Explorations into MapReduce based Parallel Reduct Computation

Pandu Sowkuntla
pandu.sowkuntla@uohyd.ac.in
University of Hyderabad
Hyderabad, Telangana, India

Abstract

Feature selection is the process of selecting a minimal subset of features that provide the same classification ability as the given set of attributes. Prof. Pawlak introduced Rough Set Theory (RST), which has emerged as an effective framework for feature selection. RST based feature selection also known as attribute reduction or reduct computation.

This research is motivated by the challenges presented by today's massive data sets: big dimensionality, variety of the data, and data partitioning strategy of MapReduce framework. Existing approaches for attribute reduction in categorical data sets (decision systems with categorical attributes (CDS)) adopted horizontal partitioning (HP) strategy for partitioning the data to the cluster of nodes. This strategy results in computational overheads for big dimensional data sets. Furthermore it presents an immense problem if the data set contains missing values (in incomplete decision systems (IDS)), or if the data set contains different types of attributes (in hybrid decision systems (HDS)).

This thesis proposes a MapReduce based approach for attribute reduction in CDS with big dimensionality, that investigates the vertical partitioning (VP) strategy. Approaches for IDS are proposed using Novel Granular Framework (NGF) (an extension to RST) and adopt HP and VP strategies. Fuzzy-rough sets (an extension to RST) based accelerator is introduced, and approaches are proposed using HP and VP strategies. The proposed approaches are implemented using Apache Spark. Experimental analysis carried out on benchmark large-scale data sets with the variance in object and attribute space. The results show that the HP based approaches perform well for the larger object space data sets with moderate attribute space. And the VP based approaches scale well for big dimensional data sets with moderate object space. In future, this research has a scope to explore approaches that can simultaneously scale in both huge object and attribute spaces.

Keywords: Rough set theory, Attribute reduction, Apache Spark, Horizontal partitioning, and Vertical partitioning

ACM Reference Format:
Pandu Sowkuntla. 2021. Explorations into MapReduce based Parallel Reduct Computation. In *Proceedings of Doctoral Symposium: First International Conference on AI-ML Systems, 21-24 October 2021 (Doctoral Symposium).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Over the last few years, there has been an exponential increase in data generation per day. Mining knowledge from the large amounts of data is a challenging task because of the uncertainty and inconsistency in the data. *Feature subset selection* is one of the approaches that helps in exploiting the data redundancy to reduce the uncertainty from large-scale data sets. Feature subset selection is the process of selecting a minimal subset of features that provide the same classification ability as the whole attributes.

1.1 Rough Set Theory for feature selection

In 1982, Prof. Pawlak [9] introduced Rough Set Theory (RST) as Soft Computing paradigm. RST is effective in dealing with uncertainty and vagueness in the data. It has become an area of great interest to the researchers for the feature selection.

Feature subset selection using rough sets principles is known as *attribute reduction* or *reduct computation*. The selected feature subset is termed

Doctoral Symposium, 21-24 October 2021, Bangalore, India

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

as *reduct*. Thus the "*reduct is a minimal subset of conditional attributes that provide the same classification ability as the set of conditional attributes in the decision system*". In RST, the reduct computation methods are primarily categorized into: (i) *Dependency measure*, and (ii) *Discernibility matrix*. In the current research, the methods are proposed based on both the dependency measure and discernibility matrix approaches which are fall under *filter* based methods of feature selection.

Due to the exponential growth of data, if the data set is large-volume and/or high dimensional, the traditional (sequential) attribute reduction algorithms can not perform well. Most of the researchers found parallel/distributed computation as the good solution for scalable attribute reduction. Therefore, researchers try to parallelize the traditional attribute reduction algorithms to achieve scalability.

