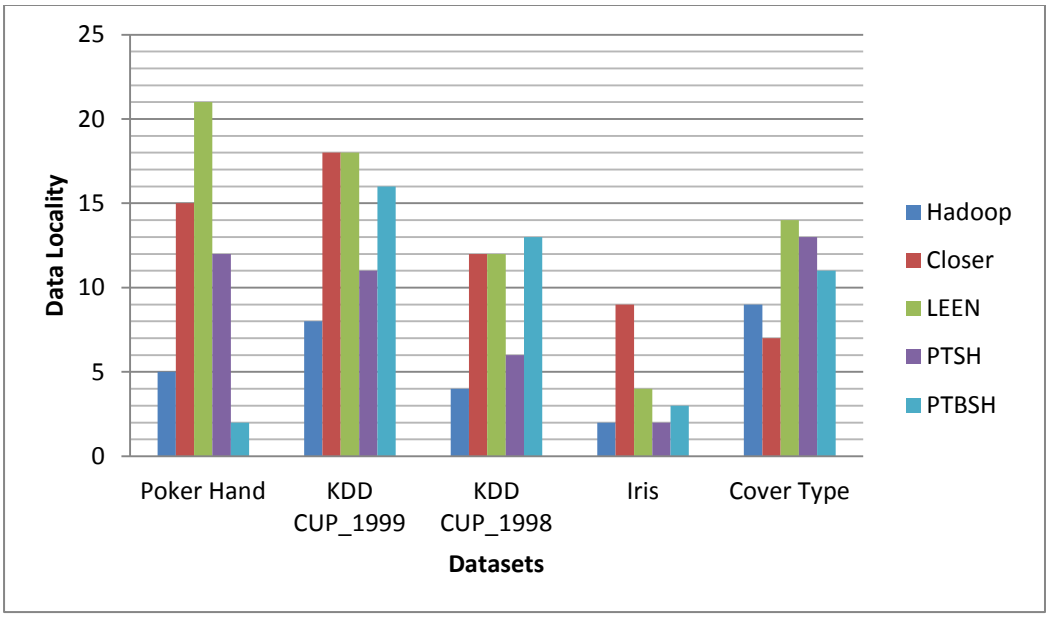

Fig 2. Shows the range of data locality

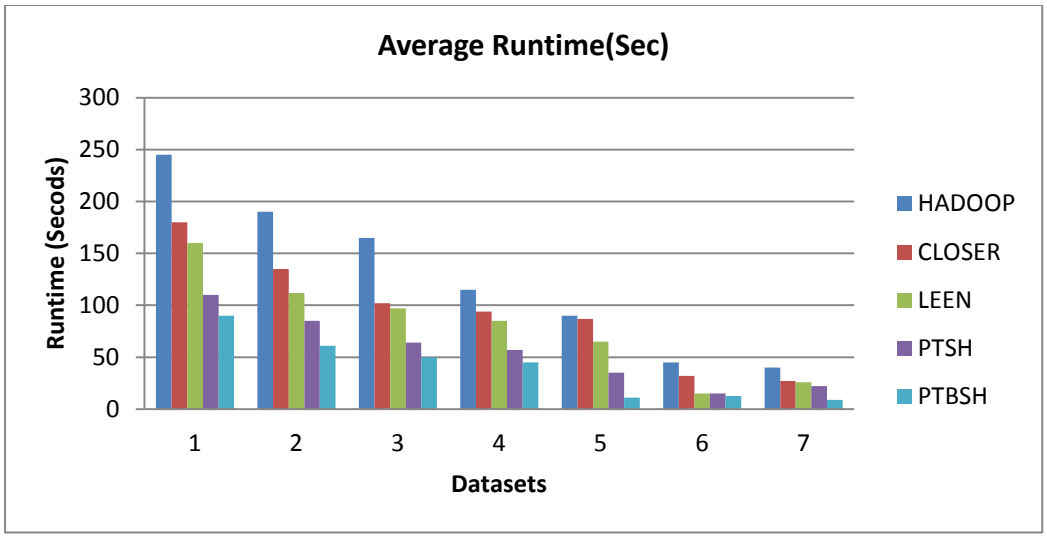

Fig 3. shows the runtime of the different partition techniques for dataset (1-5) by using table 1.

VI. CONCLUSION

This paper presents the new proposed partitioning techniques namely PTBSH. We have applied the bagging concept as well as for the better accuracy in the classification. It handle the reduce side skewness only. All the experiments performed on 9 nodes MapReduce clusters. Where, we have observed our proposed techniques performed outstanding with respect to native Hadoop, Closer, LEEN and PTSH. It was found that skewness and workload balance simultaneously influenced the efficiency of MapReduce framework. PTBSH is having very less overhead during the shuffling phase with respect to the other existing methods. It handles the both combinations of the attributes alternate and continuous. PTBSH is not very time consuming and it ensures the even distribution of the data among the data nodes as compared to the existing methods.

