ISSN 2319–7757
EISSN 2319–7765

Indian Journal of Engineering

Optimized Feature Subset selection in Big-Data Mining through evolutionary approach

Saravanakumar R[1]☼, Shashikala B[2]

1.Associate Professor, Department of Computer science & Engineering, Dayananda Sagar Academy of tech.& Mang., Bangalore, 560078, India
2.Research Scholar, Department of Computer science & Engineering, BTL IT, Bangalore, Karnataka 560099, India

☼Corresponding author:
Associate Professor, Department of Computer science & Engineering, Dayananda Sagar Academy of tech. & Mang. Bangalore, 560078, India, e-mail: saravanakumar.rsk28@gmail.com

Publication History
Received: 12 January 2017
Accepted: 3 February 2017
Published: April-June 2017

Citation
Saravanakumar R, Shashikala B. Optimized Feature Subset selection in Big-Data Mining through evolutionary approach. *Indian Journal of Engineering*, 2017, 14(36), 120-125

ABSTRACT

"Big Data" is a term for data sets that are so complex that traditional data processing applications are inadequate to deal with them as it poses the challenges in performing data analysis of huge volume of data. The central problem in handling Big Data is

identifying a representative set of features for effectively performing any data mining task like classification, prediction, etc., This paper mainly focuses on the Feature Selection algorithm to select a subset of quality features from the given input set for efficiently describing the input data while reducing effects from noise or irrelevant variables and still provide good mined results. The algorithm decomposes the original dataset in blocks of instances to perform learning in the map phase by using the meta heuristic firefly algorithm. As Feature Subset Selection with respect to Big Data Mining is a NP Hard Problem the heuristic approach using map-reduce paradigm in Hadoop framework will lead to better solution as it does processing over batches of data parallel and merges the sub-optimal solutions to get a near optimal solution.

Keywords: Big data, Feature Selection, Firefly, Particle Swarm Optimization, Genetic Algorithm

1. INTRODUCTION

"Big Data" is a new paradigm that is introduced in the field of computer science to abstract the size of data. It is the proliferation of structured and unstructured data that floods daily and if managed well, it can deliver powerful insights. It focuses on three characteristics namely the volume, velocity and variety of data which needs special processing to extract the quality information for satisfying the requirements in various fields like medical, military, stock market, meteorology, etc., Accuracy in big data may lead to more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk. For example in the field of medicine the rate at which data is generated is increasing exponentially and yearly a lot of people lost their life because of the medical mistakes.

Therefore it becomes vital to apply the data mining technique over big data to extract the knowledge and can help in making timely decisions like diagnosing and treating patient where more accurate and personalized clinical data is required to improve the quality and efficiency of care. A central problem in data mining is identifying a representative set of features from which to construct a classification model for a particular task. As the name implies Big Data will have many features, which are the measurable properties of the process being observed. It may be the combination of relevant, irrelevant and redundant data. Several techniques are developed to address the problem of reducing irrelevant and redundant variables which are a burden on challenging tasks. The focus of feature selection is to select a subset of variables from the input which can efficiently describe the input data while reducing effects from noise or irrelevant variables and still provide good mined results. In this work the problem of feature selection for machine learning is addressed and proposing a solution through a heuristic approach using Hadoop Map-Reduce technique.

2. TRADITIONAL DATABASE AND BIG DATA

A Database Management System (DBMS) is a technology of storing and retrieving users data with utmost efficiency along with safety and security features. DBMS allows its users to create their own databases i.e. collection of data which are relevant with the nature of work they want. We have database systems which are designed to manage large bodies of system. And to improve the management of data, we have several data models, among them relational model is one which serves as a formal basis for user-friendly query languages. Today, the relational model is the primary data model for commercial data-processing applications because of its simplicity which eases the job of the programmer as compared to other data models like network model or the hierarchical model. The data in RDBMS is stored in database objects called tables each of which is assigned a unique name. The table is a collection of related data entries and it consists of columns and rows. As every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data even amounts to quintillions of bytes. Hence, even efficient database systems even collapse and couldn't handle the data. It is this quantum of data which is referred to as big data.

3. CHALLENGES AND ISSUES

We live in the information-age where accumulating data is easy and storing it inexpensive. Now days, many disciplines have to deal with big datasets that additionally involve a high number of features. Unfortunately, as the amount of machine readable information increases, the ability to understand and make use of it does not keep pace with its growth.

