Journal of Theoretical and Applied Information Technology
28th February 2018. Vol.96. No 4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A FRAMEWORK FOR HANDLING BIG DATA DIMENSIONALITY BASED ON FUZZY-ROUGH TECHNIQUE

[1]MAI ABDRABO, [2*]MOHAMMED ELMOGY, [3]GHADA ELTAWEEL, [4]SHERIF BARAKAT

[1] Information Systems Department, Faculty of Computers and Information, Suez Canal University, Egypt
[2] Information Technology Department, Faculty of Computers and Information, Mansoura University, Egypt
[3]Computer Science Department, Faculty of Computers and Information, Suez Canal University, Egypt
[4]Information Systems Department, Faculty of Computers and Information, Mansoura University, Egypt
E-mail: [1]mai_abdrabo86@yahoo.com, [2*]melmogy@mans.edu.eg, [3]ghada@ci.suez.edu.eg, [4]sherifiib@yahoo.com

ABSTRACT

Big Data is a huge amount of high dimensional data, which is produced from different sources. Big Data dimensionality is a considerable challenge in data processing applications. In this paper, we proposed a framework for handling Big Data dimensionality based on MapReduce parallel processing and FuzzyRough for feature selection. This paper proposes a new method for selecting features based on fuzzy similarity relations. The initial experimentation shows that it reduces dimensionality and enhances classification accuracy. The proposed framework consists of three main steps. The first is the preprocessing data step. As for the next two steps, they are a map and reduce steps, which belong to MapReduce concept. In map step, FuzzyRough is utilized for selecting features. In reduce step, the fuzzy similarity is presented for reducing the extracted features. In our experimental results, the proposed framework achieved 86.4% accuracy by using decision tree technique, while the accuracy of the previous frameworks, which are performed on the same data set, achieved accuracy between 70 to 80%.

Keywords: *Big Data, FuzzyRough set, MapReduce, Feature selection, Decision tree.*

1. INTRODUCTION

Nowadays, it is a big challenge to process massive volume of data and discover new knowledge from it. Data are rapidly increased at a tremendous rate. Data are generated from different sources like sensors, smart devices, and social collaboration technologies in various formats. Processing Big Data makes the ability to discover knowledge efficiently [1]. In general, Big Data can be defined by its characteristics, as shown in Fig. 1. One of the data problems is Big Data dimensionality. Our research handles this issue based on how to select features and return the most significant ones.

Big Data has many difficulties. One of the most urgent problems is dimensionality because the size of data is huge. Traditional methods can be used for handling the volume of Big Data. So, we have to use parallel environment, which is provided in MapReduce. MapReduce is the most popular parallel processing method for processing huge data. MapReduce is a framework for processing and analyzing Big Data efficiently. MapReduce regards as a revolution in Big Data processing world, which is published by Google as open source software. On the other hand, one of the most efficient techniques for attribute reduction is Rough set [2,32].

Classification performance can be improved by using feature selection methods. They eliminate noisy, redundant, and irrelevant features. Relevant results can be extracted in restricted time by using traditional methods for millions of instances. The feature selection algorithm can be classified into the wrapper, filter, and embedded methods. The procedure that is independent of the learning algorithm is called filter method. It is regarded as subset selection. Filter method enhances the learning process. The main problem of this feature selection type is the resulting subset, which does not work very well with the learning algorithms.

On the other hand, wrapper methods use the significant features of minimal size, which is based on supervised learning output. Based on the training method, the features can be selected. The approaches that are laying between filter and wrapper techniques are called embedded methods [3].

Journal of Theoretical and Applied Information Technology
28th February 2018. Vol.96. No 4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

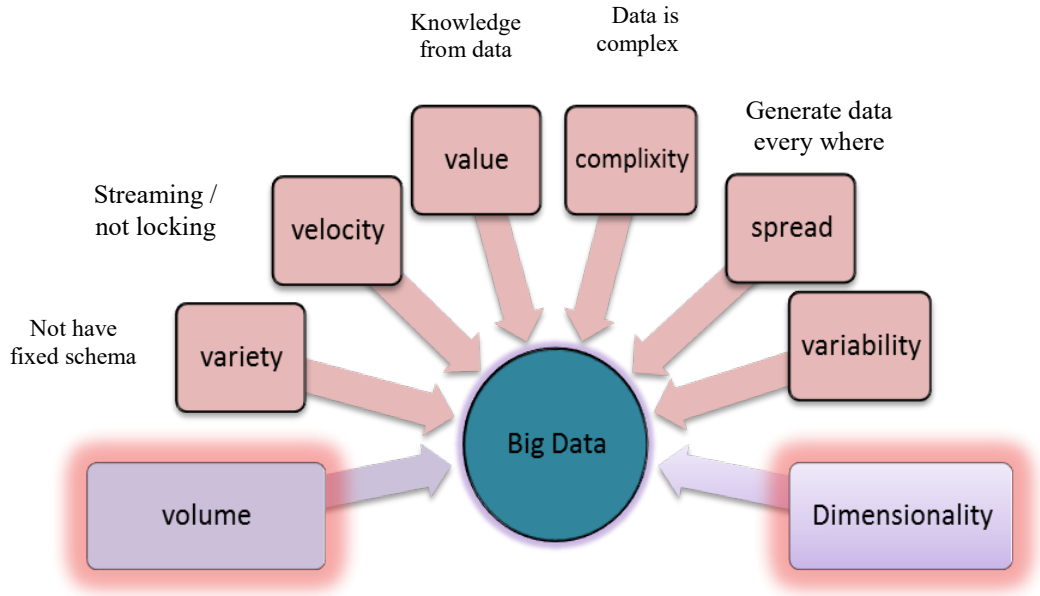

Figure 1. The Characteristics Of Big Data.

In this paper, the fuzzy rough set is proposed as feature selection method. Fuzzy-Rough Feature Selection (FRFS) gives and implies toward which discrete alternately real-valued loud information (or a mixture of both) camwood a chance to be successfully decreased without those have to user-supplied data. Additionally, this procedure could be connected with information for constant alternately ostensible choice attributes, and all things considered camwood a chance to be connected on relapse and additionally arrangement datasets [11,30].

The rest of this paper is structured as follows. Section 2 presents the techniques that are used in our proposed framework in detail. Section 3 presents some current, previous studies on removing misclassified instances for classification techniques. In Section 4, the proposed framework is described in detail. The experimental results are presented in Section 5. The conclusion and future work are discussed in Section 6.

2. BASIC CONCEPTS

2.1 Knowledge Discovery (KDD)

The KDD process consists of three main steps, which are preprocessing, attribute selection, and data mining (classification) stages, as shown in Fig. 2.

2.1.1 Preprocessing

Data preprocessing stage consists of data cleaning, data reduction, and data transformation.

Data cleaning is used to clean the data by filling missing data values and removing outliers. So, data reduction helps to reduce the representation of dataset into a small volume. Data transformation are utilized for putting data in a suitable format. Data preprocessing enhances data quality and performance of KDD process, especially classification accuracy [5].

