Detecting Changes in Statistics of Road Accidents to Enhance Road Safety

Katherina Meißner, Cornelius Rüther and Klaus Ambrosi

1 Introduction

When trying to detect changes in car accident statistics, police analysts are faced with a large amount of incidents on the one hand and many attributes with several attribute values leading to multitudinous possible combinations on the other. Of course, not all combinations and changes therein are essential to decide upon police actions to enhance road safety. But defining potentially interesting combinations to track in advance can lead to a narrow perspective on the actual situation. If there was an increase in the frequency of a particular combination, which had not manually been predefined, this increase would remain unrecognized and therefore untreated by the police for some periods. Tracking changes manually in these numerous operating figures is not possible.

We propose an automated approach based on Frequent Itemset Mining to detect significant changes in the statistical figures. It is based on the known Apriori algorithm which we apply to monthly slices of accident data to retain a sequence of monthly support values for each itemset. With these sequences, we try to classify the itemsets according to their appearance frequencies in each month. One major question to be answered within our framework is how to find changing itemsets that are worth being presented to the police analyst.

This paper is organized as follows: In Sect. 2, the related work is reviewed. The data set and the preparations made are then introduced in Sect. 3 before the algorithm and its parameters are presented in Sect. 4. With the frequent itemsets found for each month, we show how to detect changes in these data structures in Sect. 5. Finally, we discuss our conclusions and suggest future work in Sect. 6.

K. Meißner (✉) · C. Rüther · K. Ambrosi
Department of Economics and Information Systems, University of Hildesheim,
Universitätsplatz 1, 31141 Hildesheim, Germany
e-mail: meissner@bwl.uni-hildesheim.de

N. Kliewer et al. (eds.), *Operations Research Proceedings 2017*, Operations
Research Proceedings, https://doi.org/10.1007/978-3-319-89920-6_10

2 Related Work

There are some interesting approaches in change mining on the one hand and in association rule mining on accident data on the other hand. Song et al. [9] and Chen et al. [3] showed how to mine changes in customer behavior. They both focused on pattern mining in a marketing application. In a more general approach, Liu et al. [6] presented a method to distinguish between stable and trend rules. This approach is used for change detection in our framework.

Böttcher et al. [2] built a framework for change mining and defined the term itself. Baron et al. [1] divided the data mining process in two parts, mining the model and mining the changes, to speed up the mining process on evolving data.

As to the analysis of road accident data, Geurts et al. [4] made use of association rule mining to evaluate accident causes in so-called black spots, i. e. places where accidents regularly happen, and in contrast, the causes of accidents in other places in Belgium. Based on accident data from Florida, Pande et al. [8] conducted a market basket analysis to find associations between the accidents' characteristics. In 2014, Moradkhani et al. [7] did the same for UK-accident data, which is the data used for this research. The main focus of all of these approaches lies on finding the root causes for accidents. None of the above evaluated the change in accident statistics.

3 Data Preparation

The GB-accident data is openly available for the years 2005–2015. All accidents with personal injuries are provided with statistical information. The data set consists of 1.8 million accidents with 3.5 million vehicles and 2.6 million casualties, both having a one-to-many-relationship to accidents which has to be dissolved during data preparation. Some of the 55 variables utilized are e.g. date and time of the accident, weather conditions, type of vehicle, and age of casualty.

We decided to focus on the years 2014 and 2015 to build the analytics framework. With no other filters applied, we have a data set D consisting of 285,000 transactions with more than 300,000 different items (attribute-value-pair) after applying the following reduction methods. A single accident consists of 19–85 different items. Attributes containing location information were removed, as they were too detailed to find relevant frequencies. Attribute values like 'data missing' or 'none' were also not considered in order to prevent the mining algorithm from evaluating these uninteresting items. Moreover, attribute values with an occurrence level above 95% were pruned in advance.

Most of the attributes are provided as categorical data. The ones that are not, for example 'hour of accident' or 'age of vehicle', had to be discretized first. This is done automatically by building clusters with equal frequencies.

To analyze changes over time, D is finally separated into monthly data sets D_i and transformed to transactional data in order to find frequent itemsets.

4 Finding Frequent Itemsets

The Apriori algorithm for detecting frequent itemsets is performed on each $D_i \; \forall \, i = 1, \ldots, m$, resulting in $m = 24$ different sets of itemsets I_i. An itemset is considered frequent in our framework if at least 3% of the monthly data D_i support, i. e. contain, its combination of attribute values ($minsupp = 0.03$). For months with about 12,000 accidents, the algorithm detects about 55 million frequent itemsets. Because this amount is too high to find any interesting changing patterns, I_i must be condensed to a representative level. Therefore, the lossless representation of *closed frequent itemsets* is chosen. By removing supersets of itemsets with exactly the same support as the itemset itself, the amount is drastically reduced. The resulting itemsets are more general and therefore more applicable in practice than the pruned superset.

