A Parallel Random Forest Algorithm For Big Data In a Spark Cloud Computing Environment

1st Ashwini Ahire
School of Computer Science & Engineering
SandipUuniversity
Nashik, India

2nd Vipin Wani
School of Computer Science & Engineering
Sandip University
Nashik, India

Abstract— With the emergence of the big data age, the issue of how to obtain valuable knowledge from a dataset efficiently and accurately has attracted increasingly attention from both academia and industry. This paper presents a Parallel Random Forest (PRF) algorithm for big data on the Apache Spark platform. The PRF algorithm is optimized based on a hybrid approach combining data-parallel and task-parallel optimization. From the perspective of data-parallel optimization, a vertical data-partitioning method is performed to reduce the data communication cost effectively, and a data-multiplexing method is performed is performed to allow the training dataset to be reused and diminish the volume of data. From the perspective of task-parallel optimization, a dual parallel approach is carried out in the training process of RF, and a task Directed Acyclic Graph (DAG) is created according to the parallel training process of PRF and the dependence of the Resilient Distributed Datasets (RDD) objects. Then, different task schedulers are invoked for the tasks in the DAG. Moreover, to improve the algorithm's accuracy for large, high-dimensional, and noisy data, we can perform a dimension-reduction approach.

Keywords— Apache Spark, Cloud Computing, Data Parallel, Random Forest.

I. INTRODUCTION

The emergence of the big data age also poses serious problems and challenges besides the obvious benefits. Because of business demands and competitive pressure, almost every business has a high demand for data processing in real-time and validity. As a result, the first problem is how to mine valuable information from massive data efficiently and accurately. At the same time, big data hold characteristics such as high dimensionality, complexity, and noise. Enormous data often hold properties found in various input variables in hundreds or thousands of levels, while each one of them may contain a little information. The second problem is to choose appropriate techniques that may lead to good classification performance for a high dimensional dataset. Considering the aforementioned facts, data mining and analysis for large-scale data have become a hot topic in academia and industrial research. The speed of data mining and analysis for large-scale data has also attracted much attention from both academia and industry. Studies on distributed and parallel data mining based on cloud computing platforms have achieved abundant favorable achievements. Hadoop is

a famous cloud platform widely used in data mining. In some machine learning algorithms were proposed based on the MapReduce model. However, when these algorithms are implemented based on MapReduce, the intermediate results gained in each iteration are written to the Hadoop Distributed File System (HDFS) and loaded from it. This costs much time for disk I/O operations and also massive resources for communication and storage. Apache Spark is another good cloud platform that is suitable for data mining. In comparison with Hadoop, a Resilient Distributed Datasets (RDD) model and a Directed Acyclic Graph (DAG) model built on a memory computing framework is supported for Spark. It allows us to store a data cache in memory and to perform computation and iteration for the same data directly from memory. The Spark platform saves The Random Forest (RF) algorithm is a suitable data mining algorithm for big data. It is an ensemble learning algorithm using feature sub-space to construct the model. Moreover, all decision trees can be trained concurrently, hence it is also suitable for parallelization. To improve the performance of the RF algorithm and mitigate the data communication cost and workload imbalance problems of large-scale data in parallel and distributed environments, we propose a hybrid parallel approach for RF combining data-parallel and task-parallel optimization based on the Spark RDD and DAG models. In comparison with the existing study results, our method reduces the volume of the training dataset without decreasing the algorithm's accuracy. Moreover, our method mitigates the data communication and workload imbalance problems of large-scale data in parallel and distributed environments.

II. RELATED WORK

R. Yan Mo won the 2012 Nobel Prize in Literature. This is probably the most controversial Nobel prize of this category. Searching on Google with "Yan Mo Nobel Prize," resulted in 1,050,000 web pointers on the Internet (as of 3 January 2013). "For all praises as well as criticisms," said Mo recently, "I am grateful." What types of praises and criticisms has Mo actually received over his 31-year writing career? As comments keep coming on the Internet and in various news media, can we summarize all types of opinions in different media in a real-time fashion, including updated, cross-referenced discussions by critics This type of summarization program is an excellent example for Big Data processing, as the information comes from multiple, heterogeneous, autonomous sources with complex and evolving relationships, and keeps growing.

Such online discussions provide a new means to sense the public interests and generate feedback in realtime, and are mostly appealing compared to generic media, such as radio or TV broadcasting. Another example is Flickr, a public picture sharing site, which received 1.8 million photos per day, on average, from February to March 2012. Assuming the size of each photo is 2 megabytes (MB), this requires 3.6

terabytes (TB) storage every single day. Indeed, as an old saying states: "a picture is worth a thousand words," the billions of pictures on Flicker are a treasure tank for us to explore the human society, social events, public affairs, disasters, and so on, only if we have the power to harness the enormous amount of data [1].