Feature selection, by identifying the most salient features for learning, focuses a learning algorithm on those aspects of the data most useful for analysis and future prediction. The hypothesis explored in this work is that feature selection for supervised classification tasks can be accomplished through the heuristic approach, and that such a feature selection process can be beneficial to a variety of common machine learning algorithms. The feature selector should be efficient and fast to eliminate irrelevant and redundant data and help in improving the performance of learning algorithms.

Following are the objectives of the proposed research work.

- To develop a model that extracts the good feature subset from big data set containing features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other
- To improve the understanding of data
- To reduce the computation requirement
- To reduce the effect of curse of dimensionality
- To improve the efficiency of mining task like predictor performance
- To have high response time

4. MOTIVATION OF RESEARCH

The study of Big Data Analytics and Heuristic approaches motivated this research work.

- In various fields the data generation rate is tremendously increasing which will have all kinds of features namely relevant, irrelevant, redundant, noisy, etc., which will affect the performance of machine learning algorithms , hence paving a way for the need of feature selection.
- The conventional approach for processing big data to extract useful information would comprise either in the speed of execution or would produce poor results as the number of iterations is exponential in data size, hence it needs a shift towards heuristic approach.
- The Feature Subset Selection becomes a NP problem to be solved as the subset of features to be analysed increases for the big dataset i.e 2N where N is the number of features available. NP-Hard problems can be optimized effectively by using the heuristic approaches.
- The Map Reduce Technique in Hadoop Framework helps in paralleling the computation hence it can be an option to implement the Heuristic Approaches which will improve the performance in terms of response time.

The Hadoop MapReduce technique is widely used for Big Data Clustering , Classification through heuristic methods like Genetic algorithm , Bat Algorithm , etc.,

5. LITERATURE SURVEY

In data mining applications, data instances are typically described by a huge number of features. Most of these features are irrelevant or redundant, which negatively affects the efficiency and effectiveness of different learning algorithms. The selection of relevant features is a crucial task which can be used to allow a better understanding of data or improve the performance of other learning tasks.

"Big data analytics is the process of examining large datasets containing a variety of data types to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information." Learning from very large databases is a major issue for most of the current data mining and machine learning Algorithms [1].

The authors of [Daniel Peralta et al., 2015] have given an insight of applying the evolutionary approach for performing feature selection for Big Data Classification by adopting the Map Reduce Approach. In this paper the performance of feature selection method was evaluated by using the well-known classifiers namely Support Vector Machine, Logistic Regression and Naïve Bayes. The authors have concluded that Hadoop is the suitable framework to perform evolutionary feature selection, improving both the classification accuracy and its runtime when dealing with big data problems.

Nature – inspired algorithms are among the most powerful algorithm for optimization [Xin –She Yang , 2009].The author in this paper introduced a new Firefly Algorithm and provided the comparison study with Particle Swarm Optimization(PSO) and other relevant algorithms where the PSO outperformed Genetic Algorithm(GA) , while the Firefly algorithm was found to be superior to both PSO and GA.

Xin –She Yang 2009, had discussed about Firefly Algorithms for Multimodal Optimization. The results prove that firefly algorithm is superior to existing metaheuristics algorithm. Also implies that fire fly algorithm is potentially more powerful in solving NP-hard

problems. The authors of [Xin –She Yang, Xingshi He, 2013] discussed the recent advances and applications of the Firefly Algorithm. The author's findings conclude that metaheuristics such as firefly algorithm are better than the optimal intermittent search strategy.

E.Emary ,Waleed Yamany ,.et al , 2014 had proposed a new approach for feature selection on rough set using bat algorithm . BA is attractive for feature selection in that bats will discover best feature combinations as they fly within the feature subset space. A fitness function based on rough-sets was designed as a target for the optimization. The fitness function incorporated both the classification accuracy and number of selected features and hence balances the classification performance and reduction size. Experimentation carried out on UCI data sets compared with the proposed algorithm with a GA-based and PSO approaches for feature reduction based on rough-set algorithms. The results on different data sets show that bat algorithm is efficient for rough set-based feature selection. The used rough-set based fitness function ensures better classification result keeping also minor feature size.

Filomena Ferrucci , M-Tahar Kechadi 2015 , described a framework for developing parallel Genetic Algorithms (Gas) on the Hadoop platform .The framework was devised to address the Feature Subset Selection problem by exploiting the features of Hadoop in terms of scalability , reliability and fault tolerance.The preliminary performance analysis showed promising results .

