Received: April 5, 2019

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Parallel Rough Set Theory for Nonlinear Dimension-Reduction in Big Data Analysis

Amsaveni Muthusamy[1]* **Duraisamy Subramani[2]**

[1]*Department of Computer Science, AVP College of Arts and Science, Tirupur, India*
[2]*Department of Computer Science, Chikkanna Government Arts College, Tirupur, India*
* Corresponding author's Email: amsaveniphd2019@gmail.com

Abstract: Attribute reduction for big data is an important preprocessing step in the area of data mining. A multi-step dimension reduction approach was proposed for attribute reduction in big data. It addressed the non-linear relationships within the attributes. The data dimension was reduced through a parametric mapping. The mapping parameters were estimated using low-rank Singular Value Decomposition (SVD). However, the user-defined criterion in multi-step dimension reduction approach has greatly influenced the efficiency of attribute reduction. This approach was proposed for a single machine that means the entire big data must fit in the main memory and the parallelism was limited. So, in this paper, parallel rough set theory based attribute reduction approach is proposed for attribute reduction in big data. Based on two descriptions of lower approximation and upper approximation, a rough set is constructed. Then a reduct is detected using inner importance measure and outer importance measure. The rough set theory is used in MapReduce framework to achieve the parallelism for attribute reduction in big data. Hence, the computation time is reduced by using parallel rough set theory based attribute reduction approach. Finally, the experiments are carried out in Amazon customer review, REUTERS-21578 and International Cancer Genome Consortium (ICGC) on AWS datasets to prove the effectiveness of parallel rough set theory based attribute reduction in terms of accuracy, precision, recall and computation time.

Keywords: Big data, Attribute reduction, Rough set theory, MapReduce, Singular value decomposition.

1. Introduction

With an increasing amount of industrial and scientific datasets, mining useful information from big data is growing today for business intelligence. The classical data mining techniques are becoming more challenging from both the perspectives of data and computational intensives. Because the sizes of the datasets are too massive to fit the main memory and the search space sizes are too large to explore using a single machine. It is known that all the attributes in datasets are not necessary for the decision-making process. It increases the search space and makes the generalization more difficult. Hence attribute reduction process is carried out as a preprocessing step in knowledge acquisition to find a minimal subset without compromising the classification accuracy.

Attribute reduction [1] can reduce data dimensions, reduce unnecessary storage and irrelevant input and significantly improve the efficiency of data processing. It is also known as feature selection in machine learning and pattern recognition [2], it reduces the complexity of handling big data. It also helps people better understand the data by telling them which are key features, and has been attracted much attention in recent years. Attribute reduction from large data is an expensive preprocessing step. Most of the attribute reduction algorithms were proposed for a single machine that means the entire data must fit in the main memory and the parallelism was limited.

A multi-step dimension reduction approach [3] was proposed for attribute reduction in big data. In this approach, the number of dimensions in the data was divided into groups. Using the low dimension of Singular Value Decomposition (SVD) the data were transformed into this group. The dimension in each group was reduced with the consideration of non-linear relationships within these dimensions and then merged the newly extracted dimensions. After each step of attribute reduction, an average of singular values across all groups was verified this provided an idea about the amount of information retained at each step. When the average information retained at each step became less than the information threshold then the split and merge process of multi-step dimension reduction approach. However, this approach may be affected by the user-defined information threshold.

So, in this article, parallel rough set theory based attribute reduction is introduced to reduce the attributes in big data. A rough set is constructed based on upper approximation and lower approximation using information granules from a granular structure. The upper approximation is a description of the instances which possibly belongs to the subset whereas the lower approximation is a description of the domain instances which are known with certainty to belong to the subset of interest. A subset of an attribute that fully portrays the knowledge in the big data is called reduct. During the computation of information granules, all the attributes in the dataset have to be scanned to construct a granular structure. It is computationally time-consuming. So, a rough set theory based attribute reduction is used in MapReduce framework to reduce the computation time and improve the efficiency of attribute reduction. In each mapper of MapReduce, rough set theory is processed in a parallel manner and the results of attribute reduction in each mapper are collected in a reducer of MapReduce. Hence the computational time for rough set theory based attribute reduction is reduced.

The rest of the article is structured as follows: Section 2 presents the literature survey related to the attribute reduction. Section 3 describes the proposed methodology. Section 4 illustrates the experimental results of the proposed method. Finally, Section 5 concludes the research work and presents future enhancement.

2. Literature survey

For attribute reduction in big data, a Multiagent-Consensus-MapReduce-based Attribute Reduction (MCMAR) algorithm [4] was proposed. A Particle Swarm Optimization (PSO) with self-adaptive memeplexes was designed to partition the particles into different memeplexes which located global best region. Then, partition the big attribute sets by constructed four layers neighborhood radius framework with a compensatory scheme where parallelism was achieved by using MapReduce mechanism. However, this algorithm may not be adequately reliable for attribute reduction.

