Computers and Electrical Engineering 000 (2017) 1–7

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Applying spark based machine learning model on streaming big data for health status prediction☆

Lekha R. Nair*, Sujala D. Shetty, Siddhanth D. Shetty

Department of Computer Science, BITS Pilani, Dubai Campus, P. O. Box 345055, Dubai International Academic City, Dubai, UAE

ARTICLE INFO

Article history:
Received 30 December 2016
Revised 9 March 2017
Accepted 9 March 2017
Available online xxx

Keywords:
Big data machine learning
Streaming data processing
Tweet processing
Apache spark
Health informatics

ABSTRACT

Machine learning is one of the driving forces of science and commerce, but the proliferation of Big Data demands paradigm shifts from traditional methods in the application of machine learning techniques on this voluminous data having varying velocity. With the availability of large health care datasets and progressions in machine learning techniques, computers are now well equipped in diagnosing many health issues. This work aims at developing a real time remote health status prediction system built around open source Big Data processing engine, the Apache Spark, deployed in the cloud which focus on applying machine learning model on streaming Big Data. In this scalable system, the user tweets his health attributes and the application receives the same in real time, extracts the attributes and applies machine learning model to predict user's health status which is then directly messaged to him/her instantly for taking appropriate action.

1. Introduction

We live in a Big Data era where overpowering amount of data has been generated and captured from every single field, which is having untapped knowledge and significance. Applying machine learning on this big data is challenging as the traditional machine learning systems are not well equipped to handle such massive volume or varied velocity. Finding the right system and programming model for big data analytics is quite perplexing while platforms meant for the same can only deal with a fraction of machine learning algorithms at scale. According to Condie et al. [1], efficient systems support is still lacking for already established machine learning use cases in the big data scenario, while issues in data mining on big data is detailed in [2].

Hadoop, the big data processing system, together with Mahout, the machine learning platform, has been a preferred option for performing machine learning on huge data sets. Apache Spark which emerged later and deliberated as the second generation processing engine for big data [3], with its in-memory processing nature has proven to be much faster than Hadoop, especially in iterative processing as in machine learning applications. Spark has integrated libraries including Spark MLlib and Spark Streaming, meant for machine learning and data stream handling respectively.

Current big data deluge is the result of data from several sources including health care data and social network data. As per Abbas et al. [4], social media and mobile applications have opened up new pathways for health care delivery. Availability of large health care data set and advanced machine learning techniques have made computers to qualify in diagnosing health impairments with a significant level of accuracy.

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. Zhihan Lu.

* Corresponding author.
E-mail addresses: lekharnair@gmail.com (L.R. Nair), sujala@dubai.bits-pilani.ac.in (S.D. Shetty), siddaredevill@gmail.com (S.D. Shetty).

http://dx.doi.org/10.1016/j.compeleceng.2017.03.009
0045-7906/

This paper demonstrates the application of Spark based machine learning model on streaming big data with a real time health status prediction use case. In the current work, by utilizing Spark and its machine learning library MLlib, a decision tree model is formed from the available healthcare data which is then applied on streaming user data to remotely predict health status while harnessing twitter for real time data transfer. Here health status of a user is remotely analyzed on real time when his vital signs and physical conditions are tweeted in a predefined format. With Spark Streaming, tweet streams are filtered and health attributes are extracted from the tweet on which machine learning model is applied to predict the health status. A direct message is sent to the user about the health status, based on which he/she can decide on whether to seek expert medical care or not. The application can be deployed on premise or on cloud environment.

Further organization of this paper is as follows. Rest of Section 1 explains the motivation for proposed work, usage of Twitter as a communication channel and related works. Section 2 briefly explains streaming data analysis using Apache Spark. Section 3 details the application model, datasets used and system implementation. Results are presented in Section 4. Section 5 involves discussion and the paper is concluded in Section 6.

1.1. Motivation for health status prediction

The modern world individual has great concern about staying healthy. We have witnessed advancements in health care sector not only in just disease diagnosis or treatment, but also in predicting the possibility of occurrence of an ailment beforehand or an early detection of the same, based on currently available health care data. Health care data varies from electronic medical records to wearable health sensor data. Compact, cheap and accurate versions of many of the health care equipment/devices like blood sugar monitor or blood pressure monitor which were previously available only in the medical laboratories are now available at many homes. Wearable health sensors are quite popular now, wherein smart watches equipped with heart rate monitor and accelerometers are affordable and acceptable to the common man. Continuous health monitoring is gaining momentum as a lifestyle due to the growing awareness that prevention or early detection of life threatening diseases like cancer or heart disease can tremendously improve the survival rate.

Usually consultation with a health practitioner is required though vital sign values are available with the individual to learn about his health status, which is not a feasible option to be done frequently. With advanced machine learning techniques applied on the health care data already available, present computer systems can predict health status with reasonable accuracy. This can be used as an initial screening test for ailments though it is not fully fool proof and cannot be considered as a substitute for doctor's diagnosis, not ignoring the fact that human errors in diagnosis are also not uncommon.

1.2. Usage of twitter as a communication channel

Social network sites like Facebook or Twitter are becoming an inseparable part of human life which generate huge amount of data that includes opinions, feelings or general information regarding anything of interest by the community. Due to the wide popularity of social network sites, many internet service providers are providing free access to these sites with their promotional packages. In Twitter, the online microblogging site, text messages termed tweets that is having a size restriction of maximum 140 characters can be posted by a user, which is visible to his followers in real time. Twitter also allows direct message to be sent to a follower, which is not visible to others.

Twitter being text oriented and restricted in character lengths, it can be used as an effective communication tool especially in the smart phone environments where memory, bandwidth and display size are limited. Twitter being offered by a third party, its reliability and future may be questionable, but currently its usage as an efficient and free real time communication channel cannot be dismissed. Above all, Twitter is one of the few sources which offers public access to real streaming data for research and analysis and hence the same is chosen for this research work.

1.3. Related works

In recent years, research works involving machine learning on Big Data is quite active. In [5], predictive analysis of sensor data related to oil and Gas Company is performed using H_2O, the machine learning library. Usage of online logistic regression for detection of phishing URL is discussed in [6] which used Hadoop and Mahout. Another phishing URL detection system is discussed in [7] where Storm is used for streaming data processing and Weka classifiers for machine learning.

Many research works are being carried out to expose useful information by analyzing the social media data, especially twitter data, such as revealing sentiments regarding persons or products [8], filtering of spam [9, 10], finding trending topics [11], detecting real time events like earthquakes [12], or personality prediction [13, 14]. Machine learning is involved in many of these, but streaming data is handled only in a few works.

People are turning to social media for health related matters like getting information regarding health care products and services, sharing experiences and seeking expert opinions. A lot of health related messages are often communicated through these sites and quite a number of research works were done in analysis of social networks for effective health care. Dredze [15] explains the possibilities of using Twitter and other social media to provide real time public health statistics in an inexpensive way. Description of a system is given in [16] that processes twitter messages to identify information about threats to public health at an early stage which can reduce response time. The developed system collects data from twitter, filtered using keywords and relevant sentences are analyzed to obtain information on health events. This is a general system

Please cite this article as: L.R. Nair et al., Applying spark based machine learning model on streaming big data for health status prediction, Computers and Electrical Engineering (2017), http://dx.doi.org/10.1016/j.compeleceng.2017.03.009