Report on The Bremen Big Data Challenge 2019

Syed Nizam Uddin

Feven Legesse Mammo

Submitted in the requirement of the Data Mining Course

May 17/2019 B.C

Table of Contents

Executive summary

Data is simply everything our senses consider or perceive. Most of the time It is arranged in a determined way and exists in a variety of ways, such as quantitative and qualitative form. If there is a large volume of data, both structured and unstructured we call it Big data. Big data is everywhere, and we can collect and analyze whatever it is produced. What matters is what are we going to do with it. (Big Data, 2019)This report is about the classification problem which is classifying different everyday athletics movements based on a data set given from a data analytics competition named "The Bremen Big Data Challenge 2019". The Bremen Big Data Challenge invites students to a competition of analyzing a big dataset on yearly bases starting from 2016 and it is available to all federal state of Bremen. The data set will release every year on the first of March and the solution should be submitted before the 31st (end of) of March. The challenge comes with different big data set each year. (Weiner, Diener, Stelter, Externest, & K¨uhl, 2017) Basically, this year (2019) data set is a time series which is sensor data recorded on one Leg above and below the Knee. Time series data means data changes over time, or it deals with data that is ordered in time. (JAIN, 2016) Sensors produce high-frequency data that can recognize the movement of objects in their extent. In the sensor data, there are 22 movements: "Race (run). Walking (walk). Standing (standing). Sitting (sit). Getting up and sitting down (sit-to-stand, stand-to-sit). Up and down stairs (stair-up, stair-down). Jump on one or both legs (jump-one-leg, jump-two-leg). Left or right curve run (curve-left-step, curve-right-step). Turn left or right first, left or right foot first (curve-left-spin-Lfirst, curve-left-spin-Rfirst, curve-right-spin-Lfirst, curve-right-spin-Rfirst). Lateral steps to the left or right (lateral-shuffle-left, lateral-shuffle-right). Change of direction when running to the right or left, left or right foot first (v-cut-left-Lfirst, v-cut-left-Rfirst, v-cut-right-Lfirst, v-cut-right-Rfirst)". (Bremen Big Data Challenge 2019, 2019) Generally, the task is to classify the movements of the athletics based on those sensor recorded data. The first step was cleaning up the data or making ready the data in order to fit the model and predict. Then, from different kind of models, we used one of the powerful and recent machine learning algorithms called "LightGBM". The LightGMB algorithm is one of the applications of GBDTs (gradient boosting decision tree) and is a tree-based algorithm. It has high speed and better predictive performance comparing to other machine learning algorithms. (Mandot, 2017) Finally, we have got 84% accuracy in classification prediction.

Introduction

Currently, Big data is an important issue in almost every industry. For instance, in organizations, like Banking, Health care, Government, Education, Manufacturing and so on. Big data means, data sets that are enormous or multiplex that traditional data processing software can't cope with them. We can analyze them for insights that take us to better decisions and strategic business moves. Big data challenge encompasses capturing and storage of data, data analysis, searching, sharing, transferring, visualizing, querying and updating of data and information privacy. (Tuts, 2017)

The Bremen Big Data Challenge (BBDC) is a student challenge which is a yearly event beginning from March 2016. It focuses on or aims to flash curiosity in the field of big data science among students. Every year there is a new big dataset with a task to solve within the data set. The BBDC is open to all students in the federal state of Bremen and in order to participate students must form a group of one to three members and register. A new data set publishes in every first of March and within 31 days the task should be completed and submitted. Finally, the best five teams' prediction result wins and granted cash prizes. (Weiner, Diener, Stelter, Externest, & K¨uhl, 2017)