Big data are a collection of dataset consisting of massive unstructured, semi-structured, and structured data. The four main characteristics of big data are volume (amount of data), variety (range of data types and sources), veracity (data quality), and velocity (speed of incoming data). Although many studies have been done on big data processing, very few have addressed the following two key issues: (1) how to represent the various types of data with a simple model; (2) how to extract the core data sets which are smaller but still contain valuable information, especially for streaming data. The purpose of this paper is to explore the above raised issues which are closely related to the variety and veracity characteristics of big data. Logic and Ontology, two knowledge representation methodologies, have been investigated widely. Composed of syntax, semantics and proof theory, Logic is used for making statements about the world.

Although Logic is concise, unambiguous and expressive, it works with the statements that are true or false and is hard to be used for reasoning with unstructured data. Ontology is the set of concepts and relationships that can help people communicate and share knowledge. It is definitive and exhaustive, but it also causes incompatibility among different application domains, and thus is not suitable for representing and integrating heterogeneous big data [2].

Learning from distributed data (or generally, in parallel) has received in the last decade a growing amount of attention and this trend accelerates in recent years [1], [2]. Several factors contribute to this development: data sets which exceed memory and computing capacities of single computing nodes, inherent data distribution in sensor networks and mobile environments , and ubiquity of multi-core and many-core systems which permit parallel processing of data sets. There has been many excellent works on supervised learning in this domain. Among them we are interested in classification approaches which yield interpretable models. In addition high accuracy, the ability to understand a model is still one the primary requirements in real-world applications. This ability permits to detect model artifacts and to identify the key variables influencing the classification outcome. The latter is especially valuable in business domains, where models frequently serve as tools uncovering relationships and optimization. The most popular interpretable classifiers are the decision trees. Many authors have designed decision tree algorithms working on distributed. Most of them mimic the classical tree-growing approaches but use distributed versions of the split-point selection methods. While (to our knowledge) they are the only methods which yield interpretable overall models learned on distributed data, they have several weaknesses. First, they assume a "tightly connected" processing environment (such as a cluster in a data center) and do not constrain the number and size of exchanged messages. This can preclude their deployment in truly distributed scenarios such as mobile environments or wide-area networks where the bandwidth can be limited and the latency can be high. Another disadvantage is lack of interpretable intermediate models (e.g. decision tree for each data site). Such intermediate models allow for checking early whether learned models contain artifacts and "make sense". They can also expose dominant features of each data site and the differences between them [3].

The behavior of the PV system is influenced by various factors, including solar strength, the temperature of the cell, and possible shading. Similarly, wind energy systems depend upon the input wind speed, tower shadow effect, among other factors. These environmental factors along with variations in load, capacitor switching, charging of transformers, starting of induction machines, use of nonlinear loads, and welding transformers lead to PQ problems such as sag, swell, notch, harmonics, etc. In the past, many researchers had highlighted and studied the PQ problems due to the variations in linear/nonlinear load; however, the disturbances occurred due to environmental factors, such as deviations in wind speed and solar irradiance may also lead to various operational issues.

This includes mal-operation of protective devices, failure and overloading of electrical equipment, instabilities, and so on. For example, when wind/PV systems are interfaced to the grid with the help of dc/dc and dc/ac converters, and maximum power point tracking controllers are incorporated into these systems, system complexity increases further to tackle PQ problems. In the past studies, PQ indices, such as peak values, crest factor, total harmonic distortion (THD), power factor, instantaneous frequency, and energy deviation, were calculated using frequency spectrum or Parseval's theorem for monitoring of the disturbances. The techniques, such as fast Fourier transform (FFT), chirp Z-transform, Welch algorithm, and zoom FFT, have been widely used for monitoring of electrical parameter. But, sometimes these techniques lead to misclassification of the disturbances. For example, FFT is not accurate in the analysis of nonstationary disturbances including voltage notch and transients[4].

Fuzzy rule based methods for pattern classification have received considerable attention recently. These fuzzy-rule-based classifiers attempt to minimize training error (or empirical risk). However, noise attack or malice distortion usually decreases the discrimination and increases the uncertainty. Therefore, the discrimination and uncertainty are two important kernels in noisy data classification. To consider the discrimination, principal component analysis (PCA) has been applied in the optimization of classification. One study proposes a self-constructing neural fuzzy inference network (SONFIN) using PCA. In a later study, SONFIN is successfully applied to classification problems.

However, PCA lacks an analysis of the statistics among different classes, explaining why the discriminative capability of PCA is still low. To consider the uncertainty, type-2 fuzzy networks allow researchers to model and minimize the effects of uncertainties in rule-based systems. Type-2 fuzzy logic systems (FLS) outperform their type-1 counterparts in handling problems with uncertainties such as noisy data. This ability is attributed to type-2 fuzzy sets, which have 3-D membership functions (MFs). The third dimension of type-2 fuzzy sets and the footprint of uncertainty (FOU) provide an additional degree of freedom for type-2 FLS to directly model and handle uncertainties. For example, type-2 fuzzy sets have been applied to speech recognition problems.

