DOI: https://doi.org/10.24113/ijoscience.v7i7.394

Study on Multi-Objective Bio-Inspired Algorithms for Feature Selection

Rachna Kulhare
PhD Scholar
Computer Science & Engineering
Rabindranath Tagore University - Bhopal, Madhya Pradesh, India
rachna12kulhare@gmail.com

Dr. S. Veenadhari
Associate Professor,
Computer Science & Engineering
Rabindranath Tagore University - Bhopal, Madhya Pradesh, India

Neha Sharma
PhD Scholar
Computer Science & Engineering
Rabindranath Tagore University - Bhopal, Madhya Pradesh, India

***Abstract*—With the era of big data, the problems of data size and data optimization have become more diversified and complicated, thus the optimization method has become the focus of people's attention. Algorithm is used to solve practical problems in various fields. In this paper, we studied different techniques of feature selection for big data using optimization algorithm.**

***Keywords*-Optimization Algorithm, Big Data, Features Selection.**

I. INTRODUCTION

Big data is a way to extract information from data sets that are large to be dealt with by traditional data-processing application software. With the rapid development of information technology, all walks of life in the world are carrying out the information revolution, and almost every industry is trying to find and use the value of big data [11]. One of the characteristics of Big Data is value, standing for the business value Big Data offers organizations toward sustainability and growth. Many businesses are already using Big Data and analytics to boost their earnings. Big data analytics are used by companies like Google, Amazon, Netflix, to improve their client's knowledge and predict market or user trends to improve certain services or products. 5Vs data is big data having 5Vs properties (Variety, Velocity, Volume, Veracity and Value,).

1. Big data is referred to by various sources as structured, unstructured and semistructured data. Emails, PDFs, photographs, videos, audios, social media posts, and more are all examples of data. One of the most essential properties of huge data is its diversity.
2. The speed at which data is created in real time refers essentially to speed. It encompasses the rate of change, the connecting of arriving data sources at variable rates, and activity bursts in a larger sense.
3. One of the hallmarks of large data is its volume. We understand that indeed big data shows enormous amounts of data produced on a regular basis from different origins such as business processes, social media platforms, networks, human interactions, computers, etc. Data warehouses hold a massive quantity of information.

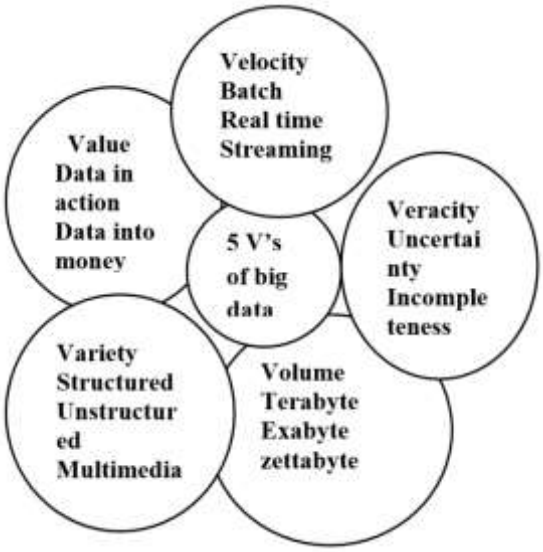


Fig 1: characteristics of big data

1. The term "veracity" refers to the data's degree of trustworthiness. Big data needs to find an alternative way in which to filter or to translate them, since most of the data is unstructured and irrelevant, as this is crucial for organizational advancements.
2. Worth is the biggest problem on which we must focus. We don't only store or process the amount of data. It is the quantity of useful, dependable, and trusted information that must be saved, processed, and evaluated in order to obtain insights.

II. TYPES OF BIG DATA

1. Structured: one of the big data types is structured and Data that can be processed, stored, and retrieved in a fixed manner is referred to as structured data. It describes to neatly ordered material which can be

kept and retrieved from a databases with ease using basic search engine methods. As an example, consider a database management system.

2. Unstructured: Unorganized information refers to data that does not have any particular shape or structure. This makes processing and analysis of unstructured data very difficult and time-consuming. An example of unstructured information is e-mail. Two important types of big data are structured and unstructured. CSV (comma value separated) example file.
3. Semi-structured: The third category of huge data is semi-structured. Semi-structured data refers to the information that has both the structured and unstructured formats stated above. To be more specific, it refers to data that, while not categorised under a certain repository (database), has essential information or tags that separate different pieces within the data. We've now reached the end of the data kinds. Let's talk about data qualities. For instance, an audio file, photographs, and so on.