This report is about a time series data classification problem based on The Big Data Challenge 2019 competition data set. Time series data is a set of observation taken at a certain time mostly at the equal interval and used to forecast the feature values based upon the earlier observed values. (JAIN, 2016) The task is to classify different every day and athletic movements. There is a sensor data documented on one leg below and above the knee and it has 22 movements. Generally, there are 19 data sets which are 15 of them for training and the rest four are to be used for evaluation of our solution. Since it is very important to pre-process (Cleaning) the data before feeding it into our model, we apply a different kind of methods. In data pre-processes stage we used standardization (calculating statistically like, mean, mode, standard deviation, variance, skewness, and Kurtosis), feature selection and feature extraction (Dimensionality reduction). After this pre-process (cleaning of data) there were assigning of all labels (movements) to digit values and split the final CSV file into training and testing data. Finally, we have got an accuracy of 84 % by fitting of the model named "LightGBM" in to "training data. "LightGBM "model is a powerful and relatively new algorithm which is lately released from Microsoft. We selected this algorithm because of its performance and high rate in a large data set. (Mandot, 2017)

Data description

The Bremen Big Data Challenge 2019 data set is a sensor time series data recorded on one leg below and above the knee. In the data set, there are 22 movements. Race (run). Walking (walk). Standing (standing). Sitting (sit). Getting up and sitting down (sit-to-stand, stand-to-sit). Up and down stairs (stair-up, stair-down). Jump on one or both legs (jump-one-leg, jump-two-leg). Left or right curve run (curve-left-step, curve-right-step). Turn left or right first, left or right foot first (curve-left-spin-Lfirst, curve-left-spin-Rfirst, curve-right-spin-Lfirst, curve-right-spin-Rfirst). Lateral steps to the left or right (lateral-shuffle-left, lateral-shuffle-right). Change of direction when running to the right or left, left or right foot first (v-cut-left-Lfirst, v-cut-left-Rfirst, v-cut-right-Lfirst, v-cut-right-Rfirst). All data are accessible as a CSV file (comma separated values). In each 19 data sets subject, there are 440 individual recorded csv data set and in each 440 data sets, there are sensor recorded data with 19 columns which represents the individual sensors" 0. EMG1 1. EMG2 2. EMG3 3. EMG4 4. Airborne 5. ACC upper X 6. ACC upper Y 7. ACC upper Z 8. Goniometer X 9. ACC lower X 10. ACC lower Y 11. ACC loewr Z 12. Goniometer Y 13. Gyro upper X 14. Gyro upper Y 15. Gyro upper Z 16. Gyro lower X 17. Gyro lower Y 18. Gyro lower Z". Out of 19 datasets, 15 of them are training data subjects and the rest 4 of the data set are used to evaluate our solution. Below figure will show the sensor and its description under that. (Bremen Big Data Challenge 2019, 2019)

