Maximilian Eibl, Martin Gaedke. (Hrsg.): INFORMATIK 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 931

Pattern Based Decision Tree Analysis for Risk Detection in Smart Cities

Matthias Scholz[1] and Gunther Piller[2]

Abstract: Increasing amounts of data on living environments and human interactions are becoming available. Their potential for valuable services improving the wellbeing of individuals is large and growing. This calls for an investigation of algorithms and system architectures that support possible use cases. In this paper we outline how pattern based decision tree analyses can be applied to the identification of risks caused by time-dependent effects from multiple influencing factors. For this purpose we apply the method to open data on car accidents and weather conditions. We also show how such systems can take advantage from up-to-date in-memory technology.

Keywords: Data Mining, In-Memory Computing, Smart City Services

1 Introduction

Detailed information on living environments and human interactions become more and more available. It spans wide areas covering e.g. transport, air quality, traffic, health, weather, municipal services, events and citizen data. Often this information is published through open data platforms like NYC Open Data [OD17] or London Datastore [LD17]. Fostering citizen engagement and opportunities for collaborations as well as smart systems and services for better living are typical goals of most offerings (for examples see e.g. [IS16, He16]).

To succeed, manifold challenges have to be overcome. This includes the collection of suitable data and their preparation, the identification of services with high added value, the development of appropriate algorithms, as well as the conception and implementation of underlying IT systems.

An interesting class of smart services aims at the identification and mitigation of risks caused by time-dependent effects from multiple influencing factors. Personal guidance for avoiding health hazards triggered by different air pollutants, weather conditions, individual factors and momentary personal constitutions are a typical example (see e.g. [Sc16]). Data for such services often come from multi-sensor networks including also people as sensors [Ci10, Co14, SR15].

In this paper we describe a promising algorithm and its implementation for such services.

[1] University of Applied Sciences Mainz, Lucy-Hillebrand-Straße 2, 55128 Mainz, matthias.scholz@hs-mainz.de
[2] University of Applied Sciences Mainz, Lucy-Hillebrand-Straße 2, 55128 Mainz, gunther.piller@hs-mainz.de

doi:10.18420/in2017_95

It builds upon sequential pattern mining for individual influencing factors. Frequent patterns are then used as features for classification. The resulting models can finally be used to evaluate the risk of momentary situations through the processing of real-time data [LC10, Sc16].

In order to examine the usefulness of the proposed method in a smart city context, it is applied to an analysis of weather and accident data obtained from NYC Open Data [OD17]. The corresponding prototype has been implemented upon the in-memory platform SAP HANA [SP17]. In addition we outline how requirements for typical applications of the discussed analysis methods in smart city or smart living environments can be satisfied through an appropriate design of the underlying IT system. For this purpose we sketch the architecture and programming model of the implemented prototype.

The paper is structured as follows: In Section 2 the data mining approach is introduced. Its application to open data from NYC is described in Section 3. Section 4 sketches the architecture of the underlying IT system. We finally conclude with a short summary.

2 Sequential Pattern Mining and Decision Tree Analysis

The presented method combines sequential pattern mining with classification through decision tree mining [LC10]. In this section the approach is described in general terms. Then a concrete application is discussed.

First, all data which might be related to the occurrence of considered incidents need to be combined into sequences. They describe the time development of potential influencing factors. The collected data are then prepared for data mining. Important steps are a possible segmentation of incidents into buckets with similar characteristics and a discretization of measurements of influencing factors.

As a next step sequential pattern mining is used to identify frequent sequences of influencing factors before moments with incidents and moments without. These frequent sequential patterns are then treated as features to characterize the original datasets. Based on these features decision trees can be derived for different segments of incidents. When all relevant influencing factors are taken into account, the resulting trees represent a set of rules with which the risk of incidents can be evaluated by means of actual measurements.

