Kapil Bastola

Feb 4, 2017

DS 670

Contribution and State of Art

**Contribution**

There are three important numerical values present in the dataset. One is the time taken to travel between two points, next is the average time it takes the vehicles to travel from one point to another and the third is the number of vehicles that travel between the two points for each reading.

First thing I would like to do with the dataset is to look at the descriptive statistics of the different fields in the dataset. In particular, for the numerical fields like Average time taken, Average speed, or Number of vehicles, I will look at the histograms of each of these field to see the distribution of values. I will also look at the scatterplots of each numerical variable to see if any correlation exists between the different fields. For descriptive fields I will look at what different categorical values exists and if there are any attributes I can work with.

Since the dataset we have is unlabeled, we would have to use unsupervised learning methods to get anything out of the dataset. In order to look at the structure of the data for this dataset I will use clustering techniques to group datasets into different clusters. There are various algorithms that can be used to accomplish this task and ones I will investigate further are k-means clustering and clustering using representatives (CURE). In order to identify what the correct number of cluster to use is I will use gap analysis or internal evaluation. I will also evaluate or assess the results I get from clustering of the dataset to see whether the results from the clustering task are useful or not.

Since the data we have is time based, we can conduct time-series analysis. In order to do the time series analysis, first we will plot the entire time series to see if there is any trend in the dataset and to check for any seasonality in the dataset. When I look at the time stamp field in the datasets, there are data points for every five seconds. We can use the velocity data to predict future velocity or number of vehicles at a given time using time series models like Auto-Regressive Integrated Moving Average (ARIMA) model. I will use the auto correlation function (ACF) and partial auto correlation function (PACF) to determine which order of auto regression and moving average is required for the model. By predicting velocity at a certain time we can make suggestions as to what time are busy and when to be careful when using the roads.

Apart from these methods of analyses, I will also keep my eye open for any other methods like neural networks or nearest neighbors that can be used with the dataset.

**State of art**

In order to look at the different group of data points or clusters of data in the entire data set we will use k-means clustering algorithm. K-means takes 2 inputs. One is parameter k the number of cluster and next is the training set. The algorithm first randomly initializes k number of clusters centroid. In next step for each of the training data point measure the distance to each of the assigned cluster centroid and assign the data point to the centroid with the least distance. After this process we will get k different clusters. The next step is to move the random centroid for each cluster to the centroid of the assigned data points. The two steps are repeated until we get same centroids in two repetitions. End goal is to produce clusters of data points that are similar to each other.

For the road traffic dataset, the velocity and the number of vehicles can be plotted in two-dimensional space. We can then try to group these into different clusters. After the algorithm produces different clusters of the data set it is importance to validate if the results we got is any useful or not. This will also determine whether the value of k we used is relevant or not. According to Steinley and Brusco, the steps of cluster validation are to first determine if the dataset consists of more than one cluster and if so, determine the number of clusters and then interpret, test and validate the results [1]. According to Steinley and Brusco we need to conduct cluster validation in order to determine whether a cluster is good or not. For any cluster analysis we need to test a portion of a dataset to diagnose whether a clustering model represents the entire data set or not. The diagnosis process can involve both subjective as well as objective evaluation. The testing, that uses external validation, are subjective because they require the expertise about the dataset to make judgement while determining a set of theoretical justified covariates and the interpretation of each cluster in the solution of the model will also be subjective because it depends on the understanding of theoretical significance of the variables included in the clustering procedure. To complete the validation process after testing and interpretation is replication. This is the method that tells us whether the clustering result is valid or not and also whether the solution of the clustering can be generalized or not. One of the most widely used replication technique, simply called replication analysis, will check to see if there are any replicable solutions or not. [1]

Clustering using representatives or CURE is a clustering technique that does not necessarily favor spherical shapes and sizes of clusters and tries to incorporate outliers as much as possible to the clusters. This helps in clustering when the clusters are not uniform size or shape. [2] According to the authors Guha, et. al., CURE method is able to incorporate the outliers by representing each cluster by generating certain number of fixed points by selecting well scattered points and then shrinking these points toward a center of the group by a fraction. This allows there to be more than one representative for a cluster unlike the k-means, where each cluster only has one representative point. When there is more than one point for a cluster, non-spherical shapes can be obtained. CURE is especially beneficial for dealing with large datasets because it uses sampling and partitioning of data. The method first draws a random sample from the dataset and partitions the data into certain part. Each of these parts is then clustered separately. These partial clusters are clustered again to get the final resulting clusters. The research shows that the clusters methods using CURE provides a better quality clustering than traditional methods like k-means. The ability to do random sampling and partitioning also lets CURE models to have better performance and gives the models ability to scale properly for large datasets without any compromise in quality of the model. [2]

