Available online at: https://ijact.in

Date of Submission	31/12/2018
Date of Acceptance	07/04/2019
Date of Publication	30/04/2019
Page numbers	3112-3116 (5 Pages)

Cite This Paper: Alzyadat, WJ., Alhroob A., Almukahel IH., Rodziah A.(2019) Fuzzy map approach for accruing velocity of bigdata, 8(4), COMPUSOFT, An International Journal of Advanced Computer Technology. PP-3112-3116.

ISSN:2320-0790

compusoft
An International Journal of Advanced Computer Technology

FUZZY MAP APPROACH FOR ACCRUING VELOCITY OF BIG DATA

Wael Jumah Alzyadat[1*], Aysh AlHroob[2], Ikhlas Hassan Almukahel[3], Rodziah Atan[4]

[1,2,3]Department of Software Engineering, Faculty of Information Technology Isra University, Amman, Jordan
[4]Department of Software Engineering and Information Systems, Faculty of Computer Science and Information Technology, University Putra Malaysia, Selangor, Malaysia.
waael.alzyadat@iu.edu.jo

Abstract: Each characteristic of Big Data (volume, velocity, variety, and value) illustrate a unique challenge to Big Data Analytics. The performance of Big Data from velocity characteristic, in particular, appear challenging of time complexity for reduced processing in dissimilar frameworks ranging from batch-oriented, MapReduce-based to real-time and stream-processing frameworks such as Spark and Storm. We proposed an approach to use a Fuzzy logic controller combined with MapReduce frameworks to handle the vehicle analysis by comparing the driving data from the new outcome vehicle trajectory. The proposed approach is evaluated via amount of raw data from the original resource with dataset after the processing of the approach using ANOVA to estimate and analyze the differences. The difference before and after using approach is a positive impact in several stages of the volume of datasets, variances, and P-value that mean significantly and contribute for two aspects i.e. accuracy and performance.

Keywords: Big Data; Velocity; Fuzzy Logic Controller; MapReduce.

I. INTRODUCTION

Big Data has become a hot topic in both academia and industry. It is defined as datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze [1]. This technical complication is due to the characteristic of big data which includes mainly 4V's (Volume, Variety, Value, and Velocity)[2]. Velocity refers to the rate of generated, processed and analyzed data. The proliferation of digital devices such as smart phones and sensors has led to an unprecedented rate of data creation; thus, driving a growing need for real-time analysis and evidence-based planning [3]. For many applications, the speed of data processing is even more important than volume. Real-time information makes it possible for a more agile decision making. The usefulness of big data technique lies in its power to optimize the outcome, improve the processing efficiency, and reduce costs [4].

The velocity of data in driving has increased due to improved technologies, increased processing power, and speed of monitoring and processing. Many amateurs illustrate approaches and technique to solve big data problem; including focus on fuzzy clustering algorithms which apply to cluster approaches for accuracy. The author investigates the parallelization and scalability of a common and effective fuzzy clustering algorithm named Fuzzy C-Means (FCM) algorithm. The algorithm is parallelized using the MapReduce paradigm outlining how the Map and Reduce primitives are implemented [5].

Data filtering will be done using ANOVA which is used to evaluate the differences between data sets. It can be used for the recorded process. The datasets need not be equal in size. Datasets suitable for the ANOVA can be small or infinitely large sets of numbers. The advantage of using the filtering method is reducing statistical complexity which provides independent criterion used for feature evaluation.

This article focuses on Fuzzy logic which is considered to be a technique where computing is based upon “degrees of truth” used to handle the random and imbalance relation of MapReduce mapping function (peer to peer). It uses the relationship on runtime and fuzzy logic determines the path and MapReduce speeds up the process. This is important to velocity big data and ensures that qualified output value is achieved [6].

MapReduce applies the concept of Hadoop. It is a programming paradigm that allows for massive scalability across hundreds or thousands of servers in a Hadoop. The MapReduce concept is fairly simple to understand because it consists of mapper function and reduces function processing

A number of previous works theoretical and technological were encouraged to accuracy and scalability terms of Big Data challenges [7] as shown in Table 1. One of vital to handling Big Data challenge focused on cluster concept; called Fuzzy set techniques can override stymie the issue of accuracy and scalability.

