mathematics

MDPI

Article

Revealing Driver's Natural Behavior—A GUHA Data Mining Approach

Esko Turunen [1,*] and Klara Dolos [2]

[1] Department of Mathematics and Statistics, Tampere University, Kalevantie 4, 33100 Tampere, Finland
[2] Central Office for Information Technology in the Security Sector (ZITiS), Zamdorfer Street 88, 81677 München, Germany; klara.dolos@ZITIsS.bund.de
* Correspondence: esko.turunen@tuni.fi; Tel.: +358-50-301-4667

Abstract: We investigate the applicability and usefulness of the GUHA data mining method and its computer implementation LISp-Miner for driver characterization based on digital vehicle data on gas pedal position, vehicle speed, and others. Three analytical questions are assessed: (1) Which measured features, also called attributes, distinguish each driver from all other drivers? (2) Comparing one driver separately in pairs with each of the other drivers, which are the most distinguishing attributes? (3) Comparing one driver separately in pairs with each of the other drivers, which attributes values show significant differences between drivers? The analyzed data consist of 94,380 measurements and contain clear and understandable patterns to be found by LISp-Miner. In conclusion, we find that the GUHA method is well suited for such tasks.

Keywords: natural driving behavior; data mining; GUHA method

Citation: Turunen, E.; Dolos, K. Revealing Driver's Natural Behavior—A GUHA Data Mining Approach. *Mathematics* **2021**, *9*, 1818. https://doi.org/10.3390/math9151818

Academic Editor: Jan Rauch

Received: 15 June 2021
Accepted: 27 July 2021
Published: 31 July 2021

1. Introduction

In this paper, we investigate the applicability and usefulness of the GUHA data mining method [1–3] and its computer implementation LISp-Miner [4] in the context of digital forensics. The focus in forensics is the identification of individuals based on different kinds of evidence found at a crime scene. We investigate whether digital data in-vehicle can be used to describe individuals' natural driving behavior. Our research is related to the publication [5], where drivers were classified by machine learning techniques in order to identify them in a forensic scenario of a hit and run accident.

The data we use is freely available data http://ocslab.hksecurity.net/Datasets/driving-dataset (accessed on 30 July 2021).

In total, ten drivers traveled between Korea University and the SANGAM World Cup Stadium in the surroundings of Seoul (South Korea) on three different road types and in-vehicle data was collected. The total driving time per individual was between 121 and 184 min. For reasons explained in [5], we do not use all the features available in the data for our analysis. Moreover, due to the nature of the GUHA method, the time dimension in the data is not relevant in our analysis.

2. The GUHA Method Briefly

Data mining is about finding interesting relations, associations, and structures in given data. At present there are several data mining methods developed for various types of data and related problems; many of them are based on statistics, either classical or Bayesian, and neural networks are used as well as machine learning approaches. GUHA (an abbreviation for Generalized Unary Hypothesis Automata) [1] differs from all the other data mining methods in that it is based on a well-defined logical formalism; dependencies on the data are labelled by truth values TRUE or FALSE (i.e., supported or unsupported by the data); however, the truth value of a statement is determined in a rather unusual way. In a GUHA context, 'data' are a flat matrix with rows and columns; their cells can, in principle, contain

any form of symbols, but in practice the data must be converted to a binary form before the data mining process can take place.

The GUHA method has several computer implementations, the most advanced of which is LISp-Miner software developed and maintained at the Prague School of Economics [4] and freely downloadable from https://lispminer.vse.cz/ (accessed on 30 July 2021). GUHA is not a black box method. To be able to use LISp-Miner software, we need to have at least a rough idea of what the data is about to be able to ask questions, called *analytic questions*, about the data. For example (relevant in this study), '*What are the characteristics in driving style that distinguish driver A from all other drivers?*' is an analytic question. LISpMiner finds all the dependencies relevant to the questions that are in the data, even though this may be a time-consuming process. The ability of LISp-Miner software to analyze such questions is based on the specific logical language of GUHA, the central part of which is *generalized quantifiers* such as 'φ is *often* followed by ψ', 'φ and ψ are *quite equivalent*', 'φ occurs *much more often than average* when ψ is present', and 'φ and ψ *almost always exclude* each other'. Here, φ and ψ are logic statements describing the data; continuing our previous example, φ could mean '*Driver A*' and ψ could stand for '*Vehicle speed is very high and horizontal acceleration is high*'. Within the user-specified boundary conditions, LISp-Miner checks contingency tables of form

	ψ	$\neg\psi$	
φ	a	b	$r = a + b$
$\neg\varphi$	c	d	$s = c + d$
	$k = a + c$	$l = b + d$	$m = a + b + c + d$

where m is the number of rows in (Boolean) data matrix, and

- a is the number of objects satisfying both φ and ψ,
- b is the number of objects satisfying φ but not ψ,
- c is the number of objects not satisfying φ, satisfying ψ,
- d is the number of objects not satisfying φ nor ψ.

