International Journal of Engineering &Technology, 7 (4) (2018) 6997-7001

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.28772
Research paper

Hybrid Approach Using Fuzzy Logic and MapReduce to Achieve Meaningful Used Big Data

Ikhlas Almukahel [1] *, Wael Alzyadat [2], Mohamad Alfayomi [3]

[1] *Teacher Assistant, Software Engineering Department, Faculty of Information Technology, Isra University, Amman, Jordan*
[2] *Assistant prof., Software Engineering Department, Faculty of Information Technology, Isra University, Amman, Jordan*
[3] *Professor, Computer Science Department, Faculty of Information Technology, Isra University, Amman, Jordan*
**Corresponding author E-mail: Ikhlas.almukahel@iu.edu.jo*

Abstract

Big data faces many challenges from different aspects; these challenges are represented in characteristics, such as volume, velocity, variety, and value. Preprocessing and analyzing big data are important issues to acquire quality information toward accurate values for correct decision making. Quality data taxonomy points to two basic actions to ensure that data is meaningful and predictive. Consequently, a hybrid approach using fuzzy logic and MapReduce is utilized to produce a new version of MapReduce which consist of four layers. Data collection is achieved in the first layer. The second layer consist of preprocessing data, where semi-structured data is treated to clean up and obtain the map function to acquire relationships. The third layer includes the application of fuzzy controller as well as classification to generate rules. Finally, in the fourth and last layer, data reduction and classification are carried out to achieve a meaningful and predictive outcome. The result showed the efficiency of the approach through Sensitivity = 80%, Specificity = 86% and F-measure= 2.5 that were validated in TREC conference website. The hybrid approach treating the 4Vs towards achieving meaningful which has positive effect support doctor to take the right decision.

Keywords: *Big Data; MapReduce; Meaningful; Predictive; Fuzzy Logic Controller.*

1. Introduction

The massive amount of data is acquired from many sources in different domains, such as industry, business, social networks, internet, health, finance, economics, and transportation. Flexible tools and techniques are needed to lead the big data era the term big data refers to deriving, collecting, and processing massive amounts of data [1] due to that the characteristics of big data have become difficult to analyze and manage with traditional data processing tools [1, 2]. The main challenge of big data is extracting value to make a decision, predict and improve services [3]. Traditional data mining techniques are almost unable to handle big data so many artificial intelligent techniques are applied to big data framework. The objective is combining fuzzy logic controller with MapReduce to extract quality data and evaluate the approach use PIMA India dataset for diabetic patients [4].

2. Related works

Big data is still a misleading definition for the concept itself. Many researchers and organizations define big data as datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze [5]. The issues of big data refer to the 4V's, which are the characteristics of volume, velocity, variety, and value [6]. Many frameworks apply MapReduce, which was originally produced by Google to solve the web search index creation problem. It has been established as a de-facto solution that deals with big data scalability [7]. MapReduce is a programming framework the main functions are automatically parallelized and executed on computing big data. Apache Hadoop is one of the most popular open-source implementations of MapReduce paradigm [8]. Recent years publications focused on approaches to the issues of big data refer to the 4V's [9]. Recent years also publications focused on approaches to improve extracting value from big data at a possible time response. Table one shows a comparison between similar approaches solving big data classification problem.

Table 1: Comparison Among Similar Approaches

Authors	Problem	Approach	Case studies	Tools
Abdrabo, M., et al. (2018) [10]	Big data dimensionality	MapReduce parallel processing and fuzzy rough	Diabetes dataset electroencephalography	WEKA
Jin, S. Peng, and D. Xie. (2017) [11]	Big data classifying problems	MapReduce approach with dynamic fuzzy inference	Six UCI datasets	JAVA
del Río S., et al. (2015) [12]	Extraction of information	The Chi et al.'s algorithm for classification	Six problems from the UCI dataset repository	JAVA

Abdrabo et al. (2018) introduced a framework that enhances the reduction of big data dimensionality. Selecting more important features helps to improve classification performance decision tree

accuracy was 86.4% while its precision was 84.3 %. In the data preprocessing step, and tried to overcome two main problems: heterogeneous data using peer to peer transformation and incomplete data based on assigning a fixed number. In map step, a fuzzy-rough set was assigned to feature selection. In reduce step, assigning fuzzy applications means clustering for identifying similar features to assign them the same key [10].