Table I Comparison among Fuzzy Techniques

Authors	***Nature of problem***	***The role of Fuzzy set***	***Advantages of using fuzzy***
del Río et al. (2017) [5]	Classification	Linguistic Fuzzy rule-based classification	Descriptive model with good accuracy
Mahmud, et al ,(2016)[8]	Health-shocks prediction	Fuzzy linguistic summarization	Provide interpretable linguistic rules to explain the causal factors
He, Q., et al. (2015) [9]	Parallel sampling Represent	Handle uncertainties of boundaries of hypersurfaces	Granules by Fuzzy boundary, the algorithm maintains identical distribution

In spite of effective Fuzzy set techniques such as Fuzzy C-Means (FCM) algorithm, there is still debate as to what types of uncertainty are captured by fuzzy logic for the best practice to apply fuzzy on parallel platforms. MapReduce paradigm consensus on parallelized [5] & consists of Mapper and Reduce functions. The mapper function dealing with sources from distributed platform across servers to set allocate through Reduce functions that existing filtering operation. The advantage of using the filtering method is reducing statistical complexity which provides independent criterion used for feature evaluation [6].

The remainder of this paper is addressed below. Section 2 presents the concept and components for the proposed Fuzzy Map approach for accruing velocity of Big Data. In section 3 we implement the approach and analysis experiment. Section 4 we illustrate experimental results. Finally, we summarize the paper in Section 5.

II. RESEARCH METHOD

In this section, we present a fuzzy map approach to accruing the big data velocity by filter technique. The approach consisting of six components; the first component is Data collection and preprocessing. Second is Data fuzzification which converts data to crisp and builds membership function. Third is extract fuzzy using if-then rule according to the dataset. The fourth component involves Map function which is important acquiring relations among dataset, separated attribute, and content. The fifth component is the response of defuzzification to produce new data as recodes. The last component is applying filter technique to calculate the matching percentage between row and new dataset. The Figure below illustrates the data following the below approach.

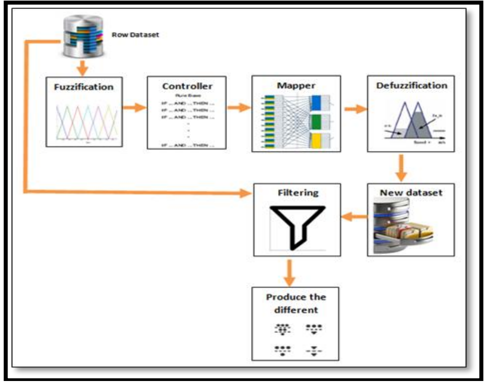


Figure 1 Hybrid approach fuzzy logic controller and MapReduce

The components from the Fuzzy Map approach for accruing velocity of Big Data interacts as follows:

A. Component One: Collecting

Component one is collecting and preprocessing the data. Our data used is Vehicle trajectory and kinematics data; it is collected using detailed vehicle trajectory data on southbound US 101 and Lanker shim Boulevard in Los Angeles, CA, eastbound I-80 in Emeryville, CA and Peachtree Street in Atlanta, Georgia. Data was collected through a network of synchronized digital video cameras and customized software application developed for the NGSIM program, the data source is https://www.kaggle.com/zhaopengyun/driving-data/home

B. Component Two: Fuzzification

Fuzzification component helps in converting numeric data to categorical data according to membership function of each attribute into fuzzy categorical as the dataset. In this case, fuzzy regions refer to intervals for the linguistic terms. Therefore, we can construct triangular membership functions.

C. Component Three: Fuzzy

Build fuzzy using if-then rule after calculating the degree of a linguistic value which is determined as the linguistic term whose membership function is maximal in this case. Then, we repeat the process for all instances in the data to construct fuzzy rules covering the data.

D. Component Four: Map Function

Divide through apply map function which takes the output from the fuzzy controller (if-then) rule to reduce the random grouping generate by MapReduce. The output of this component is the value (key value) which describes the degree of the fuzzy rule mapping to data content.

E. Component Five: Defuzzification

Defuzzification is responsible for producing a result in logic, given fuzzy sets and membership degrees. It is the process that maps a fuzzy set to a crisp set. It is typically needed in fuzzy control systems.

F. Component *Six: Filtering*

Filtering technique involves applying ANOVA where computing is based upon the distance of differentiating. The output is a match or a mismatch among old and new datasets.