For example, there is a sort of equivalence between the statements φ and ψ, denoted by

$$v(\psi \approx \varphi) = \text{TRUE if } (a + d)/m > p, \text{ where } p \in (0, 1], \text{ and } a + d \geq Base$$

It is important to note that '$\approx$' is a generalized quantifier, not a logical connective. The closer the value of the parameter p is to the value of 1, the more confidently the statements φ and ψ appear only simultaneously in the data. Clearly, if $b = c = 0$ then the relation is classical equivalence. The higher the value of the parameter *Base*, the more significant the dependence. In practice, LISp-Miner goes through up to hundreds of thousands of such contingency tables but prints only those labelled by TRUE. However, due to its strong logical and combinatorial basis, LISp-Miner does not go through all possible contingency tables but only those relevant to the question. This speeds up the calculation considerably.

To date, another ten different quantifiers have been implemented in LISp-Miner software; we will introduce them later in the appropriate section. By appropriately adjusting the parameters p and *Base*, we find small-in-number in cases but with extremely important associations if they exist in the data.

3. Presentation of the Analyzed Data

The analyzed data were obtained as follows. In total, 10 drivers traveled between Korea University and the SANGAM World Cup Stadium in the surroundings of Seoul (South Korea) on three different road types. The number of features recorded was 51 in 1 s time intervals. Total driving time per individuals was between 121 and 184 min (cf. [5]). The key research task is, from this data set, to distinguish each driver individually. In this study, the starting point of our data mining analysis is a raw data set with 94,380 rows and the following 12 columns:

- Drivers A, . . . , J, (10 in all),

- Acceleration Speed Lateral,
- Acceleration Speed Longitudinal,
- Acceleration Pedal Value,
- Acceleration Pedal Value (RN),
- Fuel Consumption,
- Master Cylinder Pressure,
- Cylinder Pressure (RN),
- Steering Wheel Angle,
- Steering Wheel Angle (RN),
- Vehicle Speed,
- Vehicle Speed (RN).

The roughness (RN) was calculated with rolling windows of size 20 s for all features with the R-function roughness {seewave}.

4. Data Preprocessing

Because LISp-Miner analysis is based on binary data, we processed the raw data as follows. Apart from driver identification, all the values in cells are numeric and linearly ordered. For simplicity, we divided each column into seven equally long sections and, to illustrate the results, we colored them as follows

Each driver is handled separately, which produces ten new columns. Thus, there are $11 \times 7 + 10 = 87$ columns in the input data in LISp-Miner and so the analyzed data is a $94{,}380 \times 87$ Boolean data matrix. There are no empty cells in this data. The task is to characterize each of the ten drivers with a maximum of 77 different characteristics or combinations of these characteristics. In GUHA logic terminology, these 77 columns are called *attributes* or (*unary*) *predicates*, for example,

- Acceleration Speed Lateral (extra low),
- Steering Wheel-Angel (higher),
- Vehicle Speed (lower),
- Driver (D).

are such attributes. A *statement* (or *open formula*), denoted by φ or ψ, is composed of these attributes by the logic conjunction. For example, if

- φ stands for Driver(D),
- ψ stands for Steering Wheel Angle(higher) & Vehicle Speed(lower).

then output of LISp-Miner procedure could be $\varphi \approx \psi$, where $\approx$ is a specified generalized quantifier, assuming of course that the data supports such an association. In GUHA language, such *closed formulas* are called *hypothesis*. This explains the name GUHA: an automaton that produces hypotheses from a given data. Thus, LISp-Miner produces *hypothesis*, i.e., associations that the data supports.

Further, we divided the data thus obtained into two parts. The rows with a sequence number not divisible by four (75% of all the rows) form the *model set*, and the remainder (25%) the *test set*. We first examined in the model set and selected a few hypotheses that are most strongly supported by the data, then we performed the same analysis among the test set and examined whether the same strongest hypotheses can be found among these results. Finally, we report some of the strongest hypothesis supported both by the model set and the test set.

5. Analytical Questions and the Most Relevant Answers to Them

In this chapter, we present the three key analytical questions we posed to the LISp-Miner software and some of the answers we received. Of the results, we have selected only the most significant.

5.1. The First Analytic Question: Which Hypotheses Distinguish Each Driver from All Other Drivers

It is natural to use the Above Average Quantifier of the 4ft-Miner procedure in LISp-Miner, because the hypotheses it produces answer to the question: in terms of which combination of attributes in ψ, does driver φ differ most clearly from the other drivers? Here the truth definition $v(\varphi \approx \psi)$ = TRUE is based on the condition

$$a/(a+b) \geq (1+p)(a+c)/m, \text{ where } p > 0 \text{ and } Base \geq a$$

in the related contingency table, and $v(\varphi \approx \psi)$ = FALSE elsewhere. For example, if $p = 4$ and the above two conditions hold, then the statement ψ is at least 5 (=1 + p) times more common for driver φ than it is on average. More generally, the larger p, the more clearly the combination of attributes in ψ distinguishes driver φ from other drivers.