In MapReduce with dynamic fuzzy inference/interpolation for big data applications, both inference and interpolation methods can work together to produce a final output. It was shown that the average accuracy of classification in terms of performance in tease two methods has been contrasted in experimental research including six different big data problems [11].

del Río, López, Benítez& Herrera (2015), the big data classification problems especially the extraction of information is uncertainty that associated with the noise inherent to available data. The classification evaluated using the accuracy obtained and the runtime spent by the models, where this research aimed to analyze the quality of the ChiFRBCS-BigData algorithm in the big data scenario [12].

Classification problems provide extraction of information with ambiguity associated with the noise inherent in availability [13]. Furthermore, evaluation is carried out using the accuracy obtained and the runtime spent by models [4].

Fuzzy logic is used to handle the random and imbalanced relation of MapReduce mapping function (peer to peer) with the runtime. Fuzzy logic determines the path and MapReduce speeds up the process, which is significant to velocity and value. The pinpoint using fuzzy technique is to efficiently process a large volume of big data within limited run times.

Del.Rio, S. et al (2015) propose the Chi-FRBCS-BigDataCS algorithm, In order to effectively deal with big data a fuzzy rule-based classification system that is able to deal with the uncertainty that is introduced in large volumes of data by using MapReduce framework to distribute the computational operations of the fuzzy model while it includes cost-sensitive learning techniques in its design to address the imbalance of big data [12].

Mahmud, et al (2016) focused on using a fuzzy rule summarization technique, which can provide stakeholders with interpretable linguistic rules to explain the causal factors affecting health-shocks [14]. He, Q, et al. (2015) present a Parallel Sampling method based on HyperSurface (PSHS) for big data with uncertainty distribution to get the Minimal Consistent Subset (MCS) of the original sample set whose inherent structure is uncertain [15]. Notice that all classifier for solving them is not understandable (black box type) that is often vital in medical diagnosis there is no explanation and discussion about the fuzzy rule. The table below shows a comparison between fuzzy techniques.

Table 2: Comparison Among Fuzzy Techniques

Authors	Nature of problem	The role of fuzzy set technique	Advantages of using fuzzy techniques
Del Río, S.et al (2015) [12]	Classification	Linguistic fuzzy rule-based classification	A descriptive model with good accuracy
Mahmud, et al (2016) [14]	Health-shocks prediction	Fuzzy linguistic summarization	Providing interpretable linguistic rules to explain the causal factors
He, Q, et al. (2015) [15]	Parallel sampling represent	Handling uncertainties of the boundaries of hypersurfaces	granules by a fuzzy boundary; the algorithm maintains identical distribution

3. Research method

The approach consists of four layers. First comes the data collection layer and the second is data preprocessing layer which converts semi-structured data into a structured format and implements the dataset to be input throughout layers that involve map function with the duty of acquiring relations among dataset, separated attributes, and content, while the function of the third layer is applying Fuzzy Logic Controller (FLC) and starting the first-round classification.

The fourth layer is responsible for reducing function with second-round classification to find an outcome that uses two equations for precision and recall to calculate F-measure for evaluating the accuracy. Each layer involves components that lead to treating with characteristics of big data; the whole process achieves 4V's as follows; the first layer achieves volume, while second and third layers achieve verity and velocity and layer four achieves value. The figure below illustrates the process of the hybrid approach with fuzzy logic controller and MapReduce.