2.1 Sequential Pattern Mining

To identify frequent sequences of influencing factors one has to consider sequences $S = \langle s_1, s_2, \ldots, s_n \rangle$ of temporally ordered values for corresponding measurements. In Section 3 we use one entry per day for all factors. For example, a sequence of length three for temperature contains the discretized temperature values for three subsequent days, e.g. $\langle cold, medium, high \rangle$. Depending on the nature of a particular influence factor, the

chosen daily entry can be determined differently, e.g. as an average, a minimum or maximum, or through accumulation.

For a particular factor all sequences of length *n*, e.g. *n = 3 days*, before days with incidents are collected within a dataset for high-risk sequences D_h. Sequences of similar length before days without incidents build a dataset D_l of low-risk sequences. This collection of sequences is carried out for all considered factors. Sequential pattern mining is then performed for all factors separately within high- and low-risk segments, i.e. based on the corresponding datasets D_h and D_l, respectively. In our approach we search for frequent sequential patterns with length from one to *n*. For sequences with length $1 < m \leq n$ we also consider patterns containing at most $m - 1$ arbitrary entries. In this way one can account for factors which are effective over a period of several days.

A common measure for the significance of a frequent sequential pattern *S* is its occurrence or *Support*, denoted by $\sigma\,(S, D)$. It amounts to the total number of input-sequences in the database *D* that contain *S* as a subsequence (see e.g. [Za01]). Further important measures are *Confidence* and *Lift* [Za01, HK11]. In Section 3 we use *Lift* for pattern pruning, i.e. the selection of patterns for the decision tree analysis. *Lift* measures to what extent the occurrence of sequences with *S* as subsequence is enhanced in the high-risk dataset D_h [HK11], i.e.

$$Lift\,(S) = \frac{\sigma\,(S, D_h)/\,|D_h|}{\sigma\,(S, D_h + D_l)/\,|D_h + D_l|}$$

Here $|D_{h/l}|$ denotes the total number of input-sequences in D_h and D_l, respectively. Further ways of efficient pattern pruning can be found in [Za01, HK11, AH14].

2.2 Decision Tree Mining

Following Lee et al. [LC10], the selected frequent sequential patterns for all factors are interpreted as features. The measured sequences before days with and without incidents are considered as transactions. These are then characterized by the presence or absence of the identified features – which can be interpreted as attributes of the transactions. In addition, one attribute describes whether an input-sequence belongs to a high- or low-risk dataset, respectively.

The transactions and their corresponding attributes are then taken as input for decision tree mining. The attribute for risk classification is set as a target for the mining process. The resulting tree itself is a binary tree. Each node queries the presence of a frequent sequential pattern.

3 Initial Analysis of Car Accidents and Weather Patterns

To study the method from Section 2 within a smart city context, we have applied it to data on car accidents from NYC Open Data [OD17] in combination with weather data from Weather Underground [We17]. The goal of this initial investigation was twofold: First, to examine the architecture and implementation of the discussed method, so that the benefits of current in-memory technology can be leveraged; second, to study framework conditions for a meaningful application of the presented method in smart city scenarios.

The relation of weather conditions and car accidents is an active field of research. In general, road traffic safety is the result of complex interactions of technical, environmental and behavioral factors (see e.g. [Pe04]). Analyses indicate that a base rate of crashes depending on non-weather factors exists while weather conditions are able to substantially push crash rates on days with bad weather. Also non-extreme weather conditions can have a contributing role – for instance in the case of driver fatigue [Pe15].