It is natural that the value a (called *support*) must be large enough. Otherwise, the result would have low general significance. Moreover, to make the distinguishing attributes as clear as possible, the value of the parameter p must be as large as possible. On the other hand, it lowers the threshold value *Base*, so a compromise must be made. After a few experiments, we came to the values $p \geq 5$–10 and $Base \geq 100$. In Figure 1, there is a screenshot from the front page of LISp-Miner when retrieving attributes describing driver D, where $p \geq 5.2$.

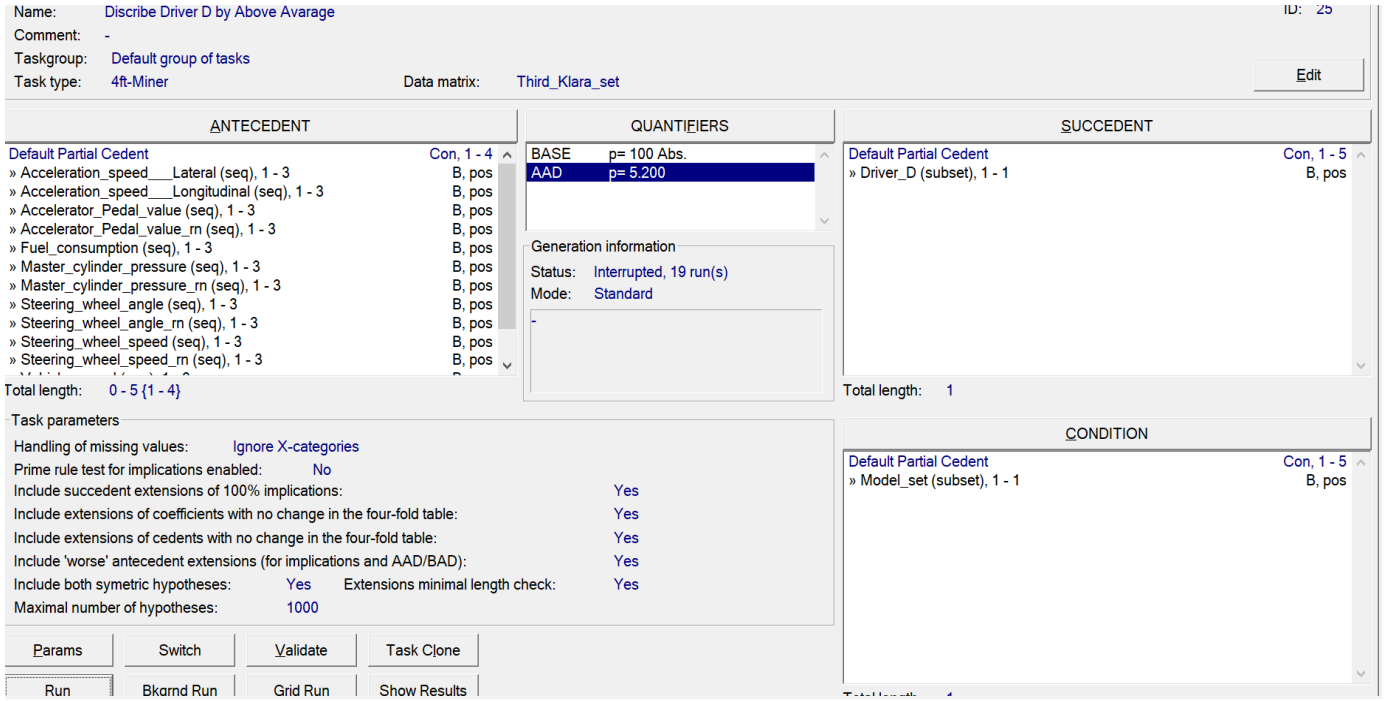

Figure 1. LISp-Miner screenshot when retrieving attributes describing driver D.

We observe that computer performance may sometimes constrain computing. If we limit the number of attributes in ψ to 12, LISp-Miner goes through about 2.5 million contingency tables. For this, a standard desktop computer takes about 25–30 min. For example, when examining the attributes characteristic of driver G, LISp-Miner went through 25,556,692 contingency tables, from which 101 fulfilled the required boundary conditions; these are the related hypothesis. This took 29 min, 28 s, see Figure 2.