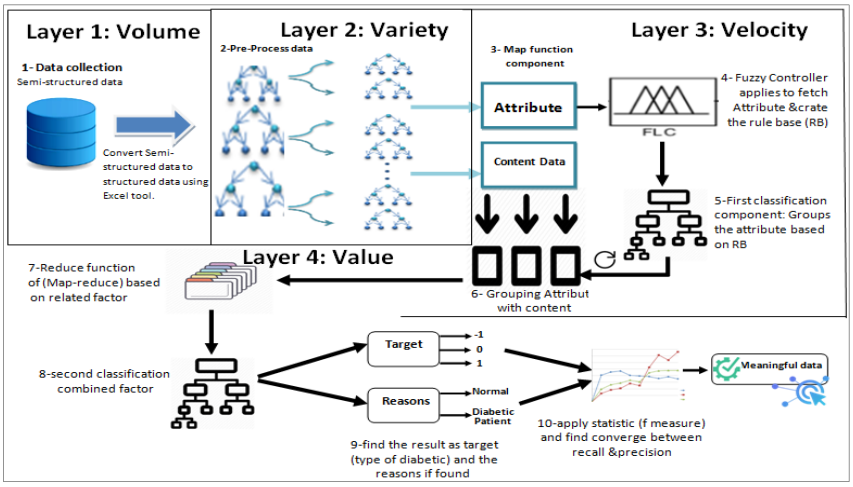

Fig. 1: Hybrid Approach Fuzzy Logic Controller and Mapreduce.

3.1. Layer1: data collection (volume)

Input dataset was obtained from the University of California Irvine (UCI) machine learning repository. The dataset originates from the National Institute of Diabetes and Digestive of Kidney Diseases. Several constraints were placed on the selection of instances from a larger database. All patients here are of the female gender, with at least 21 years of years, from PIMA Indian heritage dataset containing 8 attributes and 786 records.

3.2. Layer 2: data preprocessing (variety)

After monitoring data, all values which affect the results, such as null, outliers, and missing values, are defined to detect and prepare data for map function. Preprocessing output will achieve the variety and veracity that are related to big data introduce a degree of uncertainty that has to be handled. In addition to reducing volume and enhance velocity requirements. Preprocessing will include two main steps: removal of all missing values and data conversion.

3.3. Remove missing value

(If attribute. Value = = NULL || 0)
Then Delete record;

1) Normalization

Apply normalization to convert all string data to numerical value this important to make later process more sufficiency and quality. Data preprocessing is carrying out through rules applied in raw data.

3.4. Layer 3: map function and fuzzy rules (velocity)

The map function is applied, taking the output from the preprocessing component as structured data. In this step presenting mapping attributes and values. Then fuzzy controller (IF-THEN) rules are used to avoid random grouping generated by MapReduce; the second role is associated with the classifier rate.
Fuzzy rules are implemented as following steps

- Step 1: includes converting numeric data to categorical data according to the normal reading as shown in Table 3. Each attribute in the fuzzy categorical dataset refers to an interval for the linguistic terms. Therefore, the length of fuzzy linguistic term is defined as "low" and "high". Triangular membership function which is also constructed; e.g., in the first case, we have the corner points a = 1, b = 2 and c = 3 where b is a normal reading whose degree in the membership function equals one.
- Step 2: fuzzy IF-THEN rules are generated covering the training data, using the dataset from Step 1. First, degrees of the membership function for all values are calculated in the data. Through each instance and each variable, a linguistic value is determined as whose membership function is maximal, while the process is repeated for all instances to construct fuzzy rules covering all the data.
- Step 3: a degree is adjusted for each rule. Degrees of membership function are then aggregated.
- Step 4 a final rule is obtained base after deleting redundant rules. Considering the degrees of rules, redundant rules and those with lower degrees are deleted. Fuzzy based rule = $2^{\wedge 8}$, focusing on 12 rules that cover 90% of the diabetic patient's dataset.

