CS7CS4 Research Assignment 1

TEAM ID : 25 TASK ID : 104 - DATA AMOUNT

Bhaskar Rao	Bhavesh Mayekar	Hamid Hassani
raob@tcd.ie	mayekarb@tcd.ie	hassanih@tcd.ie
18300829	18301407	15338952

CONTRIBUTION

All the team members have contributed equally during the project. The team met every second day at the Lloyd building to brainstorm, plan project and discuss the results. Team finalized the 3 data sets to be used for the research question (available on the Github repo). Team then distributed data sets between themselves (1 for each member) & started their analysis. Individual work contribution has been detailed below:

Bhaskar Rao: He worked on the public "Bike Sharing Dataset" from UCI machine learning repository. He did EDA, data cleaning, feature engineering on the data set before implementing machine learning models to predict the count of bikes rented. He the analyzed the effect of training data size on the accuracy of the model. He discussed his results & findings with the team and asked for help whenever required.

Bhavesh Mayekar: He worked on the publicly available "Census Income Data Set" from UCI machine learning repository. He performed various data cleaning and feature engineering operation on the dataset and implemented various machine learning models. The aim was to predict the whether an individual has a salary above 50k based on various input parameters. He tested various models with varying sizes of training datasets and analyzed the accuracy. Later the results and findings were discussed with the team. The team plans to use this analysis in second phase of the project.

Hamid Hassani: He worked on the publicly available "Bank Marketing Data set" from UCI machine learning repository. He did detailed EDAs to understand the data and shared insights with other team mates. He implemented multiple classification models to predict if a customer will subscribe to a term deposit based on marketing campaigns ran by the bank. He used this data set to test the effects of data amount on accuracy of the models. He presented his results and incorporated team's feedback to further improve his analysis.

Apart from the individual analysis, team was also reviewed each other's code and methodology. Team also collaborated to complete the research paper. Each individual prepared his own points for the paper and sat together to discuss and consolidate the key points using which the paper was finalized.

WORK COUNT

Word count excluding cover sheet, title, author names, tables and figures, references, acknowledgement, and appendix is **977**

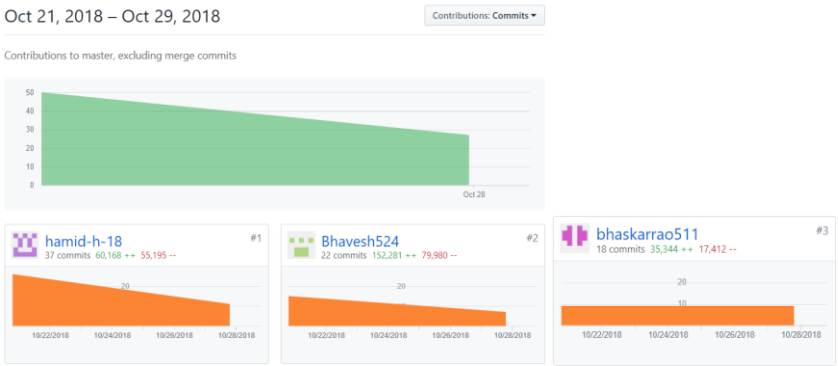
URL TO SOURCE CODE REPOSITORY

https://github.com/bhaskarrao511/CS7CS4--task-104--team-25/tree/master/sourceCode

URL TO SOURCE CODE REPOSITORY ACTIVITY

https://github.com/bhaskarrao511/CS7CS4--task-104--team-25/graphs/contributors

COMMIT ACTIVITY OF TEAM MEMBERS



We would like to thank Prof. Joeran Beel and Prof. Douglas Leith for their teachings without which this research would have been impossible

Effects of varying training data amount on Machine Learning algorithm

Bhaskar Rao
Trinity College Dublin
raob@tcd.ie

Bhavesh Mayekar
Trinity College Dublin
mayekarb@tcd.ie

Hamid Hassani
Trinity College Dublin
hassanih@tcd.ie

1 ABSTRACT

Machine learning (ML) enables computers to analyse the data and learn without any explicit programming. This task requires data and acquiring good data for training the model is one the most expensive and difficult parts. The accuracy of the model depends upon the type, quality and quantity of it. This report studies the amount of data sufficient to train models by algorithms on two multivariate datasets. The result shows that increasing the size of training data will increase the overall accuracy but for some algorithms no significant improvement was achieved.

