Design of Enhanced Machine Learning Algorithm for Optimizing Complexity and Redundancy of Big Data

A

Summary of Ph.D Research

Submitted

In fulfillment for award of the Degree of

DOCTOR OF PHILOSOPHY

In

Faculty of Sciences

(Computer Science)

Submitted by:
PRITY VIJAY
Reg. No. RS/SGVU/DeRes/Engg/11/15/66
Enrolment No. RS68001508782

Under Supervisor of
Dr. Bright Keswani
Associate Professor
Department of Computer Applications
Suresh Gyan Vihar University, Jaipur
Rajasthan, India, Jaipur

SURESH GYAN VIHAR UNIVERSITY
MAHAL, JAGATPURA, JAIPUR-302025
INDIA, Jaipur
MAY-2018

Abstract

We are living in a big data world where enormous data as a flood is brimming from all around to spawn Data Ocean. These data are fascinating if handled appropriately or else it is nothing more than trash. An ordinary algorithm is not competent in dealing out this mammoth dataset, as they are programmed to work based on the instruction. At present machine learning and data mining is gaining esteem as it is consists of a wide range of robust algorithms, which is capable of dispensation big data. ML algorithms as an input receive huge sets of data and search interrelated pattern to construct learning rule for training predictive model proficient to predict the outcome of some kind of unseen future data automatically without any supplied instruction and become better with experiences just like human training. The only necessity of machine learning algorithms for model building is vast sets of examples but messy and complex datasets are tricky for ML techniques and thus brings challenges during formulating learning rules. Almost all real-world dataset are complex which tends to produce inaccurate machine learning result. Extensive study during this research revealed various challenges possess by machine learning classification algorithms because of complex imbalance dataset. The imbalance is a term given to inappropriate target class distribution of a dataset. Clean and balance dataset is a demand for machine learning algorithms to generate a strong model. On the other hand when the data comes from different sources, consists of lots of impurities in terms of redundancy outliers, missing value etc. which is again a problem for traditional machine learning algorithms.

The main aspiration of this work is to recognize the performances hurdle of machine learning classification algorithm due to complexity added by imbalance dataset for training purpose. The main contribution of this thesis is to generate a hybridization pre-processing and resampling technique which will able to optimize the complexity due to an imbalance big datasets and thus enhances performances of ML classification algorithms during assembling a precise predictive model. The proposed algorithm, Hybridization Preprocessing and Resampling Technique (HPRT) is an enhanced technique, for reducing the complexity of Bigdata. HPRT algorithms contains several steps to clean and balance the dataset when highly imbalance and the

complex dataset are received as an input. Redundant and irrelevant records along with outliers are detected and dropped in the first step from negative (majority) sample, which solves two purpose – it reduces the redundancy of a dataset while cutting the size of a majority class to a great extent and bringing near to normal distribution to a data set while cleaning it. SMOTE technique is then applied in a next step to add a synthetic feature to negative (minority) sample and bring normal distribution between both the classes of a dataset. HPRT based neural network based model is constructed to compare its performance with traditional neural network. Neural network three-layer architecture is formed for building a model, an error occurred during the training process is observed, calculated and minimized by weight updating method to bring out an accurate and powerful model. Thus HPRT enhances the performance of the neural network algorithm and is a major contribution of this thesis. HPRT is capable of enhancing existing machine learning algorithms. Thus HPRT based neural network model is constructed for detecting credit card frauds which is selected as a sample data. Confusion matrix and other matrices show moderately high and precise results when compared to traditional existing algorithms. The HPRT is well trained to optimize the complexity of large and big dataset and observed to deliver a good result as a result performance of machine learning algorithm enhances. Thus this research is advantageous for several domain using Bigdata and ML techniques for decision making and predicting future data.

Introduction

We are persisting in an eminently hi-tech age. It is a period where lots of technologies are growing up to engender huge mass of data. These massive data in a tremendous velocity is a by-product of each domain of modern society, results in an accelerating Bigdata. Internet acceptance turns worldwide as a result many electronic devices such as smartphones, cell phones, laptops, sensors, smart kitchen, and household appliances provoked mammoth amount of digitized data. With the period of time data expanded in each domain. A domain like medical, credit card, banks etc. can be benefited from Bigdata analytics but extracting a useful pattern from these data's requires extraordinary skills. Usage of current technologies liberates huge amount of data which is helpful in many prospective for the various industries. These enormous historical transactional data produces as a by-product from several domains can be utilized for number of decision making purpose but only if it can be processed and utilize properly. The conventional old approaches of performing data analysis need to be altered with the augmentation of Bigdata. Therefore urgent necessitate of new tools and technologies for the data analytics in each domain is in demand.