Missing data handling is a big problem facing KDD process. Handling of the missing data is an essential assignment if data corrupted with huge values of missing data. It is a significant step to reach for accurate and high performance. So, managing the missing data is the main stage in KDD process. There are many techniques like deletion, C4.5 decision tree, maximization likelihood (ML), mean/mode imputation (MMI), regression methods, K-nearest neighbor (KNN), multiple imputations, and Bayesian iteration imputation [4].

Data Transformation is one of the first steps in data preparation. It transforms and handles data in a suitable format for the used algorithm. There are two widely used well-known techniques, which are numeration and discretization. Machine learning algorithms can use an enumeration for transforming nominal attributes into numeric ones. Using dummy variables can transform categorical attributes to numeric ones [6].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.1.3 Feature extraction

Feature reduction can be used for reducing the number of retrieved attributes. It is apparent that attributes with variance close to zero are not helping to separate the data in the machine learning model. Therefore, attributes with variance near zero are often removed from the dataset. Highly correlated attributes capture the same underlying information. Therefore, they can be removed without compromising the model quality.

2.1.4 Data classification

Classification is a data mining process that divides items into an accumulation of target classifications or classes. The objective of the classification is to precisely anticipate the tested objects to a class for every case in the information. There are many classification techniques, such as support vector machine (SVM), classification tree, and Naïve Bayes (NB) [7].

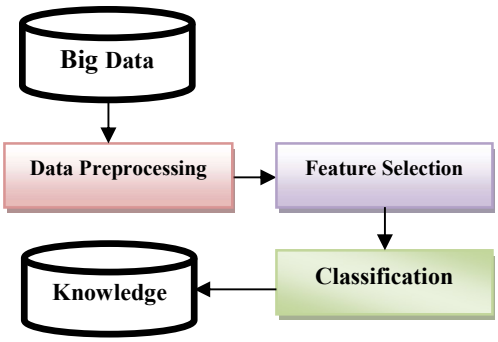

Figure 2. The Main KDD Steps.

2.2. Rough Set

Pawlak [8] proposed the rough sets theory to deal with uncertainty and fuzzy materials to simplify knowledge. In the rough sets theory, people utilize their general learning to order their general surroundings as unique or cement. The attributes are the main standard for arranging everything. The attributes can be categorized by gathering similar ones. This is called indiscernible relation, which is denoted as Ind. It is considered the basis of rough sets theory.

One of the principal focal points of the rough set theory is that it does not need bother with any preparatory or extra data about information. The main significant problems that moved toward utilizing noisy sets hypothesis like combine information decrease, detection the conditions of information, a rating of information essentialities, a period of choice calculation from data, suppose data characteristics, detection patterns in data and reveal effect relations. The rough set can be used for feature selection called Rough Set quick reduction algorithm, as shown in Fig. 3.

2.2.1 Information systems

Rough set based on many components the first one is information systems which discuss as follows.

$$IS = (U, A) \quad (1)$$

Non-empty st of objects can be represented as the universe (U), $U = \{x_1, x_2, ..., x_m\}$, and A is the set of attributes. Each attribute $a \in A$ (attribute a belonging to the considered set of attributes A) defines an information function:

$$f_a : U \to V_a \quad (2)$$

The set of values can be represented as Va, called the domain of attribute a. In all attributes, there are decision attributes and condition attributes.

2.2.2 Indiscernible relation

For every set of attributes $B \subset A$, an indiscernible relation $Ind(B)$ is defined in the following way: two objects, x_i and x_j, are indiscernible by the set of attributes B in A, if $b(x_i) = b(x_j)$ for every $b \subset B$. The equivalence class of $Ind(B)$ is called the elementary set in B because it represents the smallest discernible groups of objects. For any element x_i of A, the equivalence class of x_i in relation $Ind(B)$ is represented as $[x_i]_{Ind(B)}$.

2.2.3 Upper and Lower approximations

The upper and lower approximation are the essential ideas for analysis data based on rough sets. The idea of lower and upper approximation is to identify which element in the set surely have a place or potentially have a place.The definition is appeared as takes after:

Let X denotes the subset of elements of the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

universe U, the lower approximation of X in B, denoted as $\underline{BX}$, is defined as the union of all these elementary sets, which are contained in X. More formally:

$$\underline{BX} = \{x_i \in U \left| [x_i]_{Ind(B)} \subset X \right.\} \quad (3)$$

The lower approximation of the set X for objects xi was shown in the above Eq. (3), which belongs to the elementary sets contained in X (in the space B), $\underline{BX}$ is called the lower approximation of the set X in B. The upper approximation of the set X, denoted as $\overline{BX}$, is the union of these elementary sets, which have a non-empty intersection with X:

$$\overline{BX} = \{x_i \in U \left| [x_i]_{Ind(B)} \cap X \neq 0 \right.\} \quad (4)$$

The above statement is to be read as: the upper approximation of the set X is a set of objects x_i, which belong to the elementary sets that have a non-empty intersection with X, $\overline{BX}$ is called the upper approximation of the set X in B. The difference is called a boundary of X in U.

$$BNX = \overline{BX} - \underline{BX} \quad (5)$$

2.2.4 Core and reduct of attributes

The ideas of center and reduct are two critical ideas for rough sets theory. On the off chance that the arrangement of characteristics is reliant, one can be changed on discovering all conceivable negligible subsets of traits. These prompt an indistinguishable number of rudimentary sets from the entire arrangement of properties (reducts) in finding the arrangement of every single essential characteristic (core).

Disentanglement of the information system can be utilized to perceive a few estimations of traits, which are redundant. For instance, a few traits, which are excess, can be erased or be sifted by methods for the rearrangements strategies. If, $Ind(A) = Ind(A - a_i)$, then the attribute a_i is dispensable, otherwise, a_i is indispensable in A. In other words, if after deleting the attribute a_i, the number of elementary sets in the information system is the same, then it concludes that attribute a_i is dispensable.