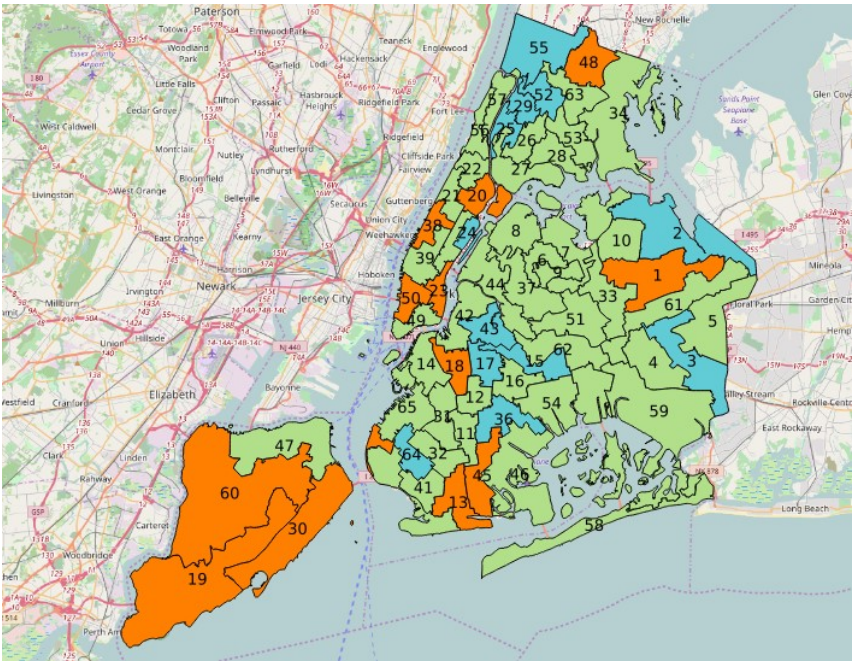

Fig. 1: Accident data from NYC on district level for a sample day classified as high-risk (orange), low-risk (blue), medium (green).

With a pattern based decision tree analysis one can study the influence of the temporal development of different weather factors on accidents. To examine this possibility we have taken data on car accidents in NYC from the year 2015. Extreme weather conditions with immediate impact, like ice or heavy snowfall, were excluded by considering only data

from spring to fall. The weather effects considered in such an approach could lead to changes of the base rate of accidents caused by other factors. Therefore we first have identified days with more than average accidents for individual district. Days with a number of accidents higher than two standard deviations from its average value were classified as high-risk, days with less than two standard deviations than average as low-risk. A typical distribution for a specific day is shown in Figure 1.

As potential influencing factors we considered hail, air pressure, humidity, fog, precipitation, visibility and temperature. Corresponding data from [We17] were captured as sequences with maximal length *n = 2 days*, before days with incidents and days without, building up the data sets D_h and D_l, respectively. With this input the sequential pattern mining and decision tree analysis from Section 2 was applied.

As a result we obtained sequential patterns for all districts and potential influencing factors. Of interest were patterns with $Lift > 1$. Several such patterns were identified. For example, averaging over all districts a frequent pattern with two sequential days of very high temperature ($Lift = 2{,}7$) and a frequent pattern with a strong change from high to very low visibility ($Lift = 2{,}3$) were found. For selected districts the lift of these and similar patterns was much larger – up to values close to 8.

Following the method described in Section 2.2, all patterns with relative high lift were taken as input for a decision tree analysis. Decision trees were trained with data from selected districts and tested against data from different ones. As a typical result, test runs for decision tree models based on patterns with $Lift > 1{,}4$ prior to days with an increase of accidents and patterns with $Lift > 1{,}2$ for days without an increase, yield the quality measures (see e.g. [HK11]): $Accuracy \approx 75\%$, $Negative\ Predictive\ Value \approx 87\%$, $Positive\ Predictive\ Value \approx 53\%$, $Specificity \approx 77\%$, $Sensitivity \approx 70\%$.

This first analysis indicates that weather patterns covering several days could have indeed an influence on car accidents. However, such effects certainly are mixed into effects driven by multiple non-weather factors, or extreme weather conditions with immediate impact. For more conclusive results analyses based on data spanning several years and multiple locations would be necessary. Also a segmentation of accidents according to the similarity of non-weather factors would be required. Characteristics relevant for this could span from regional differences in traffic infrastructure and conditions to details on drivers and involved vehicles [Pe15]. If weather patterns with significant risks for car accidents could eventually be verified, they might be used for preventing measures, like variable speed limits.