III. Feature Selection Methods for Big Data

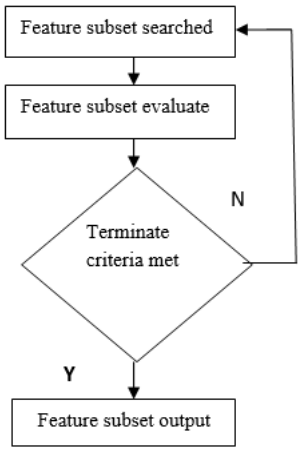
One of the most essential data analysis tools is feature selection, which is typically used to find correlated features and remove redundant or uncorrelated information from a feature collection. Arbitrary or loud data can make it difficult for a classifiers to practice good correlations, and redundant or correlated characteristics add to the classifier's complexities without offering any meaningful information. Filter, wrapper, and embedding approaches are examples of feature selection strategies. [4].

Feature selection methods for big data includes static big data, dynamic data, missing data, heterogeneous data, unreliable data and imbalanced data.[4]

IV. Feature Selection Framework

A feature selection method can be divided into two parts:

- A feature selection technique that accounts, how to select features from the original entire set.
- A feature set evaluation technique that presents how to evaluate the feature subsets.



A. *Feature Selection in Big Data Using Cooperative Co-Evolution*

A CCEA have three phases: problem decomposition, sub-problems evolution, collaboration &evaluation. CCEA-based feature selection approach for Big Data: Datasets from the UCI ML repository were collected, and preprocessed using Microsoft Excel and WEKA.2 These datasets were processed using six ML classifiers, NB, SVM, k-NN, J48, RF, and LR. The performance of these classifiers was evaluated based on precision, recall, F1 score, and accuracy. CCEAFS was applied to the datasets to reduce the number of features. The datasets with reduced dimensionality were then processed using the same six ML classifiers, and the performance of each classifier was evaluated based on the aforementioned metrics. The performance results obtained by the classifiers with and without FS were analyzed, and the effect of FS on the performance of the classifiers was investigated.[1]

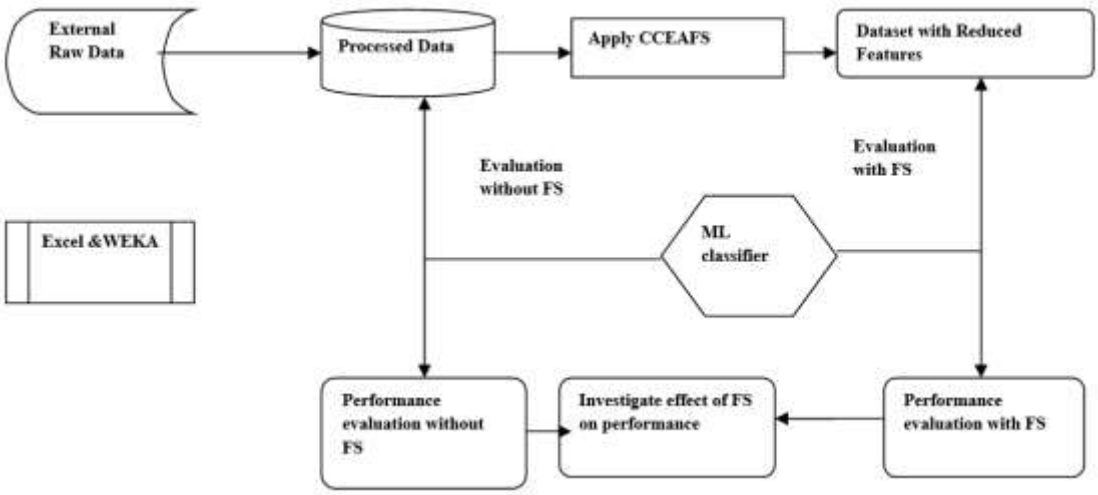


Fig 2: CCEA based FS

B. Feature Selection Modelfor Big Data Analytics

Feature selection is a multi-objective of optimization issues targeted at maximizing the accuracy of the classification and reducing the number of selected features. Whale Optimization Algorithm (WOA) is use with wrapper techniques for an ideal choice of features. The algorithm proposed for addressing Feature selection model is mainly initialized, evaluated, transformed, and iterated [2].