1.2 MapReduce programming model

According to a comprehensive review of the literature on reduct computation over the last two decades, it is observed that, researchers are more likely to adopt the MapReduce model [2] for creating parallel/distributed approaches than non-MapReduce models (traditional parallel/distributed computation models such as MPI, OpenMP and BSP). Because the MapReduce framework provides a consistent structure for deriving granular aggregated information in the *reduce* phase using constructed partial granules information in the *map* phase, which is crucial for achieving rough set based attribute reduction. Thus, in recent years (especially last decade) several MapReduce approaches were proposed for attribute reduction in large-scale data sets [11–13].

The current research also adopts MapReduce model for proposing the approaches. Hadoop, Twister, Apache Spark, etc., are some of the existing MapReduce frameworks. The proposed approaches in this research are implemented on Apache Spark framework [14].

2 Research motivation and objectives

Several researchers have been interested in attribute reduction in large-scale data sets, and many approaches have been proposed. The majority of these approaches are hindered by the challenges presented by today's massive data sets. The current research is motivated by these challenges, which include *big dimensionality*, *variety of the data* in large-scale data sets, and *data partitioning* strategy used to divide the input data set.

2.1 Big dimensionality

Similarly to big data, the term "*big dimensionality*" [1] has been invented to describe the enormous amount of attributes reaching to levels that render existing attribute reduction approaches ineffective. Some of the most prominent areas that deal with big dimensionality challenges are, *microarray analysis*, *text* and *image classification*.

To accelerate the attribute reduction in large data sets, many classical rough set-based methods have been developed using the MapReduce model. Despite the effectiveness of these methods, they are confronted with various issues and are unable to effectively and efficiently handle the large-scale data sets with big dimensionality.

2.2 Impact of data partitioning

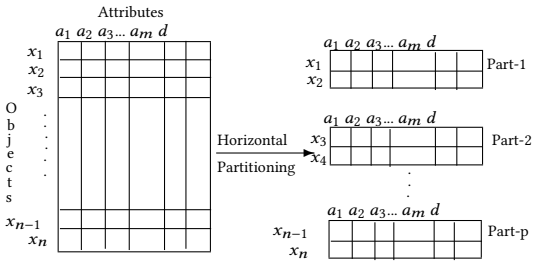

Figure 1. Horizontal partitioning of the data

All the existing rough set theory based attribute reduction approaches [4, 10, 15] for CDS using the MapReduce model adopted horizontal partitioning strategy (shown in Figure 1) for partitioning the input data to the cluster of nodes. This strategy results in computational overheads for the big dimensional data sets. Because, with this strategy, considerable amount of data to be communicated across shuffle and sort phase and a complex reduce phase is involved in any MapReduce framework. This has inspired us to look into alternative data partitioning strategy that avoid the problems of horizontal partitioning.

2.3 Variety of the data

In the current research, the *variety of the data* refers to missing object values (incomplete) of attributes in the data sets or different types of attributes (e.g., categorical, numerical,...etc.) in the data sets. The decision systems with categorical (or discrete) attributes are known as Categorical Decision Systems (CDS). The decision systems that

include objects with missing attribute values are referred to as Incomplete Decision Systems (IDS). The decision systems with different types of attributes are known as hybrid decision systems (HDS). The extensions to classical rough sets such as tolerance rough sets [7] and fuzzy-rough sets [3] are used to deal with IDS and HDS respectively. Attribute reduction in these decision systems pose much severe computational challenges and involve higher space and time complexities in building MapReduce based approaches.

From the literature, it is observed that, all the existing approaches for attribute reduction in IDS are sequential methods, parallel/distributed approaches have not proposed. Different strategies in MapReduce framework are needed to parallelize the existing extensions of classical rough sets. This has motivated us to investigate MapReduce based approaches to deal with massive incomplete data sets.

From extensive review of literature [5, 6], it is observed that, the approaches for attribute reduction in HDS involve higher space and time complexities compared to classical rough sets. It is also observed that a substantial decrease in memory usage is achieved in the discernibility matrix based approach relative to the dependency measure based approach. Further discernibility matrices are more suitable for performing parallel/distributed computation. The advantages of discernibility matrix over dependency measure, and also non availability of discernibility matrix based scalable approaches in the literature encouraged us to investigate MapReduce approaches based on discernibility matrix.