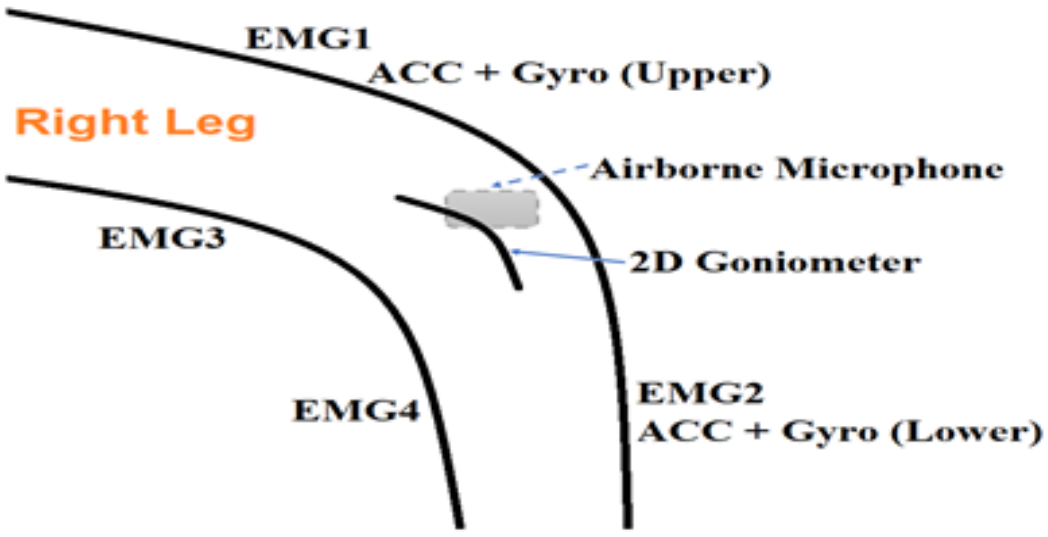

Figure: 1. Sensor position (Bremen Big Data Challenge 2019, 2019)

Number	Sensor	Muscles / Position
1	Electromyography EMG 1	Musculus vastus medialis
2	Electromyography EMG 2	Musculus tibialis anterior
4	Electromyography EMG 3	Musculus biceps femoris
5	Electromyography EMG 4	Musculus gastrocnemius
6	Accelerometer+ Gyro (upper)	Shank, distal ventral
7	Accelerometer + Gyro (Lower)	Thigh, proximal ventral
8	Electrogoniometer	Knee of the right leg, lateral

Table 1: Sensor placement and captured muscles. (Liu & Schultz, 2019)

```
Subject, Datafile, Label
Subject02,Subject02/Subject02_Aufnahme000.csv,curve-left-step
Subject02,Subject02/Subject02_Aufnahme001.csv,curve-left-step
Subject02,Subject02/Subject02_Aufnahme002.csv,stand-to-sit
```

Fig 2: sample training data (Bremen Big Data Challenge 2019, 2019)

As figure 2 above shows, every single line refers to a recording of movements and the meaning of the columns is expressed below.

Subject: - id of the subject

Datafile: - a path of the file containing the sensor data for this recording.

Label: - recorded movement.

Challenge.csv or the test data has columns which have the same meaning as a train.csv column of the label constantly contains the letter X to indicate the value is not present. So, at the time of submitting the solution should fit the test data with each X replaced by a label.

Subject, Datafile, Label

Subject01,Subject01/Subject01_Aufnahme000.csv,X

Subject01,Subject01/Subject01_Aufnahme001.csv,X

Subject01,Subject01/Subject01_Aufnahme002.csv,X

Fid3: sample testing data (Bremen Big Data Challenge 2019, 2019)

Data Preprocessing

Data Preprocessing is one of the most data mining tasks which includes preparation and transformation of data into suitable form to mining procedure. Pre-processing refers to the transformations applied to your data before feeding it to the selected models. There are many techniques to pre-process the data either by manually or by using different python libraries. Let's see the steps involved in data preparation:

Step 1: Feature Extraction

We are provided with Data file which consists of 19 subject folders, each subject folder has about 440 CSV files, each CSV file has 19 Columns, each column represents one of the sensor data. Hence, the total number of columns is equal to the total number of sensors by which there are 19 sensors. Feature extraction is the transformation of patterns into features that are considered as a compressed representation.

The time series data captured by sensors have a very high dimensionality, therefore mining such a data is a challenge because a huge number of features can be extracted from raw data (C.A. Ratanamahatana, 2010). To reduce the dimensionality of the data we have calculated many statistical features such as Mean, Median, Mode, Standard Deviation, Skewness and Kurtosis on each column of the subjects.