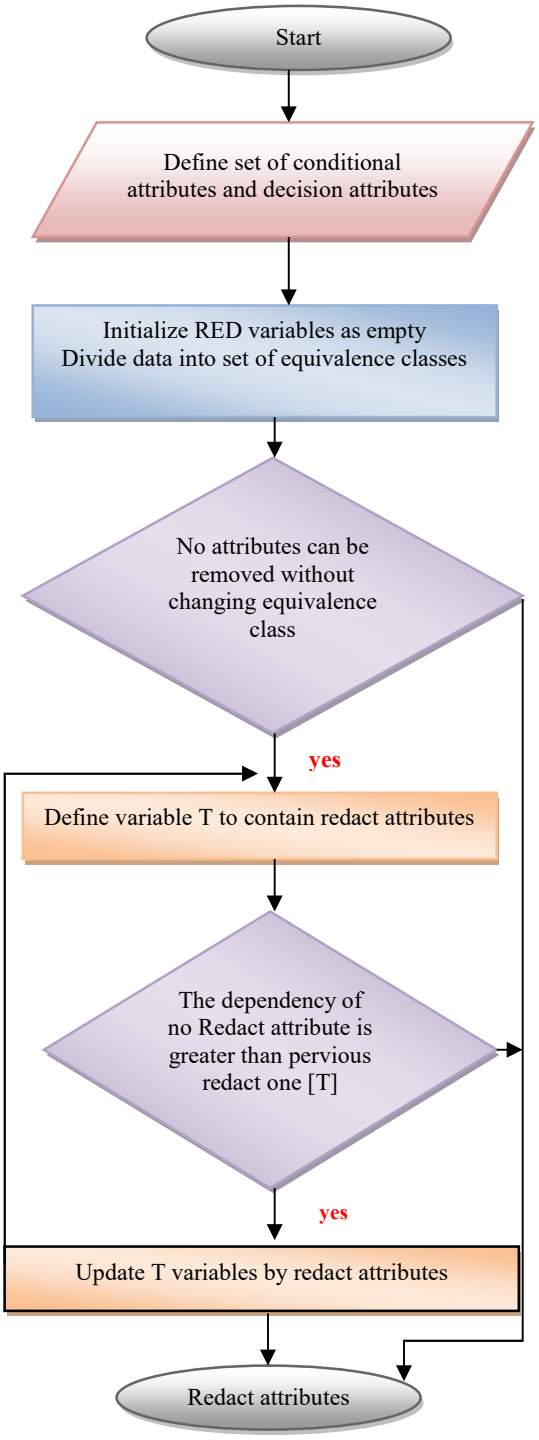

Figure 3. The Rough Set Quick Reduction Algorithm

As shown in Fig. 3, Rough Set quick reduction algorithm pseudo code is listed below:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Input: Con, the set of all the conditional attributes. D, the set of decision attributes
Output: A Reduct attributes RED

RED=ϕ
While ɣ RED(D) ≠ ɣ Con(D)
T=RED
$\forall x \in (con - RED)$
If ɣRED $\cup \{x\}$ (D) > ɣT (D)
T= RED $\cup \{x\}$
RED= T
Return RED

2.3 Fuzzy-rough sets

To improve the attributes selection, the previous rough set techniques must be extended to fuzzy rough sets. Simply because most datasets contain real-valued attributes and the rough set cannot handle noisy data. Fuzzy equivalence classes are the central concept of fuzzy-rough sets [10]. It is necessary to apply discretization for performing fuzzification step. The process to apply membership degrees of feature values to fuzzy set helps in dimensionality reduction process. FuzzyRoughset is a better guide for feature selection. The Fuzzy-Rough set is a generalization of the Rough set, derived from the approximation of a fuzzy set in crisp approximation space [9]. Fuzzy-Rough feature selection builds on the notion of fuzzy lower approximation to enable reduction of datasets contains real-valued features and crisp, positive region in traditional Rough set theory is defined as the union of the lower approximations. If the fuzzy-Rough reduction process is to be useful, it must be able to deal with multiple features, finding the dependency between various subsets of the original feature set trying to discover the smallest reduct from a dataset. Reduction performs based on calculating the similarity between features [11, 30]. There are many ways to calculate similarity, but the most obvious one is [11]:

$$\mathbf{SM\ (A,\ B)} = \frac{1}{1+DM(A,B)} \quad (6)$$

$$\mathbf{DM} = \sum_{i=1}^{n} (|Ai - Bi|) \quad (7)$$

where DM represents distance measure of two fuzzy set. DM represents the Euclidean distance to measure distance in crisp values. The Fuzzy-Rough Set, quick reduction algorithm, is listed below:

Input: C, the set of all the conditional attributes.
D, the set of decision attributes
Output: A Reduct RED

RED=ϕ, βbest=0, βprev
Do
T=RED
Bbest = βprev
$\forall x \in (c - RED)$
If β RED $\cup \{x\}$ (D) > βT (D)
T= RED $\cup \{x\}$
βbest = βT(D)
RED=T
Until βprev(D) == βbest(D)
Return RED

Fuzzy-Rough set is critical to deal with continuous values. This paper proposes implementation FuzzyRoughSet in MapReduce based on the similarity between features for feature selection. It improves the efficiency and decreases the complexity.

2.4 MapReduce

The MapReduce framework was first introduced by Google and is now widely used in large-scale data processing on distributed clusters. In the MapReduce model, computation is expressed as two functions: map and reduce as shown in figure 4. The map function takes an input pair and produces a list of intermediate key/value pairs.The intermediate values associated with the same key k2 are grouped together and then passed to the reduce function. The reduce function takes intermediate key k2 with a list of values and processes them to form a new list of values. map(k1,v1) -→ list(k 2, v2)) and reduce(k 2,list(v 2)) → (list(v3)) MapReduce jobs are executed across multiple machines: the map stage is partitioned into map tasks and the reduce stage is partitioned into reduce tasks. The underlying system automatically executes this map and reduce tasks in parallel [12]. MapReduce can be executed in Matlab platform. By parallel mode in Matlab, MapReduce processes data in five stages [13,31].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

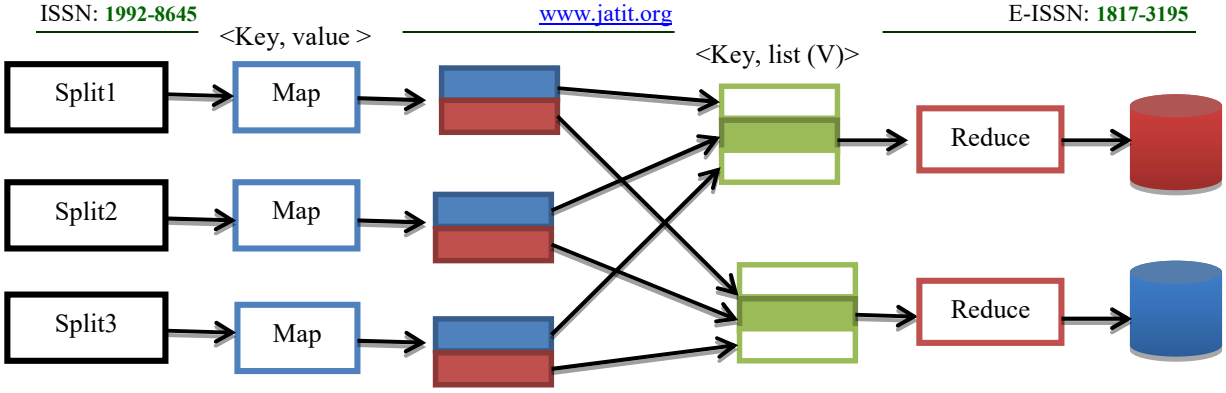

Figure 4. The Architecture Of Mapreduce Framework.

MapReduce implementation pseudo code:

Step1: Read dataset in tabular mode
Ds ← datastore('data file location path','TreatAsMissing','NA')
Step2: Map dataset
Map(Data, intermKVstore)
For each value in input:
Emit intermediate(value, 1);
Step3: reducer function
Reduce (intermkey , interm_value,outKvStore)
Initialize Result as empty;
For each v in interm_value:
Add value to result;
Emit final(intermediate_key,result);
Step4: Producing output
MapReduce←MapReduce(ds,@map,@reduce)

3. RELATED WORK

This section introduces an overview of some previous studies that related to MapReduce and Rough set. It also presents how MapReduce and the Rough set can cooperate to manage to Big data. For example, Pandy et al. [14] discussed the challenges related to big data. They focused on the difficulties that faced Big Data processing and management. A major role of MapReduce was shown in big data processing. They also presented an overview of MapReduce features and how it can eliminate effort and time. They proved that analysis of the performance of MapReduce improved the performance. Data are produced from different sources like social media, digital pictures, and videos.