Xiong et al.'s [10] approach of a hyperclique pattern miner is used to only keep potentially interesting patterns, even when using a quite low *minsupp*. By applying this approach, all itemsets in I_i with an all-confidence value below 15% (*minAconf*) are pruned, as the items within these sets tend have a poor correlation. All-confidence is defined as $\text{all-confidence}(X) = \frac{\text{supp}(X)}{\max_{x \in X}\{\text{supp}(x)\}}$, where $\max_{x \in X}$ is the maximum support of all items x within itemset X. For association rule induction this would imply that all rules derived from this itemset X have a minimum confidence of all-confidence at least.

Association rules were not considered in the final framework. The dependence of items which the rules seemed to illustrate was contradictory, since many rules with similar confidence were found having the shape $A \Rightarrow B$ and $B \Rightarrow A$. Sorting them by confidence and removing the duplicates led to difficulties when joining the rules of two different intervals, because it could not be ensured that from one itemset the same rule was kept for all months. Hence, changes within the rules support or confidence could hardly have been detected that way.

The thresholds for *minsupp* and *minAconf* are determined by trading off the huge amount of itemsets returned with low parameter values and the possible interesting itemsets being pruned when using values that are too high. The data structure with an immense amount of items but only a relatively small number of transactions per month is optimal for a depth-first search algorithm like Eclat. Surprisingly, experiments on the data sets with different parameter combinations showed that Eclat was significantly slower than Apriori in finding closed frequent itemsets while the task of finding all frequent itemsets was performed faster.

5 Detecting Changes

The preparation for the change detection process is conducted in accordance with Liu et al.'s [6] approach. The itemsets I_i found for each month i are joined to one set of itemsets I to obtain support sequences of length m for each itemset. Missing support values for parts of the sequence, which occur when the respective itemset is infrequent

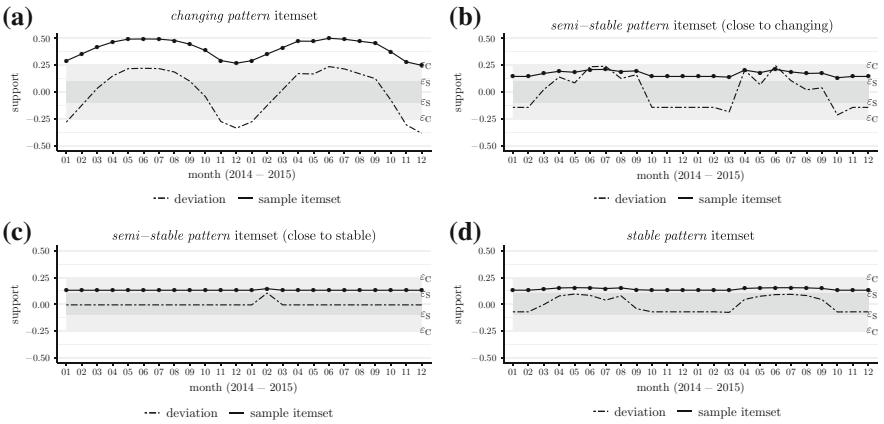

Fig. 1 Monthly progression of changing, semi-stable and stable itemsets

in one month's accident data D_i or has an all-confidence value below the threshold, are filled with the itemset's support generated on the entire transactional data D (*overall support*). This way, unintended breaks in the sequential data are avoided. Itemsets with an *overall support* or an overall all-confidence below the thresholds defined in Sect. 4 are then pruned to further reduce the amount of itemsets.

Change detection for each itemset $X_j \in I \; \forall \, j = 1, \ldots, |I|$ is initiated by calculating the relative deviation $\text{dev}_i(X_j)$ between the itemset's support $\text{supp}_i(X_j) \; \forall \, i = 1, \ldots, m$ for each month and the corresponding $\text{mean}(X_j)$ of the monthly support values as shown in Eq. (1).

$$\text{dev}_i(X_j) = \frac{\text{supp}_i(X_j) - \text{mean}(X_j)}{\text{mean}(X_j)}, \quad i = 1, \ldots, m \tag{1}$$

Based on this computation we define two thresholds ε_s and ε_c to classify all itemsets X_j according to their change level unambiguously.

Stable itemsets I_s with $|\text{dev}_i| \leq \varepsilon_s \; \forall \, i = 1, \ldots, m$.
Semi-stable itemsets I_{ss} with $|\text{dev}_i| \leq \varepsilon_c \; \forall \, i = 1, \ldots, m$ and $\exists \, i : |\text{dev}_i| > \varepsilon_s$.
Changing pattern itemsets I_c with $|\text{dev}_i| > \varepsilon_c \; \forall \, i = 1, \ldots, m$.

The thresholds ε_s and ε_c are determined by evaluating the itemset progresses visually using graphs. In particular, we examined the itemsets within these classes that have either a very high or very low sum of deviations, as the probability of misclassification is severe for these itemsets. In Fig. 1, we show some characteristic sequences for itemsets within the classes. Due to the boundary to both other classes, the class of *semi-stable* itemsets has to be evaluated for both thresholds ε_s and ε_c (cf. Fig. 1b, c).

