ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

AN EFFICIENT PARALLEL ROUGH SET APPROXIMATION AND RULE GENERATION USING MAP REDUCE TECHNIQUE

N.Sukanya
Department Of Computer Science and Engineering
United Institute Of Technology, Coimbatore,India
Mail-Id:sukanyabecse@gmail.com

ABSTRACT

Massive data mining and knowledge discovery present a tremendous challenge with the data volume growing at an unprecedented rate. Rough set theory has been successfully applied in data mining. The lower and upper approximations are basic concepts in rough set theory. The effective computation of approximations is vital for improving the performance of data mining or other related tasks. Map Reduce technique has gained a lot of attention from the scientific community for its applicability in massive data analysis. Process large-scale incomplete data with rough set theory, introduced the matrix representations of lower and upper approximations in IIS. According to the characteristics of the matrix, has proposed three parallel strategies based on MapReduce to compute approximations.

INTRODUCTION

- Roughset can be used for feature selection, feature extraction, data reduction, decision rule generation, and pattern extraction (templates, association rules) etc.

Basic Concepts of Rough Sets

- Information/Decision Systems (Tables)
- Indiscernibility
- Set Approximation
    - ❖ Upper Approximation
    - ❖ Lower Approximation

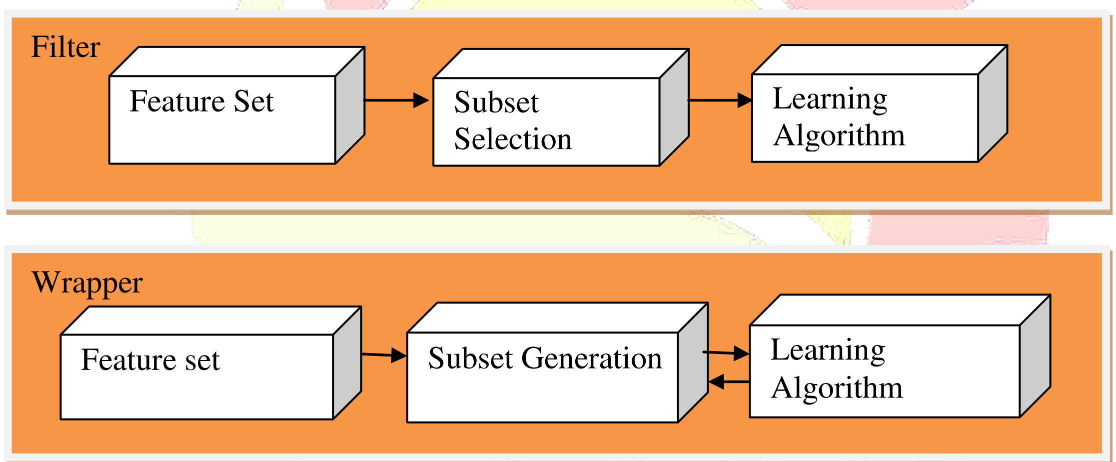
ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

FEATURE SELECTION

The main aim of feature selection (FS) is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features. In real world problems FS is a must due to the abundance of noisy, irrelevant or misleading features. For instance, by removing these factors, learning from data techniques can benefit greatly. Given a feature set size n, the task of FS can be seen as a search for an \optimal" feature subset through the competing 2n candidate subsets.