In summary, the existing components from the Fuzzy Map approach for accruing velocity of Big Data turn for forwards original datasets with modified content by two factors that are attributes and index. Instead of the data collection and preprocess component after fetch dataset job from sources as well as keeping the original structure and index to use in the filter component to find out the matching and mismatch via ANOVA; the Fuzzification component treats with attributes via Membership function to acquire the linguistic term. When done the second component, the Fuzzy rule interplay with categorical data to calculate the degree of a linguistic value which belongs to the fourth component. Defuzzification component is syndicate Fuzzy sets and membership degrees in a crisp set then forward to filter component to compute distance by ANOVA.

III. EXPERIMENT AND ANALYSIS

Fuzzy Map approach for accruing velocity of Big Data uses R language and Microsoft Excel 2016. The R package used is shown in the Table below.

Table II R Packages Used in the Experiment

R Packages	*Purposed*
Read Rectangular Text Data (Readr)[10]	Read rectangular data
Dplyr [11]	Data manipulation
Tidyr[12]	Work with Attributes (column) and Raw (Observations)
Classification and Regression Training caret[13]	Preprocess data set
HadoopStreaming[14]	Provides a framework for writing Map/Reduce
HiveR[15]	Function Map, Manager and Plots
FuzzyR[16]	Design and simulate Fuzzy logic

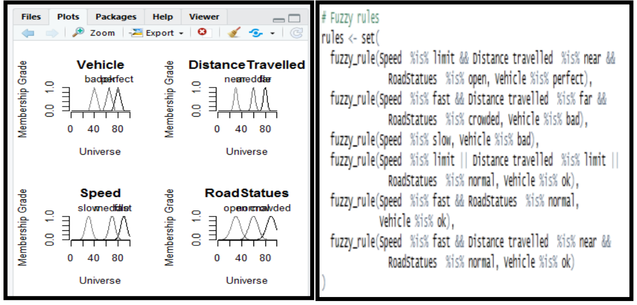
The data collection and preprocess fetch the raw dataset; it includes 82 attributes and 17897 records which are different types that is string, numeric, and Boolean. Furthermore, the quintiles description for each attribute is used in various ways which can be observed by the mean and standard deviation. After variables are defined, the next step is defining the Fuzzy rules, vehicle report which is the final state. The Fuzzy rules are the links between the "non-final" variables distance traveled, speed, road status and vehicle, as shown in the Table below.

Table III Variables Consists in Fuzzy

Variables	*Values*
Vehicle	bad, ok, and perfect
Distance travelled	near, meddle, and far
Speed	slow, limit, and fast
Road statues	crowded, normal, and open

The most effective rules are six rules; Figure 2 shows the effective rules. The rules structure must include variables and values to control the direction of interval data as well as define the Map through membership.

The longest rule involves all variables consists of four rules; they are different by values and operators (and, or). There is one rule shortest rule that involve two variables which are speed and vehicle.



```
# Fuzzy rules
rules <- set(
  fuzzy_rule(Speed  %is% limit && Distance travelled  %is% near &&
              RoadStatues  %is% open, Vehicle %is% perfect),
  fuzzy_rule(Speed  %is% fast && Distance travelled  %is% far &&
              RoadStatues  %is% crowded, Vehicle %is% bad),
  fuzzy_rule(Speed  %is% slow, Vehicle %is% bad),
  fuzzy_rule(Speed  %is% limit || Distance travelled  %is% limit ||
              RoadStatues  %is% normal, Vehicle %is% ok),
  fuzzy_rule(Speed  %is% fast && RoadStatues  %is% normal,
            Vehicle %is% ok),
  fuzzy_rule(Speed  %is% fast && Distance travelled  %is% near &&
              RoadStatues  %is% normal, Vehicle %is% ok)
)
```

Figure 2: Effective Rules to Vehicle Analysis

The roadmap of rule set presents in Fuzzy includes six pathways as shown in the Figure 2. While all rule set involve speed variables with values, the fast value of speed variables used in three rules, the limit value appears twice in different rules, and slow value used only once. The road status variables are covered in five rules in which three indexing of it is normal and one is open and crowded. Meanwhile, vehicle variables coverage in all rule sets presents the values three ok, two is bad, and one is perfect. Defuzzification the final process in the approach by removing all value associated with the bad key the raw dataset reduces the volume which directly affects the velocity of big data. Then convert data to a crisp value. An employing ANOVA to analyze the variance for both raw and new data set first determine and calculate the sum, average, and variance to calculate ANOVA for the four variables. As shown in Table IV.