Keywords – Machine Learning, Training Data, Accuracy

2 INTRODUCTION

Machine Learning is a process by which computer can make prediction through analysing the input data and it is either curve fitting or classification tasks. [1] In last few years, the use of machine learning has been increased tremendously due to increase of computational power. A report published by McKinsey Global Institute claims that ML will revolutionize the future innovation [2].

In ML, data plays an important role and in order to train the algorithm, the data is divided into training and testing datasets. Therefore, the first question that arise is that how much data is required to train the model effectively. As author's knowledge, there is no definite answer to this, but in most scenario, it depends on various factors like complexity of the algorithm, input features, correlation between data etc.

3 RELATED WORK

In [3], it was proposed that training size should be defined by specifying confidence interval widths for classification algorithm in bio spectroscopy field. As mentioned in [4], increasing the training dataset will overfit the model. It was found in [5] that how the performance of models vary with the training dataset size in biomedical applications. The investigation in [6] describe about how much training data is required to have an accurate model in medical image deep learning systems.

4 METHODOLOGY

In this research, following steps were followed: Identifying relevant datasets and their target features, data pre-processing, breaking the dataset into test and train, splitting training dataset into chunks of different length, build models upon these chunks, evaluate these models with the test data and compare the accuracy vs train data size.

4.1 Data Sets

Two multivariate datasets from the UCI machine learning library were used for the research, "Bike Sharing" and "Bank Marketing" dataset. Former dataset has 17,379 observations and 16 features recorded for two years at day-hour level. The target variable is the number of bikes rented and features are environmental conditions at the hour. The later dataset is a bank customer level data which has 45,212 observations and 17 features. Target variable is if the contacted customer subscribed to the bank term deposit or not & the features are client information.

4.2 Data Pre-processing

In both the datasets, null values were treated, and EDA was performed to understand the datasets. Label encoding was applied on categorical variables. In "Bike Sharing" dataset, new features like Sunday flag (day is Sunday or not) & day period ("noon", "evening" etc.) were created.

4.3 Train data splitting

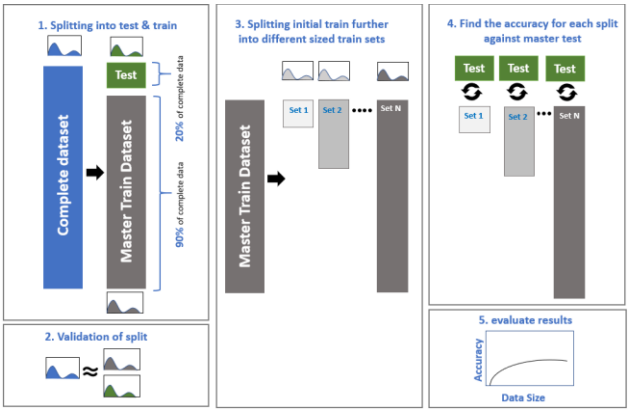


Figure 1: Data splitting approach

To analyze the impact of data amount on the performance of ML algorithms, the dataset was split into test and train parts in ratio of 20:80. (Fig. 1) Ten training datasets were further generated from the initial train set. The difference between size of consecutive training sub sets was constant, i.e 10% of initial train data size. Also, distributions of all test & train datasets were compared with that of the complete dataset to verify sample selection.

4.4 Model Building

As dependent variable in the "Bike sharing" dataset is continuous, linear regression, support vector regression and random forest algorithms were implemented. In "Bank Marketing" dataset, as dependent variable is binary true/false, logistic regression, K nearest neighbour (KNN), decision tree, random forest & Gaussian Naïve Bayes (GNB) algorithms were implemented. All the models were implemented on the individual training sub-sets generated from the initial train dataset.