A general framework called Forward Attribute Reduction (FWAR) [5] was proposed for attribute reduction in incomplete ordered information systems. This framework handled incomplete data by integrating dominance-based rough sets with α-cut sets. In this framework, the discernibility functions and judgment theorems were established by applying Boolean reasoning techniques for attribute reduction. Furthermore, near-optimal attribute reducts were determined through designing backward and forward attribute reduction algorithms for inconsistent and consistent systems respectively. However, this framework still consumes high computational time for attribute reduction.

A discernibility matrix based incremental attribute reduction algorithm [6] was proposed for attribute reduction in the dataset which contains dynamic data. This algorithm obtained all reducts including optimal reducts of dynamic data. In addition to this, another incremental attribute reduction algorithm was developed to improve the efficiency of the discernibility matrix based incremental attribute reduction algorithm. In real applications, the objects of dynamic data may vary in groups. In order to deal this, run the incremental algorithm based on their mechanism according to the number of changing objects which affect the performance of attribute reduction.

An attribute reduction method using distance measure was proposed called fuzzy rough set-based attribute reduction [7]. Initially, a fuzzy rough set model was built based on distance measure with a fixed parameter and then it was replaced by a variable one to better differentiate attribute reduction with fuzzy rough sets. In addition to this, an iterative model based on variable distance parameter was proposed and based on this a greedy convergent algorithm was designed for attribute reduction. In some cases, such as an improper selection of membership function, the proposed algorithm still has low classification accuracy.

A parallel attribute reduction algorithm [8] was proposed in the Dominance-based Neighborhood Rough Set (DNRS). Parallel computing was applied to improve the efficiency of attribute reduction. It handled a large amount of data to be processed by a single computer. Partitioning, Communication, Agglomeration and Mapping (PCAM) was a general framework which had a hybrid decision system for attribute reduction. However, this framework has the problem of maintaining a huge volume of knowledge from a hybrid decision system.

A novel fuzzy rough set model [9] was proposed for attribute reduction in multi-label learning. Based on the advantages of ensemble learning, local sampling was employed to obtain a more robust distance between the target sample and its nearest miss sample. It solved the low separability of fuzzy similarity relation on high dimensional multi-label data. Finally, a multi-label fuzzy dependency function was described, and a forward greedy attribute reduction algorithm was proposed to select optimal multi-label attribute subset. However, there is no consistent evidence to indicate the efficiency of a novel fuzzy rough set model.

An attribute reduction method [10] was proposed based on Max-Decision Neighborhood Rough Set model (MDNRS) for attribute reduction. This method focused on the boundary samples and enlarged the positive region by adding the samples whose neighborhoods have a maximal intersection with some decision classes. An attribute reduction model was designed based on this idea. It effectively removed the most redundant attributes without compromising the classification accuracy. However, the classification accuracy of attribute reduction method is low.

Hierarchical attribute reduction algorithm [11] was proposed for big data using MapReduce. In this algorithm, the hierarchical decision table was defined and then the relationships of hierarchical decision tables were discussed under different levels of granularity. Parallel computation of the equivalence classes and the attribute significance were designed for attribute reduction. Finally, MapReduce was used to design a hierarchical attribute reduction algorithm. However, the computational complexity of this algorithm is high.

3. Proposed method

In this section, the proposed parallel rough set theory based attribute reduction is discussed in detail. The computational complexity of classical rough set theory is reduced by using the rough set theory in the MapReduce framework. It handled the big data in a distributed environment. MapReduce helps to split the big data into many small blocks equal to the number of mappers and process the rough set theory in each mapper for the attribute reduction process. In the rough set theory, reduct returns the most informative attributes in the dataset.

3.1 Parallel rough set theory

The big data may consist of noise, stop words, special characters, etc. These are more challenging to classify the data. So, the collected big is preprocessed to remove the noise, stop words and special characters. In the pre-processing step, the stemming and lemmatization processes are applied. Then, the big data attributes are reduced by using rough set theory. The classical rough set theory is based on the assumption that some information is related to every instance in a dataset. It is the approximation of vague set by a pair of concepts called upper approximations and lower approximations. The upper approximation is a description of the instances which possibly belongs to the subset whereas the lower approximation is a description of the domain instances which are known with certainty to belong to the subset of interest which are calculated by information granules. Information granules are a bunch of instances drawn together by similarity, indistinguishability, connectivity, and proximity of functionality.