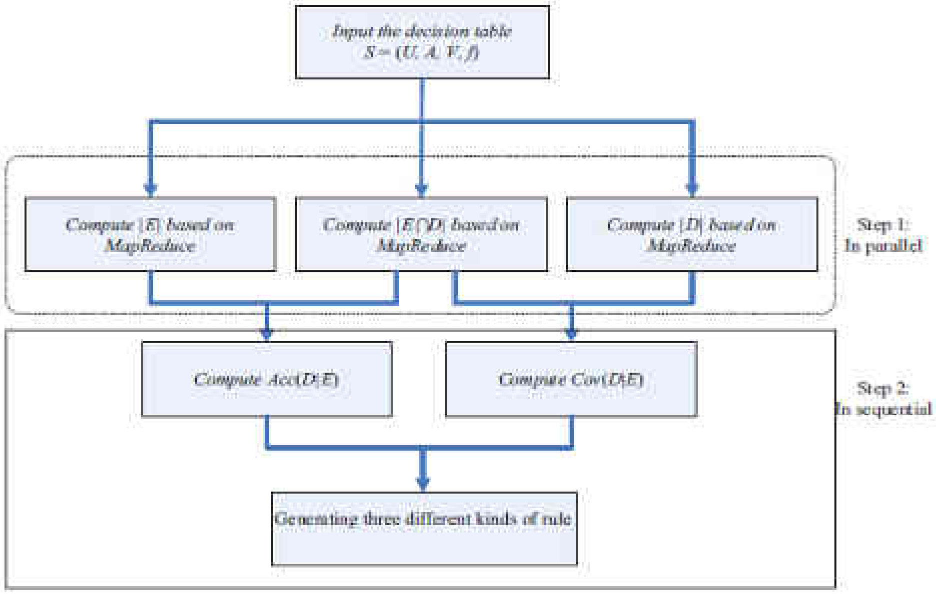
FILTER AND WRAPPER METHODS

Feature selection algorithms may be classified into two categories based on their evaluation procedure . If an algorithm performs FS independently of any learning algorithm (i.e. it is a completely separate pre-processor), then it is a filter approach. In effect, irrelevant attributes are filtered out before induction. Filters tend to be applicable to most domains as they are not tied to any particular induction algorithm.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016



MAP REDUCE FRAMEWORK

Hadoop enables resilient, distributed processing of massive unstructured data sets across commodity computer clusters, in which each node of the cluster includes its own storage. MapReduce serves two essential functions: It parcels out work to various nodes within the cluster or map, and it organizes and reduces the results from each node into a cohesive answer to a query.

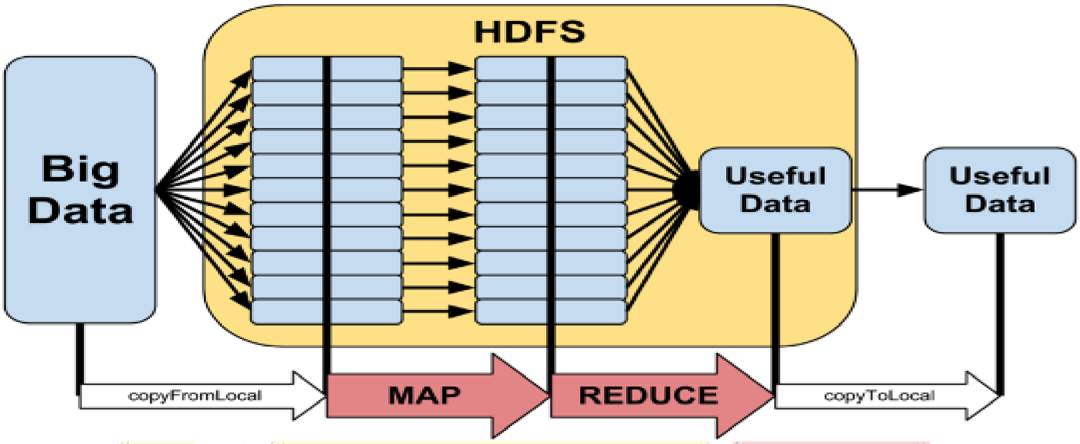
MapReduce is composed of several components, including:

- **JobTracker** -- Master node that manages all jobs and resources in a cluster
- **TaskTrackers** -- Agents deployed to each machine in the cluster to run the map and reduce tasks