2.4 Research objectives

Each of the research problems mentioned in preceding section form the objectives of this research.

- The first objective of this research is to investigate an alternative data partitioning strategy known as "*vertical partitioning*", which is used to partition the input data set to the nodes of the cluster. The applicability of this strategy is explored for rough set based attribute reduction in large-scale CDS with big dimensionality.
- The second objective is to explore MapReduce based attribute reduction approaches for large-scale IDS that uses existing Novel Granular Framework (NGF) (an extension to classical rough sets) to handle the incompleteness in the data and adopt horizontal and vertical partitioning strategies.
- The third and fourth objectives of this research are to explore discernibility matrix based attribute reduction in large-scale HDS using MapReduce with the strategies of horizontal and vertical partitioning.

The summary of aforementioned objectives of this research can be enunciated as follows:

"***This research objective is to explore MapReduce based parallel/ distributed reduct computation in Categorical, Incomplete and Hybrid decision systems, where the relevance of horizontal and vertical partitioning strategies are investigated in partitioning the data to the nodes of the cluster.***"

3 Research contributions

Contributions to this research are made in relation to the research objectives outlined in the preceding section. A brief summary of each contribution is given below.

3.1 Parallel attribute reduction in Categorical Decision Systems

A MapReduce based algorithm MR_IQRA_VP is proposed using vertical partitioning strategy for attribute reduction. The vertical partitioning is used alternative to horizontal partitioning that partitions the input data set in attribute space to the nodes of the cluster. Figure 2 shows vertical partitioning of the input data. From the figure, with the vertical partitioning strategy, all the objects' information of a subset of attributes is available in a data partition of a node in the cluster.

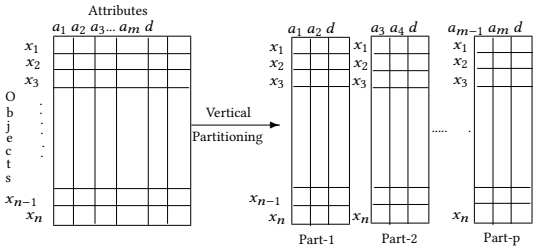

Figure 2. Vertical partitioning of the data

With vertical partitioning strategy we have managed a massive reduction in data transformation and communication in the shuffle and sort phase of Apache Spark, which is a primary bottleneck of the horizontal partitioning based reduct algorithms.

The positive region removal and granular refinement features are successfully incorporated into MR_IQRA_VP algorithm, which delivered huge computational gains.

The advantages and limitations of the proposed MR_IQRA_VP algorithm is theoretically and experimentally studied, inferences are obtained through comparative analysis with horizontal partitioning based algorithms: PLAR [15], IN_MRIQRA_IG [10] and PFSPA [4]. The efficiency of the MR_IQRA_VP is evaluated based on the *computational evaluation* (*Running time* and *reduct size* metrics are used), *performance evaluation* (*Speed up*, *scale up* and *size up* metrics are used) and *impact of the partitioning strategy*. Extensive experimental results showed that MR_IQRA_VP is a more suitable and scalable algorithm for the big dimensional data sets having moderate size object space.

The work in this contribution has been published in [13].

3.2 Parallel attribute reduction in Incomplete Decision Systems

MapReduce based parallel/distributed approaches are proposed based on the Novel Granular Framework (NGF) [7] for attribute reduction in IDS using horizontal and vertical partitioning strategies. Briefly, this contribution includes the following:

- An alternative representation of the NGF is proposed and adopted to develop the MRIDS_HP algorithm. This algorithm uses the horizontal partitioning strategy.
- Algorithm MRIDS_VP is developed by parallelizing the existing NGF based on the strategy of the vertical partitioning.

