Prediction Model to measure crowdedness within Amsterdam.
Thesis Design Proposal

Author:
Don de Lange

University of Amsterdam, Science Park 904, The Netherlands,
don.delange@student.uva.nl

Abstract. The aim of this research is to construct a model that predicts the crowdedness at selected spots within the city of Amsterdam. Historical data will be gathered from localized crowdedness measure points and may be combined with GVB data, telecom data, and weather data. Either a multivariable regression model or clustering model will be used as prediction model. The modelss will be evaluated with the actual crowdedness counts at the selected spots in the city of Amsterdam.

Keywords: Crowdedness, Open Data, Regression, Clustering

1 Introduction

The city of Amsterdam is working towards becoming a smart city, with the use of gathered open source data from different sources within the city. Crowdedness measure is one of these instances and is aimed at predicting the degree of crowdedness within the city in real-time and making these predictions available for the city to use.

Amsterdam is a tourist attraction for tourists all over the world, increasing the crowdedness of the city. As it is unclear what the level of crowdedness will be in the near future, Amsterdam can only react to the crowdedness in the city. For example, making sure there is enough space for all the travellers during rush hour or making sure there is enough security during King's day.

The aim of this research is to built a prediction model for the crowdedness in public spaces in the city of Amsterdam, which is now unevenly measured within the city with no clear method of predicting the level of crowdedness within the city. For this project, data will be gathered from "precise" localized data sources and city-wide data sources and built a prediction model that can reliably predict the crowdedness within the city of Amsterdam. The main focus will lie on correctly predicting the crowdedness at the same spots as the localized data sources and maybe generalize it within the entire city.

The main challenge of this project lies in the following points; first, combining data from different open sources in an effective manner. AS there are a lot of indicators that could contribute to the crowdedness in the city, data from all

these indicators must be gathered, processed, and used in an effective manner. And second, making the data usable for predictions. As stated in the previous point, the data from different sources in different formats will be used. Only data that has a significant influence on the accuracy of the model must be used. In addition, the model must be able to use the data, meaning that the model must be able to handle both numerical and categorical data.

1.1 Research Questions

The main research question is stated as follows:

How can an accurate and fast prediction of the level of crowdedness within the city of Amsterdam be given, based on input from city-wide available data sources?

The main research question will be answered with the following two sub questions.

- *How can data be gathered from the available city-wide data sources in a time efficient manner?*
- *What model could be used to provide an fast and accurate prediction of crowdedness at local crowdedness measure points within the city, based on historical data? No baseline*

2 Related Work

Within the specific domain of crowdedness within the city of Amsterdam, there is no research available. So this research will serve as a first step in this area.

2.1 Crowdedness

Niu et al (2017) investigated how to estimate the crowdedness level for buses. The crowdedness is measured in real-time using sensors. This research will expand upon this by predicting the crowdedness.

2.2 Machine Learning Approaches

As there is no labelled data available, this research looks at unsupervised machine learning models. In addition, the data for this research contains both categorical and numerical variables, which the model must also be able to handle.

The k-prototype algorithm is an clustering algorithm that can cluster data with mixed numerical and categorical variables (ji et al., 2012). The algorithm is less prone to falling into local optima, as the uncertainty of the data is preserved. In addition, the clusters are based on the dissimilarity between the different variables of all the data points.

Yu et al. (2018), improved upon the previous given k-medoids algorithm by improving the clustering performance, whilts perserving the original computional efficiency and simplicity of the algorithm.

Choi et al. (2016) developed a deep learning method to predict heart failure. To evaluate its performance, the model was compared to regularized logistic regression, neural network, support vector machine, and k-nearest neighbor classifier. This research aims to implement a similar evaluation method for the prediction model, by evaluating different predictive models.

Glanz et al (2018) developed a predictive model that predicts the probability that a patient will have an opioid overdose, based on five different variables. The paper offers a method for the construction and validation of a predictive model. This research aims to do use a similar method in constructing a predictive model, within the context of crowdedness within the city of Amsterdam.

Ing and Ing (2018) constructed a multivariable predictive model with ten different variables for the nanogram of GCA. This research aims the appy a similar predictive model in the context of crowdedness in the city.

3 Methodology

3.1 Methods

For this research, the following prediction one of the folloiwng methods will be used. Or two methods will be used and compared to each other in terms of performance.