The Study of classification of diseases by Genetic Algorithm for Multiclass Support Vector Machine using Hadoop has given an idea to use GA as an optimizer to find the optimal values of hyper-parameters of SVM and adopt a supervised learning approach to train the SVM model [8].

Dimensionality reduction as a pre-processing step to machine learning is effective in removing irrelevant and redundant data, increasing learning accuracy, and improving result comprehensibility. However, the recent increase of dimensionality of data poses a severe challenge to many existing feature selection and feature extraction methods with respect to efficiency and effectiveness. In the field of machine learning and pattern recognition, dimensionality reduction is important area, where many approaches have been proposed. An endeavour to analyse dimensionality reduction techniques briefly with the purpose to investigate strengths and weaknesses of some widely used dimensionality reduction methods was discussed [4].

Ahmed K. Farahat in 2011 had proposed a greedy method for unsupervised feature selection. The methodology defined an effective criterion for unsupervised feature selection which measured the reconstruction error of the data matrix based on the selected subset of features. The paper then discusses a novel algorithm for greedily minimizing the reconstruction error based on the features selected so far. The greedy algorithm is based on an efficient recursive formula for calculating the reconstruction error. Experiments on real data sets demonstrated the effectiveness of the proposed algorithm in comparison to the state-of-the-art methods for unsupervised feature selection.

Still there are many possible research directions in this area.

6. RESEARCH METHODOLOGY

In order to achieve the objectives of this research work, issues such as data storage, response time, prediction accuracy and efficiency need to be explicitly incorporated in the design. The problems need to be faced while handling big data set can be solved easily and effectively by adopting parallelization. Hadoop MapReduce is a parallel programming technique build on the framework of Google app engine mapreduce. It is used for processing large data in a distributed environment. It is highly scalable and can be build using commodity hardware.

In Hadoop , mapreduce splits the input into larger sized chunks and processes these chunks simultaneously over the cluster, thus reducing the time complexity for solving the problem by distributing the processing among the cluster nodes.Basically there are two phases namely the Map phase and Reduce Phase. In the Map phase the input dataset will be processed to get intermediate results which will be combined in some way to form the final output.

Firefly Algorithm is nature inspired meta heuristic approach used for solving search based and optimization problems. It was developed by Xin–She Yang inspired by the flashing behaviour of fireflies.

In firefly algorithm three assumptions are taken into account.

1. All fireflies are unisex
2. Brightness is determined by objective function
3. Attractiveness of flies α Brightness and Brightness α 1/distance.

In this research work the Big data Set would be the input and the reduced dataset would be the output as represented in Fig 1 which could be used for machine learning algorithms like classification , prediction ,.etc.,

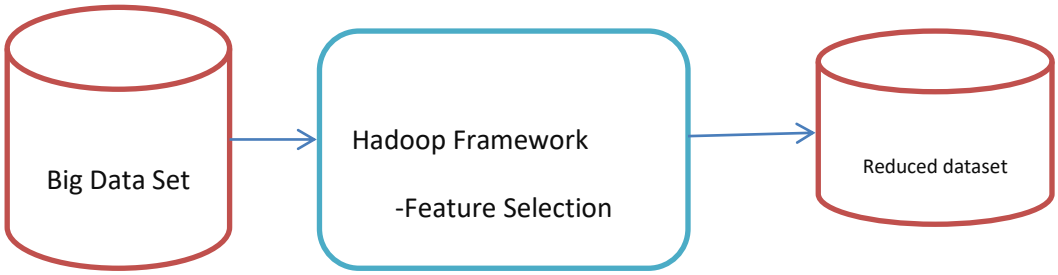

Figure 1 Proposed System

In Map Reduce, the given bigger dataset will be split into set of smaller datasets (islands)and each set will be processed in the Map Phase where the firefly algorithm will be executed. As a result at the end of Map Phase at each island, a solution of selected features would be obtained.

Those results obtained at the end of Map phase would be integrated at the Reduce phase to obtain the final result of selected features of dataset. Using the output of the Reduce phase, the given bigger dataset can be reduced to smaller size by considering those selected features only. Thereby the reduced dataset could be used further for any data mining task. The overall picture of the proposal is depicted in Fig. 6.2.