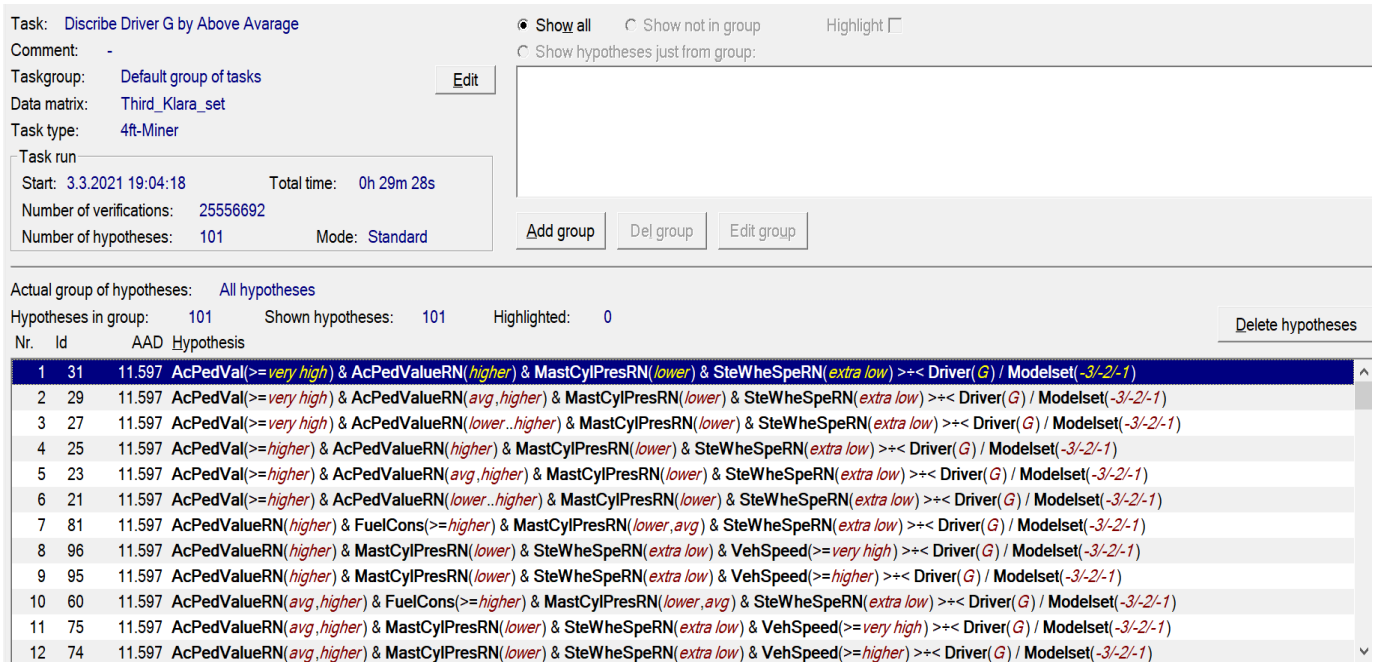

Figure 2. Screenshot on LISpMiner results for driver G.

Moreover, there are several almost identical hypotheses among the 101 (in the sense of over lapping attributes) among the outputs produced by LISp-Miner. For example, driver G is associated with the following attributes (see the first two lines on Figure 2).

AccelPedal Value & Accel Pedal ValueRN & Master Cylinder PressureRN & SteerWheSpeedRN
AccelPedal Value & Accel Pedal ValueRN & Master Cylinder PressureRN & SteerWheSpeedRN

We do not report them all but only the most apparent, the choice is more or less random. The results are on the following Figure 3.

Figure 3 is to be understood as follows. For example, in the first row (A: model), the hypothesis 'Acceleration Speed Longitudinal(higher) & Fuel Consumption(average) & Master Cylinder Pressure(extra high) & Steering Wheel Angle(lower ... higher)' is more than 12 times (exactly 1 + 11,325) more common for driver A than for all the other drivers combined. Indeed, in the model set, there are at least (*Base* $\geq$) 100 rows (in fact a = 104), where this property is associated with driver A. Moreover, there are only six (=b) rows where this property does not exist but A does. Thus, the *share* (104 out of 110) is 95%. It should be mentioned (although this is not written in Figure 3) that for the group of the rest of drivers the corresponding figures are 5326 out of 65,349; thus, the share is only 8%. Correspondingly, in the second row (A: test), the statement 'Acceleration Speed Longitudinal(higher) & Fuel Consumption(average ... very high) & Master Cylinder Pressure(extra high) & Acceleration Speed Lateral(average)' is more than 11 times (exactly 1 + 10.1) more common for driver A than for all the other drivers combined. In the test set, there are at least (*Base* $\geq$) 35 rows (in fact a = 40), where this property is associated with driver A. Moreover, there are only 7 (=b) rows where this property does not exist but A does. Thus, the share (40 out of 47) is 85%. In the group of the rest of the drivers the share is only 8%. It can also be seen from Figure 3 that the properties describing the drivers G, H, and I are exactly the same both in the model set and in the test set. The results for the other drivers are also very similar in both sets. Moreover, it is noteworthy that all the original 11 variables in row data appear in the results at least once'.

LISp-Miner includes a Bayesian statistics-based tool to assess the reliability of results. The result is presented as a graph of its distribution (see Figure 4). The tapered the distribution (Graph on the left-hand side), the more reliable the result. The theoretical basis for this interpretation is explained in the publication [6].