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

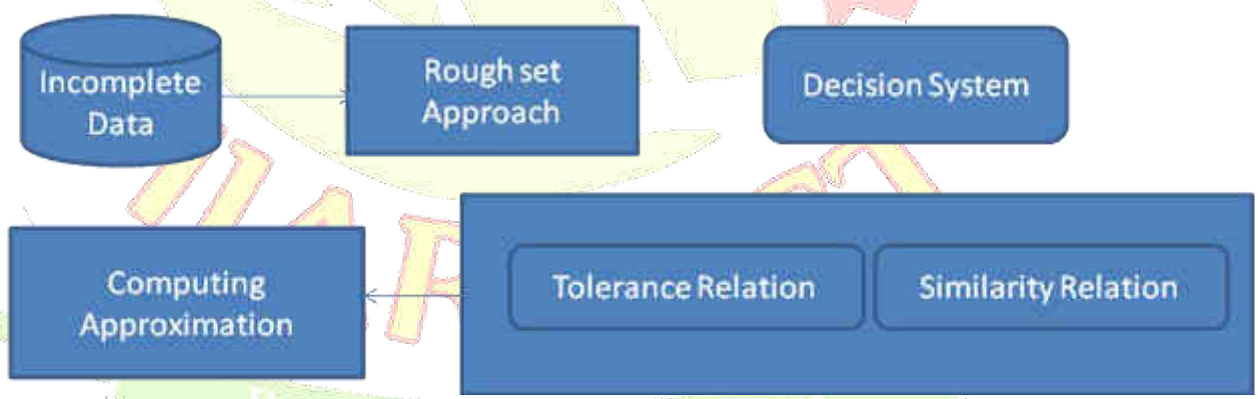
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016



PROBLEM DEFINITION

- The Existing approach for incrementally updating approximations of a concept is presented under the characteristic relation-based rough sets.
- There are several extensions for the indiscernibility relation at present, such as tolerance relation, non-symmetric similarity relation, and valued tolerance



PROPOSED SYSTEM

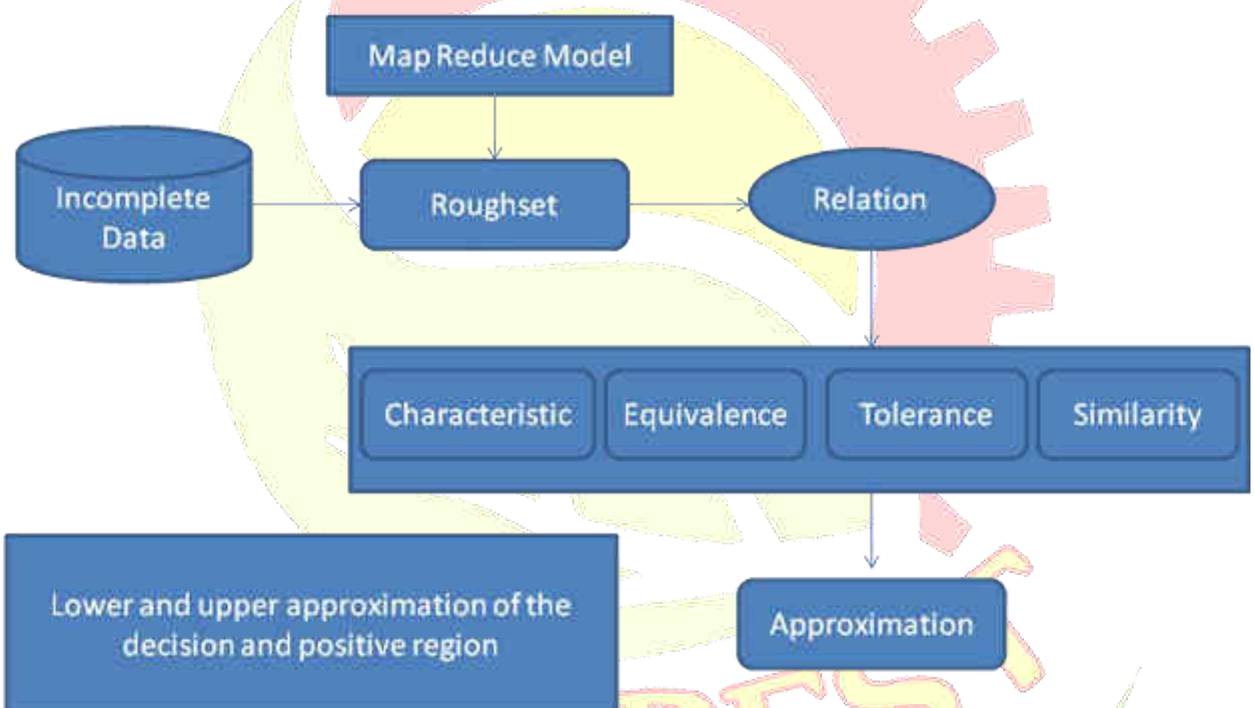
- MapReduce-based parallel method to construct the relation matrix is designed for fast computing approximations.
- Sub-Merge operation, which reduces the space requirement and accelerates the process of merging the relation matrices.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