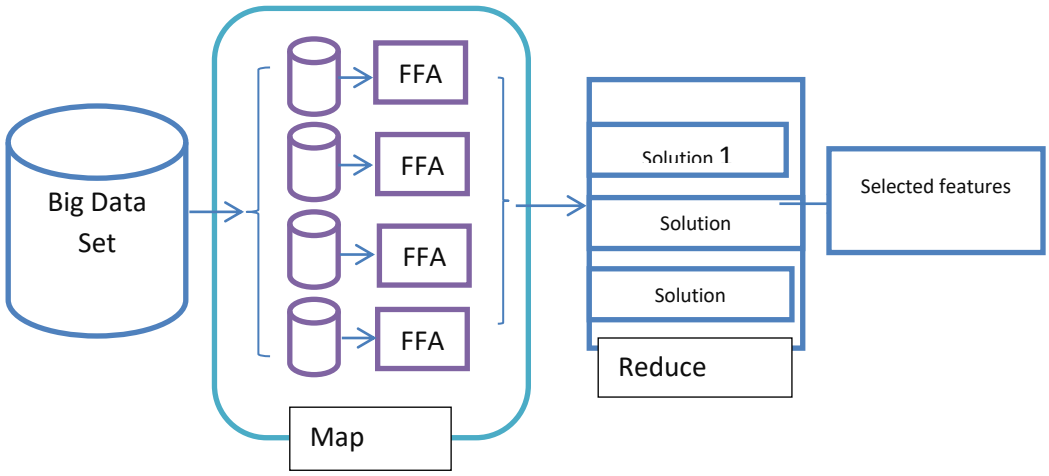

Figure 2 Overall representation of the proposed System

In the Firefly Algorithm, the objective function for each firefly will calculated by considering its accuracy in the classifier task [C4.5 algorithm]where each firefly represents the solution i.e the subset of features to be considered. The formulation of light intensity will be based on the objective function and the attractiveness will depend on the absorption coefficient and the distance between the fireflies.

The movement of fireflies depends on the light intensity i.e. the movement of flies towards the global optimal solution. At each island the fireflies are processed iteratively until the termination criteria where at the termination phase the fireflies should be ranked and the best would be taken as the solution.

7. CONCLUSION

This paper presents an Evolutionary Feature Selection algorithm designed upon the Map Reduce paradigm using Firefly Meta heuristic approach, intended to pre-process big dataset so that they become affordable for other machine learning techniques which are currently not scalable enough to deal with such dataset. The theoretical evaluation of the model highlights the full scalability of the proposed approach with respect to the number of features in the dataset, in comparison with a sequential approach. The model will be able to reduce adequately the number of features of large datasets, leading to reduced versions of them, that are at the same time smaller to store, faster to compute, and easier to classify.

REFERENCES

1. Alpaydin E, Introduction to Machine Learning , MIT Press , Cambridge, Mass , USA, 2nd edition, 2010
2. Daniel Peralta, Sara Del Rio, Sergio Ramrez-Gallego, Evolutionary Feature Selection for Big Data Classification : AMap Reduce Approach, Research Article , Mathematical problems in Engineering volume 2015.
3. Emary E, Waleed Yanmay, et al, New approach for feature selection on rough set using bat algorithm, IEEE 2014.
4. Samina Khalid, Tehmina Khalil, Shamila Nasreen, A Survey of Feature Selection and Feature Extraction techniques in machine learning, IEEE 2014.
5. Ahmed K. Farahat ; Univ. of Waterloo,; Ali Ghodsi ; Mohamed S. Kamel, An Efficient Greedy Method for unsupervised Feature Selection, IEEE 11th International Conference on Data Mining, 2011.
6. Xin – She Yang, University of Cambridge, Firefly Algorithms for Multimodal Optimization, Springer – Verlag Berlin Heidelberg, 2009.
7. Filomena Ferrucci, M-Tahar Kechadi , Federica Sarro ,A Parallel Genetic Algorithms Framework based on Hadoop MapREduce , ACM SAC'15 April 2015.
8. Ankit Deshmukh, The Study of classification of diseases by Genetic Algorithm for Multiclass Support Vector Machine using Hadoop, IJECS November 2015.

Saravanakumar R and Shashikala B,
Optimized Feature Subset selection in Big-Data Mining through evolutionary approach,
Indian Journal of Engineering, 2017, 14(36), 120-125,