Application of machine learning model on streaming health data event in real-time to predict health status using Spark

Abderrahmane Ed-daoudy*
LTTI, ESTF, Université Sidi Mohamed Ben Abdellah, Route d'Imouzzer, BP 2427, Fès 30000, Morocco
a.eddaoudy@gmail.com

Khalil Maalmi
LTTI, ESTF, Université Sidi Mohamed Ben Abdellah, Route d'Imouzzer, BP 2427, Fès 30000, Morocco
k_maalmi@yahoo.com

Abstract— **In healthcare field, a huge amount of data collected in real-time by IoT systems, remote sensing device and other data collection tools brings new challenges that focus primarily on data size and the fast growth rate of such large data. Applying machine learning model on this voluminous data having varying velocity becomes extremely complex for traditional methods of data mining. To deal with this challenge, Apache Spark, a powerful big data processing tool can be used successfully for streaming data event against machine learning through in-memory and distributed computations. This work aims at developing a real-time health status prediction system with breast cancer use case using spark streaming framework with machine learning especially Decision Tree. The system focus on applying machine learning model on streaming data coming with rapid rate to predict health status based on several input variables. Based on this, the system first preprocesses the dataset and analyzes it to create an offline model for learning system, the model then deployed on system and use it in real-time to predict health status.**

Keywords— machine learning, spark streaming, data mining, MLlib, real-time, big data

I. INTRODUCTION

Breast cancer is the most deadly and frequent cancer in women. It appears in women in the form of tumors in the breast. It can be diagnosed and detected by physical examination or image analysis [1]. So, early detection of the stage of cancer allows treatment which could lead to high survival rate. Mammography is currently the most effective imaging modality for breast cancer screening. Knowledge and mining information from large database has been recognized by many researchers as a key research topic in database system and machine learning.

Soft computing techniques are widely used for classification of disease in medical field especially data mining which have been designed to discover useful knowledge and understandable patterns from databases [2][3]. Indeed, digital data is becoming increasingly important in many domains like healthcare, the amount of data generated in real-time is becoming very important, which involves a number of problems, the main one being the processing and prediction of streaming data event coming with rapid rate. Solving this problems using traditional technologies require hardware resources and time-consuming for the analysis. To deal with this challenge, powerful distributed computing platforms are widely used including Hadoop MapReduce [4][5] and apache spark [6]. One of the main disadvantages of MapReduce is its inefficiency in the execution of iterative algorithms. MapReduce is not designed for iterative processes. The mappers read the same data again from the disk. Therefore, after each iteration, the results must be written to the disk for the next iteration, which severely degrades performance. Sometimes, MapReduce tasks are short-lived, at which the overloading of the initialization of this task becomes a significant additional cost for the task itself. MapReduce are no sufficient enough and it's not designed to real-time data handling.

Apache Spark [6] was started as a research project at UC Berkeley in the AMPLab; Spark was launched with the aim of designing a programming model that supports types of applications that MapReduce, while maintaining its functions of fault tolerance. Spark offers an abstraction called Resilient Distributed Datasets (RDDs) [7] to effectively support these applications. RDDs can be stored in memory between queries without replication. They reconstruct the lost data in case of failure; each RDD remembers how it was built from datasets (by transformations like map, join or groupBy). RDDs enable Spark to increase performance up to 100x in iterative analysis. RDDs can support a wide variety of iterative algorithms. Spark has integrated libraries including Spark machine learning library (MLlib) and Spark Streaming, meant for machine learning and data stream handling respectively, which is more suitable for real-time prediction.

The organization of this paper is as follows: Section II discusses the background and related works, and Section III explains the proposed approach, Spark framework and decision tree technique. Section IV provides experimental results and a discussion on the findings. Finally, in section V we conclude the paper and present future work.

II. RELATED WORKS

In recent decades different types of techniques and methods of CAD (Computer Aided Diagnosis) systems have been proposed especially in healthcare. Many of these techniques were centered on the use of data mining tools in predicting outcomes, recently machine learning is the modern and powerful science of discovering patterns and predicting from database based on statistics, data drilling, pattern recognition, and predictive analyzes.