3.5. Layer 4: reduce (value)

Reduce function is associated with the listing of data and put into groups of values, where each group is produced as a pair (key, value). The output is a list of attributes (keys) and all their associated values Measurement is carried out for converging between productive and perspective results to achieve meaningful data useful for doctors and patients. Classification measures used are accuracy, precision, recall, and F-measure to evaluate the classifier used to verify the effectiveness based on the confusion matrix. The hybrid approach is evaluated using the Text Retrieval Conference (TREC) [16], [17].

4. Experiment and analysis

A hybrid approach using fuzzy logic and MapReduce is applied to achieve meaningful data on both R packages [18] and WEKA [19], by the independent running of the same equipment computer properties; R integrated suite of software facilities for data manipulation using separated packages. Readr package [20]to Read Rectangular Text Data, Dplyr package [21] which purposed for data manipulation, Tidyr package [22] is important to work with Attributes-Feature and Raw-Observation, Classification and Regression Training caret package [23], PreProcess package [24] which role is preparing data while HadoopStreaming [24, 25] and HiveR [26] Provides a framework for writing map/reduce Function manager and plots them and FuzzyR [27] to Design and simulate fuzzy logic. On the other hand, WEKA is a collection of learning algorithms and data preprocessing tools applied built-in function. Experiment steps as follow:

4.1. Layer1: data collection (volume)

In WEKA dataset is imported using import data function to retrieves data from a file this depending on how to view and understand the whole data. An advantage of R packages is that they are able to treat in-memory and out-memory storage, which reflects covering the volume characteristic of big data.

4.2. Layer 2: verity (data preprocessing)

Once a dataset has been read, various data preprocessing tools, In Weka, built-in functions are used in preprocessing called filter, while in R the PreProcess package prepared data and made it suitable to analysis.

4.3. Layer 3: map function and fuzzy rules (velocity)

Map instructions divide the data to key and value. Map-function in R is involved in two packages, which are HadoopStreaming and hive to deal with the scalability of big data. In Fuzzy logic, rules are applied to predict diabetes depending on the relations among attributes and the outcome. In table 3 the main affected rules.

Table 3: Main Effected Rules That Define Diabetic and Non-Diabetic

Rule	class
R1: If (Npreg is High) and (Glu is Low) and (BP is High) and (Skin is Low) and (Insulin is L) and (BMI is High) and (PED is High) and (Age is Low)	Non-diabetic
R2: If (Npreg is Low) and (Glu is Low) and (BP is Low)) and (Skin is Low) and (Insulin is Low) and (BMI is Low) and (PED is High) and (Age is Low)	Non-diabetic
R3: If (Npreg is High) and (Glu is Low) and (BP is High) and (Skin is Low) and (Insulin is Low) and (BMI is High) and (PED is Low) and (Age is Low)	Non-diabetic
R4: If (Npreg is Low) and (Glu is Low) and (BP is Low) and (Skin is High) and (Insulin is Low) and (BMI is Low) and (PED is Low) and (Age is Low)	Non-diabetic
R5: If (Npreg is High) and (Glu is Low) and (BP is High) and (Skin is Low) and (Insulin is High) and (BMI is Low) and (PED is Low) and (Age is Low)	Non-diabetic
R6: If (Npreg is Low) and (Glu is Low) and (BP is High) and (Skin is Low) and (Insulin is Low) and (BMI is Low) and (PED is Low) and (Age is Low)	Non-diabetic
R7: If (Npreg is Low) and (Glu is Low) and (BP is Low) and (Skin is Low) and (Insulin is Low) and (BMI is Low) and (PED is Low) and (Age is Low)	Non-diabetic
R8: If (Npreg is High) and (Glu is High) and (BP is High) and (Skin is Low) and (Insulin is High) and (BMI is L) and (PED is High) and (Age is L)	Diabetic
R9: If (Npreg is Low) and (Glu is H) and (BP is Low) and (Skin is Low) and (Insulin is High) and (BMI is High) and (PED is High) and (Age is High)	Diabetic
R10: If (Npreg is Low) and (Glu is High) and (BP is L) and (Skin is High) and (Insulin is Low) and (BMI is Low) and (PED is Low) and (Age is Low)	Diabetic
R11: If (Npreg is High) and (Glu is High) and (BP is High) and (Skin is High) and (Insulin is High) and (BMI is High) and (PED is High) and (Age is High)	Diabetic