Consider U be a finite and non-empty set of attributes in dataset and R be an equivalence relation on U, (U, R) be an approximation space and $\mathcal{D}$ an inclusion degree defined on $\mathcal{P}(U) \times \mathcal{P}(U)$. Then for any instance, $I \subseteq U$, the α-lower and β-upper approximations are defined by

$$\begin{cases} \underline{R}_\alpha(I) = \left\{ i | \mathcal{D}\left(\frac{I}{[i]_R}\right) \geq \alpha, i \in I \right\} \\ \overline{R}_\beta(I) = \cup \left\{ [i]_R | \mathcal{D}\left(\frac{I}{[i]_R}\right) > \beta, i \in I \right\} \end{cases} \quad (1)$$

In the Eq. (1), $[i]_R$ is the equivalence class including i, $\underline{R}_\alpha(I)$ and $\overline{R}_\beta(I)$ are the lower and upper approximation with respect to R respectively. The pair $\left(\underline{R}_\alpha(I), \overline{R}_\beta(I)\right)$ is called a rough set. A boundary of I is denoted by $BN_R(I) = \overline{R}_\beta(I) - \underline{R}_\alpha(I)$ which is called boundary region of I.

The local rough set needs to compute only the information granules of instances within a given target concept for determination of upper or lower approximations. Generally, the main goal of every attribute reduction process is to determine the subset of attributes that maximize the margin between the classes. In rough set theory based attribute reduction, the margin is illustrated by the boundary region between the lower approximation and upper approximation of I.

Assume, $S = (U, A)$ be an information system where A be a non empty set of attributes, $I \subseteq U$ a target concept and $B \subseteq A$. If $\left|\underline{R_B}(I)\right| > \left|\underline{R_A}(I)\right|$ and $\left|\underline{R_{B'}}(I)\right| \ngeq \left|\underline{R_A}(I)\right|$ for any $B' \subset B$, where B is a local attribute reduct of S with respect to I. From the above assumption, multiple reducts may exist for a target concept in an information system. A heuristic algorithm is designed with a greedy and forward search strategy to find the attribute reduct. In this algorithm, two important measures inner importance measure and outer importance measure are used for heuristic function. The inner importance measure and outer importance measure are calculated as,

$$Sig_{\alpha}^{inner}(a, B, I, U) = \left|\underline{R_B}_{\alpha}^{U}(I)\right| - \left|\underline{R_{B-\{a\}}}_{\alpha}^{U}(I)\right| \quad (2)$$

In the Eq. (2), $I \in U, B \subset A, \forall a \in B$. It determines the significance of every attribute in the dataset.

$$Sig_{\alpha}^{outer}(a, B, I, U) = \left|\underline{R_{B\cup\{a\}}}_{\alpha}^{U}(I)\right| - \left|\underline{R_B}_{\alpha}^{U}(I)\right| \quad (3)$$

In the Eq. (3), $I \in U, B \subset A, \forall a \in A - B$. It is used in the attribute selection process.

In attribute reduction approach, starting with the attribute with the maximal inner and outer importance is taken into the attribute subset in each loop until this attribute subset satisfies the stopping criterion and then get an attribute reduct. An attribute reduction approach for finding an attribute reduct with respect to a target concept is summarized as follows.

Attribute Reduction Algorithm for Target Concept:

```
Input: $S = (U, A)$, $\alpha$ and $I \subseteq U$
Output: reduct $red$
1. Assign $red = null$
2. Calculate $Sig_{\alpha}^{inner}(a_k, A, I, U), k \leq A$
3. if $\begin{pmatrix} Sig_{\alpha}^{inner}(a_k, A, I, U) = max \\ \left(Sig_{\alpha}^{inner}(a_k, A, I, U)\right) \end{pmatrix}$
4. Put $a_k$ into $red$
5. End if
6. Assign $x = 1, R_1 = red, I_1 = I, U_1 = U$
7. while $\left(\left|\underline{R_{red}}_{\alpha}^{U_x}(I_x)\right| < \left|\underline{R_A}_{\alpha}^{U_x}(I_x)\right|\right)$
8. $I_{x+1} = \cup_{i \in I_x} [i]_{red}^{U_x} - CL_{red}^{U_x}(I_x)$
9. $x = x + 1$
10.if $\begin{pmatrix} Sig_{\alpha}^{outer}(a_0, red, I_x, U_x) = max \\ \{Sig_{\alpha}^{outer}(a_k, red, I_x, U_x), a_k \in C - red\} \end{pmatrix}$
11. $red = red \cup \{a_0\}$
12. End if
13. $R_x = R_{x-1} \cup \{a_0\}$
14. End while
15. Return $red$.
```

