Internet Activity Prediction using Decision Tree Regression

CSE 416: Mobile and Wireless Communication Sessional

Authors:

Name: Md. Moshiur Rahman
Student ID: 1602022

Name: Shabbir Ahmed
Student ID: 1602044

Name: Md. Shahriar Haque
Student ID: 1702072

Submitted to:

Md. Abu Marjan
Lecturer

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
HAJEE MOHAMMAD DANESH SCIENCE AND TECHNOLOGY UNIVERSITY,
DINAJPUR-5200, BANGLADESH

Contents

List of Figures ii

List of Tables iii

Abstract 1

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Objectives . . . 2
- 1.3 Contributions . . . 2

2 Methodology 3
- 2.1 Data Processing . . . 3
- 2.2 Normalization . . . 4
- 2.3 Train-Test Split . . . 4
- 2.4 Regression . . . 4

3 Experiments and Evaluation 5
- 3.1 Dataset Description . . . 5
  - 3.1.1 Activities in the dataset . . . 5
  - 3.1.2 Source of data . . . 7
- 3.2 Performance Measure Indexes . . . 7
  - 3.2.1 Coefficient of Determination (R^2) . . . 7
  - 3.2.2 Mean Absolute Error (MAE) . . . 8
  - 3.2.3 Mean Squared Error (MSE) . . . 8
  - 3.2.4 Root Mean Square Error (RMSE) . . . 9
- 3.3 Experimental Setup . . . 9
- 3.4 Experimental Results . . . 10
- 3.5 Conclusion . . . 11
- 3.6 Limitation and Future Scope . . . 11

References 13

List of Figures

1 Overview of the proposed methodology 3
2 Comparison between actual values and predicted values 10

List of Tables

1 Regression performance results 11

Abstract

The use of mobile devices to connect humans to other humans and to internet generates an astonishing volume of data records. Mobility management, tourist flows, urban structures, and interactions urban well-being, and a variety of other problems can all benefit from this type of data. In this thesis work, one of the richest open source dataset ever released on two geographical areas is explored. The dataset is composed of telecommunications data from the city of Milan and the Province of Trentino [4]. The main goal of this work is to experiment and provide a regression model to predict internet usage activity that can in turn helps to diagnose cellular network issues and fix them. Decision tree regressor is used to generate the regression model to achieve this goal. This regression model fits well with the dataset which eventually provides adequate prediction results and the performance evaluation of the regression model appeared in this work was satisfactory.

1 Introduction

The rapid and widespread adoption of mobile phones, as well as the exponential growth in the usage of Internet services, are producing massive amounts of data that can be leveraged to deliver new basic and quantitative insights into socio-technical systems. [4] The Call Detail Records (CDRs) can be used to extract human mobility patterns [9], social interactions [8], estimates population densities [5], models cities structures [7] and models the spread of diseases [11] [10]. Mobile data are used for the analysis of ongoing COVID-19 pandemic epidemiology. [6] Mobile data These CDRs can furthermore be used to gain insights into the health of the wireless cellular networks based on mobile phone usages and detect any potential issues.

1.1 Motivation

The current standard way of analyzing network performance and diagnosing issues are done by cellular system engineers by looking the the CDRs and manually identifying

the issues (for example the highly overloaded cells based on the number of active users). This process is very cumbersome and time consuming due to the very high volume of the data. This issue can be alleviated by providing the user friendly representation of the data that can complement (or even replace) the manual analysis. One of the easy to understand representation of the CDR data is graph where nodes represent the cell cites and the edges represent the human activities using their cell phones. This representation can help cellular system engineers to readily and visually identify the troubled cells (for example highly overloaded cell) by looking at the structure of the graph and the number of the connections of each cell cites (number of edges connected to each node). Since mobile data is huge in scale, manipulating this amount of data is challenging. This challenge might be faced with machine learning. Intrinsic insights can be extracted from the existing dataset. For example predicting internet activity based on the call-in, call-out, sms-in, sms-out, cell-id etc.