Data mining is a technique for extracting essential patterns and pulling out knowledge from huge set of records. That extracted pattern from the massive quantity of data is advantageous for many areas such as fraud detection, disease detection, market analysis, customer retention, science exploration, etc. depending upon the nature of data. Data mining uses a machine learning algorithm to discover relevant information from the massive data set. ML (Machine Learning) techniques includes a wide range of algorithms for learning predictive rules from historical data to build a model that can predict future data.

Machine Learning algorithms require huge set of clean data as a pre-requisite for training purpose but almost every dataset when gathered from different sources is unclean. Missing values, redundancy, outliers, integrity constraints violation, etc. are some of the problems which possess challenges for machine learning algorithms. Imbalance dataset is another complexity for machine learning algorithms. Imbalance dataset is consisting of the non-equal distribution of target class where majority class always dominates over minority class. ML algorithms does not go well with an imbalance data in terms of verdict patterns from it. To overcome the complexity possess by imbalance dataset, modification of data is mandatory. Sampling is a technique which deals with the modification of dataset either by Under Sampling or Over Sampling. But both of these traditional sampling techniques suffers from several disadvantages.

Overview of the research work

Rising trend had been recognized in the demands and usage of ML in the last few years with the growth of Bigdata for analyzing big chunks of data. Conventional method using statistics for data extraction and interpretation had been modified by automatic generic methods sets through machine learning. The traditional method of data analysis based on trial and error becomes difficult with the large and heterogeneous dataset. Machine learning provides a smart solution in terms of fast and efficient tools and data-driven models for processing real-time data with accurate results

Redundant, irrelevant, noisy or unreliable data causes difficulties for ML algorithms during training phase (S.B. Kotsiantis et al. 2006). Thus performance of ML algorithms are deeply affected by these factors. Therefore before mining of such dataset preprocessing is an important and time-consuming approach. Steps includes during data preprocessing are data cleaning, normalization, feature extraction, transformation, data selection, etc. The output of a data preprocessing is a clean training set, on which machine learning algorithms can apply. Real-world datasets mostly consist of noises in terms of missing or null values, outliers, redundancy, etc. Supervised machine learning algorithm performance enhances after data preprocessing but eliminating noise during the preprocessing step is the most time-consuming problem. Almost 80% of time is dedicated to data cleaning out of total data preprocessing time. Therefore there were a need of robust automated or semi-automated tools and techniques for data cleaning.

Important characteristic of imbalance class has been observed. Several complexity due to Imbalance problems along with its solution is reviewed. Sampling as an approach to deal with the imbalance problem of data. Undersampling is an approach of reducing majority class sample and oversampling is an approach of increasing minority class by adding artificial features to a minority class. None of the methods alone is capable of generating satisfactory result because Undersampling causes loss of important feature whereas Oversampling causes over fitting within a dataset. No existing approach as an optimal solution is present (Satyam Maheshwari 2007) presented and therefore combinational effect of both of these approaches can be a better approach to balance the dataset.

Many author suggested hybrid technique i.e. combination of more than one technique, for solving imbalance issue along with various other issues. During literature review it has been clear that traditional machine learning algorithms are not suitable for Bigdata. Therefore alteration of it is needed to enhance its performance. Hybrid method for enhancement of performance of ML algorithms reviewed in various research paper but nowhere a generalized hybrid technique for optimizing complexity and redundancy of Bigdata is mentioned which is capable of enhancing the performance of several machine learning algorithm.

Messy data and data imbalance is some of the major problems for ML algorithms and thus a focus area of this research. Above mentioned problems are extensively studied as a literature review and taken care of in the current study. Many approaches to solve these problems are present but each one of them suffers from challenges and treat them separately. Therefore, there is a need of single solution for messy data and imbalanced data problem capable of optimizing complexity of Bigdata. This work addressed two important parameter of data cleaning- Redundancy and outliers as because these problems exists in huge amount in almost every real world dataset making them dirty and thus degrading the performances of machine learning algorithm. Our proposed technique, Hybridization Preprocessing Resampling Technique (HPRT), proves as a single solution for both the problem and capable of removing redundancy and dropping outliers together, hence reducing the complexity of an imbalance dataset, which enhances performance of ML classification algorithm.