REFERENCES

1. W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," Health Information Science and Systems, vol. 2, no. 1, pp. 1–10, 2014.
2. D. Tomar and S. Agarwal, "Asurvey on data mining approaches for healthcare," International Journal of Bio-Science and Bio-Technology, vol. 5, no. 5, pp. 241–266, 2013.
3. Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*, *5*(5), 241-266.
4. Doulkeridis, C., & Nørvåg, K. (2014). A survey of large-scale analytical query processing in MapReduce. *The VLDB Journal*, *23*(3), 355-380.
5. K. Shim, "MapReduce algorithms for big data analysis," Proceedings of the VLDB Endowment, vol. 5, no. 12, pp. 2016–2017, 2012.
6. Gao, Y., Zhou, Y., Zhou, B., Shi, L., & Zhang, J. (2017). Handling Data Skew in MapReduce Cluster by Using Partition Tuning. *Journal of Healthcare Engineering*, *2017*.
7. Y. C. Kwon, M. Balazinska, B. Howe, and J. Rolia, "A study of skew in MapReduce applications," in Open Cirrus Summit, IEEE, Moscow, Russia, June 2011.
8. Shams, R., & Mercer, R. E. (2013, December). Classifying spam emails using text and readability features. In Data Mining (ICDM), 2013 IEEE 13th International Conference on(pp. 657-666). IEEE.
9. Lavanya, D., & Rani, K. U. (2012). Ensemble decision making system for breast cancer data. International Journal of Computer Applications, 51(17).
10. López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113-141.
11. Kwon, Y., Balazinska, M., Howe, B., & Rolia, J. (2012, May). Skewtune: mitigating skew in mapreduce applications. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (pp. 25-36). ACM.
12. Kwon, Y., Balazinska, M., Howe, B., & Rolia, J. (2012). Skewtune in action: Mitigating skew in mapreduce applications. *Proceedings of the VLDB Endowment*, *5*(12), 1934-1937.
13. S. Ibrahim, H. Jin, L. Lu, B. He, G. Antoniu, and S. Wu, "Handling partitioning skew in MapReduce using leen," Peer-to-Peer Networking and Applications, vol. 6, no. 4, pp. 409–424, 2013.

14. Y. Xu, P. Zou, W. Qu, Z. Li, K. Li, and X. Cui, "Sampling-based partitioning in MapReduce for skewed data," in ChinaGrid Annual Conference (ChinaGrid), IEEE, pp. 1–8, Beijing, China,2012.

15. S. R. Ramakrishnan, G. Swart, and A. Urmanov, "Balancing reducer skew in MapReduce workloads using progressive sampling," in Proceedings of the Third ACM Symposium on Cloud Computing, pp. 16–2012, ACM, San Jose, California, October 14–17, 2012.

16. J. Lin, "The curse of zipf and limits to parallelization: a look at the stragglers problem in MapReduce," in 7th Workshop on Large-Scale Distributed Systems for Information Retrieval, ACM, Boston, MA, USA, 2009.

17. Chen, Q., Yao, J., & Xiao, Z. (2015). Libra: Lightweight data skew mitigation in mapreduce. *IEEE Transactions on parallel and distributed systems*, *26*(9), 2520-2533.

18. Elmeleegy, K., Olston, C., & Reed, B. (2014, June). Spongefiles: Mitigating data skew in mapreduce using distributed memory. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*(pp. 551-562). ACM.

19. B. Gufler, N. Augsten, A. Reiser, and A. Kemper, "Handing data skew in MapReduce," in Proceedings of the 1st International Conference on Cloud Computing and Services Science, INSTICC, vol. 146, pp. 574–583, Noordwijkerhout, Netherlands, 2011.

20. T. White, Hadoop: The definitive guide, O'Reilly Media/Yahoo Press, CA, USA, 2012.

21. B. M. Patil, R. C. Joshi, and D. Toshniwal, "Association rule for classification of type-2 diabetic patients," in Machine Learning and Computing (ICMLC), 2010 Second International Conference on IEEE, pp. 330–334, Minneapolis, MN, USA, 2010.

22. United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Center for Behavioral Health Statistics and Quality, National.

23. Z. Liu, Q. Zhang, R. Boutaba, Y. Liu, and B. Wang, "OPTIMA: on-line partitioning skew mitigation for MapReduce with resource adjustment," Journal of Network and Systems Management, vol. 24, no. 4, pp. 859–883, 2016.

24. S. Chopra and M. R. Rao, "The partition problem," Mathematical Programming, vol. 59, no. 1–3, pp. 87–115, 1993.