- The Mean μ is the average of each column xi ={Column1, column2, … ,Column19}. It was calculated by the equation (Esmael, Arnaout, Fruhwirth, & Thonhauser, 2013)

$$\mu = \frac{1}{m}\sum_{i=1}^{m} xi$$

- Standard Deviation σ was calculated to measure the how the values are spread out {Column1, column2, … ,Column19}

$$\sigma = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(xi - \mu)^2}$$

- Kurtosis Ku is calculated to measure the peakness of the probability distribution of the data.

$$Ku = \frac{\mu_4}{\sigma^4}$$

Where $\boldsymbol{\mu_4}$ is the 4th moment about the mean, and it is given by the equation

$$\mu_4 = \frac{1}{m}\sum_{i=1}^{m}(xi - \mu)^4$$

- Skewness Sk was calculated to measure the asymmetry of the data.

$$Sk = \frac{\mu_3}{\sigma^3}$$

Where the $\boldsymbol{\mu_3}$ is the 3rd moment about the mean and is given by

$$\mu_3 = \frac{1}{m}\sum_{i=1}^{m}(xi - \mu)^3$$

Step 2: By eliminating “Lay” label from the data, we have created new train.csv file.

Step 3: On the data we have assigned labels or moments to the respective digit values.

Step 4: Splitting the train.csv file into training and testing data, the technique we used is splitting into 80:20 ratio where 80 percent is our training data and 20 percent is the testing data.

Data Analysis

Data analysis will help us to select the model based on the data behavior. A common way of visualizing the distribution of numerical value is by using Histogram. A histogram divides the values within a numerical variable into "bins" and counts the number of observations that fall into each bin. By visualizing counts of bins in columnar fashion, we can obtain a very immediate and intuitive sense of distribution values within variable. We have used pandas hist() method to plot the distribution of sensor data. As we have 19 sensors hence, we can see 19 boxplots in the below figure(Represented by X1,X2,…,X19 respectively).

By understanding the data behavior , after carrying out sanity testing on different models we have finally adhered to use LightGBM model for our final prediction