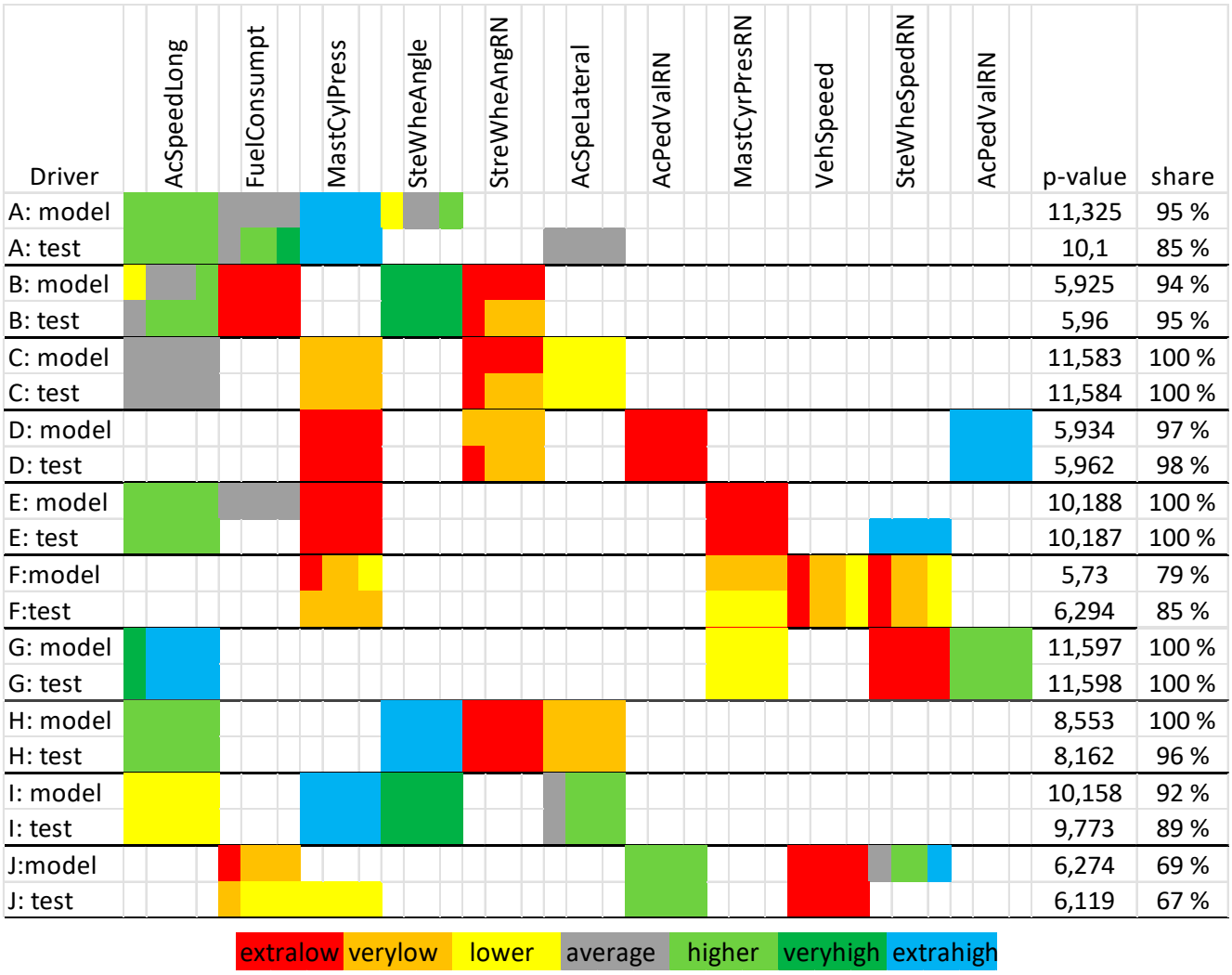

Figure 3. LISp-Miner answers to the analytic question 'Which hypotheses distinguish each driver from all other drivers?