C. Distributed Multi-Objective Cooperative Coevolution Algorithm for Bigdata

Distributed Multi objective Cooperative Co evolutionary Algorithm (DMOCCA) is based on a divide-and-conquer technique. DMOCCA is designed to address the large-scale optimization G-S-VSRP problem. This algorithm is based on cooperative co-evolution where the problem is broken down into sub problems. Each subproblem optimizes the geohashes between every two adjacent ports in the vessel voyage. DMOCCA is a distributed algorithm using Apache Spark to parallelize optimization of subproblems [3].

D. Improved Scalable Random Forest For Bigdata Analytics

SRF is developed by hyper parameter optimization and dimension technique. It consists of three layers .

Storage Layer: Hadoop distributed file system (HDFS)is used to provide highly scalable storage with fault tolerance manner.

Processing Layer: Spark processing engine is used to provide faster processing time against data of any size by utilizing in-memory caching.

Application Layer: The input workload traces are divided into training dataset and validation dataset for building prediction model and validating the generated model respectively. SRF is an ensemble of tree predictors which is based on decision trees combines with bagging and random feature sub spacing. It generates bootstrap samples from learning set and learn corresponding randomized trees. During prediction, each tree in the forest cast a unit vote for prediction and these votes are then combined using an aggregation method to compute the final ensemble prediction [5].

E. Research of improved particle swarm optimization algorithm based on big data

PSO algorithm has an individual particle with no volume and no mass in the search space. At each iteration, the particle updates itself according to the two pieces of extremum information. One is the individual optimal value pbest found by the individual particle, and the other is the global optimal value gbest found by the population. In this way, the particle population can approach the individual particle that has good adaptive value and find the optimal solution in the end [6].

F. Improved Competitive Swarm OptimizationAlgorithms for Feature Selection

CSO is an optimization algorithm based on the inter particle competition mechanism, which can effectively avoid falling into local optimum. With the improvement of adaptive confrontation-based search strategy and its ability to adjust dynamic parameters, the algorithm can maintain the diversity of entire particle swarm so the ability to generate new feature subsets is strong [7].

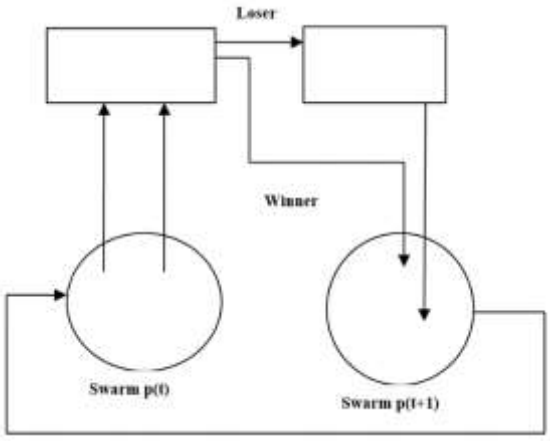


Fig 3: CSO algorithm

G. *Big data analysis and scheduling optimization system oriented assembly process for complex equipment*

A big data analysis and scheduling optimization system, which integrate the manufacturing and human factor data related to the assembly quality to realize the quality evaluation and job scheduling optimization of complex equipment assembly. Therefore, it is more effective in predicting production status and dealing with production perturbation. The system consists of several sub modules.

- The data operation module: the module collects manufacturing information through the bottom data acquisition system, including the automatic measuring instruments and other manufacturing process and resources database. At the same time, the module needs to process the data preliminarily and encode all the manufacturing resources with a unique identification.
- The evaluation and scheduling modules: these core modules provide service of multi-objective assembly quality evaluation, human error prediction of assembly workers and job scheduling optimization.
- Visual output module: It timely releases the scheduling scheme to the job site and early warning of stations with abnormal progress or resource shortages.
- Database management module: The complex assembly line requires a set of professional assembly orders, which contain assembly process description, relevant inspection quotas, as well as assembly parts, tools, materials et. al.[8].