R12: If (Npreg is High) and (Glu is High) and (BP is High) and (Skin is H) and (Insulin is Low) and (BMI is High) and (PED is Low) and (Age is High)	Diabetic

In R, Hive and HadoopStreaming packages are used to derive big data characteristics. Especially, MapReduce is a parallel distributed system. Reducing random factor affects the map, where the significance of MapReduce appears in obtaining velocity. In WEKA classification and regression algorithms applicable to the preprocessed data use classify panel.

4.4. Layer 4: reduce function (value)

Reduce function performs calculations on small chunks of data in parallel then it combines the sub results from each reduced-chunk. Patients are classified into two classes: diabetic and non-diabetic. The value is extracted from big data achieved via classifying the PIMA dataset, through calculating Precision- Sensitivity (Eq.1), Recall- Specificity (Eq.2) and F-measure (Eq.3). The results illustrated in Figure 2.

$$\text{Precision} = \frac{(tp)}{(TP+FP)} \text{ [28]} \quad (1)$$

$$\text{Recall} = \frac{(tp)}{(TP+FN)} \text{ [28]} \quad (2)$$

$$\text{F-measure} = \frac{(2*\text{precision}*\text{Recall})}{(\text{precision}+\text{Recall})} \text{ [28]} \quad (3)$$

Where TP, TN, FP, and FN indicate in the following order:

- True positives: predict Diabetic as Diabetic.
- True negatives: predict Non-Diabetic as Non-Diabetic.
- False positives: predict Non-Diabetic as Diabetic.
- False negatives: predict Diabetic as Non-Diabetic.

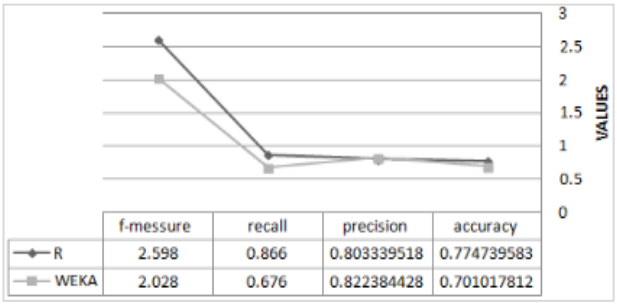

Fig. 2: F-Measure and Accuracy in R and WEKA Values and Periods for F-Measure are Shown in Graphs.

5. Results

The results obtained of the hybrid approach fuzzy logic and MapReduce in big data are very interesting and can be used confidently to help for decision making and achieve meaningful. The result evidence shows the scalability, which is able to extract process and manipulate with preserving the accuracy of classification at a satisfactory level, measured by determining the confusion matrices which contains information about actual and predicted classifications by an approach as shown in the figure below.

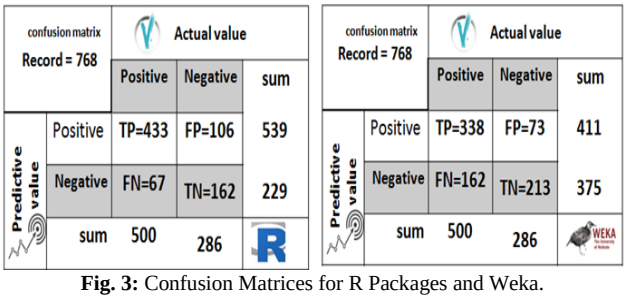

Fig. 3: Confusion Matrices for R Packages and Weka.