Traffic forecasting can be a good way for a city to manage the situation of traffic in the roads. Short term forecasting can be achieved using many parametric and non-parametric techniques but ARIMA has been found to be the most accurate method to forecast traffic flow according to the referenced paper by Kumar and Vanajashi. [3] The paper mentions that predicting of traffic flow in short term based on past traffic data is an important element of an intelligent transportation system (ITS). Short term traffic flow forecasting involves predicting traffic volume in the next time interval in the range of five minutes to one hour and this has been researched extensively in the past. The number of vehicles in the road or the traffic flow is a random process. The statistical techniques used to predict traffic flow can be both parametric and non-parametric. The non-parametric techniques like non-parametric regression and neural networks include both descriptive and inferential statistics while the parametric techniques like linear and non-linear regression, historical average algorithms, smoothing techniques, and autoregressive linear processes assume that the dataset has a distribution with stable set of parameters. The paper suggests that the time series analysis based techniques like the autoregressive integrated moving average (ARIMA) is one of the most precise methods for the prediction of traffic flow when compared to other available techniques mentioned above. [3]

In order to use the ARIMA model we need to provide which level we want to set the model at. We can utilize just the auto-regression (AR) part of the model by setting to a numerical value or we can just set the moving average (MA) portion of the model. In order to determine which part of the model to chose and at what orders, we look at the auto correlation function and partial auto correlation function. Auto correlation function (ACF) gives us the correlation between a time series and the lags of itself while the partial auto correlation function (PACF) gives us the correlation of a time series with its lag that is not explained by correlation at lower order lags. [6] If the ACF shows any significance at any of the lags we use that order of lag for the moving average and if PACF shows any significant spikes at any of the lags we use that order of lags for the auto regression. After we set the order for the model we can generate the model using R. We can then predict next series of data points based on the model. In order to predict the accuracy of the model we can use mean average predicted error [2]. Looking at the value of MAPE will tell us whether our model is highly accurate, good, reasonable or inaccurate. [2] Kumar and Vanajakshi give us the scale of what percentage of MAPE refers to the labels mentioned before. Generally, forecasts where the MAPE value is less than 10 percent can be considered highly accurate, 11-20 percent can be considered good, 21-50 percent can be considered reasonable while the forecasts with MAPE value of 51 percent or more is considered inaccurate. For traffic flow the data varies from time to time and a MAPE value that is between 10 percent and 20 percent can be considered acceptable for this study. [3]

The paper called Short-term traffic flow prediction models-a comparison of neural network and nonparametric regression approaches by Smith and Demetsky affirm the notion by Kumar and Vanajashi that predicting traffic flow plays an important part in any intelligent highway or transportation system. [4] Smith and Demetsky confer that nearest neighbor models are well suited for this type of traffic flow prediction and will provide models that are very accurate and portable. They also argue that the non-parametric models using nearest neighbors can be easily understood by field personnel who will be using the output from the model in day to day basis. [4]

In conclusion, I will be using clustering and time series analysis, specifically ARIMA, in this project based on the characteristics of the data set. Since there are no labels in the data that can be used, using unsupervised learning to cluster the dataset is a good way to look at the structure of the data. The time series analysis allows us to create a prediction model to find the velocity of traffic after certain period of times.

Reference:

1. Steinley, D., & Brusco, M. J. (2011). Choosing the number of clusters in Κ-means clustering. *Psychological Methods*, *16*(3), 285-297. doi:10.1037/a0023346
2. Guha, S., Rastogi, R., & Shim, K. (2001). Cure: an efficient clustering algorithm for large databases. *Information Systems*, 26(1), 35-58. doi:10.1016/s0306-4379(01)00008-4
3. Kumar, S. V., & Vanajakshi, L. (2015). Short-term traffic flow prediction using seasonal ARIMA model with limited input data.*European Transport Research Review, 7*(3), 1-9. doi:http://dx.doi.org/10.1007/s12544-015-0170-8
4. Smith, B., & Demetsky, M. (n.d.). Short-term traffic flow prediction models-a comparison of neural network and nonparametric regression approaches. *Proceedings of IEEE International Conference on Systems, Man and Cybernetics.* doi:10.1109/icsmc.1994.400094
5. Zhao, Z., Chen, W., Yue, H., & Liu, Z. (2016). A novel short-term traffic forecast model based on travel distance estimation and ARIMA. *2016 Chinese Control and Decision Conference (CCDC)*. doi:10.1109/ccdc.2016.7532126
6. Business, F. S. (n.d.). Identifying the orders of AR and MA terms in an ARIMA model. Retrieved February 07, 2017, from http://people.duke.edu/~rnau/411arim3.htm