Nandgaonkar et al. [15] identified how the knowledge discovery is becoming a challenge because of data explosion. They discussed how to implement rough set approximation based on MapReduce. They presented an overview of parallel methods in MapReduce. They discussed what the benefits of using MapReduce are. It helps for processing large data and reduces redundant in processing. Their framework divided into Map function that receives input data and assigns set of keys then transforms similar keys to reducing function. Reduce function receives key and values for sorting and shuffling to create a possibly smaller set of values. MapReduce uses for reducing the number of input dimensions and process for saving time. Proposed method enhances feature selection for large data and mine in Big Data for new knowledge. Their proposed framework recommending computing Rough set equivalence classes and decision classes in parallel using MapReduce technique. Lower and upper approximation computed in parallel also. They discussed some challenges that are faced when using MapReduce for handling Big data.

Zhang [16] discussed how to implement Rough set for discovery knowledge from Big Data in a parallel method based on MapReduce. They proposed parallel methods because traditional methods have difficulties to handle Big Data. They proposed Rough set in parallel mode and evaluated performance. They aimed to work with unstructured data processing by using Rough set based on MapReduce.

Zhai et al. [17] proposed new algorithm. The proposed algorithm consists of four stages. The four stages are finding over sample P times between positive class instances and negative class instances using KNN. Balanced data subsets based on the generated positive class instances, train ELM (Extreme Learning Machine) classifier and integrate ELM classifier with the voting approach. ELM is an algorithm for training single

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

hidden layer feed-forward neural networks. It enhances Rough set performance to redact attributes of Big Data. So Rough set has been developed. The proposed algorithm has good speed up and scales up performance, but this algorithm cannot handle imbalanced large data sets with multiple classes.

Min et al. [18] proposed comparison Rough set tradition reduction algorithm. Based on conditional entropy, mutual information and discernibility matrix. Their traditional methods based on divided decision table S into si with 1000 records. However, it proved that traditional algorithm does not afford the computation when data more than 100.000 records. So they added MapReduce parallel processing concept to be a programming mode. So using parallel computation of MapReduce above Rough set in random records granularity. The proposed algorithm can reduce attribute faster in relatively small number. By decreasing the ratio that important attribute has to remove the redundant attribute.

Yang et al. [19] discussed data processing and knowledge discovery concepts for massive data. The Rough set is used as attribute reduction technique for massive data. They focus on the parallel mode of MapReduce and how to implement Rough set in it. Their Experiment result for five data sets proved that parallel mode is more efficient for Big Data than traditional. The proposed method is more efficient for massive data mining, but it cannot handle the complete data.

Kyong et al. [20] discussed that Google's MapReduce technique makes possible to develop the large-scale distributed applications in a simpler manner and with reduced cost. The main characteristic of MapReduce model is that it is capable of processing large data sets parallel which is distributed across multiple nodes. The novel Map-Reduce software is a proprietary system of Google, and therefore, not available for open use. Although the distributed computing is largely simplified with the notions of Map and Reduce primitives, the underlying infrastructure is non-trivial to achieve the desired performance. A key infrastructure in Google's MapReduce is the underlying distributed file system to ensure data locality and availability. MapReduce and classified its improvements is simple but good scalability and fault-tolerance for massive data processing. Costs of MapReduce still need to be addressed for successful implementation.

Yang et al. [21] discussed computing attribute core for massive data based on Rough set theory and MapReduce is studied. Two algorithms for parallelized computing positive region and attribute core are proposed. A case study verifies the correctness of the proposed algorithm, and comparative experiment results show the effectiveness and high efficiency of the proposed method for data mining on the massive dataset. Attribute reduction and knowledge acquisition based on Rough set theory and MapReduce will be studied continually in the future.

To delete misclassification Smith and Martinez [22,23] proposed PRISM. It helps for enhancing classification performance over outlier identification strategies. PRISM helps for increasing classification accuracy from 78.5% to 79.8% on a set of 53 datasets, What's more, may be statistically critical. In addition, the accuracy on the non-outlier instances increments from 82.8% to 84.7%. They showed how machine learning algorithms handle noise and outlier for generating better models. Our research based on this point at preprocessing step to improve classification performance.

To organize hyperglycemia in the hospitalized inpatient [24] performed a project. It helps for most non-ICU (Intensive Care Unit) patients; they say that inpatient administered economy is optional What's all the more regularly every one of the prompts perhaps no solution on the other hand differences already, glucose at standard managed economy philosophies need help used. Consequently, traditions are proposed. Convincing expectation investigating readmissions engages recuperating offices to recognize Furthermore target patients amid a high peril. Accordingly, the targets are discovering genuine elements helping on mending focuses readmissions and also finding that capable framework for envisioning the kind of readmissions. The last exactness using boosting tree classifier is around 70% precision. That best execution is around 70%~80% precision.

The rough set is used for feature selection. All previous studies discussed how to use the rough set for feature selection in parallel mode using MapReduce. But Big Data has many problems which are challenges for the rough set. Big Data collects from different sources, so data almost has missing data. Missing data is the main problem for the rough set. The rough set takes a lot of time to perform feature selection because of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

calculating of equivalence classes and decision classes. MapReduce costs still problem needs to be addressed.

Because of problems of the rough set, we proposed to implement parallel fuzzy rough set for feature selection. The Rough set also cannot handle any crisp values. We proposed fuzzy rough set can help to remove misclassified instances to improving classification performance and time. Our proposed framework used Matlab implemented MapReduce services to overcome costs problem. The main reasons lead us to proposed previous solutions are real-valued attributes, and the rough set cannot handle noisy data.

4. THE PROPOSED FRAMEWORK

We proposed a framework for handling heterogeneous data in a parallel mode based on MapReduce. Data preprocessing is proposed as the first step. In this step, we remove irrelevant attributes, transform data into the suitable form, and handle missing data by imputing it based on KNN imputation. In the second step, we implemented FuzzyRough set for attribute selection based on MapReduce. After we applied map function, there is a problem to assign key to similar output in reduce function. For solving the previous problem, we cluster dataset for choosing higher membership in clusters and reduact data. Applying clustering to results of map function to select the nearest features to the centroid. In the fourth step, we implement classification techniques to measure evaluation and efficiency of the proposed model.

