Erschienen in: 2019 IEEE Conference on Visual Analytics Science and Technology (VAST) : Proceedings / Chang, Remco; Keim, Daniel A.; Maciejewski, Ross (Hrsg.). - Piscataway, NJ : IEEE, 2019. - S. 136-137. - ISBN 978-1-72812-284-7

Earthquake Investigation and Visual Cognizance of Multivariate Temporal Tabular Data Using Machine Learning

Arjun Majumdar Gent Ymeri Sebastian Strumbelj Juri Buchmüller Udo Schlegel
Daniel A. Keim

University of Konstanz, Germany *

ABSTRACT

This paper presents our tool for the Vast Challenge 2019 Mini Challenge 1 (MC1). It will give an overview of the approach of data preprocessing techniques used for the given dataset and it will introduce our application which is built considering the requirements and questions to be answered for the MC1. This application consists of Machine Learning techniques and Information Visualization techniques such as Integrated Spatial Uncertainty Visualization as shown in this paper [1] to convey the needed information to the end users. To show the usefulness of this application we give examples of analysis.

Keywords: Machine Learning, Visual Analytics

1 INTRODUCTION

The Vast Challenge 2019 Mini Challenge 1 presents data from the city of St. Himark which has been hit by an earthquake that damaged a nuclear power plant. The data comes from an app that allows citizens of this city to provide information on damages for 6 different attributes such as Buildings, Roads and Bridges, Power, Medical conditions, Sewer and Water and lastly for Shake Intensity. Responses from this app come in the form of a ranking from 1 to 10 according to how people judge the conditions for these different attributes mentioned above. The tasks for the MC1 require dynamic prioritization of neighborhoods for response such as which parts of the city are hardest hit. Furthermore, uncertainty in the data, the reliability of neighborhood reports has to be shown. Lastly, changes in the conditions over time need to be shown.

2 DATA PREPROCESSING

The first task in MC1 is about prioritizing neighborhoods according to the damage taken by each location which changes dynamically according to time. The dataset had a lot of missing values for different attributes which we had to deal with.

In order to perform data imputations, we tried classical approaches such as using mean, median and 0 values which yielded unsatisfactory results. We then progressed to using different matrix completion and imputation algorithms which also gave mediocre results. Finally, we treated data imputation problem for the different attributes as a supervised learning classification task by splitting the dataset into features, containing all other attributes and target, containing the attribute which has to be imputed. Further, the features and target are divided into training and testing sets by dividing the available data as training set, and the missing data are used as testing set. The model is trained on the training set and then predicts for the missing values in the testing set. We also observed that the data imputation being done for each individual attribute should be saved as different files which then should be finally merged into an imputed dataset. If we used an imputed attribute to predict missing values for another attribute, this caused bias to creep in the final dataset.

We chose LightGBM classifier for the attributes listed below where the target attribute had 11 possible values (0, 1, 2, ... 10), which are the ratings provided by the users. The metrics achieved for the different attributes are as follows:

Table 1: Original dataset:

Target attribute	Accuracy	Precision	Recall
sewer_and_water	0.488	0.4756	0.4517
power	0.4818	0.4873	0.468
roads_and_bridges	0.4845	0.4747	0.4505
medical	0.7693	0.4224	0.4618
buildings	0.4611	0.4017	0.384
shake_intensity	0.5238	0.3701	0.3621
location	0.842	0.8371	0.7881

After applying different machine learning models, the top 3 models we got are as follows:

Table 2: Original dataset - location attribute

Classifier	Accuracy	Precision	Recall
LightGBM	0.842	0.8371	0.7881
XGBoost	0.8268	0.8122	0.7652
Random Forest	0.8342	0.8191	0.7882

We then perform data imputation as mentioned above and obtain the imputed dataset, the results for the different attributes along with the metrics are as follows:

Table 3: Imputed dataset

Target Attribute	Accuracy	Precision	Recall
sewer_and_water	0.5109	0.502	0.4784
power	0.5163	0.515	0.5031
roads_and_bridges	0.5223	0.5048	0.4846
medical	0.657	0.6685	0.6312
buildings	0.4879	0.4281	0.4146
shake_intensity	0.5768	0.4928	0.5051
location	0.8297	0.7986	0.7675

Table 4: Imputed dataset - location attribute

Classifier	Accuracy	Precision	Recall
LightGBM	0.8297	0.7986	0.7675
XGBoost	0.8026	0.7757	0.727
Random Forest	0.8258	0.7978	0.7604

*e-mail for all authors: firstname.lastname@uni-konstanz.de

We therefore have the following conclusions: Location is the best target attribute, giving highest metrics and LightGBM classifier is the best Machine Learning model.

3 Application

3.1 Visualization

In the frontend, the system consists of four main components:

Slider: With the time slider you can jump at any time and analyze values over an adjusted time window. In contrast to a stream-based system, several hours or whole days can be analyzed, which could lead to further indications on an hourly or daily level. In addition, the visualization can be animated with a play button so that the temporal development of the data becomes visible. By optimizing the calculations using indexing in the database, this can also be done with a huge amount of data coming from a big time window.