In [8] a classification model is performed based on support vector machine. In this paper the training algorithm of support vector machine is based on quadratic programming which combines caching and decomposition as optimization techniques. Pauline and Santha kumaran [9] used Feed Forward Artificial Neural Networks and back propagation algorithm to train the network. Wisconsin breast cancer dataset is used to evaluate the performance of the network for various training algorithms. Karabatak [10] developed a weighted Naïve Bayesian classifier for the application breast cancer detection. Akay [11] proposed a support vector machine

00978-1-5386-7328-7/18/$31.00 ©2018 IEEE

combined with F-score method for detection of breast tumor as benign or malignant. The F-score method is applied to extract the important feature that contains five attributes and also remove unnecessary information. M.Vasantha et al. [12] proposed hybrid algorithm for mammogram classification using GLCM features. Şahan et al. [13] proposed a hybrid method using machine learning techniques for breast cancer detection using combined fuzzy artificial immune system along with k-nearest neighbor classifier in the model. Wisconsin breast cancer dataset is used to evaluate the performance. Wang et al. [14] proposed an automatic image processing system for classification of breast cell nuclei. The authors utilized breast cell histopathology (BCH) images for breast cancer prediction. In [15] Dora et al. proposed a new expert system based on Gauss-Newton representation based algorithm for breast cancer classification. In [16] Real-time disease surveillance is described by mining twitter data which was meant for flu and cancer.

In today ' s, research works involving machine learning using big data analytics system is quite active . In Rallapalli and Suryakanthi [17] a predictive model related to the risk of diabetes is performed using a scalable Random Forest classification algorithm. Real-time medical emergency response system that involves internet of things based medical sensors deployed on the human body is discussed in Rathore et al. [18]. An overview of big data architectures and machine learning algorithms in healthcare is provided in Manogaran and Lopez [19]. Machine learning is involved in many of these, but streaming data is handled only in a few works. The importance and effectiveness of big data tools in healthcare field is explained in [20]. The authors demonstrate that the effective way on healthcare delivery costing and achieve good healthcare outcomes is by integrating big data tools with data mining, big data analysis and medical informatics.

In healthcare field, large amounts of data are being generated everywhere and every day, where state of-the-art technologies tools are inadequate to deal with them. In contrast, it is increasingly necessary to extract information from operational data in real-time, which is crucial in a rapidly changing situation. Apache spark based machine learning and streaming is an effective way to develop a real-time system without expensive programs and considerable amount of time and money. This motivated to apply it to breast cancer prediction.

III. HEALTH STATUS PREDICTION IN THE PROPOSED SYSTEM

The proposed system is a data processing, monitoring application combining socket streams and Spark streaming. This application will process real-time data sent by connected devices as socket streams and store that data for real-time analytics. Fig. 1 shows the model of the healthcare analytics system. Firstly, socket streams continuously produce data stream message. Then are sent to the spark streaming application, where the real-time processing is performed. These data can be generated by different medical devices such as medical data feeds, patient records, mobile application data, sensing data, etc. The spark streaming receive healthcare data from socket streams and apply machine learning model to predict healthcare status, then store data in NoSQL Cassandra database [21]. This system has some characteristics which distinguish it from others traditional data analytics approaches. The main idea here is that there is a need for methods to analyze data stream coming from different sources each second, in a short periods of time. Also, the system should be independent of imported data volume.

A. Spark streaming data processing

Apache spark [6] is an open source big data processing framework built around speed that runs on clusters, which is much faster than the popular Hadoop, designed for fast computation and ease of use. It provides the promise of better reliability and fault tolerance along with in-memory processing features. Spark supports many Application Programming Interface (API), like spark streaming and machine learning library to build scalable, high throughput, and fault-tolerant data streaming applications.

Spark streaming is built on top of core Spark API, which allows processing of live data streaming from various sources like TCP sockets, twitter and Kafka. Input live data stream is grouped into batches of interval less than one second and processed by the batch processing at regular time intervals, called batch intervals. Spark engine integrating the powerful features to near real-time processing. Spark implements an extension through the Spark streaming module, providing a high level abstraction called discretized stream or DStream which contains mini-batches of data created from the real-time streaming source, where each mini-batch is represented as a Spark RDD.