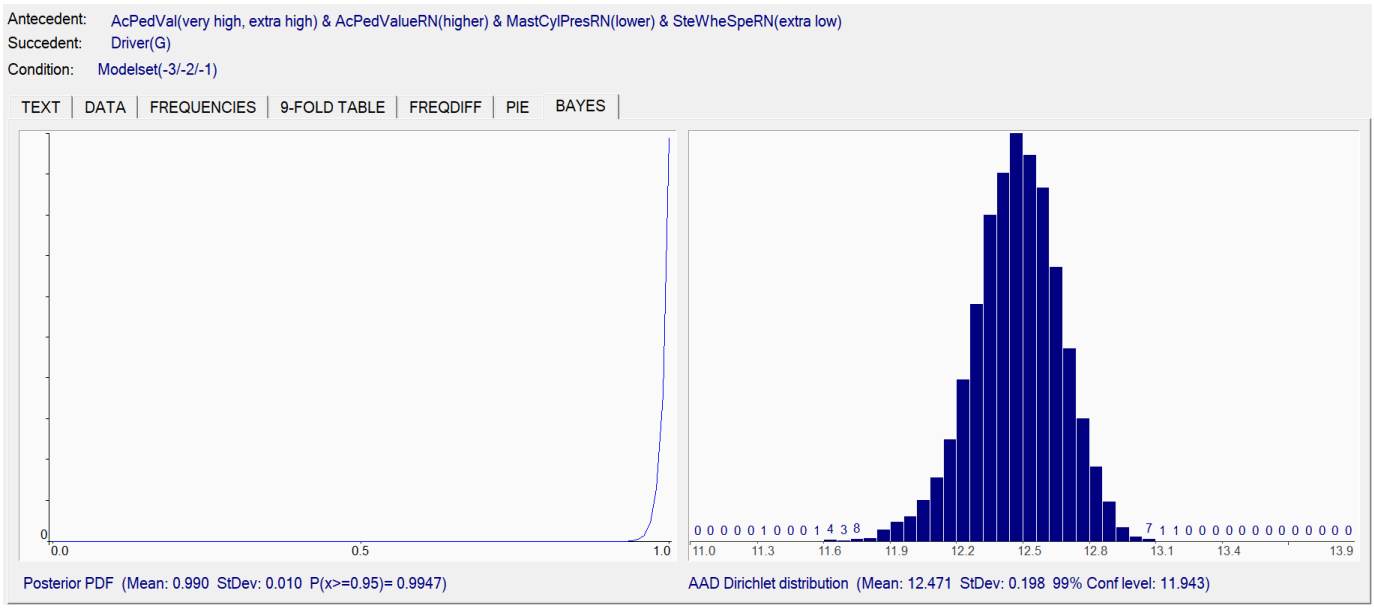

Figure 4. Bayesian theory-based interpretation of the statistical significance of the hypothesis G: model.

5.2. The Second Analytic Question: Comparing One Driver Separately in Pairs with Each of the Other Drivers (up to 3), Which Are the Most Distinguishing Attributes

There are several ways in the LISp-Miner software to perform this kind of task. One of them is SD4ft-miner procedure [7], a handy tool for finding remarkable differences between two separate subsets of the data. The truth value depends on *two* contingency tables, and there are six possible quantifiers. The simplest one is based on the condition $\left|\frac{a}{a+b} - \frac{A}{A+B}\right| \geq p$, where a and b refer to the contingency table defined by the first set and A and B refer to the contingency table defined by the second set, respectively. Obviously, $0 \leq p \leq 1$, the closer the value is to 1, the more different the sets are with respect to that property. Conditions for related *Base* values can also be set.

As an example, we examine which (up to 3) attributes distinguish driver F from the other drivers. *Base* values are 1% of the total (both in the model set and in the test set) and $p \geq 0.2$.

In addition, we limit the number of attributes to a maximum of three. In Figure 5, we have collected 1 to 3 such hypotheses (where the p-value is the highest) that are produced both in the model set and the test set.

