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.

**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:

“Conceptually, the process of determining whether a clustering is “good” falls under the rubric of cluster validation. Thus, a responsible implementation of a cluster analysis must include a diagnostic component that tests the partition to determine whether it adequately represents a cluster structure or is an arbitrary division of the data set. The three steps of the validation process contain a mixture of subjective and objective assessment. Specifically, the interpretation and testing (e.g., also termed external validation and usually consisting of a multivariate analysis of variance using the cluster solution as a grouping variable and a set of covariates that were not included in the analysis as dependent variables) are subjective in the sense that they both require domain-specific knowledge: the former because the interpretation of individual cluster solutions will depend on the theoretical significance of the variables included in the clustering procedures and the latter because domain experts will have to determine a set of theoretical justified covariates.

The final component of the cluster validation process is replication, and it is also considered to be a method for determining the validity and generalizability of the final cluster solution. In the present article, we discuss the most widely used type of replication, simply termed replication analysis, and show that replication analysis can succeed regardless of whether the other steps in the process were correct. Specifically, it is possible to get replicable solutions when an incorrect number of clusters are postulated. As such, we propose reformatting the steps outlined to combine choosing the number of clusters within the context of cluster validation.” [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. [4] According to the authors Guha ET. Al.,

“CURE achieves this by representing each cluster by a certain fixed number of points that are generated by selecting well-scattered points from the cluster and then shrinking them toward the center of the cluster by a specified fraction. Having more than one representative point per cluster allows CURE to adjust well to the geometry of non-spherical shapes and the shrinking helps to dampen the effects of outliers. To handle large databases, CURE employs a combination of random sampling and partitioning. A random sample drawn from the data set is first partitioned and each partition is partially clustered. The partial clusters are then clustered in a second pass to yield the desired clusters. Our experimental results confirm that the quality of clusters produced by CURE is much better than those found by existing algorithms. Furthermore, they demonstrate that random sampling and partitioning enable CURE to not only outperform existing algorithms but also to scale well for large databases without sacrificing clustering quality.” [4]

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. [2] The paper mentions:

“Traffic forecasting, the process of predicting future traffic conditions in short-term or near term future, based on current and the past traffic observations is an important component of any of the Intelligent Transportation Systems (ITS) applications. Short-term traffic flow forecasting, which involves the prediction of traffic volume in the next time interval usually in the range of five minutes to 1 h, is one of the important research problem in the field of ITS addressed by many researchers in the last two decades. Traffic flow or the number of vehicles crossing a particular point per unit time period is a point process or in other words, it is a type of random process, which consists of a set of isolated points collected over time… In general, the statistical techniques used for the problem of traffic flow prediction can be classified as non-parametric or parametric statistical techniques. The nonparametric techniques include nonparametric regression and neural network. The parametric techniques include linear and nonlinear regression, historical average algorithms, smoothing techniques, and autoregressive linear processes. It is reported 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 as mentioned above.” [2]

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.

“According to Lewis’ scale of interpretation of estimation accuracy, any forecast with a MAPE value of less than 10 % can be considered highly accurate, 11–20 % is good, 21–50 % is reasonable and 51 % or more is inaccurate. In most of the studies on flow prediction, a MAPE in the range of 10–20 % was reported. Since traffic flow observations vary from a few hundred vehicles per hour in off peak to several thousand vehicles during peak periods, MAPE in the range of 10–20 % is generally acceptable. Based on this, it can be seen that the results are highly accurate with MAPE less than 10 % and within acceptable limits.” [2]

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 is 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. 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
3. 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
4. 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
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