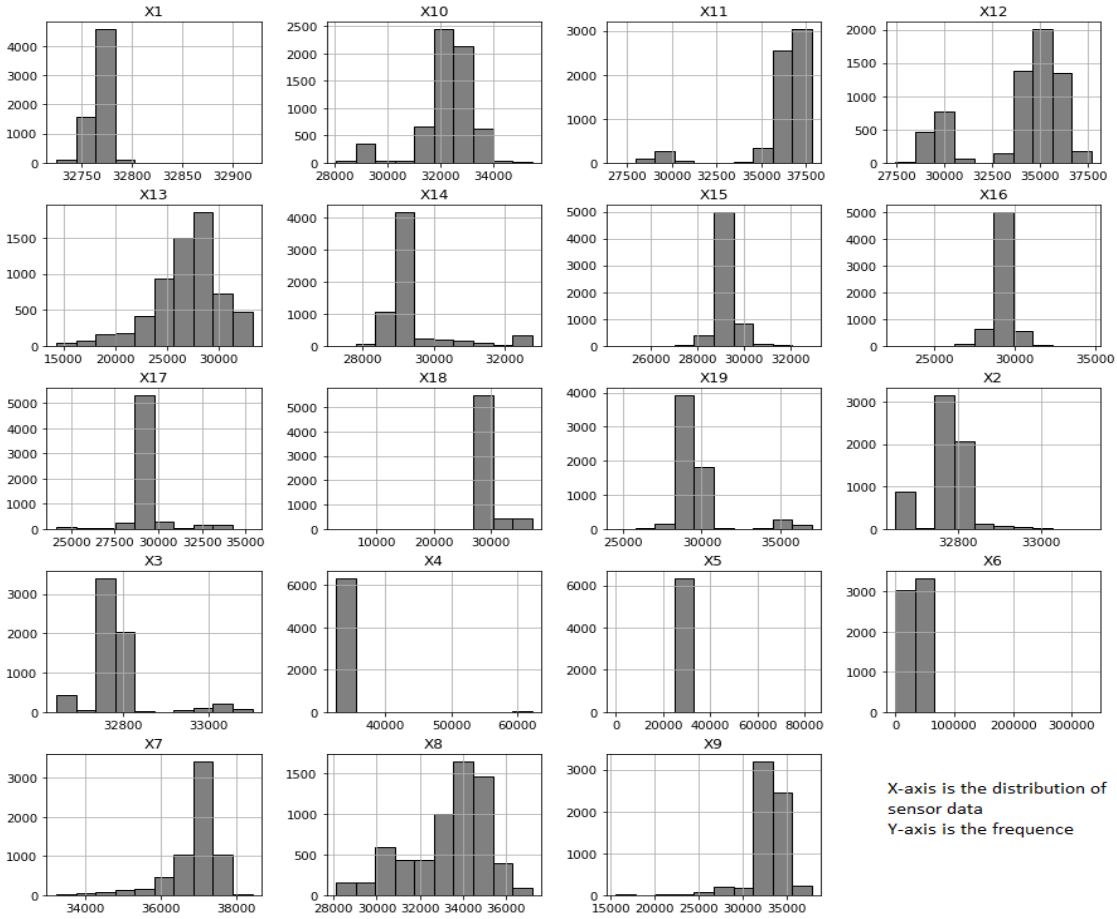

Figure 2: Distribution of Sensor data by Histogram

LightGMB model

The LightGMB algorithm is one of the implementations of GBDTs (gradient boosting decision tree) which is a tree-based algorithm and recently released from Microsoft. It has more efficient (high speed) and better predictive performance compared to other machine learning algorithms. This algorithm is treated as one of the representatives of Gradient boosting decision tree this is because:

1. It has easy-to-use open source implementation
2. Fast and accurate
3. Very important because of it improve upon the basic idea of GBDTs

When growing the decision tree, it uses the leaf-wise growth (grow trees horizontally) strategy. (Mandot, 2017)

Result and Conclusion

In this report, we have predicted a time series classification problem which is the data set provided from a yearly competition of "The Bremen Big Data Challenge 2019". Nowadays, the size of data is increasing from time to time and it becomes very difficult to deal with it in traditional data science algorithms. Seeing that, LightGBM can handle the large size of data and it is very important to use it to get a fast and better result. (Mandot, 2017) So, when we came to our data it is a huge data set so we choose this algorithm.According to our result, we have got 84% accuracy of prediction. The challenge we faced is: since it is our first time working on a very big data set the pre-processing part was long and it wasn't easy to understand how to predict (analyse) a time series classification problem and again a selection of the best algorithm. But, through the whole process, we have grasped a new skill and knowledge.

References

Big Data. (2019). Retrieved from https://www.sas.com/en_us/insights/big-data/what-is-big-data.html#dmworld

Bremen Big Data Challenge 2019. (2019). Retrieved from https://bbdc.csl.uni-bremen.de/

C.A. Ratanamahatana, J. L. (2010). Data Mining and Knowledge Discovery Handbook. *Springer*, 1049-1077.

Esmael, B., Arnaout, A., Fruhwirth, R. K., & Thonhauser, G. (2013). A Statistical Feature-Based Approach for. *International Journal of Computer Information Systems and Industrial Management Applications., 5*, 454-461.

JAIN, A. (2016, FEBRUARY 6). A comprehensive beginner's guide to create a Time Series Forecast (with Codes in Python). Retrieved from https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/

Liu, H., & Schultz, T. (2019). A Wearable Real-time Human Activity Recognition System using Biosensors Integrated into a Knee Bandage., (pp. 47-55). doi:10.5220/0007398800470055

Mandot, P. (2017, August 17). What is LightGBM, How to implement it? How to fine tune the parameters? Retrieved from https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc

Syed. (2019, May 17). *Github*. Retrieved from https://github.com/syed05/BBDC/tree/master

Tuts, B. (2017). *You tube*. Retrieved from https://www.youtube.com/watch?v=WGml5af8bJo

Weiner, J., Diener, L., Stelter, S., Externest, E., & K¨uhl, S. (2017). Bremen Big Data Challenge 2017:Predicting University Cafeteria Load.