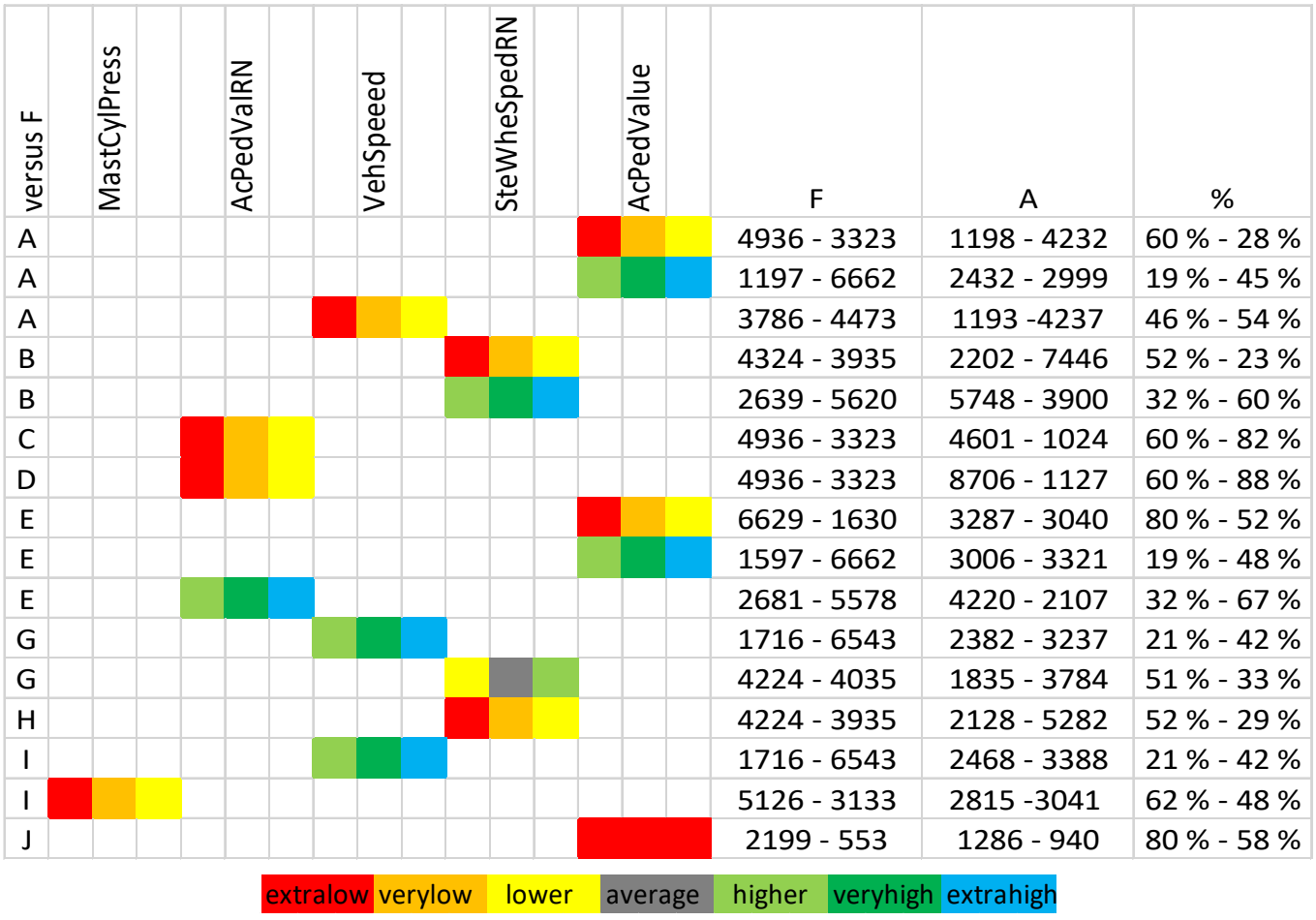

Figure 5. Hypotheses that distinguish driver F in pairs from other drivers.

Figure 5 is to be understood as follows. For example, driver A differs significantly (at least) in three different ways from driver F, two of them are related to Acceleration Pedal Value, and one to Vehicle Speed (lines 1, 2, and 3 in Figure 5). Indeed, (see the first line) there are a = 4936 rows in the model data where φ is driver F and ψ is Acceleration Pedal Value(extra low . . . lower) and b = 3323 rows where φ is present but ψ is not. Thus $\frac{a}{a+b} = 0.596$, corresponding to 60% (rounded up) of the cases. On the other hand, there are A = 1198 rows in the model data where φ is driver A and ψ is Acceleration Pedal

Value(extra low ... lower) and $B = 4323$ rows where φ is present but ψ is not. Thus $\frac{A}{A+B} = 0.277$, corresponding to 28% of the cases. This gives $\left|\frac{a}{a+b} - \frac{A}{A+B}\right| = 0.32 > p$.

It is noteworthy that of all the 11 original measurement classes, only 6 occur in Figure 5, and of all the 77 predicates, i.e., the columns that possibly characterize drivers, only 33 emerge. Fisher's exact test tested the statistical significance of the results. In all the above 16 cases, the difference is statistically significant if the limit $p = 0.01$ for significance is used. Thus, it can, with high probability, be considered certain that the results produced by LISp-Miner are not due to chance but actually describe the differences between driver F and other drivers.

5.3. The Third Analytic Question: Comparing One Driver Separately in Pairs with Each of the Other Drivers, Which Attribute (up to 3) Values Show Significant Differences between Drivers

Ac4ft-Miner procedure is well suited for investigating such an issue. The key idea is as follows: consider some background factors constant but change some others and see where this change leads. The definition of truth is the same as in the SD4ft-miner context.

Consider, as an example, the effects on Fuel Consumption, Vehicle Speed, or Vehicle SpeedRN of driver change from F to X, when one of the other attributes remains constant.

Figure 6 is to be understood as follows. For example, comparing drivers F and A (the first line in Figure 6), the model set contains $a = 2660$ rows with statement ψ is Acceleration Speed Longitudinal(lower ... higher) & Fuel Consumption(very low ... average), and φ is driver F. The corresponding b value (φ is present but ψ is not) is 776. Thus $\frac{a}{a+b} = 0.774$, corresponding to 77% of the cases.

Driver						
F	Acceleration Speed Long is	&	Fuel Consum is	77 %	2660 - 776
A	Acceleration Speed Long is	&	Fuel Consum is	63 %	1371 - 804
F	Master Cylinden Press is	&	Fuel Consum is	49 %	2666 - 2751
B	Master Cylinden Press is	&	Fuel Consum is	41 %	1984 - 2856
F	Accelaration Pedal Value is	&	Fuel Consum is	62 %	4078 - 2528
C	Accelaration Pedal Value is	&	Fuel Consum is	49 %	2495 - 2578
F	Master Cylinden Press is	&	Fuel Consum is	56 %	1665 - 1330
D	Master Cylinden Press is	&	Fuel Consum is	68 %	2850 - 1324
F	Master Cylinden Press is	&	Fuel Consum is	47 %	2408 - 2718
E	Master Cylinden Press is	&	Fuel Consum is	39 %	1710 - 2719
F	Master Cylinden Press is	&	Vehicle Speed is	51 %	2590 - 2536
H	Master Cylinden Press is	&	Vehicle Speed is	35 %	1508 - 2597
F	Accelaration PedValueRN is	&	Fuel Consum is	52 %	3495 - 3164
G	Accelaration PedValueRN is	&	Fuel Consum is	34 %	1514 - 2993
F	Accelaration Pedal Value is	&	Vehicle Speed is	46 %	3055 - 3574
I	Accelaration Pedal Value is	&	Vehicle Speed is	22 %	1002 - 3456
F	Steering Wheel SpeedRN is	&	Fuel Consum is	49 %	2064 -2160
J	Steering Wheel SpeedRN is	&	Fuel Consum is	38 %	1374 - 2214

extralow verylow lower average higher veryhigh extrahigh

Figure 6. Hypotheses that distinguish driver F in pairs from other drivers (by Ac4Ft-Miner).