Since, the algorithm MRIDS_HP is using horizontal partitioning strategy, the positive region removal feature is incorporated but granular refinement [7] is not incorporated. But MRIDS_VP algorithm incorporates both the features.

It is worth to mention that, to the best of our knowledge, the proposed approaches are the first research of its kind on parallel/distributed attribute reduction in IDS. The efficiency of the proposed approaches is proved based on computational and performance evaluation. Different benchmark data sets with variance in incompleteness percentage are used in experimental analysis. With extensive experimental analysis and theoretical validation, the proposed MRIDS_HP algorithm has been proven to be efficient and more suitable for the incomplete data sets with massive number of objects and moderate number of attributes. Similarly, the MRIDS_VP algorithm has been shown to be effective and ideal for the big dimensional data sets having modest object space. The computational and performance evaluation demonstrated that the proposed methods are efficient in attribute reduction even if we have huge number of missing values in the data.

The work in this contribution has been published in [11].

3.3 Parallel attribute reduction in Hybrid Decision Systems

In this contribution Fuzzy Discernibility Matrix (FDM) based approaches are proposed for scalable fuzzy-rough attribute reduction in HDS. In summary, the contribution include the following.

1) A fuzzy discernibility matrix based attribute reduction accelerator (DARA) is introduced. Based on this accelerator, a sequential algorithm IFDMFS (Improved Fuzzy Discernibility Matrix based Feature Selection) is developed for attribute reduction in HDS.
2) Further, MR_IFDMFS algorithm is developed using horizontal partitioning strategy. This algorithm is MapReduce based version of IFDMFS.

We carried out the experiments in two stages to evaluate the proposed sequential IFDMFS and parallel MR_IFDMFS algorithms. In the first stage of the experiments, the comparative results of the IFDMFS algorithm are presented by comparing with the existing PARA [6] algorithm, which is an accelerator for fuzzy-rough reduct computation. In the second stage of the experiments, the efficiency of MR_IFDMFS is provided by comparing with the existing state-of-the-art algorithms: MR_FRDM_SBE [8] and DFRS [5].

The merits and limitations of the IFDMFS and MR_IFDMFS algorithms are proved through extensive computational and performance evaluation. The comparative study of IFDMFS and PARA has shown experimentally that the proposed algorithm is useful in achieving shorter length reducts with substantial computational gains. The significant computational gains achieved by MR_IFDMFS algorithm over the existing approaches on all the data sets strongly establishes the role of the proposed accelerator DARA in imparting space reduction as the algorithm progresses and there by aiding in reduction of computational time.

The work in this contribution has been published in [12].

4 Conclusions and Future work

In this research work, we explored MapReduce based reduct computation in categorical, incomplete and hybrid decision systems, where the relevance of horizontal and vertical partitioning strategies was investigated in partitioning the input data to the nodes of the cluster.

The proposed approaches were implemented in Apache Spark. Since the proposed approaches are parallel/distributed, this research concentrated on the computational evaluation, performance evaluation and impact of the partitioning strategy. No experiments were performed to measure the quality of downstream tasks after using reduced attributes. Extensive experimental study was performed on several benchmark data sets with variations in object and attribute space. The experimental results, as well as theoretical validation, demonstrated that the horizontal partitioning-based methods perform effectively for huge object space data sets. Vertical partitioning-based methods were relevant and scale well for data sets with big dimensionality. And, the proposed approaches were found to have better computational gains over existing state-of-the-art approaches.

This research's fourth contribution, *Parallel attribute reduction in hybrid decision systems utilising vertical partitioning strategy*, will be finished shortly. In future, this research has the potential to look at viable rough set-based MapReduce approaches that can simultaneously scale in both huge object space and attribute space. From the proposed algorithms, it can be observed that, they dealt with *volume* and *variety* characteristics of big data. Therefore, this research offers the opportunity to investigate suitable MapReduce methods that can deal with the *velocity* characteristic of big data by the creation of streaming-based incremental approaches.