The first prediction method is a *Multivaraite Linear Regression Model*, where the model seeks to predict the numerical value of crowdedness at a given time on a given day. The categorical values will be transformed into binary values 0 and 1.

The second method is a *Clustering model with K-Medoids*, where the model makes clusters based on a combination of categorical and numerical variables. As the distance between points of categorical values is meaningless, k-means cannot be used. The K-medoids algorithm used centre points as arbitrary distance points and seeks to minimize the sum of dissimelarity between the points and the centre point. The model will cluster historical occurences of similar crowdedness together.

And the third method is a *Neural Network*, where the model attempts learn the most significant reasons for high crowdedness measures in the city, using Deep Learning. The data of the variables will be transformed using one-hot encoding and the final layer will assign a label of crowdedness to the given input data.

3.2 Data

The dataset consists of the following parts.

First, a daily count of the number of people at given data gathering points. The data is gathered in two different ways. First, by a camera that uses images

recognition to classify objects. And second, by a router that counts the number of points it can connect to.

Second, Telecom data of certain points within the city. Everytime someonse sends a requests over the internet, this request is picked up and saved. Think a person sending an app, posting a picture on Instagram, or searching dogs on google.

Third, wheather data of Schiphol from the last few months. There is no KNMI data available for Amsterdam, but Schiphol is close enough to count as an estimation.

And fourth, public transport data of the GVB subway within Amsterdam of the number of passengers. This data is anomalyzed and aggregated, meaning that only the number of passengers at a certain station is visible. Given that the number was higher than five.

3.3 Evaluation

As there is no similar research nor model available to compare the results with, this research has no baseline. The models will be evaluated in the following ways:

First, the *Linear Regression Model* will be evaluated by comparing the predicted crowdedness number with the actual crowdedness number.

Second, the *Clustering Model* will be evaluated by checking whether the instances in the given clusters are similar to each other in terms of given crowdedness.

And third, the *Neural Network* will be evaluated by checking whether the given crowdedness score fits the actual crowdedness number.

4 Risk Assessment

There are three main foreseeable risks with this project:

The first risk is that the data is not delivered on time. To make sure I'll be able to start if this happens, I have acquired a dataset with older, incomplete and less relevant data. This will enable me to start with my research on time and incorporate the real dataset later.

The second risk is that my computer lacks the computational power to run the model. If this happens, I will be able to run my model on one of the computers within the company. These computers are used to run their high computational models, so it will be able to run mine.

And the third risk, is that the data preperation takes too long with the large dataset. In this case, I will only use a small subset of my data to run the model on. This will give me some results needed to be able to finish my research on time.

Week Number	Date	Achievement
Week 1	01/04 - 05/04	Related Literature
Week 2	08/04 - 12/04	Data Gathering
Week 3	15/04 - 19/04	Data Preparation
Week 4	22/04 - 26/04	Methodology
Week 5	29/04 - 03/05	Evaluation
Week 6	06/05 - 10/05	Initial Results
Week 7	13/05 - 17/05	-
Week 8	20/05 - 24/05	-
Week 9	27/05 - 31/05	Final Evaluation results
Week 10	03/06 - 07/06	Draft Version Thesis
Week 11	10/06 - 14/06	-
Week 12	17/06 - 21/06	Project Finished

5 Project Plan

References

1. Ji, Jinchao, et al. "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data." Knowledge-Based Systems 30 (2012): 129-135.
2. Niu, Xiaoguang, et al. "A hierarchical-learning-based crowdedness estimation mechanism for crowdsensing buses." 2017 IEEE 36th International Performance Computing and Communications Conference (IPCCC). IEEE, 2017.
3. Glanz, Jason M., et al. "Prediction model for two-year risk of opioid overdose among patients prescribed chronic opioid therapy." Journal of general internal medicine 33.10 (2018): 1646-1653.

4. Ing, Edsel B., and Royce Ing. "The use of a nomogram to visually interpret a logistic regression prediction model for giant cell arteritis." Neuro-Ophthalmology 42.5 (2018): 284-286.
5. Choi, Edward, et al. "Using recurrent neural network models for early detection of heart failure onset." Journal of the American Medical Informatics Association 24.2 (2016): 361-370.
6. Yu, Donghua, et al. "An improved K-medoids algorithm based on step increasing and optimizing medoids." Expert Systems with Applications 92 (2018): 464-473.