If, on the other hand, for φ is the driver A, the corresponding figures are A = 1371, B = 804, therefore $\frac{A}{A+B} = 0.630$, corresponding to 63% of the cases. This gives $\left|\frac{a}{a+b} - \frac{A}{A+B}\right| = 0.14$.

The statistical significance of the results was tested by Fisher's exact test. In all the 10 cases, the difference is statistically significant if the limit $p = 0.01$ for significance is used. The results can therefore be considered statistically reliable. In Figure 6 we summarize some of the most significant results produced on the model set, which are also produced on the test set. The parameter *Base* = 1% of all the rows in the related (model or test) set and $p \geq 0.1$.

6. Observations

We list some of the strengths and weaknesses of GUHA and the LISp-Miner software that we have identified during this study. To begin with, LISp-Miner has the right tools to analyze large data sets; we have not even used every possible one here. The problem we solved was obvious; making analytical questions was therefore easy. In general, this may not always be that simple; using LISp-Miner and interpreting the results requires practice. There are plenty of statistically significant differences in the data that differentiate the drivers; we have listed only the most significant, definitely not all of them. If the data size increases or there is a need to study really small subsets (say, less than 0.05% of the total data), the computation time will extend; however, this problem can be alleviated by introducing more powerful computers. LISp-Miner is a freely downloadable software [7]; ın any case, since GUHA logic is a well-defined logic, users can write their own software for their own needs if they do not want to use LISp-Miner.

7. Conclusions

In this work, we have investigated the suitability of the GUHA data mining method to find statistically significant features that differentiate between drivers in data on drivers' driving behavior. The result is unequivocally positive; since those differences in the data exist, LISp-Miner, the GUHA method implementation, finds them. Indeed, using the GUHA approach, it is possible to characterize drivers by combinations of feature values very special for them. This differs from the approach of using classification models largely targeting mean values and typical ranges to characterize classes. Rare values or value combinations are usually treated as outliers; some algorithms are even promoted because of their robustness towards outliers. This was found to be one important reasons for weak results in an attempt using a one-class classification (cf. [8]). In the forensic context, a combination of classical classification and an approach based on individual patterns in driving behavior could help to gain more reliable and explainable results. Another approach worth being investigated, instead of time series analysis and machine learning, is to take a closer look at specific patterns in the data. Such patterns are value combinations that are rare, e.g., only a few seconds each hour but occur with one person only.

Author Contributions: Methodology, E.T. and K.D.; formal analysis, E.T.; writing—review and editing, E.T. Both authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This study has been conducted within the framework of COST Action CA17124 DigForASP.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rauch, J.; Šimůnek, M. An Alternative Approach to Mining Association Rules. In *Foundations of Data Mining and Knowledge Discovery*; Lin, T.Y., Ohsuga, S., Liau, C.-J., Hu, X., Tsumoto, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2005; pp. 211–231. ISBN 978-3-540-26257-2. Available online: http://www.springer.com/engineering/book/978-3-540-26257-2 (accessed on 30 July 2021).
2. Rauch, J. Observational Calculi and Association Rules. In *Artificial Intelligence: Foundations, Theory, and Algorithms*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 469, pp. 154–196.
3. Rauch, J.; Šimůnek, M. Apriori and GUHA—Comparing two approaches to data mining with association rules. *Intell. Data Anal.* **2017**, *21*, 981–1013. [CrossRef]
4. Šimůnek, M. Academic KDD Project LISp-Miner. In *Advances in Soft Computing—Intelligent Systems Design and Applications*; Abraham, A., Franke, K., Koppen, K., Eds.; Springer: Berlin/Heidelberg, Germany, 2003; pp. 263–272. ISSN 1434-922/3-540-40426-0.
5. Dološ, K.; Meyer, C.; Attenberger, A.; Steinberger, J. Driver identification using in-vehicle digital data in the forensic context of a hit and run accident. *Forensic Sci. Int. Digit. Investig.* **2020**, *35*, 301090. [CrossRef]
6. Piché, R.; Järvenpää, M.; Turunen, E.; Šimůnek, M. Bayesian analysis of GUHA hypotheses. *J. Intell. Inf. Syst.* **2013**, *42*, 47–73. [CrossRef]
7. The Official Site of the LISp-Miner Project. Available online: https://lispminer.vse.cz/ (accessed on 30 July 2021).
8. Dološ, K.; Meyer, C.; Attenberger, A.; Steinberger, J. Driver C. Driver identification considering an unknown suspect. *AMCS*. submitted.