1.2 Objectives

Such a work generally objectify to to help to diagnose cellular network issues and debug them. There is interrelations among the incoming and outgoing calls and SMSs and internet usage, so it seems to be possible to predict one of them based on others.

In this initiative our key objective is to make predictions of the internet usage of Milan based on the cellID. countrycode, received SMS, sent SMS, incoming calls and outgoing calls on the day of 1st November, 2013. And we also aim to produce the prediction accuracy based on some standard performance measure indexes.

1.3 Contributions

In this work, the internet activity is predicted, based on the mobile communication data of the city, Milan, Italy. Decision Tree Regressor (DTR) is used for the purpose of regression. The performances are measured based on mean absolute error (MAE), mean

squared error (MSE) and root mean square error(RMSE).

2 Methodology

In this work, separate methods are instigated. The overview of the proposed methodology is shown in Figure 1.

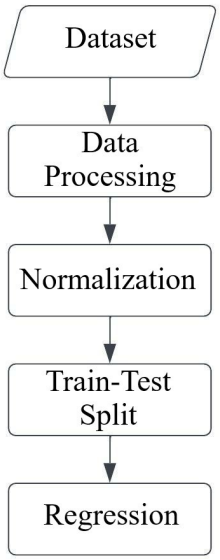


Figure 1: Overview of the proposed methodology

2.1 Data Processing

The shape of the dataset is (1891928×8). Which means- the dataset comprises 1891928 row entries which denotes the activity in a particular cell. It also comprises 8 columns which denotes the attributes of the dataset. The dataset has 1658462 row entries which have at least one *null* entry. Huge amount of sample have null values, which may cause bad impact on the regression model. To avoid this problem, the entities which have at least one null entry is deleted from the dataset. After null value deletion, the shape of the dataset is (233466×8).

2.2 Normalization

Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g. Gaussian with 0 mean and unit variance). *StandardScaler* from Sci-kit learn python package is used as normalization method. Standardize features by removing the mean and scaling to unit variance. The standard score of a sample x is calculated as:

$$z = \frac{(x - u)}{s} \tag{1}$$

where, u is the mean of the training samples or zero if $with\_mean = False$, and s is the standard deviation of the training samples or one if $with\_std = False$.

Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set. Mean and standard deviation are then stored to be used on later data using transform. For instance many elements used in the objective function of a learning algorithm (such as the RBF kernel of Support Vector Machines or the L1 and L2 regularizers of linear models) assume that all features are centered around 0 and have variance in the same order. If a feature has a variance that is orders of magnitude larger that others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected.

2.3 Train-Test Split

train_test_split from Sci-kit learn python package is used for splitting the data into training dataset and testing dataset. 30% of the dataset is split for training dataset and 70% of the dataset is split for testing dataset.

2.4 Regression

To perform the regression operation, Decision Tree Regressor is used as the regressioin model. This regression model is used from the Sci-kit learn python package. Four per-

formance measure indexes are taken into account for evaluating the performance of the regression model.

3 Experiments and Evaluation

We experimented on the SMS-call-internet dataset which is a part of the Telecom Italia Big Data Challenge 2014 and evaluated our methods with several evaluation metrics. The following represent the details of evaluation metrics, experiment setup, experimental result and a comparison of our result with the actual data.

3.1 Dataset Description

The dataset used in this work is a Mobile Phone Activity dataset which is composed by one week of Call Details Records (CDRs) from the city of Milan and the Province of Trentino (Italy) [2]. The Mobile phone activity dataset is a rich and open multi-source aggregation of telecommunications which is a part of the Telecom Italia Big Data Challenge 2014, weather, news, social networks and electricity data from the city of Milan and the Province of Trentino (Italy). The original dataset has been created by Telecom Italia associated by EIT ICT Labs, SpazioDati, MIT Media Lab, Northeastern University, Polytechnic University of Milan, Fondazione Bruno Kessler, University of Trento and Trento RISE. The dataset used in this work is a subset of this telecommunications data. The complete version of the dataset can be found online [3].