B. Use case dataset

In our experimental purpose, the Wisconsin Breast Cancer Database (WBCD) was used for training and testing the machine learning algorithm which predicts Breast Cancer [22]. It was used in many machine learning research works. For each cancer observation, we have constructed a labelled dataset with attributes, where class label attribute labelled with two classes, benign and malignant. The class label attribute values modified to just 0 and 1, where value 1 indicates malignant and value 0 indicates benign, turning it to a binary class dataset, the others features are Clump thickness, Uniformity of cell Size, Uniformity of cell Shape, Marginal adhesion, Single epithelial cell Size, Bare nuclei, Bland chromatin, Normal nucleol, Mitoses. The dataset consist of 699 records, 241 malignant and 458 benign. In this module we load the data from the csv file into an RDD of Strings. Then we use the map transformation on the RDD, which will apply the ParseRDD function to transform each String element in the RDD into an RDD of Labeled Point, and use it for training and testing the machine learning model which predicts breast cancer.

C. Machine learning model

The prediction of health status coming from the different streams needs to build a classification model, which is capable to classify the attributes of each stream in benign or malignant. The classification consists of examining the characteristics of a newly introduced element in order to assign it to a class of a predefined set. Decision Tree learning is a powerful method for pattern categorizations. It used extensively in machine learning because they are easy to use, easy to interpret, easy to operationalize, and extend to the multiclass classification setting and is chosen to perform the binary partitioning on the feature space where each partition is

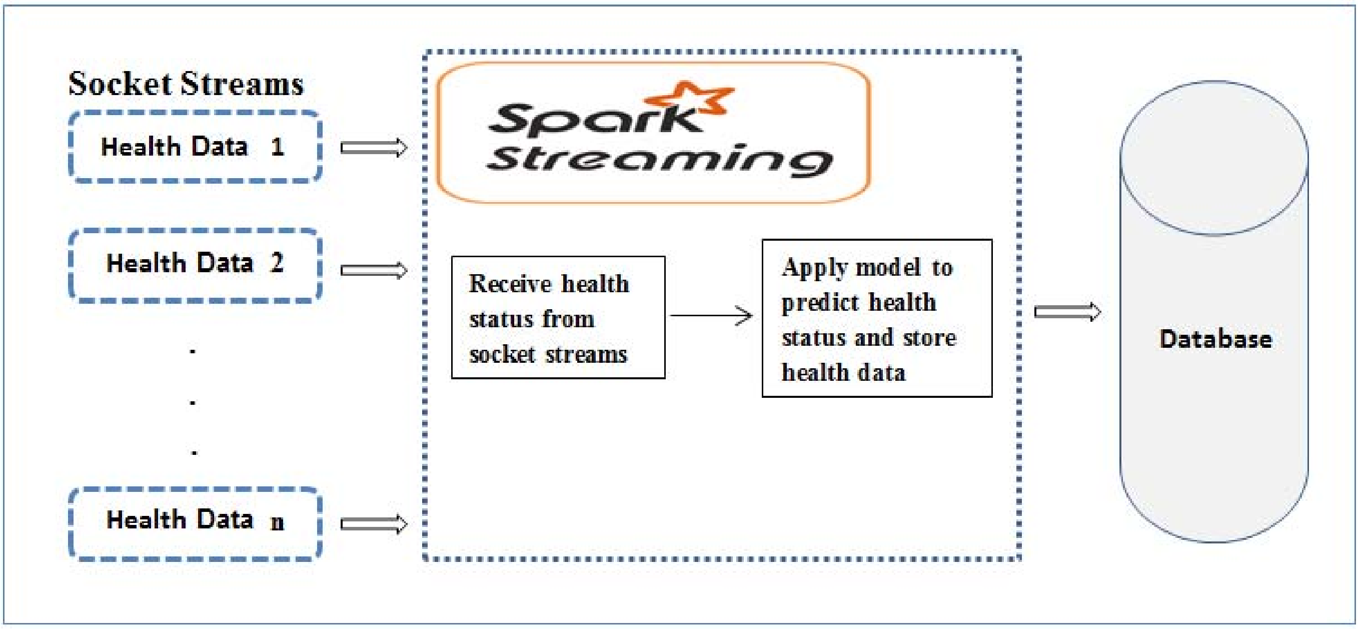
Socket Streams
Health Data 1
Health Data 2
Health Data n
Spark Streaming
Receive health status from socket streams
Apply model to predict health status and store health data
Database