Table 1. Comparative Analysis of Exiting Works

Author	Description	Results

Bazlur Rashid et al. (2020)	A Novel Penalty-Based Wrapper Objective Function for Feature Selection in Big Data Using Cooperative Co-Evolution	SVM performed best in most cases, and LR in some of the cases. However, when the CCEAFS is applied, in most NB outperformed the other classifiers.
Ibrahim et al. (2020)	Improved Feature Selection Model for Big Data Analytics	Excellent and acceptable results through the accuracy of the classification.
Fatemeh Cheraghchiz, et al. (2020)	Distributed Multi-Objective Cooperative Co-evolution Algorithm for Big-Data-Enabled Vessel Schedule Recovery Problem	Reduce the difficulty of problem
Myat Cho Mon Oo et al. (2019)	Hyperparameters Optimization in Scalable Random Forest for Big Data Analytics	the optimization of hyper parameters in SRF recommends the reasonable prediction performance for big data
Wang Yanmin et al.(2019)	Research of improved particle swarm optimization algorithm based on big data	Effectively and improve the ability of data processing
Jingyi Liu et al.(2018)	Improved Competitive Swarm Optimization Algorithms for Feature Selection	Reducing the number of particle fitness calculation times
XinyeWuetal(2017)	Big data analysis and scheduling optimization system-oriented assembly process for complex equipment	Discovered a relationship between the assembling big data and comprehensive performance of assembly task execution in the complex equipment manufacturing.
HUANG He et al.(2018)	Optimization of Renewable Energy Big Data Transactions Based on Vector Evaluated Genetic Algorithm	It is used to solve this multi-objective optimization problem and find the pareto Optimal solution.
Sanja Cviji et al.(2017)	Reliable adaptive Optimization Demonstration Using Big Data	NETSS Works adaptive optimization as a more robust

V. CONCLUSION

With the advance of computational techniques, the amount of data has risen exponentially, with a rapid rate making it hard to utilize such data in the any field without appropriate pre-processing, which in turn leads to more complexity and accuracy issues eventually creating multiple complications such as storage, analysis, privacy and security. Learning from such large databases is a major issue for most of the current data mining and machine learning algorithms. Therefore, large datasets is not easy to handle as it actually requires quite a complicated process due to the complexity and heterogeneity of its features. In this paper, a study is presented on optimization algorithm for feature selection for big data.

REFERENCES

[1] A. N. M. Bazlur Rashid, Mohiuddin Ahmed, Leslie f. SIKOS ,and Paul Haskell-Dowland, "A Novel Penalty-Based Wrapper Objective Function for Feature Selection in Big Data Using Cooperative Co-Evolution," in IEEE, August 2020.

[2] Ibrahim M. EL-Hasnony, Sherif I. Barakat, Mohamed Elhoseny, and Reham R. Mostafa," Improved Feature Selection Model for Big Data Analytics" IEEE ACCESS April 2020.

[3] Fatemeh Cheraghchiz, Ibrahim Abualhaoly, Rafael Falcony, Rami Abielmonay, Bijan Raahemiz and Emil Petriu," Distributed Multi-Objective Cooperative Coevolution Algorithm for Big-Data-Enabled Vessel Schedule Recovery Problem",IEEE, Nov 2020.

[4] Miao Rong, Dunwei Gong, Xiaozhi Gao,"Feature selection and its use in big data: challenges, methods and trends",IEEE ACCESS,,jan 2019.

[5] Myat Cho Mon Oo, Thandar Thein," Hyperparameters Optimization in Scalable Random Forest For Big Data Analytics", IEEE,2019

[6] Wang Yanmin," Research of improved particle swarm optimization algorithm based on big data", IEEE, 2019.

[7] Jingyi Liu et al," Improved Competitive Swarm Optimization Algorithms for Feature Selection," IEEE, 2018.

[8] Xinye Wu1, Jianbo Zhao2, and Yifei Tong1," Big data analysis and scheduling optimization system oriented assembly process for complex equipment," IEEE, 2017.

[9] HUANG He, TANG Haibo, QI Hu1, FENG Wei, DUAN Xiaofeng1," Optimization of Renewable Energy Big Data Transactions Based on Vector Evaluated Genetic Algorithm", IEEE, 2018.

[10] Sanja Cviji et al," Reliable adaptive Optimization DemonstrationUsing Big Data," IEEE, 2017.

[11] Xiaotao Huang etl, Experimental Teaching Design and Practice on Big Data Course, The 12th International Conference onComputer Science & Education (ICCSE 2017) August 22-25, 2017.