Table IV Comparison Among Source of Variation

Source of Variation	**SS**	**DF**	**MS**	**F**	**P-value**
Raw Data	156595	1467554	17663554	1.46E-09	2.76
New Data	169329	480381	550180	0.9	1

The Table above illustrates as follows:

I. The Degrees of Freedom (DF) using the two mechanisms is vertical by the number of columns and horizontal by rows. As equatio**n 1.**

$$(Number\ of\ Columns - 1) \times$$
$$(Number\ of\ Rows - 1) \quad (\mathbf{1})$$

E.g., (2*1) = 2.

II. **F**-ratio determines how far the data are scattered from the mean raw data.

III. **P**-value refers to probability value.

IV. Results

The result shows evidence and proof that the Fuzzy logic controller plays an important role in the enhancement and the performance of Big Data in term of velocity by reducing the random relation result from Map-reduce. The result illustrated in the Table 5 shows the main difference between two datasets; before and after applying the Fuzzy Map approach.

Table V: Difference between the Two Datasets

Steps	Raw Data	New Data (after apply approach)
Preprocess	Dimensional 17897 rows 82 attributes	Dimensional 16867 rows 33 attributes Removing all missing value
Fuzzification	Dimensional 17897 rows 82 attributes	Dimensional 14557 rows 33 attributes Remove the content associated to bad key
Filtering	Dimensional 17897 rows 82 attributes	Applying ANOVA single factor

We have applied ANOVA to analyse the variance for raw dataset which shows the result in Figure 3.

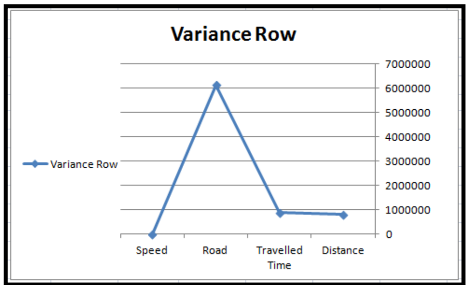


Figure 3: ANOVA Applied to Raw Dataset

The ANOVA applying to new data shows that the variables relative to F-ratio and the variance value decreased which is evidence that Fuzzy Map approach reduce all unnecessary relation and focused on the effected relations. It is also shown in the Figure below:

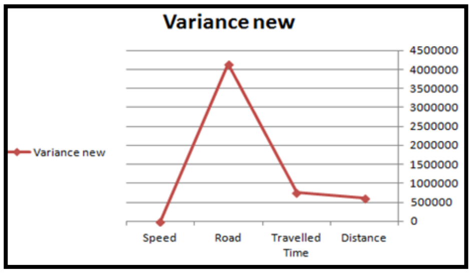


Figure 4: ANOVA Applied to New Datasets

Based on previous results we compare between variance in raw and new data as shown Table below.

Table VI Comparing the Variances between Raw and New Datasets

Variable	*Variance row*	*Variance new*	*Difference*
Distance	827682.5	620768.5	206914
Travelled Time	897190.1	767453.1	129737
Road	6132315	4136715	1995600
Speed	3299.542	1355.542	1944

Distance in raw dataset variance is 827682.5 and in new dataset variance is 620768.5. The difference between both of them is 206914 that means the new dataset reduce the distance as a result of removing all content of bad key. Travelled time raw dataset variance is 897190.1 and the new dataset variance is 620768.5. The difference between both of them is 206914 that mean the new dataset reduce is better than the raw one on travelled time. Road raw dataset variance is 6132315 and the new dataset variance is 4136715. The difference between both of them is 1995600 that mean the new dataset chose the shortest road as a result of Fuzzy rule controller. Speed from raw dataset variance is 3299.542 and the new dataset variance is 1355.542. The difference between both of them is 1944 that means the new dataset optimize speed to ideal driving behavior.