Through the evaluation results are observed as follows:

- Precision: in Weka, the result is 0.822 and in R (Readr, Dplyr, Tidyr, PreProcess, HadoopStreaming, HiveR, and FuzzyR) packages the result is 0.803, where the difference between both results is 0.01905. that means Weka is effect than R in positive prediction value.
- Recall: in Weka, the result is 0.676 and in R (Readr, Dplyr, Tidyr, PreProcess, HadoopStreaming, HiveR, and FuzzyR) packages, the result is 0.866 where the difference between both results is 0.19. that means R is effect than Weka in sensitivity measure.
- F-measure: Weka result is 2.028, while R result is 2.598, that means R is more harmonic then Weka with a difference of 0.57 between two results.
- Accuracy: in Weka, the result is 0.701, while in R the result is 0.774. The difference between both results is 0.073, which mean R (Readr, Dplyr, Tidyr, PreProcess, HadoopStreaming, HiveR, and FuzzyR) improved than Weka in Accuracy measure. The results are summarized in the figure below.

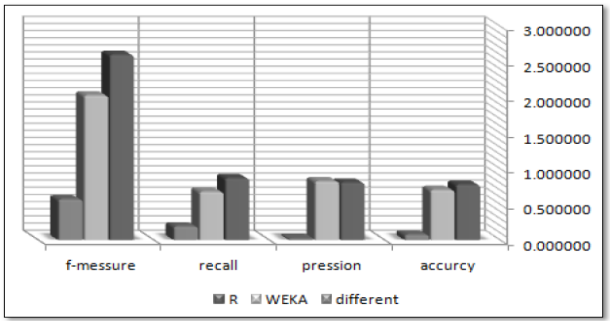

Fig. 4: Result Differences between R Packages and Weka.

6. Conclusion

This paper addressed two main challenges, the first challenge is extracting meaningful data from big data, the second challenge is combining big data technique with fuzzy logic artificial intelligence techniques through hybrid approach consisting of four layers to treat with 4V's of big data. The experiment with R (packages) and Weka are implemented MapReduce and apply it for classifying and predicting meaningful data. The effectiveness of the approach has been demonstrated through PIMA Indian diabetes dataset. The pinpoint aspects handled are management, acquisition and acting with big data.

The aspect of management of big data uses fuzzy logic controller and MapReduce dynamically to treat manipulate data, such as data updating, adding, deletion, and insertion reflecting the value of recall measure. Furthermore, the main impact in terms of big data. The significance of the approach based on F-measure to acquire meaningful data by mean of MapReduce. The contribution of this research based on the results of experiment approach supports healthcare domain to make accurate decisions. On the other hand, the precision measure is negatively affected by the random values generated in MapReduce which will be the future work.

References

[1] Manyika, J., et al., Big data: The next frontier for innovation, competition, and productivity. 2011.

[2] Xiaofeng, M., C.J.J.o.c.r. Xiang, and development, Big data management: concepts, techniques, and challenges [J]. 2013. 1(98): p. 146-169.

[3] Jin, X., et al., Significance and challenges of big data research. 2015. 2(2): p. 59-64. https://doi.org/10.1016/j.bdr.2015.01.006.

[4] Fernández, A., et al., Fuzzy rule-based classification systems for big data with MapReduce: granularity analysis. Advances in Data Analysis and Classification, 2017. 11(4): p. 711-730. https://doi.org/10.1007/s11634-016-0260-z.

[5] Chen, C.P. and C.-Y.J.I.S. Zhang, Data-intensive applications, challenges, techniques, and technologies: A survey on Big Data. 2014. 275: p. 314-347. https://doi.org/10.1016/j.ins.2014.01.015.

[6] Tidke, B. and R. Mehta, A Comprehensive Review and Open Challenges of Stream Big Data, in Soft Computing: Theories and Applications. 2018, Springer. p. 89-99. https://doi.org/10.1007/978-981-10-5699-4_10.

[7] del Río, S., et al., A MapReduce approach to address big data classification problems based on the fusion of linguistic fuzzy rules. 2015. 8(3): p. 422-437. https://doi.org/10.1080/18756891.2015.1017377.