Every time a mobile phone user engages in a telecommunication interaction (sms, voice call, Internet session), a Radio Base Station/Cell (RBS) is assigned by the operator and delivers the communication through the network. Then, a new CDR is created recording the time of the interaction and the RBS which handled it.

3.1.1 Activities in the dataset

The following activities are present in the dataset:

- **Time Interval:** Start interval time expressed in milliseconds. The end interval

time can be obtained by adding 600,000 milliseconds (10 min) to this value.

- **Datetime:** Date in yyyy-mm-dd HH:ii format.
- **CellID/Square ID (Source):** Identification string of a given square of Milan GRID.
- **Countrycode (Target):** The phone country code of the target destination.
- **Received SMS:** Activity proportional to the amount of received SMSs inside a given square id and during a given Time interval. The SMSs are sent from the nation identified by the country code.
- **Sent SMS:** Activity proportional to the amount of sent SMSs inside a given square id during a given Time interval. The SMSs are received in the nation identified by the country code.
- **Incoming calls:** Activity proportional to the amount of received calls inside the square id during a given Time interval. The calls are issued from the nation identified by the country code.
- **Outgoing calls:** Activity proportional to the amount of issued calls inside a given square id during a given Time interval. The calls are received in the nation identified by the country code.
- **Internet activity:** Number of CDRs generated inside a given square id during a given time interval. The Internet traffic is initiated from the nation identified by the Country code.

In particular, Internet activity is generated each time a user starts an Internet connection or ends an Internet connection. Moreover, during the same connection a CDR is generated if the connection lasts for more than 15 min or the user transferred more than 5 MB. The dataset includes both domestic (from Milan to other provinces of Italy) and international (from Milan to other countri es) data. The domestic data only includes

the incoming/outgoing calls while international data has all the aforementioned activity records.

The dataset is spatially aggregated in a square cells grid. The area of Milan is composed of a grid overlay of 10,000 (squares with size of about 235×235 meters. This grid is projected with the WGS84 (EPSG:4326) standard [4]. The data provides CellID, CountryCode, Province name and all the aforementioned telecommunication activities aggregated every 60 minutes.

3.1.2 Source of data

For the ease of understanding the country name from the dataset, the countrycode provided in the dataset had to be converted to the country name. This was done by merging the node and edges with the country code dataset in [1]. Furthermore, for the Great Circle visualizations, the latitude and longitude of the countries, Italy provinces, and the cells located in Milan are needed. The latitude and longitude of the countries and the Italian provinces can be derived from the maps library in Python. The latitude and longitude of the cell grids in Milan is provided as a Geojson file in [2]. These two data sources are merged with nodes and edges to get the (lat,long) pairs.

3.2 Performance Measure Indexes

The following performance measure indexes are used in this work to measure the prediction accuracy:

3.2.1 Coefficient of Determination (R^2)

The coefficient of determination (R^2) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, R^2 shows how well the data fit the regression model. R^2 can take any values between 0 to 1. The following formula is used to calculate the R^2 value.

$$R^2 = \frac{SS_{\text{Regression}}}{SS_{\text{Total}}} \tag{2}$$

Where,

$SS_{\text{Regression}}$: The sum of squares due to regression (explained sum of squares).

SS_{Total} : The total sum of squares.

3.2.2 Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is a model evaluation metric used with regression models. The mean absolute error of a model with respect to a test set is the mean of the absolute values of the individual prediction errors on over all instances in the test set. Each prediction error is the difference between the true value and the predicted value for the instance. In simpler words , It measures the average of the residuals in the dataset. Therefore, the for calculating would be:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{3}$$

Where,

N = number of data points

$\hat{y}_i$ = expected or predicted values

y_i = actual observed values (known results).

3.2.3 Mean Squared Error (MSE)

The mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.