In attribute reduction algorithm for target concept, red is an attribute to conserve the selected attributes, $Sig_{\alpha}^{inner}(a_k, A, I, U)$ is the inner significance of attributea_k, $Sig_{\alpha}^{outer}(a_k, red, I_x, U_x)$ is the outer significance of attribute a_k, $CL_{red}^{U_x}(I_x)$ is a certain set of red, C is the conditional attribute set. In order to find the attribute reduct of a target decision, the same inner importance measure and outer importance measure are calculated. Consider $S = (U, C \cup D)$ be a decision table (class label) and $B \subseteq C$. If $|POS_B(D)| \geq |POS_C(D)|$ and $|POS_{B'}(D)| \ngeq |POS_C(D)|$ for any B', then B is called a local attribute reduct of S. C is a condition attribute set, D is a decision attribute set, $POS_B(D)$is the positive region of B with respect to D and $POS_C(D)$ is the positive region of C with respect to D. From the above consideration, multiple attribute reducts for a decision table are induced with class labels. Hence, the reducts are obtained by using the inner importance measure and outer importance measure which are given as follows:

$$Sig^{inner}(a, B, D, U) = \delta_B(D) - \delta_{B-\{a\}}(D) \quad (4)$$

In the Eq. (4), $Sig^{inner}(a, B, D, U) = \delta_B$ is the inner significance measure of a in B, $S = (U, C \cup D)$ be a decision table, $B \subseteq C$ and $\forall a \in B$.

$$\delta_B(D) = \frac{|POS_B(D)|}{|U|} \quad (5)$$

$$Sig^{outer}(a, B, D, U) = \delta_{B\cup\{a\}}(D) - \delta_B(D) \quad (6)$$

In the Eq. (6), $Sig^{outer}(a, B, D, U)$ is the outer significance measure of a in B, $B \subseteq C$ and $\forall a \in C - B$.

Attribute Reduction Algorithm for Target Decision:

```
Input: $S = (U, C \cup D)$ and $\alpha$.
Output: reduct $red$
1. Assign $red = null$
2. Calculate $Sig^{inner}(a_k, C, D, U), k \leq |C|$
```

```
3.if $\left(Sig^{inner}(a_k, C, I, U) = \max\left(Sig^{inner}(a_k, C, U, D)\right), a_k \in C\right)$
4. Put $a_k$ into $red$
5. End if
6. Assign $x = 1, R_1 = red, I_1 = I, U_1/D = \{I_x^y, y \leq r\}$
7. while $\left(\left|POS_{red}^{U_x}(D)\right| < \left|POS_{C}^{U_x}(D)\right|\right)$
8. $U_{x+1} = \bigcup_{y=1}^{r}[i]_{red}^{U_x}, i \in I_x^y, I_x^y \in \frac{U_x}{D} - \bigcup_{y=1}^{r} CL_{red}^{U_x}(I_x^y), I_x^y \in \frac{U_x}{D}$
9. $I_{x+1}^y = I_x^y - CL_{red}^{U_x}(I_x^y), I_x^y \in \frac{U_x}{D}$
10. $x = x + 1$
11.if $(Sig^{outer}(a_0, red, U_x, D) = max\{Sig^{outer}(a_k, red, U_x, D), a_k \in C - red\})$
12. $R_x = R_{x-1} \cup \{a_0\}$
13. End if
14. End while
15. Return $red$.
```

In attribute reduction algorithm for target decision algorithm, red is an attribute to hold the selected attributes, $Sig^{inner}(a_k, C, D, U)$ is the inner significance of attribute a_k, $Sig^{outer}(a_k, C, D, U)$ is the outer significance of attribute a_k and r is the mutually exclusive crisp subsets. Step 7 provides a stopping criterion. In STEP 8, the reduced universe is computed and in step 9, the update every gradually reduced target concept. Finally, red returns the reduced attribute for target decision. The time complexity for computing all granules is high which cannot satisfy the requirement of efficient computation of big data. So, rough set theory is used in MapReduce framework to reduce the computational complexity.

A decision table $S = (U, C \cup D)$ is divided into m-sub decision tables which satisfies $U = \bigcup_{h=1}^{o} U_h$, $U_g \cap U_m = \emptyset, \forall g, m \in \{1,2, \ldots o\}$ and S_h is a sub-decision table of S. Here, a parallel method is designed for the calculation of attribute reduction based on MapReduce. It includes three steps are Map, Reduce and Merge. Initially, the input big data are divided into h blocks as $(U_1, B \cup D), (U_2, B \cup D), \ldots (U_h, B \cup D)$ they have a similar size and stored in different computing nodes.

Map phase: Each Map worker partitions $U_h, h = \{1,2, \ldots o\}$ by the condition attribute set B.

Reduce phase: Each Reduce worker collects the data which own the same key $\overrightarrow{I_{xB}}$ and partitions I_x by the decision D. Then red is collected using attribute reduction algorithm for target decision.

Merge phase: Collect the red from each reducer. This step will be done in a master worker.