In a type-2 fuzzy hidden Markov model (T2 FHMM) is proposed to handle not only the randomness but also the uncertainty of speech data. In a type2 fuzzy Gaussian mixture model (T2 FGMM) is proposed to model the uncertainty distribution of noise data. Experimental results indicate that T2 FHMM and T2 FGMM outperform traditional HMM and GMM, respectively [5].

III. SYSTEM ARCHITECTURE

The random forest algorithm is an ensemble classifier algorithm based on the decision tree model. It generates *k* different training data subsets from an original dataset using a bootstrap sampling approach, and then, *k* decision trees are built by training these subsets. A random forest is finally constructed from these decision trees. Each sample of the testing dataset is predicted by all decision trees, and the final classification result is returned depending on the votes of these trees.

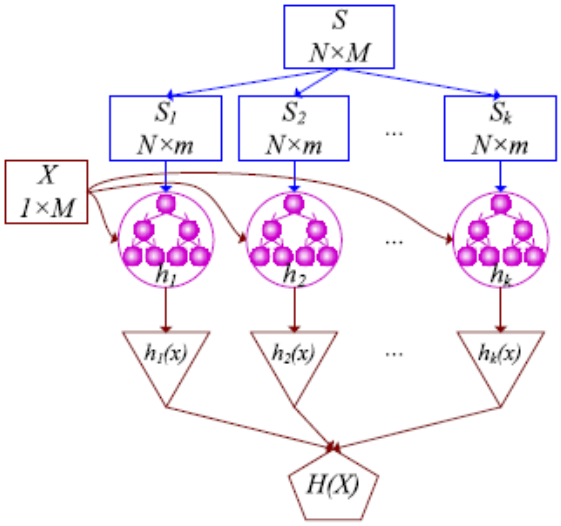

FIGURE 1: PROCESS OF THE CONSTRUCTION OF THE RF ALGORITHM

Step 1. Sampling *k* training subsets. In this step, *k* training subsets are sampled from the original training dataset *S* in a bootstrap sampling manner. Namely, *N* records are selected from *S* by a random sampling and replacement method in each sampling time.

Step 2. Constructing each decision tree model. In an RF model, each meta decision tree is created by a C4.5 or CART algorithm from each training subset *Si*. In the growth process of each tree, *m* feature variables of dataset *Si* are randomly selected from *M* variables. In each tree node's splitting process, the gain ratio of each feature variable is calculated, and the best one is chosen as the splitting node.

Step 3. Collecting *k* trees into an RF model. The *k* trained trees are collected into an RF model, which
is defined in Eq. (1):

$$H(X,\Theta j)=\sum_{i=1}^{k} h_i(x,\Theta j),(j=1,2,...,m), \qquad (1)$$

A. DIMENSION REDUCTION FOR HIGH DIMENSIONAL DATA

To improve the accuracy of the RF algorithm for the high dimensional data, we present a new dimension-reduction method to reduce the number of dimensions according to the importance of the feature variables. In the training process of each decision tree, the Gain Ratio (GR) of each feature variable of the training subset is calculated and sorted in descending order. The top *k* variables ($k \ll M$)
in the ordered list are selected as the principal variables, and then, we randomly select ($m - k$) further variables from the remaining ($M - k$) ones. Therefore, the number of dimensions of the dataset is reduced from *M* to *m*.

IV. CONCLUSION

A parallel random forest algorithm has been proposed for big data. The accuracy of the PRF algorithm will optimized through dimension-reduction and the weighted vote approach. Then, a hybrid parallel approach of PRF combining data-parallel and task-parallel optimization is performed and implemented on Apache Spark. Taking advantage of the data-parallel optimization, the training dataset is reused and the volume of data is reduced significantly. Benefitting from the task-parallel optimization, the data transmission cost is effectively reduced and the performance of the algorithm can improved.

References

1) X. Wu, X. Zhu, and G.-Q. Wu, "Data mining with big data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26, no. 1, pp. 97–107, January 2014.

2) L. Kuang, F. Hao, and Y. L.T., "A tensor-based approach for big data representation and dimensionality reduction," *Emerging Topics in Computing, IEEE Transactions on*, vol. 2, no. 3, pp. 280–291, April 2014.

3) Andrzejak, F. Langner, and S. Zabala, "Interpretable models from distributed data via merging of decision trees," in *Computational Intelligence and Data Mining (CIDM), 2013 IEEE Symposium on*. IEEE, 2013, pp. 1–9.

4) P. K. Ray, S. R. Mohanty, N. Kishor, and J. P. S. Catalao, "Optimal feature and decision tree-based classification of power quality disturbances in distributed generation systems," *Sustainable Energy, IEEE Transactions on*, vol. 5, no. 1, pp. 200–208, January 2014.

5) G. Wu and P. H. Huang, "A vectorization-optimization method- based type-2 fuzzy neural network for noisy data classification," *Fuzzy Systems, IEEE Transactions on*, vol. 21,no. 1, pp. 1–15, February 2013.