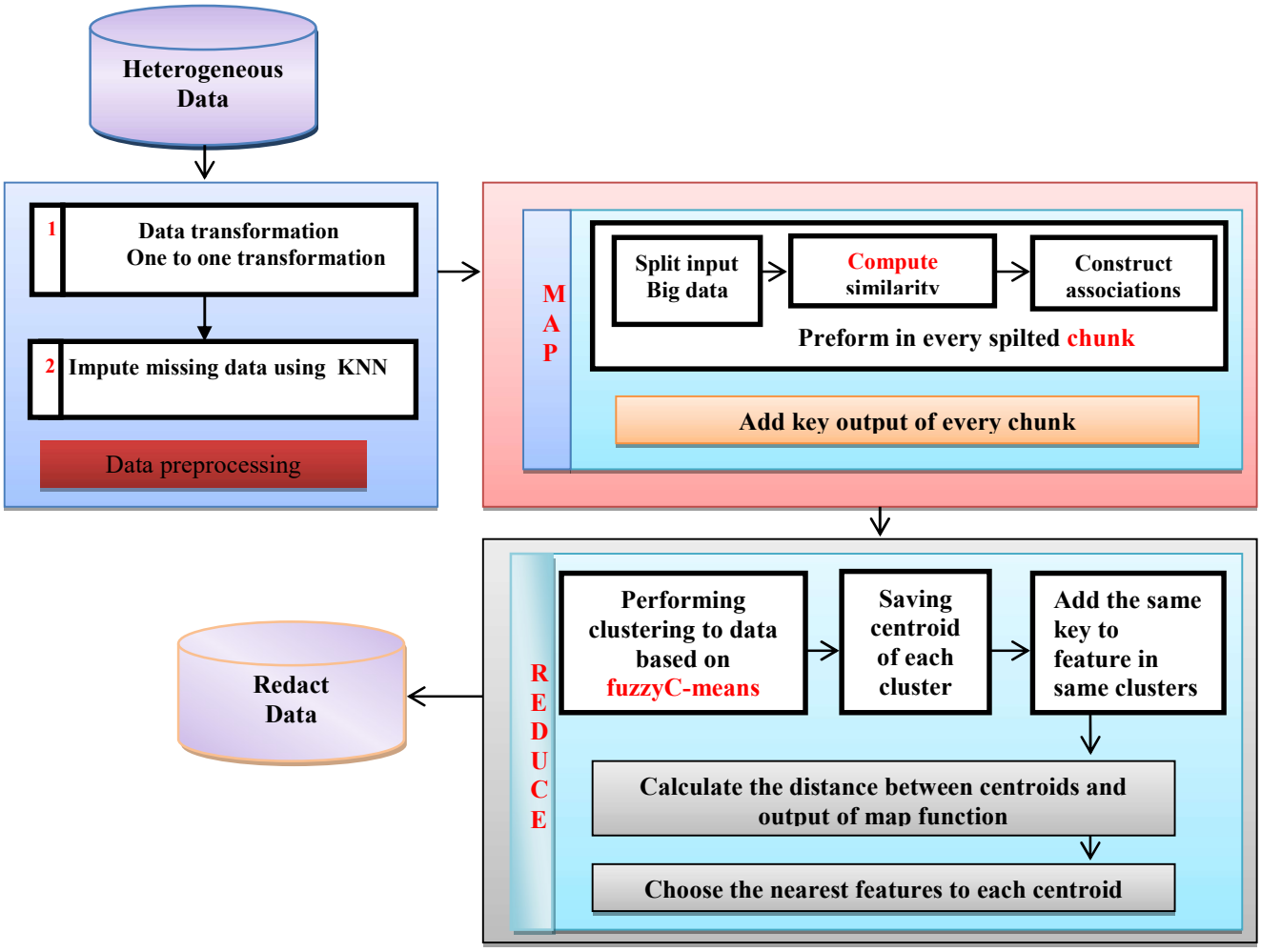

Figure 5. The Block Diagram For The Proposed Framework.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The pseudo code of proposed model:
// **input is heterogeneous data**
Step1: processing data
Transform data using one to one transformation;
Implement missing data using KNN;
Step2: mapping data
Calculate fuzzy similarity between features;
Add key for every output;
Step3: Reducing data
1. Perform data clustering;
2. Save centroid of each cluster;
3. Add the same key to similarity feature and remove redundant;
4. Calculate distance between centroid and features;
5. Assign threshold and choose nearest features in each clusters;
// **output is redact data**

4.1 Data preprocessing

For transform data, we assigned for every nominal value a numeric value. It also called Numerization that aims to transform non-numerical attributes into numeric ones. In mathematical f: X→Y be a function where f is one-to-one if and only if for every y∈Y there is at most one x∈X such that f(x)=y; equivalently, if and only if f(x1) = f(x2) implies x1=x2. It is also called linear transformations. Let T(x)= Ax be a linear transformation. T(x) is one-to-one if the columns of A are linearly independent. T(x) is onto if every row of A has a pivot [25]. MapReduce cannot handle missing data Sowe propose to impute missing data using KNN imputation, as shown in Fig. 5.

4.2 map step

In map step, we perform Rough set concept to apply for every chunk. The central concept in the Fuzzy Rough set is indiscernibility relation or fuzzy equivalence class. All values identical to a subset of considered attributes. Indiscernibility relation is an essential step in the Rough set, but also approximation is the most important phase. It divided into compute lower and upper approximation and boundary region [13,26]. It calculates the fuzzy equivalence class and then decision class. Moreover, at the last approximation of decision class calculated. The existing method calculates fuzzy similarity to measure the similarity between features as shown in equations(1) and (2). Improve the quality and speed of calculating an approximation. Parallel mode is one way where we have lots of opportunities to achieve speed and accuracy [11].

4.3 Reduce step

We perform clustering based on K-means and saving centroids as separated data then cluster the output and choose the nearest features to the centroid. Add a key to the selected features. Clustering proposes because there is a problem to assign to results from map function. We proposed clustering for making it easy to identify similar features and assign then the same key. We calculate Fuzzy C-means centroids based on equation (8).

$$C_j = \frac{\sum_{i=1}^{n}(uij).xi}{\sum_{i=1}^{n}(uij)} \qquad (8)$$

Fuzzyrough map algorithm

```
function mapper(Data,intermKVStore)
num_data=load('dataset')
[m,n]=size(num_data)
similarity1=zeros()
 Initialize intermKeys,intermVals
Initialize r as number of attributes
Foreach k in attributes(1:n)
  Foreach i in attributes(2:n)
     Initialize sim=0
     Initialize sumation=0
     Foreach j in instances(1:m)
       %calculate fuzzy similarity between
       objectives
       sumation=
sumation+abs((num_data(j,k)-num_data(j,i)))
        sim=1/(1+(sumation))
     end
              similarity1(i,k)=sim

  end
   %Assign key to each output
[intermKeys,~,idx] =
    unique(similarity1, 'stable')
    intermVals= similarity1
end
add key
addmulti(intermKVStore,'key',{intermKeys})

end
end
```

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Fuzzy-rough reducer algorithm

function reducer (key, clusters, center, selected features)

initialize number of clusters
perform fuzzycmeans clustering
fcm(data, n_clusters)
return features cluster index
[~,features_clustering_solution] = max(abs(center))
calculate the center of clusters
for j=1:numberof clusters
for i=1:sizeof(data,1)
Center= $\frac{\sum_{i=1}^{n}(uij).xi}{\sum_{i=1}^{n}(uij)}$
endfor
endfor
for i=1:sizeof(data,1)
if(clusterindex==first_cluster)
Calculate fuzzy distance between centers and cluster features using following equation $\frac{1}{1+DM(A,B)}$
elif (clusterindex==second_cluster)
Calculate fuzzy distance between centers and cluster features using following equation $\frac{1}{1+DM(A,B)}$
else
break
endif
assign a threshold for each first cluster 0.17 and second cluster 0.17
return features in the first and second cluster that minimum than the threshold.
Emit(key,selectedfeatures)
end for
End function

4.4 Classification Step

Classification is a critical data mining method with expansive applications to classify Different sorts of data utilized for each field in our existence. Classification is used to classify data based on attributes of data set of the predefined classes. We computed classification performance dependent upon correct and incorrect instances of data. The most important classification techniques are Naïve Bayes, J48, and SVM. A 10-fold cross-validation might have been utilized within training dataset [27].