Parallel Rough Set Theory based Attribute Reduction Algorithm

```
Input: $S = (U, C \cup D), red = \emptyset$
Output: reduct $red$
// Map(big data, $(U_h, C \cup D)$)
1. for each $i \in U_h$
2. Calculate $\frac{U_h}{B} = \{I_{h1}, I_{h2}, \ldots I_{ht}\}$
3. End for
4. Collect $(\overrightarrow{\iota_B}, \overrightarrow{\iota_D})$
5.End Map
//Reduce $\left(\overrightarrow{I_{xB}}, S_x'\right)$ , where $S_x' = (I_x, D), I_x = \bigcup_{h=1}^{o} I_{hx}$
6. $\frac{I_x}{D} = J_{x1}, J_{x2}, \ldots J_{xn}$
7. $J_{xy} = I_x \cap J_y$
8. Collect $(red_h)$ using attribute reduction algorithm for target decision
9. End Reduce
//Merge $(red_h)$
10. for each $h$ to $o$
11. $red = red \cup red_h$
12. End Merge
```

Hence the attributes in the dataset are reduced by the above parallel rough set theory based attribute reduction algorithm. The reduced attributes are given to three different classifiers are Support Vector Machine (SVM), AdaBoost and Random Forest (RF) to classify the data.

4. Experimental results

In this section, the efficiency of existing MCM-AR [4], FW-AR [5] and SVD based Attribute Reduction (SVD-AR) [3] is compared with proposed a Parallel Rough set Theory based Attribute Reduction (PRT-AR) in terms of accuracy, precision, recall and computation time. For the experimental purpose, Amazon customer review dataset, REUTERS-21578 text dataset and International Cancer Genome Consortium (ICGC) on AWS dataset are used. An Amazon customer review is a collection of reviews written in Amazon.com marketplace. It consists of 130Million+ customer reviews from 5 different countries. The REUTERS-21578 text dataset contains 21578 Reuters news documents from 1987. From 21578 documents, 9603 documents are used for training, 3299 documents are used for testing and 8676 documents are unused. The ICGC on AWS consists of data about cancer. It consists of 2178 donors and 40152 files.

Table 1. Comparison of Accuracy

<table>
<tr><th colspan="8">Amazon customer review</th></tr>
<tr><td rowspan="2">MCM-AR-SVM</td><td rowspan="2">FW-AR-SVM</td><td colspan="3">SVD-AR</td><td colspan="3">PRT-AR</td></tr>
<tr><td>AdaBoost</td><td>SVM</td><td>RF</td><td>AdaBoost</td><td>SVM</td><td>RF</td></tr>
<tr><td>0.51</td><td>0.6</td><td>0.64</td><td>0.8</td><td>0.84</td><td>0.71</td><td>0.86</td><td>0.9</td></tr>
<tr><th colspan="8">REUTERS-21578 text</th></tr>
<tr><td rowspan="2">MCM-AR-SVM</td><td rowspan="2">FW-AR-SVM</td><td colspan="3">SVD-AR</td><td colspan="3">PRT-AR</td></tr>
<tr><td>AdaBoost</td><td>SVM</td><td>RF</td><td>AdaBoost</td><td>SVM</td><td>RF</td></tr>
<tr><td>0.54</td><td>0.62</td><td>0.7</td><td>0.85</td><td>0.9</td><td>0.76</td><td>0.89</td><td>0.94</td></tr>
<tr><th colspan="8">ICGC on AWS</th></tr>
<tr><td rowspan="2">MCM-AR-SVM</td><td rowspan="2">FW-AR-SVM</td><td colspan="3">SVD-AR</td><td colspan="3">PRT-AR</td></tr>
<tr><td>AdaBoost</td><td>SVM</td><td>RF</td><td>AdaBoost</td><td>SVM</td><td>RF</td></tr>
<tr><td>0.5</td><td>0.53</td><td>0.6</td><td>0.74</td><td>0.79</td><td>0.67</td><td>0.8</td><td>0.86</td></tr>
</table>

4.1 Accuracy

Accuracy is the measure of correctly classify the data based on the reduced attributes in all instances. It can be calculated by

$$Accuracy = \frac{(True\ Positive + True\ Negative)}{(True\ Positive + True\ Negative + False\ Positive + False\ Negative)} \quad (7)$$

In the Eq. (7), True Positive is actual positive data which are exactly classified as positives, True Negative is the actual negative data which are classified exactly as negatives, False Positive is known negative data which are wrongly classified as positives and False Negative is known positive data which are wrongly classified as negatives.

Table 1, shows the comparison of accuracy between MCM-AR-SVM, FW-AR-SVM, SVD-AR-AdaBoost, SVD-AR-SVM, SVD-AR-RF, PRT-AR-AdaBoost, PRT-AR-SVM, and PRT-AR-RF for three different datasets.