5 Acknowledgments

Author is grateful to Digital India Corporation (MeitY, Govt. of India) for providing the fellowship under Visvesvaraya Ph. D. Scheme during the years 2016-2021 [Unique id: MEITY-PHD-1039]. This work is funded by DST, Govt. of India under ICPS project [Grant Number: File No. DST/ICPS/CPS-Individual/2018/579] and by AICTE, Govt. of India under RPS project [grant number: File no. 8-47/RIFD/RPS/POLICY-1/2016-17]. This research is also supported through UoH-IoE by MHRD Govt. of India [Grant Number: F11/9/2019-U3(A)].

References

[1] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. 2015. Recent advances and emerging challenges of feature selection in the context of big data. *Knowledge-Based Systems* 86 (2015), 33–45.

[2] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. *Commun. ACM* 51, 1 (jan 2008), 107.

[3] Didier Dubois and Henri Prade. 1992. Putting rough sets and fuzzy sets together. In *Intelligent Decision Support*. Springer, 203–232.

[4] Qing He, Xiaohu Cheng, Fuzhen Zhuang, and Zhongzhi Shi. 2014. Parallel feature selection using positive approximation based on mapreduce. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2014 11th International Conference on*. IEEE, 397–402.

[5] L. Kong, W. Qu, J. Yu, H. Zuo, G. Chen, F. Xiong, S. Pan, S. Lin, and M. Qiu. 2020. Distributed Feature Selection for Big Data Using Fuzzy Rough Sets. *IEEE Transactions on Fuzzy Systems* 28, 5 (2020), 846–857.

[6] Peng Ni, Suyun Zhao, Xizhao Wang, Hong Chen, and Cuiping Li. 2019. PARA: A positive-region based attribute reduction accelerator. *Information Sciences* 503 (nov 2019), 533–550.

[7] P. S. V. S Sai Prasad and Raghavendra Rao Chillarige. 2012. Novel Granular Framework for Attribute Reduction in Incomplete Decision Systems. In *International Workshop on Multi-disciplinary Trends in Artificial Intelligence*. Springer, 188–201.

[8] Neeli Lakshmi Pavani, Pandu Sowkuntla, K. Swarupa Rani, and P. S. V. S. Sai Prasad. 2019. Fuzzy Rough Discernibility Matrix Based Feature Subset Selection With MapReduce. In *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*. IEEE, 389–394.

[9] Zdzisław Pawlak. 1982. Rough sets. *International journal of computer & information sciences* 11, 5 (1982), 341–356.

[10] P. S. V. S. Sai Prasad, H. Bala Subrahmanyam, and Praveen Kumar Singh. 2016. Scalable IQRA_IG Algorithm: An Iterative MapReduce Approach for Reduct Computation. In *Distributed Computing and Internet Technology*. Springer International Publishing, 58–69.

[11] Pandu Sowkuntla, Sravya Dunna, and PSVS Sai Prasad. 2021. MapReduce based parallel attribute reduction in Incomplete Decision Systems. *Knowledge-Based Systems* 213 (2021), 106677.

[12] Pandu Sowkuntla and PSVS Sai Prasad. 2021. MapReduce based parallel fuzzy-rough attribute reduction using discernibility matrix. *Applied Intelligence* (2021), 1–20.

[13] Pandu Sowkuntla and P. S. V. S. Sai Prasad. 2020. MapReduce based improved quick reduct algorithm with granular refinement using vertical partitioning scheme. *Knowledge-Based Systems* 189 (feb 2020), 105104.

[14] Matei Zaharia, Reynold S Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J Franklin, et al. 2016. Apache spark: a unified engine for big data processing. *Commun. ACM* 59, 11 (2016), 56–65.

[15] Junbo Zhang, Tianrui Li, and Yi Pan. 2016. Parallel large-scale attribute reduction on cloud systems. *arXiv preprint arXiv:1610.01807* (2016).