Fig. 1. Real-time health status prediction overview

selected by finding the best of all possible divisions based on some criterion such as Entropy given by formula:

$$\sum_{i=1}^{C} -f_i \log(f_i) \qquad (1)$$

where f_i is the frequency of label i at a node and C is the number of unique labels [6]. In this work, Spark streaming handles the socket data streams using Spark streaming library, while the decision tree implementation is performed using the spark machine learning library, MLlib.

The breast cancer dataset has been randomly split into a training data set and a test data set, 70% of the data is used to train the model, and 30% will be used for testing. Decision tree has been trained using the training set which utilize the Entropy to predict breast cancer. For large distributed datasets sorting feature values is expensive, in this implementation an approximate set of split candidates are calculated over a sampled fraction of data and the ordered splits create bins and maxBins parameter specify the maximum number of such bins, maxDepth parameter specifies the maximum depth of the decision tree. Using the Entropy with varying maxBins and maxDepth, different decision tree model has been tested then the classification accuracy values are calculated in each case as the following table shows (table I).

- Classification accuracy

In this work, classification accuracy for the datasets are measured using the equation:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2)$$

where TP, TN, FP, and FN denote true positives, true negtives, false positives, and false negatives, respectively.

TABLE I. CLASSIFICATION ACCURACY (%) VALUES FOR VARIED TREE DEPHT AND MAXBINS PARAMETER VALUES.

maxDepth	**maxBins**			
	50	***100***	***150***	***200***
4	94.58	94.58	94.58	94.58
6	95.56	95.56	95.56	95.56
8	95.56	95.56	95.56	95.56

In this study several tools are used to create our real-time health status prediction system:

TABLE II. EXPERIMENT ENVIRONMENT.

Software environment	**Node environment**
Spark-2.0	Single node
Netcat server	Memory 8GB
Zeppelin-0.7	OS Ubuntu Desktop 14:04 LTS, 64 Bits
Scala 2.11 Cassandra-3.11	CPU 2.7 GHz, i7

IV. Results and Discussion

Initially, the breast cancer dataset has been processed in order to construct a labelled dataset. Then decision tree model was built and tested separately by varying parameters such as maxDepht and maxBins whose results are given in table I, the minimum model error is taken into account based on the classification accuracy of the model. An offline model has been created and saved in order to use it in real-time.

In our case study, input live data stream is performed by using a TCP message sent to Netcat server where we will continuously feed the health status attributs in predefined format. All these data events will be captured by spark streaming application in real-time, which performs a series of transformation on Dstreams and apply decision tree on extracted health attributes to predict breast cancer, and save the result to Cassandra database for historical data analysis.

Our study focuses on application of distributed machine learning model on streaming data event coming from different sources, with spark serving for processing the events streams and making prediction in real-time. On the other hand the database used in this work is not very big, resulting in small accuracy that can be better with the availability of reliable big healthcare datasets.

This work presents a real-time data analytics framework for analyzing healthcare data. The basic difference between this study and other researches is that the proposed system can not only perform the basic processing tasks, but makes an infrastructure for performing more complicated analytical tasks like machine learning algorithms on the streaming big data.

V. Conclusion

Healthcare data is constantly growing time by time at alarming rate in different and inconsistent data sources. Improving the patient outcome and making scalable real-time health status prediction, distributed stream computing platform is needed. Using traditional information technology (IT) solutions is not suitable in dealing with growing data. In today's real-time healthcare data processing is possible with the help of spark engine as it supports automated data streaming and analytics through iterative processing on large dataset.

The main contribution of our work is the creation of a real-time health status prediction system using big data tools, data mining techniques and streaming data, in which machine learning model is applied on received relevant health data events to predict health status and store the details in a distributed database. The use of big data tools especially spark, significantly improved the performance and the effectiveness of the proposed health status prediction system, especially in terms of system development time, complexity of programs and processing time, in comparison with traditional analytics tools, which requires a variety of skills, intensive and more expensive programs and considerable amount of time and money. With slight modification, the same application can predict others diseases. In future works, we aim to integrate other big data tools to our system with others application for interacting user's requests.