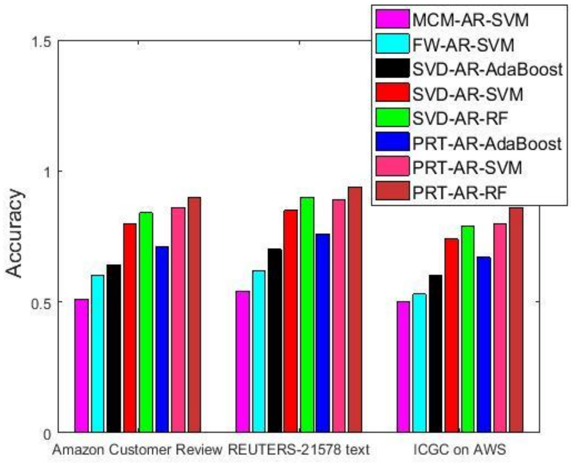

Figure.1 Comparison of accuracy

Fig. 1 shows the comparison of accuracy between existing MCM-AR-SVM, FW-AR-SVM, SVD-AR with different classifiers and proposed PRT-AR with different classifiers. By constricting proper lower approximation and upper approximation, rough set theory reduces the attributes more effectively which increase the accuracy of classifiers. The accuracy of proposed PRT-AR-RF method for amazon customer review dataset is 76.5% greater than MCM-AR-SVM, 50% greater than FWAR-SVM, 40.6% greater than SVD-AR-AdaBoost, 12.5% greater than SVD-AR-SVM, 7.1% greater than SVD-AR-RF, 26.8% greater than PRT-AR-AdaBoost, 4.7% greater than PRT-AR-SVM. From this analysis, it is proved that the proposed PRT-AR method has high accuracy than the other attribute reduction methods.

4.2 Precision

Precision value is evaluated according to the relevant information of classification at true positive prediction and false positive prediction.

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \quad (8)$$

Table 2, shows the comparison of precision between MCM-AR-SVM, FW-AR-SVM, SVD-AR-AdaBoost, SVD-AR-SVM, SVD-AR-RF, PRT-AR-AdaBoost, PRT-AR-SVM, and PRT-AR-RF for three different datasets.

Fig. 2 shows the comparison of precision between existing MCM-AR-SVM, FW-AR-SVM, SVD-AR with different classifiers and proposed PRT-AR with different classifiers. The precision of PRT-AR method is high because it has ability to deal with all kinds of irregular data and uncertain data. The precision of proposed PRT-AR-RF method for amazon customer review dataset is 69.2% greater than MCM-AR-SVM, 42.8% greater

Table 2. Comparison of Precision

Amazon customer review							
MCM-AR-SVM	FW-AR-SVM	SVD-AR			PRT-AR		
		AdaBoost	SVM	RF	AdaBoost	SVM	RF
0.52	0.57	0.63	0.78	0.83	0.69	0.85	0.88
REUTERS-21578 text							
MCM-AR-SVM	FW-AR-SVM	SVD-AR			PRT-AR		
		AdaBoost	SVM	RF	AdaBoost	SVM	RF
0.55	0.61	0.68	0.84	0.89	0.75	0.86	0.92
ICGC on AWS							
MCM-AR-SVM	FW-AR-SVM	SVD-AR			PRT-AR		
		AdaBoost	SVM	RF	AdaBoost	SVM	RF
0.5	0.54	0.61	0.75	0.78	0.68	0.81	0.85

Table 3. Comparison of recall

Amazon customer review							
MCM-AR-SVM	FW-AR-SVM	SVD-AR			PRT-AR		
		AdaBoost	SVM	RF	AdaBoost	SVM	RF
0.54	0.58	0.62	0.76	0.83	0.68	0.86	0.87
REUTERS-21578 text							
MCM-AR-SVM	FW-AR-SVM	SVD-AR			PRT-AR		
		AdaBoost	SVM	RF	AdaBoost	SVM	RF
0.59	0.62	0.66	0.82	0.87	0.74	0.85	0.90
ICGC on AWS							
MCM-AR-SVM	FW-AR-SVM	SVD-AR			PRT-AR		
		AdaBoost	SVM	RF	AdaBoost	SVM	RF
0.51	0.55	0.60	0.73	0.77	0.65	0.82	0.84

than FW-AR-SVM, 39.7% greater than SVD-AR-AdaBoost, 12.8% greater than SVD-AR-SVM, 6% greater than SVD-AR-RF, 27.5% greater than PRT-AR-AdaBoost, 3.5% greater than PRT-AR-SVM. From this analysis, it is proved that the proposed PRT-AR method has high precision than the other attribute reduction methods.