4 Implementation upon SAP HANA

In this section we describe the software system which has been used for the analysis explained in Section 3. Its architecture and implementation enables high-performance pattern based decision tree analyses for many different scenarios – in particular in the area

of smart systems for the improvement of individual living conditions. For example, it also has been applied to the analysis of bio signals and environmental data to provide personal guidance for individuals, who suffer from asthma and need to reduce their exposure to air pollutants [Sc16].

Starting point is the integration and processing of data from multiple different sources. Most relevant are environmental data from sensor networks or from generally available data services as well as data on conditions and activities of individuals and their interaction. Then data mining, i.e. feature identification and model building, has to be performed for large and scalable data volumes. Finally, the identification of risks needs to be carried out in real-time based on tested models and actual data.

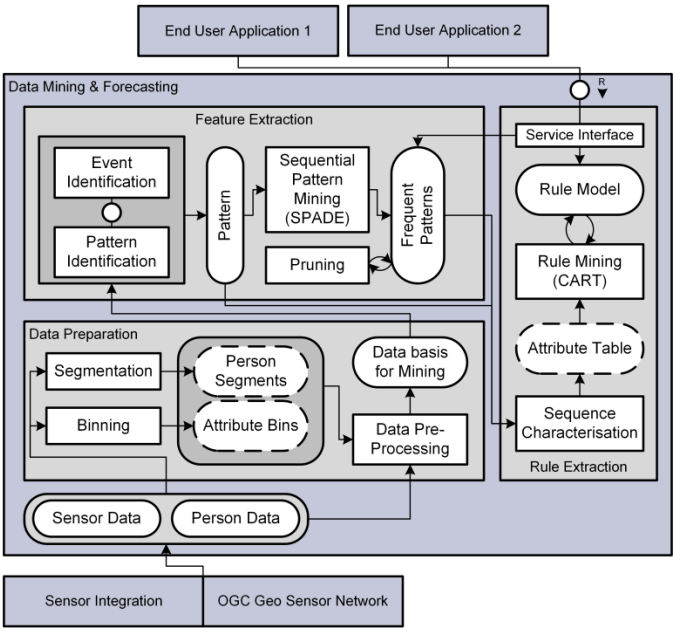

Fig. 2: Architecture of the component Data Mining and Forecasting.

Figure 2 shows the architecture of the system with focus on the data mining and forecasting components. The integration of data and their projection onto appropriate spatial-temporal references are carried out by the components Sensor Integration and OGC Geo Sensor Network which are described elsewhere [Sc16].

To enable high-performance data mining analyses of large data volumes, the in-memory platform SAP HANA is used. Feature identification and model building is carried out asynchronously, i.e. independent of the real-time evaluation of risks. Nevertheless short response times are needed also here, in particular for initial explorative analyses which are necessary for the identification of optimal mining parameters, like binning and segmentation, as well as for iterative improvements of forecasting models.

A light weight application for overall data processing and interactive mining steps, e.g. data binning and pattern pruning, has been implemented with SAP HANA Extended Application Services and SAP UI5. Data intensive calculations and data querying are handled through appropriate interfaces in the database using the SQL engine and the Application Function Library with the Predictive Analytics Library (PAL) [SP17]. As general guideline de-normalized data models have been chosen; write operations were avoided; data intensive application logic has been largely embedded into the database; stored procedures have been parallelized wherever applicable.

Examples for virtual tables are indicated in Figure 2 for segmentation, attribute bins and attribute tables. Also several opportunities for parallelization exist: The functional components for binning, segmentation, data pre-processing, pattern identification and sequential pattern mining can be executed independently in parallel for different types of data and segments.