References

[1] Hung Nguyen, W.T. Hung, B.S. Thornton, E. Thornton and W. (1998) "Lee,Classification of Microcalcifications in Mammograms Using Artificial Neural Networks." IEEE Engineering in Medicine and Biology Society, 20 (2): 88–97.

[2] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From data mining to knowledge discovery in databases," AI magazine, 1996, 17(3), pp. 37.

[3] R. Agrawal, T. Imielinski, A. Swami,"Mining association rules between sets of items in large databases," In: Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD), 22 (2), 1993, pp. 207–216.

[4] Available from: http://hadoop.apache.org/ Online, accessed December 2017

[5] D. Jeffrey, G. Sanjay, "MapReduce Simplified data processing on large clusters," Proceedings of the 6th Conference on Symposium on Operating Systems Design and Implementation, (OSDI'04); Berkeley, CA, USA: USENIX Association, 2004, pp. 137–150.

[6] Available from: http://spark.apache.org/ Online, accessed December 2017

[7] M. Zaharia, Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. Franklin, S. Shenker, and I. Stoica,"Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," Technical Report UCB/EECS, EECS Department, University of California, Berkeley, 2011, pp. 2–2.

[8] E. D. Übeyli, ,"Implementing automated diagnostic systems for breast cancer detection," Expert Systems with Applications, 33 (4), 2007, pp.1054–1062.

[9] F. Paulin,"Classification of breast cancer by comparing backpropagation training algorithm." Intenational Journal on Computer Science and Engineering, 3 (1), 2011, pp.327–332.

[10] M. Karabatak, "A new classifier for breast cancer detection based on Naïve Bayesian," Measurement, 72, 2015, pp. 32–36.

[11] M. F. Akay,"Support vector machines combined with feature selection for breast cancer diagnosis," Expert Systems with Applications, 36 (2), 2009, pp. 3240–3247.

[12] M. Vasantha, DR, V. subbiah bharathi and R. Dhamodharan, "Medical Image Feature, Extraction, Selection and Classification," International Journal of Engineering Science and Technology, 2(6), 2010, pp. 2071–2076.

[13] S. Şahan, K. Polat, H. Kodaz, and S. Güneş,"A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis," Computers in Biology and Medicine, 37(3), 2007, pp. 415–423.

[14] P. Wang, X. Hu, Y. Li, Q. Liu, and X. Zhu, "Automatic cell nuclei segmentation and classification of breast cancer histopathology images," Signal Processing 122, 2016, pp. 1–13.

[15] L. Dora, S. Agrawal, R. Panda, and A. Abraham,"Optimal breast cancer classification using Gauss–Newton representation based algorithm," Expert Systems with Applications 85, 2017, pp. 134–145.

[16] K. Lee, A. Ankit, C. Alok,"Real-time disease surveillance using twitter data: demonstration on flu and cancer," In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, 2013, pp. 1474-1477.

[17] S. Rallapalli, and T. Suryakanthi,"Predicting the risk of diabetes in big data electronic health records by using scalable random forest classification algorithm," International Conference on Advances in Computing and Communication Engineering (ICACCE) IEEE, 2016, pp. 281–284.

[18] M. M. Rathore, A. Ahmad, A. Paul, J. Wan, and D. Zhang, "Real-time medical emergency response system: exploiting iot and big data for public health," Journal of medical systems 40(12), 2016, 283.

[19] G. Manogaran, and D. Lopez,"A survey of big data architectures and machine learning algorithms in healthcare," International Journal of Biomedical Engineering and Technology 25(2-4), 2017, pp. 182-211.

[20] L. Mallu, R. Ezhilarasie,"Live migration of virtual machines in cloud environment: A survey," Indian J Sci Technol 8(9), 2015, pp. 326-332.

[21] Available from: http://cassandra.apache.org Online, accessed December 2017

[22] Available from: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/ Online, accessed December 2017