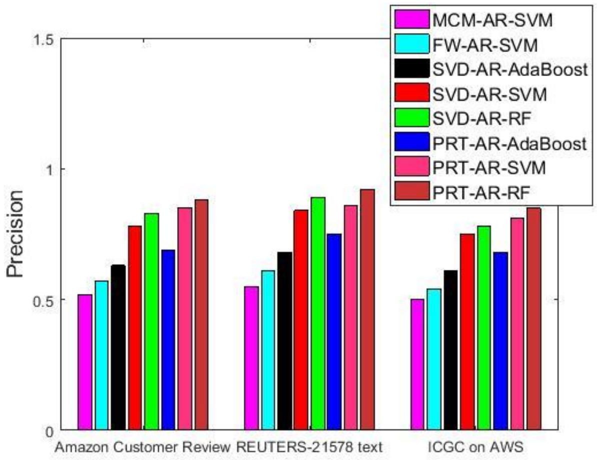

Figure.2 Comparison of precision

4.3 Recall

Recall is evaluated according to classification at true positive and false negative predictions.

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)} \quad (9)$$

Table 3, shows the comparison of recall between MCM-AR-SVM, FW-AR-SVM, SVD-AR-AdaBoost, SVD-AR-SVM, SVD-AR-RF, PRT-AR-AdaBoost, PRT-AR-SVM, and PRT-AR-RF for three different datasets.

Fig. 3 shows the comparison of recall between existing MCM-AR-SVM, FW-AR-SVM, SVD-AR with different classifiers and proposed PRT-AR with different classifiers. The recall of PRT-AR method is high because it identified and evaluated the dependencies among the data by finding reducts. The recall of proposed PRT-AR-RF method for amazon customer review dataset is 61.1% greater than MCM-AR-SVM, 50% greater than FW-AR-SVM, 40.3% greater than SVD-AR-AdaBoost, 14.5% greater than SVD-AR-SVM, 4.8% greater

Table 4. Comparison of computation time (S)

Amazon customer review							
MCM-AR-SVM	FW-AR-SVM	SVD-AR			PRT-AR		
		AdaBoost	SVM	RF	AdaBoost	SVM	RF
39.56	34.14	28.32	22.29	18.01	17.31	13.67	10.11
REUTERS-21578 text							
MCM-AR-SVM	FW-AR-SVM	SVD-AR			PRT-AR		
		AdaBoost	SVM	RF	AdaBoost	SVM	RF
30.79	24.42	20.32	14.56	10.24	11.75	9.17	7.01
ICGC on AWS							
MCM-AR-SVM	FW-AR-SVM	SVD-AR			PRT-AR		
		AdaBoost	SVM	RF	AdaBoost	SVM	RF
37.84	33.98	30.24	24.45	20.01	20.24	15.45	12.21

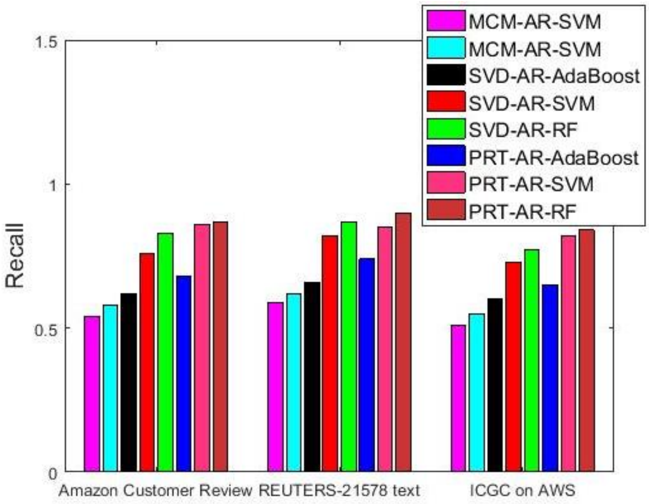

Figure.3 Comparison of recall

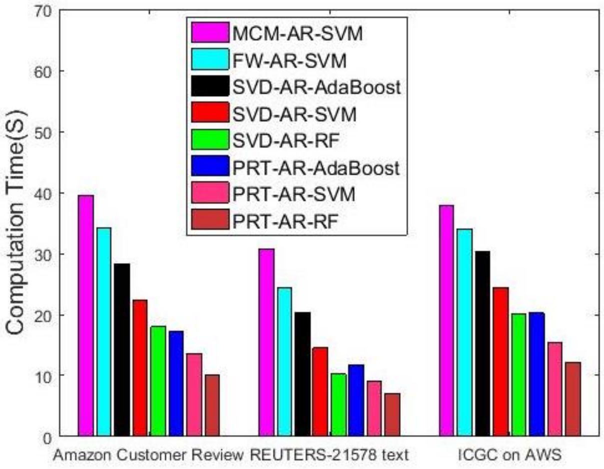

Figure.4 Comparison of computation time

than SVD-AR-RF, 27.9% greater than PRT-AR-AdaBoost, 1.2% greater than PRT-AR-SVM. From this analysis, it is proved that the proposed PRT-AR method has high recall than the other attribute reduction methods.