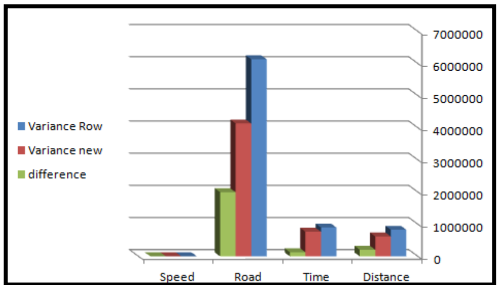


Figure 5 Variance Value Difference in Raw and New Dataset

Figure 5 illustrates that the new data set variance is more efficient in reducing the cost of process in autonomic vehicles. This reduction is due to the converge relation in data after applying to map the fuel used and distance, also the travelling time will reduce.

V. Conclusion

This research addresses the challenge in big data era about velocity, by combing big data technique with artificial intelligence technique. The approach handles velocity and assures the performance of big data analysis by reducing the processing time. This research presents an efficient approach using Fuzzy Logic and MapReduce to achieve performance data. It uses a Fuzzy controller and MapReduce to create an optimal dataset. The approach proposed in this research can be applied to other real-world applications to verify their merits and discover and solve any shortcomings.

VI. References

[1] S. Ramírez-Gallego, A. Fernández, S. García, M. Chen, and F. Herrera, "Big Data: Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce," Inf. Fusion, vol. 42, pp. 51–61, 2018.

[2] R. Kune, P. K. Konugurthi, A. Agarwal, R. R. Chillarige, and R. Buyya, "The anatomy of big data computing," Softw. - Pract. Exp., vol. 46, no. 1, pp. 79–105, 2016.

[3] Jin, Xiaolong, Benjamin W Wah, Xueqi Cheng, and Yuanzhuo Wang. 2015. 'Significance and challenges of big data research', Big Data Research, 2: 59-64.

[4] Kaisler, S.H., Armour, F., Espinosa, J.A., & Money, W.H. (2013). Big Data: Issues and Challenges Moving Forward. *2013 46th Hawaii International Conference on System Sciences*, 995-1004.

[5] Fernández, Alberto, ara del Río, Abdullah Bawakid, and Francisco Herrera. 2017. 'Fuzzy rule based classification systems for big data with MapReduce: granularity analysis', Advances in Data Analysis and Classification, 11: 711-30

[6] Faisal Y.Alzyoud and Wa'el Jum'ah Al_Zyadat., The classification filter techniques by field of application and thecresults of output. Aust. J. Basic & Appl. Sci., 10(15): 68-77, 2016

[7] J. A. Benediktsson, Y. Zhu, M. Chi, Z. Sun, A. Plaza, and J. Shen, "Big Data for Remote Sensing: Challenges and Opportunities," Proc. IEEE, vol. 104, no. 11, pp. 2207–2219, 2016.

[8] S. Mahmud, R. Iqbal, and F. Doctor, "Cloud enabled data analytics and visualization framework for health-shocks prediction," Futur. Gener. Comput. Syst., vol. 65, pp. 169–181, 2016.

[9] Q. He, H. Wang, F. Zhuang, T. Shang, and Z. Shi, "Parallel sampling from big data with uncertainty distribution," Fuzzy Sets Syst., vol. 258, pp. 117–133, 2015.

[10] H. Wickham, J. Hester, R. Francois, J. Jylänki, and M. Jørgensen, "readr: read rectangular text data. R package version 1.1. 1." R Foundation for Statistical Computing, 2017.

[11] Wickham, H. and Francois, R. (2015) dplyr: A Grammar of Data Manipulation. R Package Version 0.4.3. http://CRAN.R-project.org/package=dplyr.

[12] Wickham, H. (2017), tidyr: Easily Tidy Data with spread and gather Functions. R package version 0.6.1. URL: https://CRAN.R-project.org/package=tidyr

[13] M. Kuhn, "Classification and Regression Training (Caret)," R Program. Lang. Packag., 2015.

[14] Rosenberg DS (2012). HadoopStreaming: Utilities for Using R Scripts in Hadoop Streaming. R package version 0.2, URL http://CRAN.R-project.org/package=HadoopStreaming.

[15] J. Chung and M. B. A. Hanson, "Package ' HiveR ,'" 2017.

[16] T. R., Jon Garibaldi, Chao Chen, "Package ' FuzzyR ,'" https://cran.r-project.org/web/packages/FuzzyR/FuzzyR.pdf.