The mean-squared error is determined by the residual sum of squares resulting from comparing the predictions $\hat{y}$ with the observed outcomes y:

$$MSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{4}$$

Where,

N = number of data points

$\hat{y}_i$ = expected or predicted values

y_i = actual observed values (known results).

3.2.4 Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) stands for the standard deviation of the residuals (prediction errors). Where, residuals are measures of how far the data points are from the regression line; RMSE is a measure of how spread out these residuals are. In other words, it indicates how the data is concentrated around the line of best fit.

The formula for calculating the RMSE would be:

$$RMSE = \sqrt{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2} \tag{5}$$

Where,

N = number of data points

$\hat{y}_i$ = expected or predicted values

y_i = actual observed values (known results).

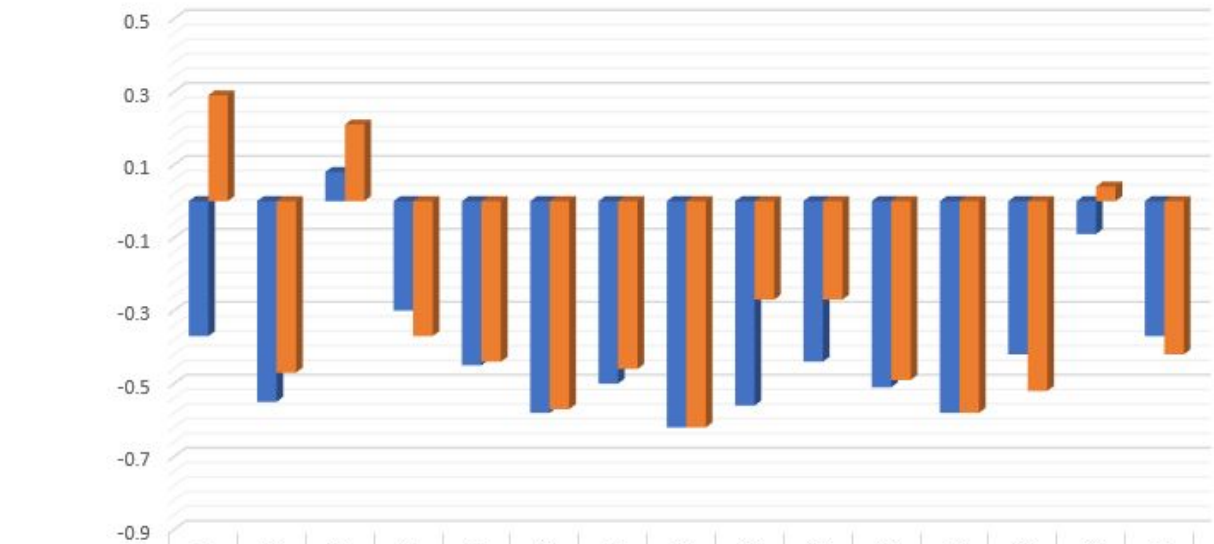
3.3 Experimental Setup

The SMS-call-internet dataset used in this work is a Mobile Phone Activity dataset including the activities i.e., datetime, cellID, countrycode, smsin, smsout, callin, callout and internet usage. The dataset initially included (1891928×8) values of these activities altogether. But there are 1658462 null values which are deleted. We also discarded the datetime column from the dataset as it is inconvenient to our experiment. Finally our used dataset is reduced to (233466×7) shape. Before regression, we split the dataset for

training and testing, we used 30 : 70 as the train-test split ratio. The column internet is used as the target label and other columns as our data.