4.4 Computation time

Computation time is the amount of time taken for the classification using the reduced attributes.

Table 4, shows the comparison of computation time between MCM-AR-SVM, FW-AR-SVM, SVD-AR-AdaBoost, SVD-AR-SVM, SVD-AR-RF, PRT-AR-AdaBoost, PRT-AR-SVM and PRT-AR-RF for three different datasets.

Fig. 4 shows the comparison of computation time between existing MCM-AR-SVM, FW-AR-SVM, SVD-AR with different classifiers and proposed PRT-AR with different classifiers. By using MapReduce function, the big data are processed at different systems hence the computation time for attribute reduction is reduced effectively. The computation time of proposed PRT-AR-RF method for amazon customer review dataset is 74.4% less than MCM-AR-SVM, 70.4% less than FW-AR-SVM, 64.3% less than SVD-AR-AdaBoost, 54.6% less than SVD-AR-SVM, 43.9% less than SVD-AR-RF, 41.6% less than PRT-AR-AdaBoost, 26% less than PRT-AR-SVM. From this analysis, it is proved that the proposed PRT-AR method has less computation time than the other attribute reduction methods.

5. Conclusion

The main contribution of this paper is to reduce the number of attributes which improves the further process in big data. In PRT-AR method, the big data is given as input to the MapReduce framework, where the mappers process the rough set theory for attribute reduction. Then, reducers combine the results of each mapper which returned a final reducts. Finally, the experimental result proved that the proposed PRT-AR method has high accuracy, precision, recall and less computation time than other attribute reduction methods for Amazon customer review, REUTERS-21578 dataset and ICGC on AWS datasets. For instance, during the analysis of attribute reduction methods in Amazon customer review dataset, the accuracy, precision, recall of PRT-AR-RF is 40.6%, 39.7% and 40.3% greater than SVD-AR-AdaBoost respectively. The computation time of PRT-AR-RF is 64.3% less than SVD-AR-AdaBoost. It proves the effectiveness of the proposed PRT-AR method. However, this work is fully concentrated on attribute reduction in big data, in future other processes such as big data clustering and big data classification will be concentrated for big data processing.

References

[1] H. Zhang, F. Kong, and H. Wang, “Research on Power Big Data Processing Attribute Reduction Method Based on Cloud Computing Technology”, *IOP Conf. Series on Materials Science and Engineering,* Vol.452, No.4, pp.1-6, 2018.

[2] J. Zhang, T. Li, and Y. Pan, “Parallel large-scale attribute reduction on cloud systems”, *arXiv preprint arXiv:1610.01807,* pp.1-14, 2016.

[3] R. Krishnan, V. A. Samaranayake, and S. Jagannathan, “A Multi-step Nonlinear Dimension-reduction Approach with Applications to Bigdata”, *Procedia Computer Science,* Vol.144, pp.81-88, 2018.

[4] W. Ding, C. T. Lin, S. Chen, X. Zhang, and B. Hu, “Multiagent-consensus-MapReduce-based attribute reduction using co-evolutionary quantum PSO for big data applications”, *Neurocomputing,* Vol.272, pp.136-153, 2018.

[5] W. Qian and W. Shu, “Attribute reduction in incomplete ordered information systems with fuzzy decision”, *Applied Soft Computing*, Vol.73, pp.242-253, 2018.

[6] W. Wei, X. Wu, J. Liang, J. Cui, and Y. Sun, “Discernibility matrix based incremental attribute reduction for dynamic data”, *Knowledge-Based Systems*, Vol.140, pp.142-157, 2018.

[7] C. Wang, Y. Huang, M. Shao, and X. Fan, “Fuzzy rough set-based attribute reduction using distance measures”, *Knowledge-Based Systems*, Vol.164, pp.205-212, 2019.

[8] H. Chen, T. Li, Y. Cai, C. Luo, and H. Fujita, “Parallel attribute reduction in dominance-based neighborhood rough set”, *Information Sciences*, Vol.373, pp.351-368, 2016.

[9] Y. Lin, Y. Li, C. Wang, and J. Chen, “Attribute reduction for multi-label learning with fuzzy rough set”, *Knowledge-Based Systems*, Vol.152, pp.51-61, 2018.

[10] X. Fan, W. Zhao, C. Wang, and Y. Huang, “Attribute reduction based on max-decision neighborhood rough set model”, *Knowledge-Based Systems*, Vol.151, pp.16-23, 2018.

[11] J. Qian, P. Lv, X. Yue, C. Liu, and Z. Jing, “Hierarchical attribute reduction algorithms for big data using MapReduce”, *Knowledge-Based Systems*, Vol.73, pp.18-31, 2015.