Map: The map visualizes two important components of our concepts: With the color the value estimated_damage is encoded. The scale goes from a bright blue to green, yellow and red. While blue/green colors appear less dangerous in perception, yellow/red colors have a warning effect, which also corresponds to the scale from a low estimated_damage to very high estimated_damage. With the size of the superimposed boxes the uncertainty of the data is encoded. It is shown that three types of uncertainty (bins) contain enough information (uncertain, medium certain, certain) not to lose any important details in the bins and are visually very well distinguishable from each other. This does not hold for more or less bins. You can also see a Radial Bar Chart by hovering over locations. This represents the mean value of the pure data in the given time window and thus allows the exploration of the individual components.

Ranking: The animated ranking visualization shows the locations sorted by the estimated_damage. This helps the viewer to identify the locations that have been hit the most, regardless of their position in the map.

Feature Importance Bar Charts: In this plot, the individual features (through an extended calculation) of the five most hit locations are shown. Because in most cases it turns out that five or fewer locations are more affected than the rest. For these locations in particular, a comparison can be made between pure data from the radial bar chart and calculated values in this plot in order to obtain double confidence in the evaluation of the individual features.

3.2 Machine Learning

After imputing the dataset, we do feature engineering to create estimated_damage attribute which uses location as target attribute and ranks the different features in the dataset according to feature importance of the model.

The estimated_damage for each data instance is obtained as follows:

$$\sum_{r \in rows} \sum_{f \in features} feature\_score(f,r) * feature\_importance(f) \quad (1)$$

Where, r or row is a row within the dataset or a data instance, f or feature are different features viz., power, medical, shake_intensity, etc. estimated_damage is then used for ranking of the neighborhoods according to being hardest hit (for a given time frame).

3.2.1 Mathematical model for quantifying and capturing uncertainty in the dataset:

The coefficient of variation (relative standard deviation) is a statistical measure of the dispersion of data points around the mean and is calculated by $cv = \frac{\sigma}{\mu}$. The metric is commonly used to compare the data dispersion between distinct series of data.

And the formula used to quantify the uncertainty in the dataset is: $Unc(t,loc) = |entries(t,loc)| * \sum_{f \in features} median(t,f,loc) * CV(t,f,loc) * feat\_imp(f,loc)$, where, number of entries for a given time frame t and location loc, median of feature f's score according to time frame t location loc. Median is more robust against outliers and is thereby used, CV is Coefficient of Variation for a given time frame t, feature f location loc, feature importance (from ML model) for a feature f location l.

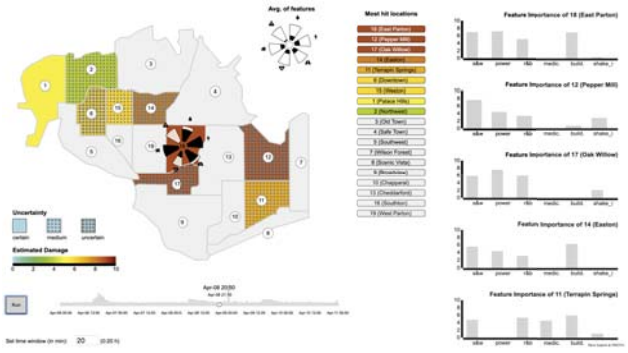

Figure 1: Illustration of our tool - MC1 VAST Challenge 2019

4 Analysis

A 20 minute time window starting on April 8, 20:50 is considered in Figure 1. It can be observed that a large proportion of regions do not provide data, which is encoded with a gray color. This can either mean that there is absolutely nothing going on or that people are more busy with the earthquake and won't get the idea to use an app on their mobile phone (except for the emergency call). Region 18, 12, 17 are the hardest hit, with the boxes at region 12 and 17 visualizing that the system is not quite sure about the prediction. The next regions to be hit hard are 11, 6 and 15, where the system is uncertain in Region 11 and 15 but very uncertain in Region 6. For this reason, it is now up to the analyst to continue dealing with these regions or not. Let us focus in the following on the most affected region, Location 18: By means of details on demand through the radial bar charts, the mean value of the pure data can be viewed. The attributes "buildings" and "power" have the highest mean value, followed by "roads_and_bridges" and "sewer_and_water". The attribute "medical" has a rather smaller value. The attribute "shake_intensity" has a very small value. A look at the feature-importance-charts on the right side confirms these values except for medical. Using intensive pre-processing and machine learning methods, an importance of the features could be calculated (see Application section). This shows that the attribute "medical" has no importance in the current scenario. The double comparison between pure values and the results of our models thus provides an even more reliable result than if only the pure data were considered. Thus, the Rescue Team can now send out emergency personnel who are responsible in particular for buildings and power supply and, if necessary, also for infrastructure and water supply. This makes it possible for a rescue team to send out the right emergency forces quickly and reliably.

5 Conclusion

In this paper we presented our tool which was built to tackle the MC1 for the Vast Challenge 2019. We gave an overview of the approach of data preprocessing techniques used, the Application and Analysis. This tool can be used to rank most affected locations and convey information for the most uncertain locations.

References

[1] D. Jäckle, H. Senaratne, J. Buchmüller, and D. A. Keim. Integrated spatial uncertainty visualization using off-screen aggregation. In *EuroVA@ EuroVis*, pp. 49–53, 2015.