5 RESULTS

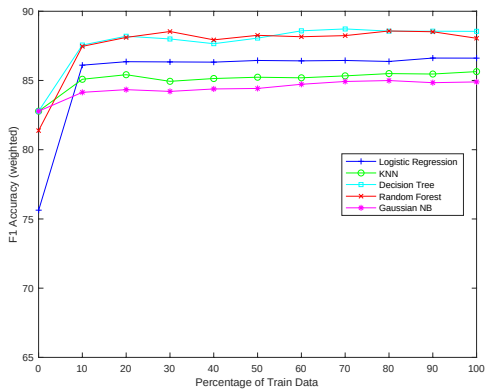


Figure 2: Results of Bank Marketing dataset

As shown in Fig. 2, in Bank Marketing dataset, the impact of changing training dataset size is related to the selected classifier algorithm. In this figure, the first point on x-axis is 0.02% and the accuracy is low because of the underfitting phenomenon but by increasing the train dataset size, the accuracy starts to improve. However, after increasing the size of dataset beyond 10%, there is no significant enhancement in the performance of the model. It can be seen that more complex models like Decision Tree and Random Forest have higher performance compared to the simple models like Logistic Regression, GNB and KNN. In addition, it is clear that after increasing the size of training dataset to 10%, all the applied methods in this setup behave in a same fashion without any obvious change in the overall accuracy.

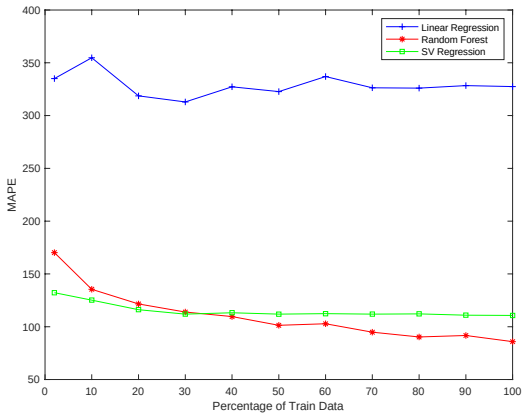


Figure 3: Results of Bike Sharing dataset

For the Bike Sharing dataset (Fig. 3), all the three algorithms behave differently when size of training dataset is increased. For linear regression model, mean absolute percentage error (MAPE) reduces till 30% of train data size and then the error increase eventually becoming nearly constant after 80%. In case of support vector regression model, error decreases till the train size is 30% of overall training data but beyond this, error becomes almost constant. For more complex model like random forest model's error decrease constantly when size of training data increase. Random forest also has least amount of error (MAPE 85.9) when compared with other two models.

The size of dataset has a significant impact of machine learning models up to a certain level. Complex models will have better accuracy compared to linear/simple models.

6 LIMITATIONS AND OUTLOOK

To further investigate about the impact of training dataset size on accuracy, the current models can be improved by using feature engineering. Next plan would be to implement algorithms on Census Income Data Set and a few more datasets related to different applications.

ACKNOWLEDGMENTS

This analysis was conducted as part of the 2018/19 Machine Learn- ing module CS7CS4/CS4404 at Trinity College Dublin).

REFERENCES

[1] P. Domingos. 2012. A few useful things to know about machine learning. *Communications of ACM*, vol. 55, no. 10, 78–87

[2] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Byers. 2011. Big data: The next frontier for innovation, competition, and productivity. Technical report. McKinsey Global Institute.

[3] C. Beleites, U. Neugebauer, T. Bocklitz, C. Krafft, J. Popp. 2013. Sample size planning for classification models. *Analytica Chimica Acta*, Volume 760, 25-33.

[4] Friedman, J., Hastie, T., and Tibshirani, R. 2001. *The elements of statistical learning*, (2nd. Ed.). Springer Series in Statistics, New York.

[5] Hajian-Tilaki, K. 2014. Sample size estimation in diagnostic test studies of biomedical informatics. *Journal of biomedical informatics*, Vol 48, 193–204.

[6] Cho, J., Lee, K., Shin, E., Choy, G., and Do, S. 2015. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? arXiv: 1511.06348 . https://arxiv.org/abs/1511.06348

[7] Sun, C., Shrivastava, A., Singh, S., and Gupta, A. 2017. Revisiting unreasonable effectiveness of data in deep learning era. arXiv:1707.02968. Retrieved from https://arxiv.org/abs/1707.02968

[8] Fanaee-T, Hadi, and Gama, Joao, 'Event labeling combining ensemble detectors and background knowledge', Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg.

[9] [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014