Research Gaps

With the advent of Bigdata nearly all the real world dataset posed several challenges due to which machine learning algorithm performances degraded. Messy data and imbalanced data are two such complexities exists almost in each real world dataset and hence focus area of this research. Several approaches to overcome above mentioned problem premeditated and recorded but none of the technique proves to be a perfect solution .The point to note is , data cleaning and data sampling is the process which is applied as initial but as a different activity for almost every datasets to deal with each problem individually. In data cleaning redundancy, outliers, missing values, etc. is taken care of and in data resampling data is modified in such a way to reduce complexity given by data imbalance problem. Again, current approaches to balance the data suffer from several challenges. Both of these problems should be solve in initial phase before feeding dataset to ML classification algorithm. But there is no such single solution which solves all the discussed problem efficiently. Many researchers suggested that combination or hybridization of more than one technique or process will provide proficient solution for data imbalance problem. Many of them developed hybridization techniques by combining one or more ML algorithm with some powerful approach to enhance it result but none of them developed a separate generalize hybridization tool as a solution which is applicable to enhance the performance of almost all ML techniques. Therefore, a single hybrid solution is needed with the capability of solving acknowledged problem competently, which automatically enhances performance of several ML classification algorithm while decreasing the complexity of Bigdata.

Objectives

1) To review several machine learning classification algorithms being used to handle Bigdata and understand performance hindrances challenges of machine learning algorithms during mining of complex Bigdata.
2) To understand the need of good quality data as an input for machine learning algorithms
3) To understand which technology tools or components are competent for optimizing complexity of an imbalance Bigdata.
4) To identify and compare among best techniques for handling an imbalanced big dataset.
5) To provide an enhanced solution for managing the complexities posed by Bigdata.
6) Development of enhanced version of data management solution, for improving Bigdata quality in terms of cleanliness (through redundancy and outlier removal) while optimizing its complexity. So that several machines learning classification algorithm can work in their own way.

Mapping the exposition to synopsis

Present data management tools and techniques are either expensive or unsuitable to accept challenges caused by the complex Bigdata. Therefore, data mining and machine learning is emerging technique, capable to extract decision making information from Bigdata and helps various industries using it to gain profit from those extracted ideas. During mining data, for revealing the intuitive power of prediction, huge amount of data example is required by machine learning algorithm, for which Bigdata is making sense. In this way, Bigdata and machine learning are helping each other.

One of the contribution of this study is to investigate the challenges of machine learning algorithm in extracting rules during classification problem. During the study lots of literature has been scrutinized and examined a number of factors affecting the performance of machine learning due to complex and messy dataset. Out of many such complexities, data imbalance is a focus area of this research. It has been observed in an experiment during this study, that if the dataset is an imbalance, machine learning techniques could not fabricate accurate result. Therefore, before feeding dataset to the machine learning algorithm pre-processing step should be applied to it, to reduce its complexity. Various disadvantages of current popular approaches of data sampling such as Undersampling and Oversampling has been premeditated and observed that in order to get benefits from Bigdata and machine learning algorithm those issue should be resolved. Therefore, HPRT, a new Hybridization Preprocessing and Resampling Technique, has been develop during this research, which solves several issues of messy data and automatically balances the dataset, thus optimizing its complexity and enhancing the performance of ML classifier. Several outcomes of this research are as follow:

1. HPRT, a new Hybridization Preprocessing and Resampling Technique, act as a perfect single data management solution having a capability of cleaning the data simultaneously, while reducing the complexity of an imbalanced dataset.

2. HRPT, act as preprocessing technique, able to enhance the performances of several traditional machine learning algorithms, which has been proved through the result of several experiments performed during study.

3. HPRT based Neural Network model for detection of credit card frauds which is selected as a case study is outcome of this study which is capable to classify frauds and non-fraud transaction and hence can rescue the financial industry from mammoth losses.

   During the study lot of literature has been scrutinized to conclude a number of factors affecting the performance of machine learning, if a complex dataset is provided to it. Out of many such complexities, data imbalance is a focus area of this research. It has been observed in an experiment during the study, that if the dataset is an imbalance, machine learning techniques could not fabricate accurate result. In spite of showing high accuracy confusion matrix shows lots of misclassification. Data preparation is another activity which need to be perform in order to clean messy dataset before feeding it to the ml algorithms because unclean data is again a threat for ml techniques. Therefore, before constructing rule out of ml algorithm pre-processing step should be applied to it, to reduce its complexity.

The solution provided here will provide a lot of benefits to credit card domain in detecting future frauds beforehand and thus will reduce the losses through fraud. A single solution is easy to understand and use and therefore can be proved as easily accepted tools among the users. HPRT implemented with big amount of data will solves numbers of issues related to it. The algorithm with little modification can be applied to various domain starting from health care to education. Thus a single solution corroborate to be a perfect data management tool to optimize complexity of a complexity of a Bigdata.

This covers all the above research objectives as mentioned in the synopsis proposal titled as "Design of Enhanced Machine Learning Algorithm for Optimizing Complexity and Redundancy of Big Data" signifies that the research title justifies the research objectives.