- An incremental method is applied to the process of merging the relation matrices. With this feature, the relation matrix is updated in parallel and incrementally to efficiently accelerate the computational process.
- A sparse matrix method is employed to optimize the proposed matrix-based method and further improve the performance of the algorithm.



CONCLUSION

- Proposed three parallel strategies based on MapReduce to compute approximations. All of them were implemented on the MapReduce runtime system kdd99.
- The extensive experimental evaluation demonstrates that our proposed parallel methods are more efficient in analyzing data with large amount of attributes.
- As we can see, computing lower and upper approximations is the key to rule induction and feature selection when utilizing rough set-based methods.
- To expand the applications of rough sets in the field of data mining and knowledge discovery from big data.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

- The rough set based parallel methods for knowledge acquisition. Based on MapReduce, design corresponding parallel algorithms for knowledge acquisition on the basis of the characteristics of the data.

REFERENCES

[1] J. B. Zhang, D. Xiang, T. R. Li, and Y. Pan, "M2M: A simple Matlab- to-MapReduce translator for cloud computing," Tsinghua Sci. Technol., vol. 18, no. 1, pp. 1–9, Feb. 2013.

[2] J. Y. Liang, F. Wang, C. Y. Dang, and Y. H. Qian, "A group incremental approach to feature selection applying rough set technique," IEEE Trans. Knowl. Data Eng., vol. 26, no. 2, pp. 294–308, Feb. 2014.

[3] Q. H. Hu, W. Pedrycz, D. R. Yu, and J. Lang, "Selecting discrete and continuous features based on neighborhood decision error minimization," IEEE Trans. Syst., Man, Cybern., Part B: Cybern., vol. 40, no. 1, pp. 137–150, Feb. 2010.

[4] Q. H. Hu, Z. X. Xie, and D. R. Yu, "Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation," Pattern Recognit., vol. 40, no. 12, pp. 3509–3521, Dec. 2007.

[5] Q. H. Hu, D. R. Yu, J. F. Liu, and C. X. Wu, "Neighborhood rough set based heterogeneous feature subset selection," Inf. Sci., vol. 178, no. 18, pp. 3577–3594, Sep. 2008.

[6] Y. H. Qian, J. Y. Liang, W. Pedrycz, and C. Y. Dang, "An efficient accelerator for attribute reduction from incomplete data in rough set framework," Pattern Recognit., vol. 44, no. 8, pp. 1658–1670, Aug. 2011.

[7] Y. H. Qian, J. Y. Liang, W. Pedrycz, and C. Y. Dang, "Positive approximation: An accelerator for attribute reduction in rough set theory," Artif. Intell., vol. 174, no. 9/10, pp. 597–618,Jun. 2010.