3.4 Experimental Results

After setting up the experiment, we implemented regression on our processed dataset to obtain the predicted values of the internet usage for each cellID of each countrycode. In our experiment, Decision Tree Regressor is used as the regression model from the Sci-kit learn python package. Implementation of this regression model has provided the predicted values. Figure 2 illustrates the comparison between some of the actual values from the dataset and their corresponding predicted values as provided by our regression model. Notice that the values represented in the chart may be negative as standard Scaler is applied.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Actual Value	-0.37	-0.55	0.08	-0.3	-0.45	-0.58	-0.5	-0.62	-0.56	-0.44	-0.51	-0.58	-0.42	-0.09	-0.37
Predicted Value	0.29	-0.47	0.21	-0.37	-0.44	-0.57	-0.46	-0.62	-0.27	-0.27	-0.49	-0.58	-0.52	0.04	-0.42

Figure 2: Comparison between actual values and predicted values

The performance of our regression model is evaluated by performing operations of four different performance measure indexes: the coefficient of determination (R^2), the mean absolute error, the mean squared error and the root mean square error. The outcomes of the operations of these performance measure indexes on our experimental regression

model is illustrated in the Table 1.

Performance Measure	**Score**
R^2	0.66263
MAE	0.24954
MSE	0.34539
$RMSE$	0.58769

Table 1: Regression performance results

From all the comparisons and the results demonstrated above, it can be said that the performance of our experimental regression model is quite satisfactory. This regression model fits well with the original dataset that was processed earlier and makes adequate predictions about the internet usage activity.

3.5 Conclusion

In this work we have only experimented on the cellID. countrycode, received SMS, sent SMS, incoming calls and outgoing calls on the day of 1st November, 2013 of the city Milan and generated a regression model. Subsequently, the regression model is used for the purpose of regression which predicts the internet usage activity based on the other aforementioned data columns. Our experimental regression model's performance can be described as adequate. We can conclude from the analization of some standard performance measures that this regression model is well-suited to the original dataset that was previously processed, and the internet usage activity is predicted satisfactorily.

3.6 Limitation and Future Scope

We have performed our experiment specifically on the Mobile Phone activity dataset of the day of 1st November, 2013 of the city Milan. Addionally, this work predicts only the internet activity based on the cellID. countrycode, received SMS, sent SMS, incoming calls and outgoing calls.

Therefore, we can extend this work by carring out experiments on the data over the whole week that is available in the dataset. Furthermore, other dataset with similar

information of other cities can be included within its future scope. Consequently, it is possible to conduct predictions for different columns using the probable interrelations of data in individual columns.

References

[1] Country phone code dataset. https://www.worlddata.info/downloads/.

[2] Mobile phone activity dataset. https://www.kaggle.com/marcodena/mobile-phone-activity/.

[3] Mobile phone activity (full) dataset. http://go.nature.com/2fz4AFr.

[4] G. Barlacchi, M. De Nadai, R. Larcher, A. Casella, C. Chitic, G. Torrisi, F. Antonelli, A. Vespignani, A. Pentland, and B. Lepri. A multi-source dataset of urban life in the city of milan and the province of trentino. *Scientific data*, 2(1):1–15, 2015.

[5] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45):15888–15893, 2014.

[6] K. H. Grantz, H. R. Meredith, D. A. Cummings, C. J. E. Metcalf, B. T. Grenfell, J. R. Giles, S. Mehta, S. Solomon, A. Labrique, N. Kishore, et al. The use of mobile phone data to inform analysis of covid-19 pandemic epidemiology. *Nature communications*, 11(1):1–8, 2020.

[7] T. Louail, M. Lenormand, O. G. Cantu Ros, M. Picornell, R. Herranz, E. Frias-Martinez, J. J. Ramasco, and M. Barthelemy. From mobile phone data to the spatial structure of cities. *Scientific reports*, 4(1):1–12, 2014.

[8] G. Miritello, R. Lara, M. Cebrian, and E. Moro. Limited communication capacity unveils strategies for human interaction. *Scientific reports*, 3(1):1–7, 2013.

[9] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[10] A. Wesolowski, C. O. Buckee, K. Engø-Monsen, and C. J. E. Metcalf. Connecting mobility to infectious diseases: the promise and limits of mobile phone data. *The Journal of infectious diseases*, 214(suppl_4):S414–S420, 2016.

[11] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270, 2012.