Decision tree regards as the most usable classification technique for modeling classification problems. One of the simplest decision tree algorithms is j48 classifier which creates a binary tree. Naïve Byes is a simple classification algorithm that builds on the probabilities concept, so it is known as a probabilistic classifier. A set of probabilities is calculated by Naïve Bayes for counting frequency and calculations of given data set values. One of the most accurate supervised machine learning algorithms is SVM (SMO) sequential minimal optimization algorithm for training a support vector classifier. Classification and regression problems can use SVM. Decision tree, Naïve Bayes, and SVM have mostly used classification algorithms which ignore missing values [27]. The equations (9-14) illustrate how to measure classification performance.

Our proposed model is evaluated using six parameters: accuracy, sensitivity, specificity, precision, recall, and f-measure. Accuracy means what is the percentage of instances correctly classified. So, we can calculate error rate by 1-accuracy. Precision means a measure of correctness in instances that positive prediction. The recall is the measure of actual instances that correct prediction. F-measure is a ratio to measure importance of either recall or precision. Specificity and sensitivity are the measures to evaluate Receiver Operating Characteristics (ROC) on any distribution accuracy.

$$\text{Accuracy}= \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

$$\text{Specificity}= \frac{TN}{FP+TN} \quad (10)$$

$$\text{Sensitivity}= \frac{TP}{TP+FN} \quad (11)$$

$$\text{Precision}= \frac{TP}{TP+FP} \quad (12)$$

$$\text{Recall}= \frac{TP}{TP+FN} \quad (13)$$

$$\text{F-measure}= \frac{2*Percision*Recall}{Percision+Recall} \quad (14)$$

where TP (True Positive) is the number of examples correctly classified to that class. The FP (False Positive) is the number of examples incorrectly rejected from that class. Finally, the FN (False Negative) is the number of examples incorrectly classified to that class. The TN (True Negative) is the number of examples correctly rejected from that class.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. EXPERIMENTAL RESULT

We applied our proposed model to two different case studies. They are Diabetes dataset and EEG dataset. Waikato Environment for Knowledge Analysis (Weka) version 3.7.10 and MATLAB R2015a had been used to carry out the experiments.

5.1 Data Description

5.1.1. DATASET1: Diabetes dataset

It is data for 130 US hospitals that collected delivery networks. Dataset covered the full decade of time between 1999 and 2008. It represents the patient's basic data, data log, procedures conducted by the hospital during his/her residence time. It consists of 50 features, and 101.766 instances information was extracted from the database for diabetic encounters. The dataset is too large, which has many features with missing values for example race (2% missing), weight (97% missing), payer code (40% missing), medical especially (50% missing) and diagnosis 3 (1% missing). Therefore, this data needs a lot of preprocessing. The dataset contained some feature to describe patients like race, gender, age, and patient number. Some features describe hospitalization, such as encounter number, admission type, discharge disposition, and admission source. There are many features to indicate multiple inpatient visits and their observations. There are also 24 features to describe some types of medicine proceed to patients. The dataset contains 80% features are nominal, and remaining ones are numeric [24,28].

5.1.2. DATASET 2: EEG dataset

It contains results of EEG (electroencephalography) test, which used to identify the electrical activity of human brain and disorders. It includes 15 features and 14980 instances. It does not have a missing value. All attributes are in real data type. EEG measurement performs with the Emotiv EEG Neuroheadset. The duration of the measurement was 117 seconds. The eye state was detected via a camera during the EEG measurement and manually added later to the file after analyzing the video frames. '1' indicates the eye-closed and '0' the eye-open state. All values are in chronological order with the first measured value at the top of the data [29].

5.2 Model Evaluation

We preprocessed it according to previously showed steps in the framework. We perform classification. A 10-fold cross-validation might have been utilized within training dataset. We divide our dataset into 33% testing and 67% training. Decision tree regards as the most usable classification technique for modeling classification problems. One of the simplest decision tree algorithms is a j48 classifier, which creates a binary tree. Naïve Byes is a simple classification algorithm that builds on the probabilities concept, so it is known as a probabilistic classifier. A set of probabilities is calculated by Naïve Bayes for counting frequency and calculations of given dataset values. One of the most accurate supervised machine learning algorithms is SVM. Classification and regression problems can use SVM. Decision tree, Naïve Bayes, and SVM have mostly used classification algorithms, which ignore missing.

As for our proposed framework, we have to prepare data by putting it in a suitable format using one to one transformation. We remove useless attributes and non-predictive ones. Missing data is an urgent problem, so our model using KNN to impute it. In the second stage, we measure the similarity between dataset features using fuzzy similarity. In the third stage, we proposed to use fuzzy means clustering to clustering data then calculate centroids. We proposed to calculate the distance between each feature and centroid. Then we proposed a threshold for selecting features so three thresholds were assigned 0.1, 0.15, 0.17. As for 0.1, there were 20 features satisfied. As for 0.15, there were 17 features satisfied. As for 0.17, there were 15 features satisfied. But we chose 0.1 as a threshold based on Pawlak's study that discussed many threshold.

In the fourth stage, we proposed to evaluate our proposed model using classification techniques like j48, SVM and Naïve Bayes. We proposed to use two attribute selection techniques traditional GenRSAR implemented in Weka and our proposed fuzzyRoughset to implement based on MapReduce. Our proposed model proved that it effects on classification performance. However, result in table 1 is bad because of the misclassified instance, so we proposed to remove misclassified instance using fuzzyRoughNN. The result improves like in Table 2. Finally, it is shown that FuzzyRough for feature selection can save time to build the model. We measure the performance of classification techniques, and it is obvious that decision tree is the best techniques for performance 86.4 % as shown in Table 3. As for the result of implementing our framework to the second dataset shown in Table 4. Results proved

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

that fuzzy-rough feature selection improves classification performance as shown in figure 6,7.

Table 1. The effect of Attribute Reduction on dataset for classification. performance

	Num of selected features	*SVM*		*NAÏVE BAYES*		*J48*	
		TP	*TN*	*TP*	*TN*	*TP*	*TN*
GenRSAR	27	64.6%	35.4%	64.3%	35.7%	60%	40%
FuzzyRough	17	65%	35%	64.6%	35.4%	64.64%	35.35%

Table 2. The effect of removing misclassified instance using FuzzyRoughNN.