[8] Hashem, I.A.T., et al., MapReduce: Review and open challenges. 2016. 109(1): p. 389-422. https://doi.org/10.1007/s11192-016-1945-y.

[9] Jovanovič, U., et al., Big-data analytics: a critical review and some future directions. 2015. 10(4): p. 337-355. https://doi.org/10.1504/IJBIDM.2015.072211.

[10] ABDRABO, M., et al., A Framework For Handling Big Data Dimensionality Based on Fuzzy-Rough Technique. Journal of Theoretical & Applied Information Technology, 2018. 96(4).

[11] Jin, S., J. Peng, and D. Xie. Towards MapReduce approach with dynamic fuzzy inference/interpolation for big data classification problems. in 2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC). 2017. IEEE. https://doi.org/10.1109/ICCI-CC.2017.8109781.

[12] del Río, S., et al., A MapReduce approach to address big data classification problems based on the fusion of linguistic fuzzy rules. International Journal of Computational Intelligence Systems, 2015. 8(3): p. 422-437. https://doi.org/10.1080/18756891.2015.1017377.

[13] Al_Zyadat, W.J. and F. Y.Alzyoud, The classification filter techniques by field of application and the results of output. Australian Journal of Basic and Applied Sciences (AJBAS), 2016. 10(15): p. 10.

[14] Mahmud, S., R. Iqbal, and F. Doctor, Cloud-enabled data analytics and visualization framework for health-shocks prediction. Future Generation Computer Systems, 2016. 65: p. 169-181. https://doi.org/10.1016/j.future.2015.10.014.

[15] He, Q., et al., Parallel sampling from big data with uncertainty distribution. Fuzzy Sets and Systems, 2015. 258: p. 117-133. https://doi.org/10.1016/j.fss.2014.01.016.

[16] Haruna, K. and M.A. Ismail. Evaluation Datasets for Research Paper Recommendation Systems. in Data Science Research Symposium 2018. 2018.

[17] The Text Retrieval Conference (TREC). 2018; Available from trec.nist.gov/evals.html.

[18] Venables, W.N., D.M. Smith, and R.C. Team, An introduction to R-Notes on R: A programming environment for data analysis and graphics. 2018.

[19] Holmes, G., A. Donkin, and I.H. Witten, Weka: A machine learning workbench. 1994.

[20] Wickham, H., J. Hester, and R.J.U.h.C.R.-p.o.p.r.R.p.v. Francois, readr: Read Rectangular Text Data, 2017. 1(0).

[21] Wickham, H., et al., dplyr: A grammar of data manipulation. 2015. 3.

[22] Wickham, H.J.U.h.C.R.-p.o.p.t.R.p.v., tidyr: Easily Tidy Data with'spread ()and gather ()'Functions, 2017. 2017. 1: p. 248.

[23] Denniston, K.J., J.J. Topping, and R.L. Caret, General, organic, and biochemistry. 2004: McGraw-Hill New York.

[24] Coombes, K.R., K.A. Baggerly, and J.S. Morris, Pre-processing mass spectrometry data, in Fundamentals of Data Mining in Genomics and Proteomics. 2007, Springer. p. 79-102. https://doi.org/10.1007/978-0-387-47509-7_4.

[25] Verma, C. and R. Pandey, Statistical Visualization of Big Data Through Hadoop Streaming in RStudio, in Handbook of Research on Big Data Storage and Visualization Techniques. 2018, IGI Global. p. 549-577. https://doi.org/10.4018/978-1-5225-3142-5.ch019.

[26] Sadhana, S.S., S.J.I.J.o.E.T. Shetty, and A. Engineering, Analysis of diabetic data set using hive and R. 2014. 4(7): p. 626-9.

[27] Bondarenko, I., et al., IDAS: a Windows-based software package for cluster analysis. 1996. 51(4): p. 441-456. https://doi.org/10.1016/0584-8547(95)01448-9.

[28] Powers, D.M., Evaluation: from precision, recall and F-measure to ROC, informedness, markedness, and correlation. 2011.