We get similar class sizes for the *changing* and *stable* class with about 70,000 itemsets each, while the *semi-stable* class contains nearly twice the number of itemsets. Since the *changing* itemsets are most important for our purpose, these will be

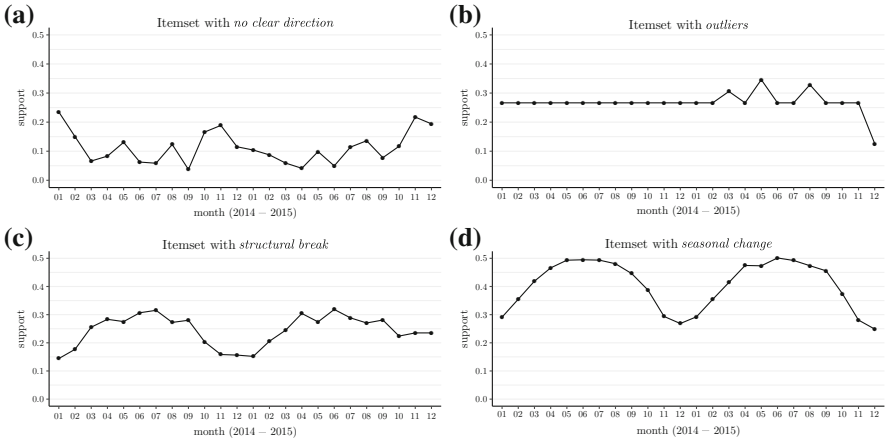

Fig. 2 Sample changing itemsets with particular progresses

presented to a police analyst first, while *stable* and *semi-stable* itemsets do not need to be investigated in the first place. The remaining class size is still too large to monitor all changes. Therefore, we propose to rank the *changing* itemsets by their amplitude of support values.

Based on our purpose to detect changes in accident characteristics, Fig. 2 displays some typical sequences for *changing* itemsets. Itemsets with no clear direction as in Fig. 2a or providing outliers in Fig. 2b are not following any trend and can therefore hardly be predicted. They have to be detected by measuring the support growth between two subsequent months for example, and presented to police analysts for further investigation. Itemsets with seasonal change as in Fig. 2d, are mostly depending on weather conditions and are therefore not surprising, which is why they can be neglected. Itemsets with structural breaks, as Fig. 2c shows, should be detected as fast as possible, since they point out a major change in the underlying data.

6 Conclusion and Future Work

Conclusion We were able to present a basic framework to detect change patterns in road accident statistics. Our assumptions were evaluated using the road safety data set for Great Britain. With a low *minsupp* and a condensed representation of itemsets, we were able to find the most interesting itemsets. We then divided the itemsets in three classes according to their dispersion from the mean support over the whole sequence and ranked the itemsets by their amplitude.

Research Agenda To utilize our framework in police practice the approach requires further research. For instance, the classification of change levels could not only be based on basic thresholds for the deviation from mean but also on growth rates

for different time intervals. A time series analysis for each itemset sequence would also be conceivable to detect seasonal changes as well as linear trends. As can be seen in Fig. 1, many itemsets have the same shape of progression of the monthly support. Here, the approach of fundamental rule changes [5] could further reduce the number of itemsets without any information loss. With an approach to cluster these sequences, we could however refrain from using manual thresholds for detecting changing sequences at all.

The geographical aspect has not been considered yet. Taking the accident location into account, e. g. by geographical clustering, will lead to even more useful results for police forces, as they will be enabled to act preventative on local black spots and, even more important, on geographically shifting black spots.

References

1. Baron, S., Spiliopoulou, M. & Günther, O. (2003). Efficient monitoring of patterns in data mining environments. In *7th East-European Conference on Advances in Databases and Informations Systems* (pp. 253–265).
2. Böttcher, M., Höppner, F., & Spiliopoulou, M. (2008). On exploiting the power of time in data mining. *ACM SIGKDD Explorations Newsletter*, *10*(2), 3.
3. Chen, M. C., Chiu, A. L., & Chang, H. H. (2005). Mining changes in customer behavior in retail marketing. *Expert Systems with Applications*, *28*(4), 773–781.
4. Geurts, K., Thomas, I., & Wets, G. (2005). Understanding spatial concentrations of road accidents using frequent item sets. *Accident Analysis & Prevention*, *37*(4), 787–799.
5. Liu, B., Hsu, W.& Ma, Y. (2001). Discovering the set of fundamental rule changes. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001)*.
6. Liu, B., Ma, Y. & Lee, R. (2001). Analyzing the interestingness of association rules from the temporal dimension. In *Proceedings 2001 IEEE International Conference on Data Mining* (pp. 377–384).
7. Moradkhani, F., Ebrahimkhani, S., & Sadeghi Begham, B. (2014). Road accident data analysis: a data mining approach. *Indian Journal of Scientific Research*, *3*(3), 437–443.
8. Pande, A., & Abdel-Aty, M. (2009). Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool. *Safety Science*, *47*(1), 145–154.
9. Song, H. S., Kim, J., & Kim, S. H. (2001). Mining the change of customer behavior in an internet shopping mall. *Expert Systems with Applications*, *21*(3), 157–168.
10. Xiong, H., Tan, P. N., & Kumar, V. (2006). Hyperclique pattern discovery. *Data Mining and Knowledge Discovery*, *13*(2), 219–242.