Data preparation as well as the identification of input-sequences before days with and without incidents is carried out upon HANA itself. For sequential pattern mining an R implementation of the SPADE algorithm is applied [Ha15]. After identified frequent sequential patterns have been obtained, pattern pruning is currently carried out interactively. For decision tree mining the CART algorithm of PAL [SP17] has been used. As outlined in Section 2, a table with transactions and corresponding attribute values is used as input. It is obtained by comparing frequent sequential patterns with input-sequences. Target attribute for the CART algorithm is the attribute value for the occurrence of an incident – a car accident for the use case from Section 3. It describes whether an input sequence belongs to a high- or low-risk dataset. As output the algorithm provides a table containing a PMML tree model. The model can be transmitted via an OData interface to the applications of end-user, e.g. mobile applications, for real-time execution based on actual data.

5 Summary

In this paper we have presented a data mining method that can be used for the identification of risks caused by time-dependent effects from multiple influencing factors. We have shown how pattern based decision tree analyses can be used for new services in the context of smart cities by applying it to open data on car accidents and weather information. Finally a suitable system architecture leveraging the advantages of current in-memory technology has been presented. We look forward to use the described method and application for further scenarios within the context of intelligent systems for better living environments.

Supported by the German Federal Ministry for Economic Affairs and Energy and the HPI Future SOC Lab.

References

[AH14] Aggrawal, C. C.; Han J.: Frequent Pattern Mining. Springer, 2014.

[Ci10] CitiSense: Project CitiSense 2010, sosa.ucsd.edu/confluence/display/CitiSensePublic/CitiSense, visited on 2017/03/30.

[Co14] Copenhagen Wheel: Project The Copenhagen Wheel 2014, senseable.mit.edu/copenhagenwheel, visited on 2017/03/30.

[Ha15] Hahsler, M.: Package 'arules'. https://cran.r-project.org/web/packages/arules/arules.pdf, 2015, visited on: 24.04.2017.

[He16] Helfert, M. et al.: Smart Cities, Green Technologies, and Intelligent Transport Systems. Proc. 4th International Conference, SMARTGREENS 2015, and 1st International Conference VEHITS 2015, Lisbon, Springer, 2016.

[HK11] Han J., Kamber M.; Pei J.: Data Mining: Concepts and Techniques. Elsevier, 2011

[IS16] ISC2: Proceedings Smart Cities Conference (ISC2) 2016 IEEE International. IEEE, 2016.

[LC10] Lee, C. H.; Chen, J. C.; Tseng, V. S.: A Novel Data Mining Mechanism Considering Bio-Signal and Environmental Data with Applications on Asthma Monitoring. Computer Methods and Programs in Biomedicine 101 (1), pp. 44-61, 2010.

[LD17] London Datastore, data.london.gov.uk, visited on 2017/03/30.

[OD17] Open Data NewYork, opendata.cityofnewyork.us, visited on 2017/03/30.

[Pe04] Peden, M. et al.: World Report on Road Traffic Injury Prevention. World Health Organisation, Geneva, 2004.

[Pe15] Perrels, A. et al.: Weather Conditions, Weather Information and Car Crashes. ISPRS International Journal of Geo-Information 4.4, pp. 2681-2703, 2015.

[Sc16] Scholz M. et al.: Capture and Analysis of Sensor Data for Asthma Patients. Proc. 24th European Conference on Information Systems (ECIS 2016), Istanbul, 2016.

[SP17] SAP PAL: SAP HANA Predictive Analysis Library. https://help.sap.com/viewer/'2cfbc5cf2bc14f028cfbe2a2bba60a50/2.0.01/en-US, visited on 24.04.2017.

[SR15] Sagl, G.; Resch, B.; Blaschke, T.: Contextual Sensing: Integrating Contextual Information with Human and Technical Geo Sensor Information for Smart Cities. Sensors 15, pp. 17013-17035, 2015.

[We17] Weatherunderground 2017, wunderground.com/weather/api/. visited on 2017/03/30.

[Za01] Zaki M. J.: SPADE: An Efficient Algorithm for Mining Frequent Sequences. Machine learning 42 (1-2), pp. 31-60, 2001.