	Num of selected features	*SVM*		*NAÏVE BAYES*		*J48*	
		TP	*TN*	*TP*	*TN*	*TP*	*TN*
GenRSAR	27	84.83%	15.16%	82.05%	17.94%	86.36%	13.64%
FuzzyRough	17	84.83%	15.16%	83.37%	16.6%	87.2%	12.8%

Table 3. Evaluation model for first dataset using classification technique performance j48, SVM and Naïve. Baves

	Accuracy	*Sensitivity*	*Specificity*	*Precision*	*Recall*	*F-Measure*
Naïve Bayes	83.4 %	4.8%	97.4%	76.1 %	83.4 %	78.3%
SVM (SMO)	84.8 %	0%	100%	72%	84.8 %	77.9%
J48	86.4%	31.6%	96.14%	84.3 %	86.4%	84.5%

Table 4. Evaluation classification model for second dataset using classification technique performance j48, SVM and NaïveBaves

FuzzyRoughset						
	Accuracy	*Sensitivity*	*Specificity*	*Precision*	*Recall*	*F-Measure*
Naïve Bayes	42.4%	97.9%	1.9%	55.6%	1.9%	3.7%
SVM (SMO)	57.8%	0%	100%	33.4%	57.8%	42.3%
J48	91.8%	90.1%	93.1%	91.8%	91.8%	91.8%
GenRSAR						
Naïve Bayes	43.3%	96.3%	4.7%	54.5%	43.3%	29.9%
SVM (SMO)	57.7%	0%	99.9%	33.4%	57.7%	42.3%
J48	90.5%	87.8%	92.4%	90.5%	90.5%	90.5%

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

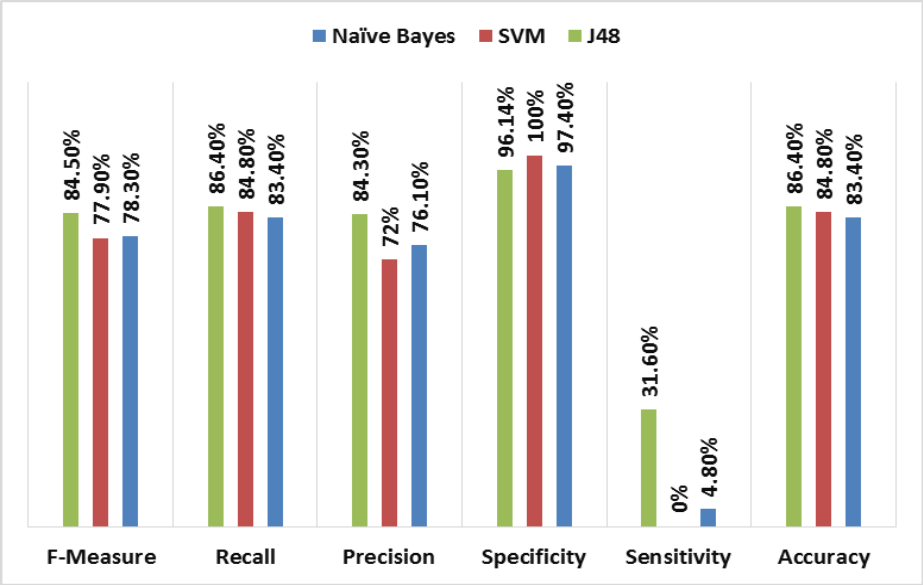

Figure 6. The Effect Of Fuzzy Roughset On Classification Algorithms For First Dataset

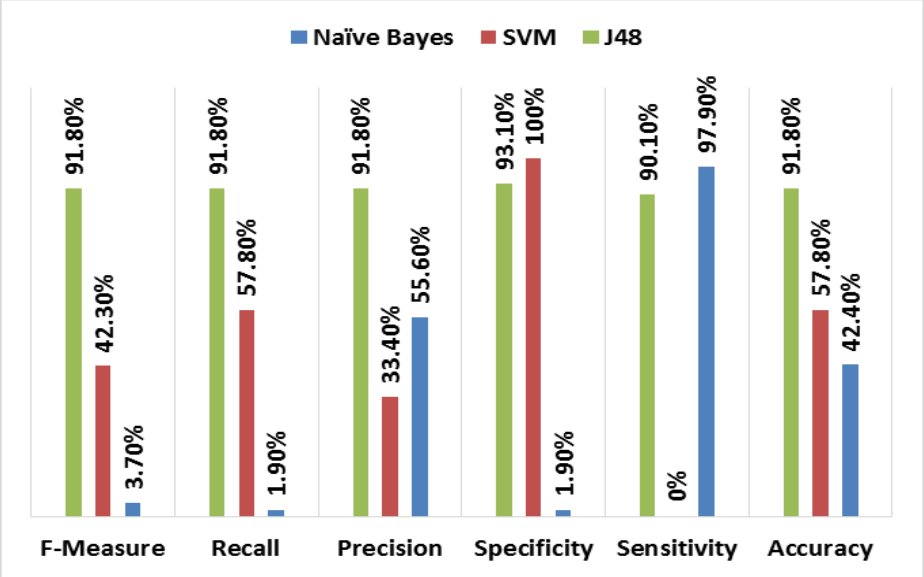

Figure 7. The Effect Of Fuzzyroughset On Classification Algorithms For Second Dataset

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. DISCUSSION

The first dataset has been used for a project of Diabetes Readmission Rate Prediction. The last exactness utilizing boosting tree classifier is around 70% precision, which is the best among all models we utilized. Since haphazardly allocate classmark will yield approximately 30% exactness, 70% precision is a colossal change. There are heaps of examines utilizing this dataset for grouping. The best execution was around 70%~80% accuracy [24]. As for our proposed framework results, decision tree accuracy reaches 86.4% and 84.3 % precision. According to many studies, it is known that FuzzyRough is better than RoughSet. According to our experimental results, it is showed that FuzzyRough set saves time more than Rough set. When we use GenRSAR for feature selection SVM built in 557.3 seconds, NaïveBayes built in 0.92 second and decision tree built in 69 seconds. When we use FuzzyRough for feature selection SVM built in 318 seconds, NaïveBayes built in 0.89 second and decision tree built in 46 seconds.

In the experiment, our framework enhances reduction of Big Data dimensionality. Selecting more important features helps to improve classification performance. After implementation our proposed framework, we gain correctly classified instances. By using different classification techniques on the reduced data set, our proposed model has been achieved the highest accuracy compared to other studies performed on the same dataset. But we hope to improve the result by more hybrid techniques for feature selection.

6. CONCLUSION

Big Data is a huge collection of data from different sources. The big dataset contains a large number of features. Not all data set features are predictive or help in knowledge discovery process, so it is the main step in knowledge discovery process to select features named feature selection. In this paper, we proposed a framework based on MapReduce concept and Rough set for feature selection. Our proposed framework has three main stages which are data preprocessing, map stage, reduce stage. In data preprocessing stage, we try to overcome two main problem heterogeneous data using one to one transformation and incomplete data based on assign fixed number. In map stage, we try to apply FuzzyRough set concepts assigned to feature selection. In reducing stage, we apply fuzzy means clustering for identifying similar features to assign the same key. The aim of handling data without transformation by divides it into a nominal data set and numerical one to avoid the difficulties in assigning a key. We also aim to use more clustering techniques for reducing data and especially hybrid ones that mixed advantages of clustering techniques.

REFERENCES

[1] Chandarana, Parth and M. Vijayalakshmi (2014), "Big Data Analytics Frameworks." *2014 International Conference on Circuits, Systems*, Communication and Information Technology Applications (CSCITA), DOL: 10.1109/CSCITA.2014.6839299.

[2] Dachao Xi, Guoyin Wang, Xuerui Zhang, and Fan Zhang (2014), Parallel Attribute Reduction Based On MapReduce."." *Rough Sets and Knowledge Technology, Springer International Publishing Switzerland 2014*, DOL: 10.1007/978-3-319-11740-9_5.

[3] Asir Antony Gnana Singh, Jebamalar Leavline, E. Priyanka, C.Sumathi,"Feature Selection Using Rough Set for Improving The Performance Of The Supervised Learner." *International Journal of Advanced Science and Technology* 87.01 (2016): 1-8.

[4] Devi Priya R and Sivaraj R (2015), "A Review Of Missing Data Handling Methods.", *International Journal On Engineering Technology and Sciences – IJETS*, ISSN (P): 2349-3968, ISSN (O): 2349-3976 Volume 2 Issue 2.

[5] Han Jiawei, Micheline Kamber, and Jian Pei. Data Mining: Concepts And Techniques. 3rd ed. USA: *Elsevier*, 2000.

[6] Atzmueller, Martin, Andreas Schmidt, and Martin Hollender. (2016), "Data Preparation for Big Data Analytics: Methods & Experiences." In: *Enterprise Big Data Engineering, Analytics, and Management, IGI Global (In Press).*

[7] Satyanarayana, N, CH Ramalingaswamy, and Y Ramadevi. (2014),"Survey of Classification Techniques In Data Mining." *IJISET - International Journal of Innovative Science, Engineering & Technology*, Vol. 1 Issue 9.

[8] Z. Pawlak, "Rough Sets," Int. J. Compute Inf. Sci., vol. 11, 1982, pp. 341–356.

[9] D. Chen, Q. Hu, and Y. Yang, "Parameterized attribute reduction with Gaussian kernel based fuzzy rough sets," *Information Sciences,* vol. 181, 2011, pp. 5169–5179.

[10] X.D. Liu, W. Pedrycz, T.Y. Chai, and M. L. Song, "The development of fuzzy rough

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

sets with the use of structures and algebras of axiomatic fuzzy sets," *IEEE Trans. Knowl. Data Eng.,* vol .21, no. 3, 2009, pp. 443–462.

[11] Richard Jensen. (2005)" Combining Rough and fuzzy sets for feature selection," Doctor of Philosophy, School of Informatics, University of Edinburgh.

[12] Zhuoyao Zhang, Boon Thau Loo, Insup Lee, Val Tanne. (2014)," Performance Modeling and Resource Management for MapReduce Applications."

[13] CSE, ME, HVPM COET, and Amravati Maharashtra.(2015) "Parallel Rough Set Approximation Using MapReduce Technique in Hadoop." *International Journal of Advanced Technology in Engineering and Science (IJATES)* 3.01.

[14] Pandey, Shweta and Vrinda Tokekar. (2014),"Prominence Of MapReduce In Big Data Processing." *2014 Fourth International Conference on Communication Systems and Network Technologies.*

[15] Nandgaonkar, Suruchi, and Raut. (2015), "A Survey On Parallel Method For Rough Set Using MapReduce Technique For Data Mining." *International Journal Of Engineering And Computer Science 4.9.*

[16] Zhang, Junbo, Tianrui Li, and Yi Pan. (2012), "Parallel Rough Set Based Knowledge Acquisition Using MapReduce From Big Data." Published in: · *Proceeding BigMine '12 Proceedings of the 1st International Workshop on Big Data, Streams, and Heterogeneous Source Mining: Algorithms, Systems, Programming Models, and Applications.*

[17] Zhai, Junhai, Sufang Zhang, and Chenxi Wang. (2015), "The Classification Of Imbalanced Large Data Sets Based On MapReduce And Ensemble Of ELM Classifiers." *International Journal of Machine Learning and Cybernetics.*

[18] Su-min, Yang.(2014),"An Improved Attribute Reduction Algorithm Based On Mutual Information With Rough Sets Theory." *Journal of Chemical and Pharmaceutical Research 6.3.*

[19] Yang, Yong.(2010),"Attribute Reduction For Massive Data Based On Rough Set Theory And MapReduce." Lecture Notes in Computer Science.

[20] Kyong, Lee, Choi, Chung, Moon.(2011), "Parallel Data Processing with MapReduce: A Survey." *Published in SIGMOD Record,* Vol. 40, No. 4.

[21] Yang, Yong and Zhengrong Chen. (2012), "Parallelized Computing Of Attribute Core Based On Rough Set Theory And MapReduce." *Springer-Verlag Berlin Heidelberg 2012.*

[22] M. Smith and T. Martinez,(2011) "Improving Classification Accuracy by Identifying and Removing Instances that Should Be Misclassified," Neural Networks (IJCNN), *The 2011 International Joint Conference on*, pp. 2690 – 2697.

[23] M. Smith and T. Martinez.(2013)." An Extensive Evaluation of Filtering Misclassified Instances in Supervised Classification Tasks," vol. 1.

[24] S. Meng, (2015), "Data Mining for Diabetes Readmission Rate Prediction." [Online]. Available at: https://github.com/siyuan1992/EE660_Project (Accessed 20- May- 2016).

[25] Arturo Magidin (2012), Is. "Is A Linear Transformation Onto Or One-To-One?". Available at: http://math.stackexchange.com/questions/26371/is- a-linear-transformation-onto-or-one-to-one , Math.stackexchange.com,(Accessed 27 Jan 2017).

[26] Zhi, Huilai. (2014), "Realization Of Rough Set Approximation Topological Operations Based On Formal Concept Analysis." *International Journal of Intelligence Science".*

[27] T. Patil and S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", *International Journal of Computer Science And Applications,* vol. 6, no. 2, 2013.

[28] UCI,[online].(2014),"Diabetes 130-US Hospitals For Years 1999-2008 Data Set".Available_at:https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008, (Accessed 22 May 2016).

[29] "EEG Eye State Data Set," UCI, 2013, https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State [Last accessed 10 May 2016].

[30] Shangzhu Jin, Jun Peng, Dong Xie, "Towards MapReduce approach with dynamic fuzzy inference/interpolation for big data classification problems," *Cognitive Informatics & Cognitive Computing (ICCI*CC), 2017*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IEEE 16th International Conference, Oxford, United Kingdom, November 2017.

[31] Ibrahim Adel Ibrahim; Mostafa Bassiouni, "Improving MapReduce Performance with Progress and Feedback Based Speculative Execution," *Smart Cloud (SmartCloud), 2017 IEEE International Conference,* New York, USA, 23 November 2017.

[32] JunboZhang, Jian-syuanWong, TianruiLi, YiPan, "A comparison of parallel large-scale knowledge acquisition using rough set theory on different MapReduce runtime systems,"*International Journal of Approximate Reasoning(Elsevier),* https://doi.org/10.1016/j.ijar.2013.08.003, 2014.